Atten Percept Psychophys (2017) 79:1993–2006
DOI 10.3758/s13414-017-1347-0

Detecting and identifying offset gaze

Tarryn Balsdon[1] · **Colin W. G. Clifford**[1]

Published online: 14 June 2017

Abstract A number of experiments have demonstrated that observers can accurately identify stimuli that they fail to detect (Rollman and Nachmias, 1972; Harris and Fahle, 1995; Allik et al. 1982, 2014). Using a 2x2AFC double judgements procedure, we demonstrated an analogous pattern of performance in making judgements about the direction of eye gaze. Participants were shown two faces in succession: one with direct gaze and one with gaze offset to the left or right. We found that they could identify the direction of gaze offset (left/right) better than they could detect which face contained the offset gaze. A simple Thurstonian model, under which the detection judgement is shown to be more computationally complex, was found to explain the empirical data. A further experiment incorporated metacognitive ratings into the double judgements procedure to measure observers' metacognitive awareness (Meta-d') across the two judgements and to assess whether observers were aware of the evidence for offset gaze when detection performance was at and below threshold. Results suggest that metacognitive awareness is tied to performance, with approximately equal Meta-d' across the two judgements, when sensitivity is taken into account. These results show that both performance and metacognitive awareness rely not only on the strength of sensory evidence but also on the computational complexity of the decision, which determines the relative distance of that evidence from the decision axes.

Keywords Social vision · Metacognition · Psychophysics · Double judgements

✉ Tarryn Balsdon
t.balsdon@unsw.edu.au

[1] School of Psychology, UNSW Sydney, Mathews building, Sydney, NSW, Australia 2052

Introduction

Of particular interest in visual psychophysics is the sensitivity of a participant to a specific stimulus or property of that stimulus. Often it is unclear, when an observer is asked to make a single decision about a stimulus (present/absent, category A or B, et cetera), what evidence he or she is using to make the decision. "Double judgement" experiments were originally developed to tackle this issue. Participants are asked to make both a detection judgement and an identification judgement about a single stimulus, for example, to detect which interval a Gabor patch is presented in, and to identify its phase (Huang, Kingdom and Hess, 2006). It has been reasoned that if performance on the two judgements is roughly equal, the observer must be basing their detection judgement on the same information used to make their identification judgement. If identification performance is worse, then the property of the stimulus being identified must not be integral in making the detection judgement.

Most experiments show detection performance to be at least equal to identification performance and that identification performance becomes closer to detection performance with decreased similarity between the stimuli to be identified (Thomas, 1985). However, certain experiments reveal that participants can show superior performance in the identification judgement relative to the detection judgement (e.g. vernier acuity, Harris and Fahle, 1995; colour judgements, Rollman and Nachmias, 1972; bull's eye acuity, Allik et al., 1982, 2014). This result has been treated in a number of ways; some researchers calculated identification performance based on only those trials in which detection was correct (Tolhurst and Dealy, 1975). This treatment rests on the assumption that an identification decision made when the detection decision is incorrect would be a guess and that any correct response in this instance would result from chance. However, it is difficult

to justify this assumption when performance in the identification decision is significantly above chance on trials where detection is incorrect. However, if observers are not simply guessing on incorrect detection trials, it begs the question of how they can successfully identify a property of a stimulus they cannot detect.

To clarify, the use of the term "detection" in these experiments does not necessarily mean that observers were detecting the presence of a stimulus versus no stimulus at all, but rather in some cases, observers were detecting some property of the stimulus. Detection is a special type of discrimination where observers are essentially discriminating one stimulus from another, except that one stimulus is "null," and the analysis of detection and discrimination judgements can be undertaken in the same manner (Macmillan and Creelman, 2004). In keeping with previous literature, we used the term detection to encompass designs where observers are detecting the presence of a property of a stimulus, such as spatial offset, from the absence of that property, such as no spatial offset, or an offset of 0.

That identification might be better than detection may not be as mysterious as it seems. In a recent experiment, Allik et al. (2014) replicated previous findings (Allik et al. 1982) showing superior identification compared with detection of spatial offset in a bull's eye acuity task. A 2x2AFC design was employed in which participants were presented with two annuli: one of which contained a dot in the exact centre, and one with a dot offset horizontally to the left or to the right of centre, in a single trial. Participants were asked to make two decisions: which direction the dot was offset in (the identification judgement) and which annulus contained this dot (the detection judgement). Allik et al. showed that performance across both judgements could be effectively explained by a simple Thurstonian model in which the "strengths" of the stimulus property (spatial offset) are represented on a single internal continuum of spatial positions (Thurstone, 1927). Let x and y denote the (signed) evidence in the first and second intervals respectively, such that, for example, a positive value of x denotes evidence that the spatial offset in the first interval was rightwards. The identification judgement can be made on the basis of whether $x + y > 0$. The detection judgement is more complex, because the observer does not simply need to know whether there was overall more evidence for left vs. right spatial offset but also the relative position of the stimuli in the two intervals. Therefore, observers also must test whether $x - y > 0$ to know in which interval the evidence was presented (thus adding a decision axis), so the detection judgement becomes based on whether $(x + y)(x - y) > 0$ (or, equivalently, $(x^2 - y^2) > 0$). This description of the two judgements is shown in Fig. 1, where the evidence of spatial offset in each interval, x and y, are shown intersected. For the identification decision (Fig. 1a), evidence for left versus right offset increases away from the negative diagonal, whereas in the detection decision (Fig. 1b), evidence for one interval over another increases along that axis (x or y) away from the intersection, bounded by both the positive and negative diagonals. The detection judgement is more computationally complex, as a result of the additional decision axis. Therefore, the evidence for the detection judgement is more likely to be closer to a decision axis than in the identification judgement, meaning that more evidence is required for comparative certainty in the detection decision.

Under a Thurstonian model, the evidence on the internal continuum for spatial offset is the addition of signal from the stimulus with internal noise drawn from a Gaussian probability distribution. The theoretical intersection of the evidence from each interval shown in Fig. 1 forms a two-dimensional Gaussian that can be used to predict performance across both judgements. Figure 2a shows a top-down view of the two dimensional Gaussian distribution of internal evidence for a stimulus pair. The area under the Gaussian above the negative diagonal describes the probability of

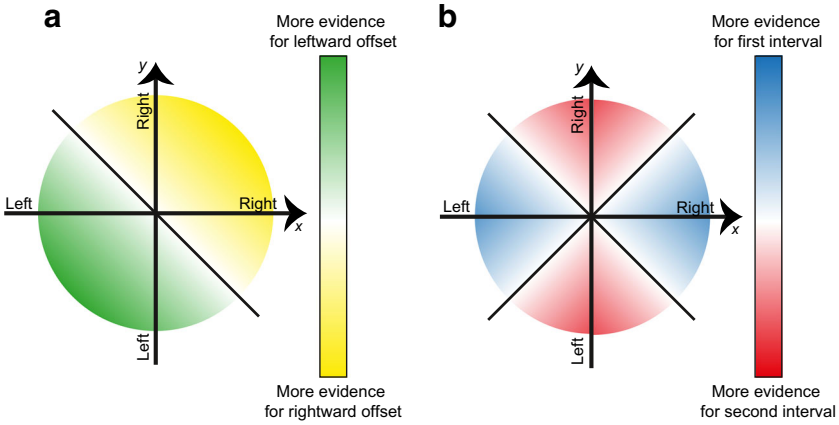

Fig. 1 Decision space for the identification judgement (a) and the detection judgement (b). The axes x and y represent the (signed) evidence in each interval whilst the diagonals represent ideal decision axes for each judgement. For the identification judgement (a) only the negative diagonal ($x - y > 0$) is necessary to decide whether the evidence over the two intervals indicates more leftward or rightward offset, whereas in the detection judgement (b) the positive diagonal ($x + y > 0$) is also necessary. In each case the further the evidence is from the decision axes, the more evidence there is for that decision

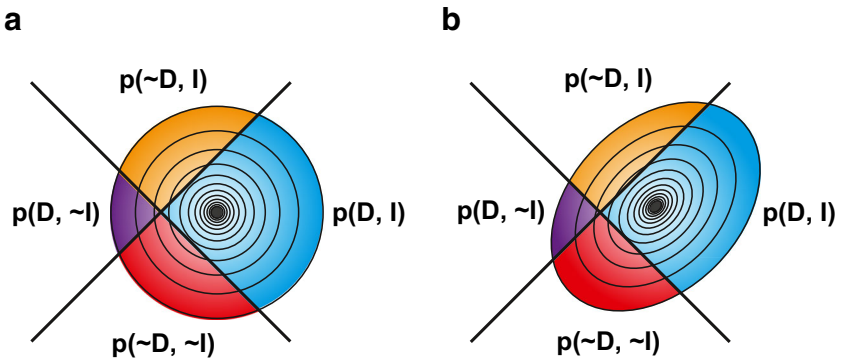

Fig. 2 Top down view of two-dimensional Gaussian distributions representing the probability distributions of evidence for a stimulus pair, when evidence is added to internal noise. The diagonals correspond to the decision axes in Fig. 1; however, here the quadrants represent the outcome of the decision made on the basis of where the evidence on a single trial sits relative to the axes. The area under the Gaussian in each quadrant therefore describes the expected proportion of responses according to accuracy across both decisions. In (a) the area under the Gaussian above the negative diagonal is equal to that below the positive diagonal, meaning that p(~D, I) = p(~D, ~I). (b) If noise in the estimated spatial offset is correlated between the two intervals, for example, due to a drift in the subjective centre, then the area above the negative diagonal is no longer equal to the area below the positive diagonal and p(~D, I) ≠ p(~D, ~I)

correct performance in the identification judgement (I) irrespective of performance in the detection judgement (D), $p(\mathrm{I}) = p(\mathrm{D}, \mathrm{I}) + p(\sim\mathrm{D}, \mathrm{I})$. Similarly, the overall probability of correct performance on the detection task is given by the sum of the areas of the Gaussian to the left and right of the pair of decision axes, $p(\mathrm{D}) = p(\mathrm{D}, \mathrm{I}) + p(\mathrm{D}, \sim\mathrm{I})$. Note that, under Allik et al.'s implementation of this model (Fig. 2a), $p(\sim\mathrm{D}, \mathrm{I}) = p(\sim\mathrm{D}, \sim\mathrm{I})$, meaning that identification performance is at chance for trials on which detection performance is incorrect. The model proved effective in explaining experimental double judgment data with a single free parameter relating to the width (standard deviation) of the Gaussian noise distribution (Allik et al., 2014).

The model implemented by Allik et al. (2014) was also outlined by Klein (1985). However, Klein also suggested a more complex version. Klein questioned the assumption that the centre of the stimulus could be considered a natural zero point that was fixed for each observer. He argued that slight changes in an observer's subjective centre would cause their perceived x and y values to drift in a correlated fashion, resulting in an artificial expansion of the distribution of evidence along the positive diagonal shown in Fig. 2b, thereby elongating the two-dimensional Gaussian in this direction. To account for this "correlated noise" a second parameter is needed, because the Gaussian is now elliptical rather than circularly symmetric. In comparing the fit of the two models to Allik et al.'s 1982 data, Klein found the model accommodating for correlated noise to be the better fit. However, Klein did not account for the fact that the winning model was more complex, and so his conclusion could have been an artifact of a more complex model inevitably having more explanatory power. Henceforth these models will be referred to as H_{A1} and H_{A2} (for the simple Thurstonian model and the two parameter version incorporating correlated noise, respectively).

Two "null" models also are introduced. Similar to models H_{A1} and H_{A2}, these models assume that perceptual evidence for making judgements increases on an internal continuum of spatial offset with additive Gaussian noise. The first (H_{01}) addresses whether the data could be explained by a single parameter corresponding to the rate of increasing perceptual evidence for offset with stimulus offset (the standard deviation of the Gaussian distribution) across the two judgements, with any differences in performance arising from chance differences in guesses. This is essentially a high-threshold model; if the evidence exceeds a certain threshold, the observer is able to make both judgements correctly, but below the threshold the observer is merely guessing, and incorrect responses only arise from guessing. The second null model (H_{02}) takes the opposite line of reasoning, with two completely separate parameters corresponding to the rate of increasing perceptual evidence for offset with stimulus offset in each judgement. This model assumes that the two judgements are based on completely separate evidence, and therefore any difference in performance across the two judgements corresponds to a difference in the amount of evidence for each decision.

One further question is whether the pattern of performance over the two judgements might be a peculiarity of lower order visual mechanisms. Findings of better identification than detection typically have been shown using simple stimuli, and for only a few properties of these stimuli. But in the real world, we are confronted with far more complex stimuli, for which much higher-order visual processing areas are activated. The fact that there are multiple sources of evidence from complex stimuli makes it unclear whether the simple distribution of evidence described in Figs. 1 and 2 could apply.

The perception of gaze direction is one example of a set of stimuli considered to recruit dedicated higher-order visual processing mechanisms. For example, activity in the human superior temporal sulcus has been demonstrated to show selectivity for the observed direction of gaze (Calder et al., 2007; Carlin et al., 2011). Not only have observers been shown to be

generally very good at making decisions about the direction of gaze offset (Cline, 1967, Jenkins and Langton, 2003), but accuracy in these decisions is integral to social interactions and higher order thought concerning others' beliefs and desires (Friere et al., 2004, Baron-Cohen, 1992). The processing of gaze direction has been suggested to involve calculating the spatial offset of the iris relative to the sclera (Anstis et al., 1969), a judgement similar to that of Allik et al.'s experiment. A coarser spatial cue to gaze direction recruits luminance changes across the eye, such that when the polarity of a gaze stimulus is reversed, the perceived direction of gaze is also reversed (Sinha, 2000; Riccardelli et al., 2000). Yet, in comparing detection and discrimination of Gabors of different phases (a similar judgement to the coarse spatial cue) Huang, Kingdom, and Hess (2006) found certain stimulus pairs could not be identified at detection threshold, the opposite of the effect shown by Allik et al. Furthermore, it is highly likely that a number of cues external to the eye region contribute to perceived gaze direction, such as the orientation of the head (Wollaston, 1824; Otsuka et al., 2014, 2015), the emotional expression of the face (Ewbanks, Jennings and Calder, 2009), and even prior beliefs about where someone is looking (Mareschal, Calder, & Clifford, 2013; Mareschal, Otsuka, & Clifford, 2014). Thus, it is unclear whether the pattern of performance over detection and identification of spatial offset demonstrated in the simple stimuli of Allik et al. will generalize to these more complex and ecologically valid stimuli.

Experiment 1 seeks to address whether the pattern of performance in detecting offset gaze and identifying its direction will mirror that of previous experiments, with superior performance in identification judgements compared to detection judgements. The data will then be used to compare the four models outlined above; the single parameter Thurstonian model (H_{A1}) that shows detection to be the more computationally complex judgement; and the two parameter Thurstonian model (H_{A2}) that also incorporates an unstable subjective direct gaze; as well as the two null models (H_{01} and H_{02}). Using the double judgements procedure, we can thereby explore the role that evidence of spatial offset plays in identifying and detecting offset gaze and how this evidence might be used differently across the two judgements.

Whilst the results of experiment 1 can be used to understand how perceptual evidence may be used differently to make detection and identification judgements, it is important to understand how perceptual evidence is incorporated into a phenomenal experience of the world. There are numerous examples in the literature of dissociations between phenomenal experiences and the accuracy of perceptual judgements, such as in blindsight (Weizkrantz et al., 1974, 1995, Azzopardi and Cowey, 1998), priming (Carr et al., 1982), and backwards masking (Marcel, 1980; 1983), where observers are able to make accurate judgements despite claiming to be unaware of the stimulus. In these cases, the detection threshold is frequently used as the threshold for awareness. Given that previous findings have shown the identification threshold is lower than the detection threshold, this may suggest that the observer is not aware of the evidence used to make their identification judgements. Experiment 2 extends the double judgements procedure to test this hypothesis. A metacognitive rating is incorporated into the procedure in order to test the metacognitive awareness of observers across the two judgements. If perceptual evidence is being used differently across the two judgements in a way that affects the phenomenal experience of the observer, this should be evident from the observer's metacognition of the identification and detection judgements.

Experiment 1 Methods

Participants

Participants were three naïve observers and one author; three were male, two were left-handed, mean age was 25.75 (range 24-29) years, and all had normal or corrected-to-normal vision. Ethical approval for the experiment was granted by the UNSW Human Research Ethics Committee and adheres to the Code of Ethics of the World Medical Association (Declaration of Helsinki). Participants were recruited through a web-based facility called SONA-P and were reimbursed for their time after giving informed consent.

Apparatus and Stimuli

Stimuli were presented on a gamma corrected 18-inch Diamond Digital DV998FD CRT monitor with a background luminance of 33.2 cd/m^2 (resolution 1024 × 768) running at 85 Hz, using Matlab and the Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997; Kleiner et al., 2007). Throughout the experiment, participants sat 57 cm from the screen, with their chin on a chin rest. Responses were entered via the keyboard.

Stimuli were identical grey scale faces created with Daz software (http://www.daz3d.com/). The face was female, with cropped hair and a neutral expression. The face was mirrored vertically through the centre to control for any artifacts that might bias observers based on asymmetry. On half the trials, the face also was flipped left to right to further control for any effect of asymmetry. Gaze direction was controlled by removing the original eyes from the face (using Gimp software) and replacing them with realistic counterparts, generated in Matlab. This allowed for the eyes to be moved according to precise angular coordinates. The original image was resampled to be presented at a realistic size—with an interocular distance of 6.2 cm.

Procedure

A 2x2AFC double judgement procedure was used. On each trial participants were presented with two faces in temporal succession. In one face, gaze was directed at the participant, whereas in the other, it was directed away, either to the left or the right. For half the trials, the face with direct gaze was presented first, and in the other half, the face with averted gaze was presented first. The participant was asked to detect which face they thought was looking away from them—the first or the second (by pressing the 1 and 2 keys)—and to identify which direction the face was looking (by pressing the left and right arrows). The order of these questions was counterbalanced between blocks, within participants. Participants were informed of the order of the judgements before each block and were cued as to the first response immediately after the stimuli were presented, and then the second response was cued immediately after their first response. Cues were simply the text "left or right?" and "1 or 2?" for the identification and detection judgements respectively. No feedback was given. Each face was immediately cued with a fixation cross for 250 ms and immediately followed by 500 ms of noise, spatially filtered to match the spatial frequencies of the face image. The stimuli were presented for 300 ms each (Fig. 3).

Six different gaze deviations were tested (0.25, 0.5, 1, 2, 4, and 8 degrees) in pseudo-randomly mixed trials, by applying Matlab's randperm function to an ordered list of trials, for each participant independently. Participants were given unlimited time to respond and were not given any feedback. Each participant performed 192 trials per offset angle (96 in each direction), making 1,152 trials in total, over 2 sessions of approximately 1 hour each.

Analysis

Data analysis was conducted in two stages. First, the proportion correct responses to the detection and identification tasks were calculated for each observer at each gaze offset to compare whether the current pattern of performance across the two tasks mirrored previous results, with superior performance in the identification task compared to the detection task (Harris and Fahle, 1995; Rollman and Nachmias, 1972; Allik et al., 1982, 2014). This also can be quantified as a difference in performance thresholds, which are calculated as the gaze offset required to achieve 75% correct (where 50% correct represents floor performance, or what could be achieved by guessing alone). Second, the proportion of responses based on accuracy across both tasks (both responses correct p(D, I), both incorrect p(~D,~I), only identification correct p(~D, I), and only detection correct p(D, ~I)) was calculated for each observer, at each gaze offset. These data were then used to compare four hypotheses by modeling how these proportions of responses would be predicted to change with increasing evidence for offset gaze under each hypothesis.

To test these hypotheses, four models were defined to describe performance across the two judgements as evidence increases. All models assumed that evidence increased with increasing spatial offset of the pupils from the sclera in the form of a cumulative Gaussian function. All models therefore describe the proportion of responses expected with respect to accuracy across the two judgements by the rate of increasing evidence for each decision, the standard deviation, σ, of the cumulative Gaussian function. In all models, the mean of the cumulative Gaussian function was set to 0, such that at a spatial offset of 0 there is no evidence for offset gaze (increase in evidence was empirically symmetrical—increasing both left and right—but treated as unidimensional in the data, that is, gaze offset to the left and right were treated the same). The difference between the models is merely the way in which the evidence for each decision relates, which is formularized in Table 1.

Model H_{01} assumes the same evidence was used to make the detection and identification decisions. The model therefore fits a single parameter (σ_{01}) to describe the rate at which evidence increases, where the incorrect responses are divided equally between the three possibilities (both incorrect, detection incorrect, and identification incorrect) as they result from guessing alone. Model H_{A1} similarly assumes the same evidence is used for both judgements; however, each decision is based on different transformations of the evidence: $x + y > 0$ for identification and $x^2 - y^2 > 0$ for detection, where x and y refer to the evidence from each of the two stimuli presented on each trial, as explained in the introduction, and described by

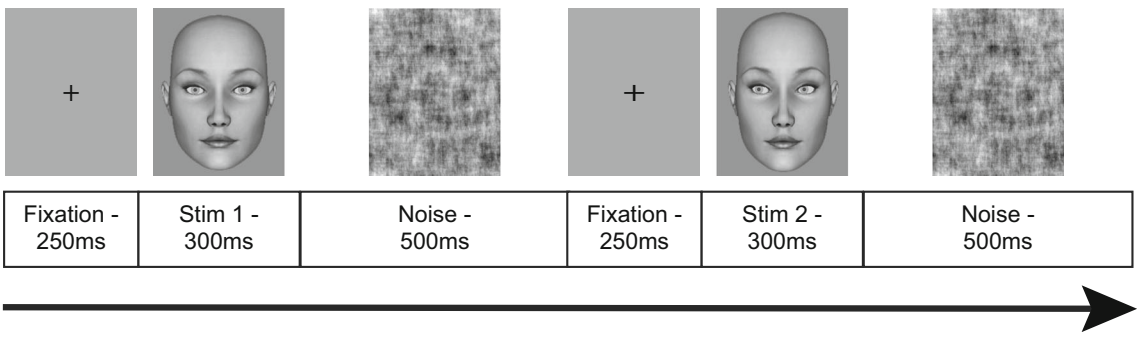

Fig. 3 Presentation and timing of stimuli on each trial

Table 1 Expected proportion of responses according to accuracy across both judgements for each model

H_{01}		Detection	
		Correct	Incorrect
Identification	Correct	$\frac{1}{2}(3p_{01} - 1)$	$\frac{1}{2}(1 - p_{01})$
	Incorrect	$\frac{1}{2}(1 - p_{01})$	$\frac{1}{2}(1 - p_{01})$
H_{A1}		Detection	
		Correct	Incorrect
Identification	Correct	p_{A1}^2	$p_{A1}(1 - p_{A1})$
	Incorrect	$(1 - p_{A1})^2$	$p_{A1}(1 - p_{A1})$
H_{A2}		Detection	
		Correct	Incorrect
Identification	Correct	$p_{A2a}.p_{A2b}$	$P_{A2a}(1 - p_{A2b})$
	Incorrect	$(1 - p_{A2a})(1 - p_{A2b})$	$P_{A2b}(1 - p_{A2a})$
H_{02}		Detection	
		Correct	Incorrect
Identification	Correct	$p_{02a}.p_{02b}$	$p_{02b}(1 - p_{02a})$
	Incorrect	$p_{02a}(1 - p_{02b})$	$(1 - p_{0b})(1 - p_{02a})$

p_{01}, p_{A1}, p_{A2a}, p_{A2b}, p_{02a}, and p_{02b} refer to the probability of making that decision, which increases with spatial offset of gaze as a cumulative Gaussian function. For H_{A1} and H_{A2}, these correspond to the schematics in Fig. 2a and b respectively

Klein (1985) and Allik et al. (2014). Model H_{A2}, also described by Klein (1985), requires two parameters describing the rate of increase of evidence with spatial offset, as it takes into account the possibility that there might be correlated noise between the decisions (due to drift in the participant's subjective direct gaze). This noise can be represented as stretching the two-dimensional Gaussian along the positive diagonal; thus, the cumulative Gaussians corresponding to increased evidence with spatial offset for each decision must have different standard deviations (σ_{A2a} and σ_{A2b}). Finally, model H_{02} tests the second null hypothesis that each decision is based on different evidence. The evidence for each decision increases with increasing gaze offset, but at different rates for each decision, as described by the two parameters σ_{02a} and σ_{02b}.

The equations in Table 1 describe the expected proportion of responses according to accuracy across both decisions as evidence for each decision increases. These expected proportions were compared to the actual proportions of responses each observer made across all gaze offsets tested. The models were fit to each individual observers' responses separately by modifying the parameters (using the Nelder-Mead simplex algorithm (Lagarias et al., 1998) implemented in custom Matlab software) until the sum of squared error (SSE) between the data and the model's predictions was minimized.

Results

The proportion of correct responses of each observer at each spatial offset for each decision are displayed in Fig. 4, and hit and false alarm rates are displayed in Table 2. Table 2 also displays the 75% correct thresholds for each task, which were calculated by fitting a Quick function to the data (Quick, 1974). Each observer demonstrates a lower threshold in the identification task compared with the detection task, indicating superior performance in the identification task compared with the detection task, as found in previous studies (Allik et al., 1982; 2014). This was confirmed with a 2x2x6 repeated measures ANOVA (question order x judgement x offset), which revealed a significant effect of offset ($F(1,5)$ =

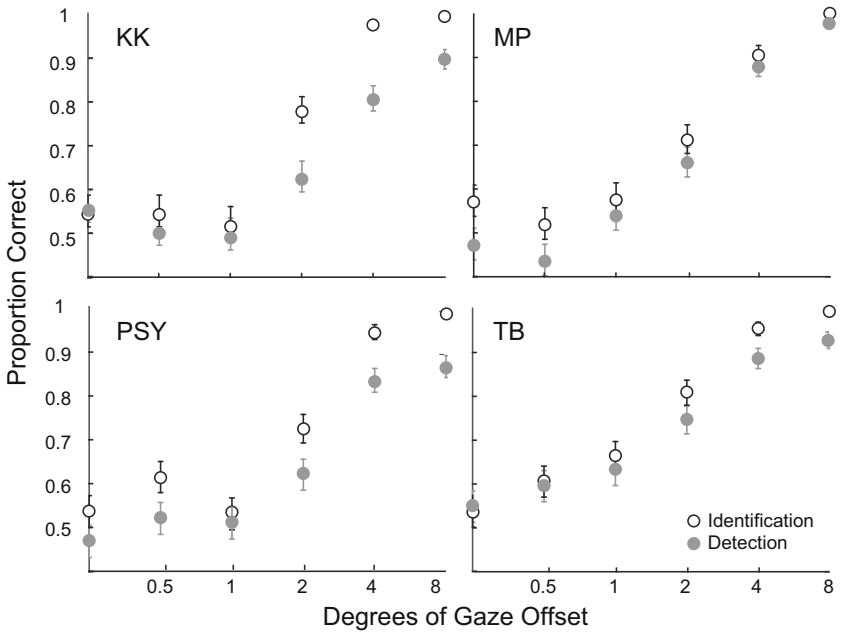

Fig. 4 Performance in the identification and detection judgements for each subject. Error bars represent the standard error assuming a binomial distribution

Table 2 Hit and false-alarm rates for each task for each observer and associated 75% correct thresholds

Observer	Judgement	Measure	Offset of gaze						75% correct threshold
			0.25	0.5	1	2	4	8	
KK	Detection	Hit Rate	0.71	0.61	0.65	0.76	0.85	0.91	3.75
		False Alarm Rate	0.58	0.59	0.65	0.50	0.24	0.11	
	Identification	Hit Rate	0.54	0.65	0.57	0.77	0.97	1.00	1.95
		False Alarm Rate	0.44	0.54	0.52	0.21	0.02	0.01	
MP	Detection	Hit Rate	0.66	0.56	0.73	0.79	0.94	0.97	2.85
		False Alarm Rate	0.71	0.69	0.65	0.47	0.18	0.01	
	Identification	Hit Rate	0.33	0.23	0.34	0.54	0.81	1.00	2.35
		False Alarm Rate	0.19	0.19	0.19	0.11	0.00	0.00	
PSY	Detection	Hit Rate	0.47	0.49	0.48	0.65	0.85	0.86	3.85
		False Alarm Rate	0.53	0.45	0.46	0.41	0.19	0.14	
	Identification	Hit Rate	0.69	0.76	0.66	0.82	0.98	0.97	2.25
		False Alarm Rate	0.61	0.53	0.59	0.38	0.09	0.00	
TB	Detection	Hit Rate	0.56	0.72	0.67	0.83	0.86	0.89	2.15
		False Alarm Rate	0.47	0.53	0.41	0.34	0.09	0.03	
	Identification	Hit Rate	0.48	0.59	0.57	0.81	0.96	1.00	1.15
		False Alarm Rate	0.41	0.39	0.25	0.20	0.05	0.01	

134.84, $p < 0.001$, $\eta_p^2 = 0.978$) and a significant effect of judgement ($F(1,1) = 31.949$, $p = 0.011$, $\eta_p^2 = 0.914$), but no significant interaction between the two ($F(1,5) = 1.743$, $p = 0.186$, $\eta_p^2 = 0.367$). There was no significant difference in performance based on question order ($F(1,1) = 0.923$, $p = 0.408$, $\eta_p^2 = 0.235$) and no interaction between question order and offset ($F(1,5) = 1.763$, $p = 0.181$, $\eta_p^2 = 0.37$).

In the detection task, the hit rate is taken as the proportion of trials in which the stimulus with offset gaze is presented in the first interval and the observer responds that it is in the first interval, whereas the false-alarm rate describes the proportion of trials in which the stimulus with offset gaze is presented in the second interval, but the observer responds that it is in the first interval. In the identification task, the hit rate describes the proportion of those trials where a stimulus with leftward gaze is presented on which the observer responds that gaze is directed to the left, whilst the false alarm rate describes the proportion of those trials where a stimulus with rightward gaze is presented but on which the observer responds that gaze is directed to the left. The proportion of correct responses is calculated as (HR+(1-FAR))/2, where HR is the hit rate and FAR is the false-alarm rate (Macmillan and Creelman, 2004). The threshold is the gaze offset at which it is predicted the observer will score 75% correct, which is estimated by fitting a Quick function to the data (Quick, 1974).

The proportion of responses corresponding to accuracy across both judgements is shown in Fig. 5 for each participant. The lines correspond to the predictions based on each model for each participant. The best fitting parameter values are displayed in Table 3, and the individual SSEs are displayed in Table 4. The proportion of variance explained by each model was calculated for each observer in order to compare the models. Models H_{01} and H_{02} explained on average less variance than H_{A1} and H_{A2} (90.32% and 86.13% vs. 92.05% and 93.52% respectively). These null models therefore do not explain the data as well as the alternative models.

That model H_{A2} explains on average more variance than model H_{A1} has previously been taken as evidence for preferring model H_{A2} (Klein, 1985). However, model H_{A1} actually represents a nested model of H_{A2}, because (as shown from the equations in Table 1) H_{A2} is equivalent to H_{A1} in the case where parameter p_{A2a} is equal to parameter p_{A2b}. H_{A2} is therefore expected to describe the data at least as well as H_{A1}, but to accept it as the better model it must be tested as to whether it explains enough variance to justify the extra parameter using an F-test (Dobson, 1990):

$$F = \frac{(S_0 - S_1)/1}{S_1/(K-L)}$$

Where S_1 and S_0 represent the sum of squared error for H_{A2} and H_{A1}, respectively, K is the number of independent data points fitted and L is the number of parameters of H_{A2}, and the degrees of freedom is equal to K - L. Group data did not provide evidence for preferring the more complex model ($F(1,10) = 2.26$, $p > 0.05$). When testing individual participants' data, only participant KK's data showed significant evidence for preferring the more complex model ($F(1,16) = 7.80$, $p < 0.05$). Individual F values also are displayed in Table 4.

The names of the parameters refer to the calculations. The values in each case describe the standard deviation of a cumulative Gaussian function with mean 0. Applying the equations in Table 1 to the cumulative Gaussians described by these parameters gives the lines in Fig. 5.

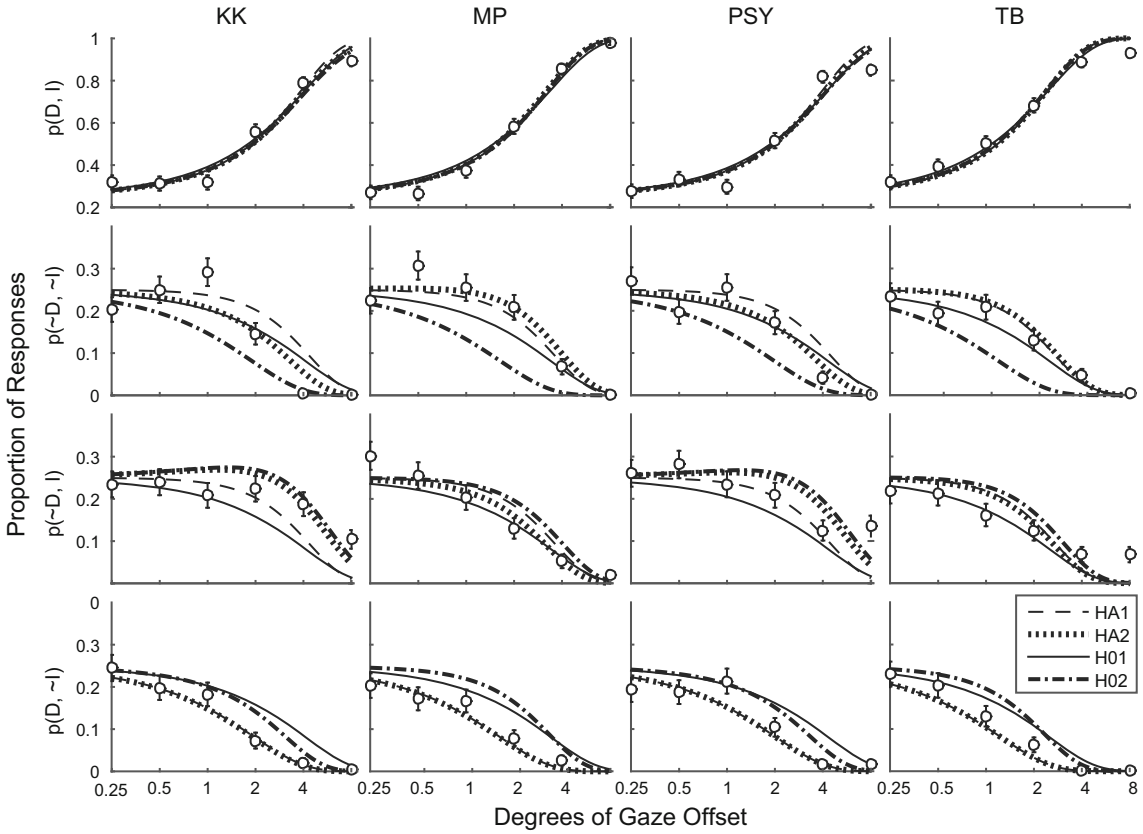

Fig. 5 Proportion of responses predicted by each model according to accuracy across both judgements. p(D,I) refers to the proportion of responses where both decisions are correct, p(~D, ~I) refers to both decisions incorrect, p(~D, I) refers to only identification correct, and p(D, ~I) refers to detection only correct. Lines show the fit of each model compared to empirical data (points). Error bars represent the standard error of the mean assuming a binomial distribution

Discussion

The results mirrored the previous findings of Allik et al. (1982, 2014) with superior performance in identification judgements than detection judgements across all observers. Thresholds measured for these stimuli were far lower than those measured for the Bull's-eye stimuli used by Allik et al. (1982, 2014): Allik et al. 2014 found thresholds between approximately 7.5 and 20 min of arc, whereas thresholds measured here were approximately between 1 and 2 min of arc (calculated from the thresholds presented in Table 2). This may be a result of expertise for judging the direction of gaze (thresholds for judging gaze direction have been found to be as low as 0.6 and 0.71 min; Jenkins and Langon, 2003; Cline 1967) but also is likely related to the fact that gaze stimuli have what is effectively a smaller annulus diameter, which has been shown by Legge and Campbell (1981) to produce lower thresholds (also down to less than 1 min of arc). This evidence suggests that this pattern of performance across identification and detection judgements is not a quirk of low-level vision but exists in complex and ecologically valid stimuli. In some sense, this is a surprising result; given the importance of

Table 3 Best fitting parameters for each model for each observer

Model	H_{01}	H_{A1}	H_{A2}		H_{02}	
Parameter	σ_{01}	σ_{A1}	σ_{A2a}	σ_{A2b}	σ_{02a}	σ_{02b}
KK	4.17	3.46	2.68	4.74	2.55	5.14
MP	3.45	2.87	3.18	2.56	2.63	3.21
PSY	4.37	3.56	2.92	4.61	2.72	5.06
TB	2.53	2.17	2.23	2.11	1.86	2.43

Table 4 The sum of squared error of the model predictions to each observer's data

	H_{01}	H_{02}	H_{A1}	H_{A2}	F
KK	0.066	0.059	0.052	0.035	7.801
MP	0.041	0.091	0.020	0.017	2.767
PSY	0.068	0.070	0.058	0.049	3.038
TB	0.021	0.061	0.030	0.029	0.074
Total	0.195	0.280	0.160	0.131	2.265

SSE is the sum of the squared distance of each point in Fig. 5 to the line that describes the model prediction. The F statistic compares H_{A1} and H_{A2} where larger values indicate greater evidence for rejecting model H_{A1}, and the total F statistic is performed on group data.

gaze perception in social interactions and the recruitment of dedicated neural mechanisms, why might perception of gaze offset not be optimized for detection over and above identification?

The simple Thurstonian model employed by Allik et al. (2014) was a better fit to the proportions of correct responses across both judgements than the hypothesis that assumed that identification performance was in fact equal to detection performance. This model explained more variance in the data than the two null models, even though H_{02} contained an extra parameter (of note, it is not expected that H_{02} explain more variance than H_{01} as these models are independent). The two parameter model proposed by Klein (1985) explained slightly more variance in the data than the simpler Thurstonian model. However, this model would inevitably explain more variance than the simpler model, given that H_{A1} is a special case of H_{A2} where $p_{A2a} = p_{A2b}$, the additional parameter in H_{A2} means that this model must explain at least as much variance as H_{A1}. Our data suggest that the more complex model is not a significantly better fit. Observers' responses were not, therefore, significantly impacted by correlated noise between the two intervals. Previous studies have found correlated noise only when identifying very similar stimuli (Hirsch et al., 1982, Thomas, 1985), so perhaps the extra variance explained by the model occurred at small spatial offsets, whilst the simpler model gave a more parsimonious explanation of the data over the entire performance function. This is not necessarily to say that direct gaze forms a natural zero point. In fact, observers typically tend to judge slightly leftward gaze as direct (Calder et al. 2008). It does suggest that, possibly through expertise for these stimuli, observers' perception of direct gaze is quite stable, although this may vary between individuals.

The simple Thurstonian model explains that the computational complexity of the detection judgement means it requires more evidence than the identification judgement, as evidence for the detection judgement is more likely to be closer to a decision axis. Superior performance in the identification judgement, therefore, is less surprising. Observers were simply better at making the simpler judgement, and their performance was reliant on how certain they could be given the distance of the internal evidence from their decision axes.

Despite this seemingly simple explanation, this pattern of performance is still intuitively perplexing. It suggests that on certain trials observers were able to make an accurate identification of the direction of gaze offset, without being able to judge which face was actually looking in that direction. If detection of gaze offset were the criterion for awareness of the direction of gaze, this would suggest that the identification judgement could be made in the absence of awareness, based on evidence not available to consciousness. Conceptually, the detection threshold is a valid measure of awareness, and many authors have used the detection threshold just so (Marcel, 1980,1983; McCauley et al. 1980; Carr et al., 1982), especially in a two-interval, two-alternative, forced-choice design (as we used). However, there is some evidence to suggest that detection judgements can be made in the absence of awareness, for instance in the case of blindsight (Weizkrants et al., 1974, 1995; Azzopardi and Cowey, 1998), suggesting accuracy in detection does not equate to awareness of a stimulus. Furthermore, chance performance in the detection judgement in this task does not mean participants could not detect the stimulus as a whole but rather that they were at chance in detecting a property of that stimulus – direct gaze. Thus, the assumption that performance in the detection judgement marks conscious awareness of that property of the stimulus needs to be tested.

One way to test this assumption is to use metacognitive ratings (Timmermans and Cleeremans, 2015; Barrett et al., 2013; but see also Jachs et al., 2015). If observers were making accurate guesses based on evidence inaccessible to consciousness, then they would not be able to discriminate their correct responses from their incorrect responses. Metacognitive ratings ask observers to rate how certain they are that their response is correct, thereby giving a measure of how well they are able to discriminate their correct and incorrect responses. However, these ratings are highly subject to bias; observers could be extremely liberal and rate that they are certain that many of their responses are correct based on internal evidence that more conservative observers may rate as far less certain. A Signal Detection Theory (SDT, Green and Swets, 1966) approach can be used to tackle this issue. Under SDT, d' gives a measure of an observer's sensitivity to a stimulus by measuring the distance between the means of the distributions of internal evidence for the presence of that stimulus and the distribution of noise, or internal evidence for a comparison stimulus. This measure is impervious to how an observer might choose to treat the evidence or where they place their criterion for responding one way or the other. Maniscalco and Lau (2012, 2014) have extended this approach to metacognitive ratings, producing a measure, meta-d', of sensitivity to one's own correct and incorrect responses, irrespective of any bias in the criteria for confidence ratings. Meta-d' is calculated in a similar manner to d', with the exception that a 'hit' is taken as a high confidence rating on a correct response, and a false alarm is taken as a high confidence rating on an incorrect response. Since meta-d' is a dimensionless measure in the same way as d', sensitivity to the stimulus and sensitivity to one's own correct and incorrect responses can be directly compared.

To test whether observers were aware of the direction of gaze when they were unable to accurately detect which face was looking in that direction, we modified the traditional double judgements procedure to include metacognitive ratings. In Experiment 2, we asked participants to return to complete more trials with this modified procedure.

Experiment 2 Methods

Participants, apparatus, and stimuli were the same as in Experiment 1.

Procedure

Three spatial offsets were chosen based on participants' previous performance. The smallest corresponded to approximately 75% correct performance in the identification judgement (corresponding to the greatest hypothesized difference in performance between the two judgements). The second corresponded to approximately 75% correct performance in the detection judgement to match the performance of the smaller offset in the identification judgement, and finally, a large offset of 10 degrees to produce ceiling performance in both detection and identification judgements. These thresholds values were chosen as they are frequently used as thresholds for awareness. Participants were asked to give their responses to both judgements in the form of a pseudo-type-2 "metacognitive" rating (Galvin et al. 2003). Table 5 shows the new responses. Different keys were used for the identification and detection judgements (which were performed as a 2x2AFC as before), which were labelled with stickers. Participants were instructed to place their index, middle, and ring fingers of each hand on the keys corresponding to the first response at the beginning of each trial. The keys D – K were used for the detection judgement and X – M for the identification judgement, so that participants could easily move their hands up or down to make the second response.

As in experiment 1, the order of responses was counterbalanced between blocks, within participants. Participants performed 192 trials per offset condition, giving 576 trials completed in a single session of approximately 1 hour.

Results

Responses to the large gaze offset of 10 degrees were not included in the analysis as ceiling performance was expected; these trials were only included so that the experiment was not too difficult, given the other offsets were predicted to elicit near-threshold performance. Performance in trials containing the other two offsets was first compared against performance in the first experiment to ensure the extra cognitive load of the task did not cause performance to suffer. Hit and false-alarm rates are shown in Table 6. In each case, participants performed at a similar level to what was predicted by data in the previous experiment, with an average of a 1% increase (with a standard deviation of 4%) from the 75% correct performance predicted. The models of experiment 1 were not fit to the data, because there were not enough gaze offsets tested to constrain the model.

Responses were then sorted into contingency tables based on the rating/response, henceforth only the two smaller spatial offsets, corresponding to approximately 75% correct performance in identification and detection judgements, respectively, are reported. These data were used to calculate d' and meta-d' using the HMM toolbox (Hierarchical meta-d' model, created by Dr Steve Fleming, UCL). The toolbox uses a hierarchical Bayesian framework to fit Maniscalco and Lau's (2012, 2014) meta-d' model to the data, utilising MCMC (Markov Chain Monte Carlo) inference on arbitrary Bayesian models implemented in MATLAB and JAGS (Just Another Gibbs Sampler). Meta-d' provides a SDT approach to estimating metacognitive sensitivity based on how well participants are able to distinguish between their correct and incorrect responses. The ratio of meta-d' to d' measures the degree to which the participant is able to make metacognitive use of the information used in their type-1 decision. Meta-d' ratio was roughly equal across the two tasks, with a significant correlation of 0.91 ($p < 0.01$; Fig. 6).

Discussion

Observers' ability to discriminate their correct from incorrect responses across both tasks was measured by asking observers to rate how certain they were that they were correct on each trial. These data were used to calculate meta-d', an index of the observers' metacognitive sensitivity. It was hypothesized that if observers were more aware of the perceptual evidence used to make the identification judgement than the detection judgement, especially when detection performance was below threshold, then the ratio of meta-d' to d' would be greater in the identification task than the detection task.

However, inspection of Fig. 6 indicates that the ratio of meta-d' to d' across judgements does not deviate systematically from equality, meaning that when sensitivity was taken into

Table 5 Correspondence between the sign of the identification and detection responses, the certainty of the response and the rating form it is given in

Rating	*1*	*2*	*3*	*4*	*5*	*6*
Certainty	Highly certain I am correct	Moderately certain I am correct	Low certainty/I don't know if I am correct	Low certainty/I don't know if I am correct	Moderately certain I am correct	Highly certain I am correct
Identification	Left	Left	Left	Right	Right	Right
Detection	1st Interval	1st Interval	1st Interval	2nd Interval	2nd Interval	2nd Interval

Table 6 Hit and false-alarm rates for each participant in each task for the two gaze offsets measured

Participant	Judgement	Measure	Gaze offset no.	
			1	2
KK	Detection	Hit rate	0.69	0.79
		False-alarm rate	0.48	0.18
	Identification	Hit rate	0.75	0.94
		False-alarm rate	0.38	0.15
MP	Detection	Hit rate	0.82	0.83
		False-alarm rate	0.36	0.28
	Identification	Hit rate	0.58	0.70
		False-alarm rate	0.03	0.01
PSY	Detection	Hit rate	0.79	0.85
		False-alarm rate	0.42	0.24
	Identification	Hit rate	0.85	0.93
		False-alarm rate	0.32	0.03
TB	Detection	Hit rate	0.58	0.77
		False-alarm rate	0.35	0.27
	Identification	Hit rate	0.76	0.85
		False-alarm rate	0.28	0.27

Hit and false-alarm rates are measured as described in Table 2. The 'gaze offset no.' refers to the smaller gaze offset (75% correct identification threshold) and larger gaze offset (75% correct detection threshold) respectively, which are shown for each participant in Table 2

consideration, awareness was roughly equal. There was nothing particularly special about the identification judgement that meant that observers were more sensitive in discriminating their correct and incorrect judgements. It is therefore worthwhile considering how participants were making their ratings based on performance across both judgements.

Figure 7 shows the average ratings given by each participant according to their performance across both judgements, collapsed across the two spatial offsets. The pattern of ratings tends to follow the theoretical distance of the evidence from decision criteria. The most confident ratings are given in the case where both responses are correct (and where the evidence can be furthest from both criteria, as can be seen in Fig. 2). Perhaps counterintuitively, more confident ratings are given when both responses are incorrect compared with when the identification response is incorrect and the detection response is correct. If the metacognitive ratings were based solely on the strength of the perceptual evidence, then one would expect that observers would report roughly equal metacognitive confidence when the identification response is incorrect irrespective of detection performance, but this is not the case. Figure 2 shows that the evidence when both decisions are incorrect tends to be further from the decision criteria than when only the identification decision is incorrect. It thus appears likely that participants were using the distance of the evidence from their decision criterion as a factor in making their metacognitive ratings. Furthermore, the pattern is the same for identification and detection ratings, indicating that both ratings were made on the bases of the distance of the evidence from both decision axes, despite the fact the identification decision could be made on the basis of just one decision axis (the negative diagonal of Fig. 1). This may be because observers' confidence ratings were not completely independent;

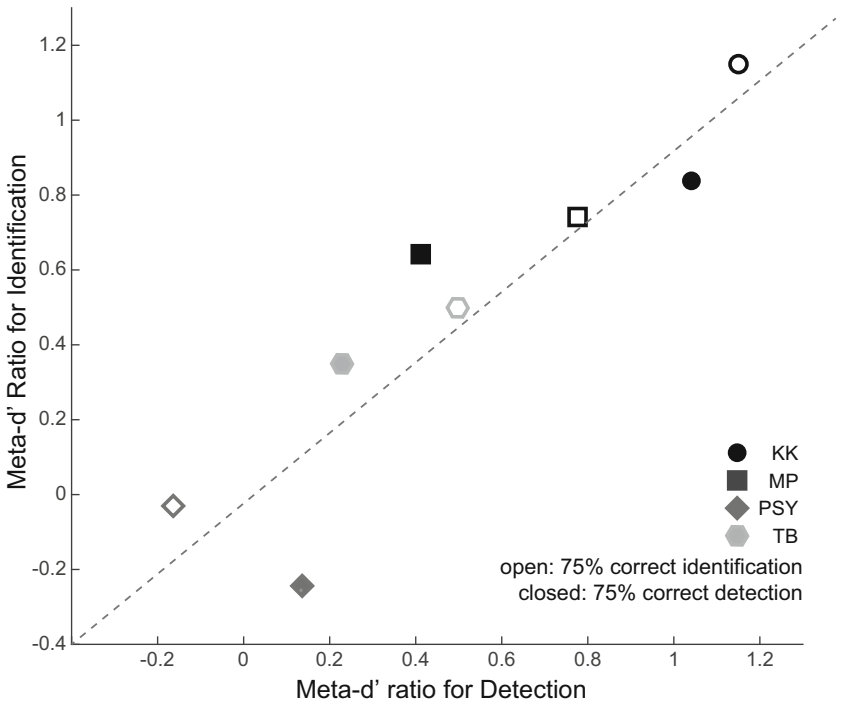

Fig. 6 Ratio of meta-d' to d' for identification compared with detection. The line shows where equal ratios would sit. Actual correlation is 0.91 ($p < 0.01$)

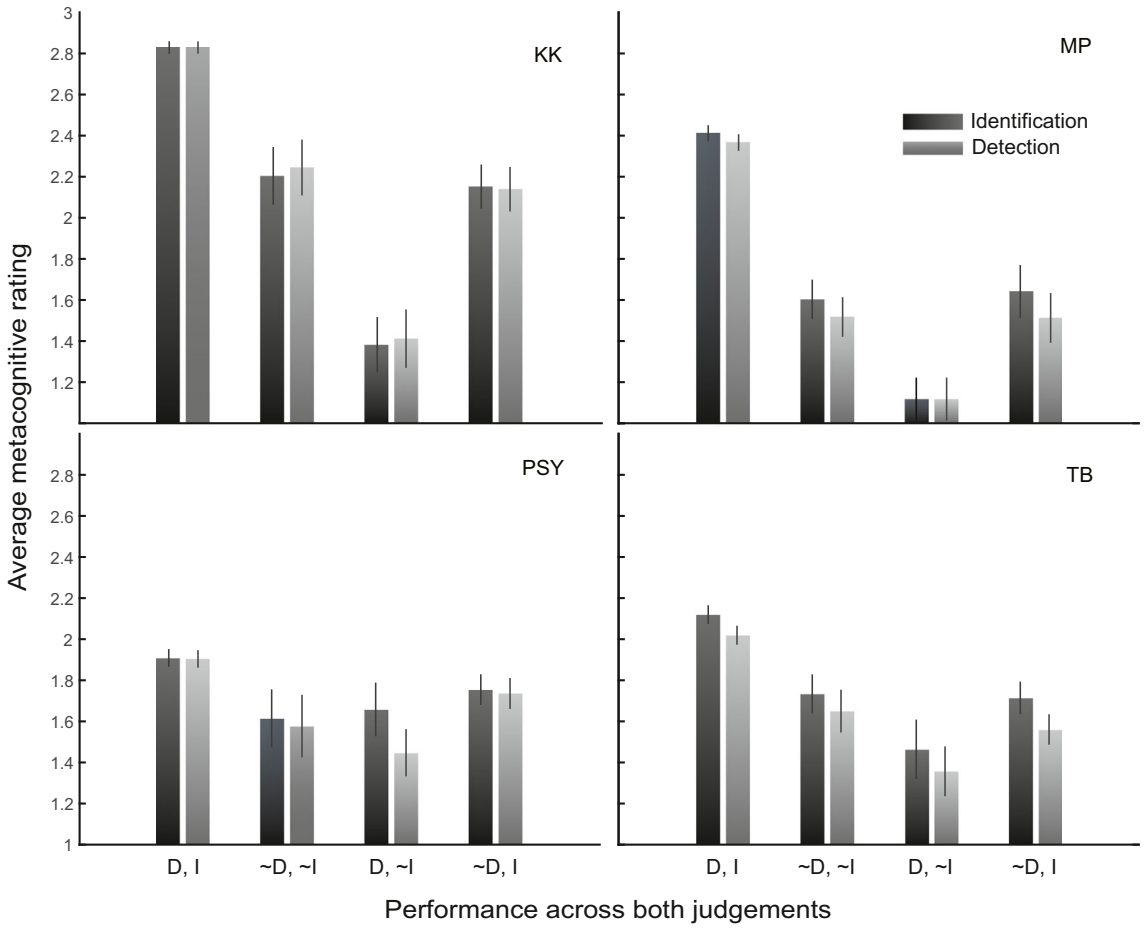

Fig. 7 Average ratings (across *both* spatial offsets) for identification and detection judgements according to performance across both judgements. Ratings have been converted so that a rating of 3 corresponds to high certainty, 2 to moderate certainty, and 1 to low certainty. D, I refers to both decisions correct, ~D, ~I refers to both decisions incorrect, ~D, I refers to only identification correct, and D, ~I refers to detection only correct

they reported less confidence in one decision when they were less confident in the other.

Conclusions

In these experiments, we observed superior performance in identifying the direction of gaze offset than detecting offset gaze. Of the four models examined, a simple Thurstonian model (H_{A1}) proved the most parsimonious explanation of the empirical data. Under this model, the detection judgement is actually more computationally complex than the identification judgement, since the observer needs to account for both $x - y > 0$ and $x + y > 0$ (where x and y denote the signed evidence of spatial offset in each interval). The decision space depicted in Fig. 1 also reveals that evidence for the detection judgement will always be of equal or lesser distance from the nearest decision axis than in the identification judgement, making the evidence for this judgement more uncertain.

By incorporating metacognitive ratings into the double judgements procedure, we were able to analyse not only how participants were using the evidence to make their different judgements but also how this evidence might be affecting their sensory experience of the stimuli. Observers' ability to discriminate their own correct and incorrect responses was roughly equal when their sensitivity to the stimuli was taken into account. Furthermore, the average ratings given based on accuracy across both judgements mirrored the theoretical distance of the evidence from the decision axes based on the simple Thurstonian model. This indicates that metacognitive awareness may be reliant not only on the strength of the internal evidence, but also on the relative distance of this evidence from decision criteria. At least in this case, performance and metacognitive ratings were closely related.

That this pattern of performance was shown for judgements about the direction of gaze is interesting for two reasons. First, it means that a performance advantage for identification over detection of spatial offset is not a quirk of low-level vision, as the advantage is evident even when many properties of a stimulus are bound together, as in determining another's gaze direction. However, it is worthwhile questioning whether in this case observers were only using spatial offset information—

perhaps if they were forced to incorporate other evidence, for example, in making the same judgements but with rotated heads, the results may turn out differently. Furthermore, evidence used in perceiving gaze direction may affect identification and detection judgements separately. For instance, observers have been shown to accept greater gaze offsets as direct when the face stimulus has an angry expression (Ewbank et al., 2009), and this may affect the detection judgement more than the identification judgement. Alternatively, the identification judgement might be affected more than the detection judgement by the presence of contextual objects, which have been shown to bias the perceived direction of gaze (Lobmaier et al., 2006).

The second reason why superior performance on the identification task is especially interesting is that, given the importance of detecting gaze direction as a social cue, one could argue that detection should be optimized over identification performance. Perhaps detection of gaze offset is not so important in the temporal domain—given that our stimuli were presented in quick temporal succession, a scenario that is unlikely to occur often in the natural world. To investigate this issue, one could run experiments analogous to those presented but using a spatial 2AFC, presenting the two faces side by side. Of note, Allik et al (2014) presented their annuli side by side and replicated their 1982 results in which the annuli were presented temporally. Another possibility is that accuracy in the identification judgement is actually more important than accuracy in the detection task. Evidence suggest that observers can perceive a large "cone of direct gaze" (the extent to which increasingly indirect gaze is accepted as direct), which becomes larger with social anxiety (Gamer et al., 2011; Jun et al., 2013) and with threatening emotions (Ewbank et al., 2009). It has been suggested that we may have a prior expectation for direct gaze (Mareschal, Calder and Clifford, 2013; Mareschal, Otsuka, & Clifford, 2014) that evolved to minimize missing direct gaze, for example, by interpreting direct gaze as slightly offset, which incurs greater cost than the false alarm of perceiving slightly offset gaze as direct. In contrast, accurate identification of gaze direction is important for attentional cueing by gaze stimuli, which has been suggested to be implicit and resistant to top-down suppression (Friesen and Kingstone, 1998), as well as for accurately understanding what someone else is looking at, where both of these cues incur a similar cost for any mistake (misses and false alarms).

The double judgements procedure proved a powerful method for examining the relationship between performances across two different judgements. By analysing accuracy across both judgements simultaneously, we were able to show that differences in performance could be the result of the same evidence being used in different ways. Furthermore, the procedure was adapted to examine how the relative strength of evidence across the two judgements might differentially affect observers' perceptual awareness of offset gaze. These results show that both performance and metacognitive awareness rely on more than just the strength of sensory evidence, but also the computational complexity of the decision, which determines the relative distance of that evidence from decision axes.

Acknowledgements This work was supported by a grant from the Australian Research Council Discovery Project (DP160102239). CC is supported by Australian Research Council Future Fellowship (FT110100150). The authors thank Matthew Patten for his advice on the graphical representations in this manuscript and Kiley Seymour for her helpful comments on the manuscript.

References

Allik, J., Dzhafarov, E., & Rauk, M. (1982). Position discrimination may be better than detection. *Vision Research, 22*(8), 1079–1081.

Allik, J., Toom, M., & Rauk, M. (2014). Detection and identification of spatial offset: Double-judgment psychophysics revisited. *Attention, Perception, & Psychophysics, 76*(8), 2575–2583.

Anstis, S. M., Mayhew, J. W., & Morley, T. (1969). The perception of where a face or television'portrait' is looking. *The American Journal of Psychology*, 474-489.

Azzopardi, P., & Cowey, A. (1998). Blindsight and visual awareness. *Consciousness and Cognition, 7*(3), 292–311.

Baron-Cohen, S. (1992). Out of sight or out of mind? Another look at deception in autism. *Journal of Child Psychology and Psychiatry, 33*(7), 1141–1155.

Barrett, A. B., Dienes, Z., & Seth, A. K. (2013). Measures of metacognition on signal-detection theoretic models. *Psychological Methods, 18*(4), 535.

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

Calder, A. J., Beaver, J. D., Winston, J. S., Dolan, R. J., Jenkins, R., Eger, E., & Henson, R. N. (2007). Separate coding of different gaze directions in the superior temporal sulcus and inferior parietal lobule. *Current Biology, 17*(1), 20–25.

Calder, A. J., Jenkins, R., Cassel, A., & Clifford, C. W. (2008). Visual representation of eye gaze is coded by a nonopponent multichannel system. *Journal of Experimental Psychology: General, 137*(2), 244.

Carlin, J. D., Calder, A. J., Kriegeskorte, N., Nili, H., & Rowe, J. B. (2011). A head view-invariant representation of gaze direction in anterior superior temporal sulcus. *Current Biology, 21*(21), 1817–1821.

Carr, T. H., McCauley, C., Sperber, R. D., & Parmelee, C. (1982). Words, pictures, and priming: On semantic activation, conscious identification, and the automaticity of information processing. *Journal of Experimental Psychology: Human Perception and Performance, 8*(6), 757.

Cline, M. G. (1967). The perception of where a person is looking. *The American Journal of Psychology*, 41-50.

Dobson, A. J. (1990). *An introduction to generalized linear models*. London: Chapman and Hall.

Ewbank, M. P., Jennings, C., & Calder, A. J. (2009). Why are you angry with me? Facial expressions of threat influence perception of gaze direction. *Journal of Vision, 9*(12), 16.

Freire, A., Eskritt, M., & Lee, K. (2004). Are eyes windows to a deceiver's soul? Children's use of another's eye gaze cues in a deceptive situation. *Developmental Psychology, 40*(6), 1093.

Friesen, C. K., & Kingstone, A. (1998). The eyes have it! Reflexive orienting is triggered by nonpredictive gaze. *Psychonomic Bulletin & Review, 5*(3), 490–495.

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), 843–876.

Gamer, M., Hecht, H., Seipp, N., & Hiller, W. (2011). Who is looking at me? The cone of gaze widens in social phobia. *Cognition and Emotion, 25*(4), 756–764.

Green, D., & Swets, J. (1966). *Signal detection theory and psychophysics*. New York.

Harris, J., & Fahle, M. (1995). The detection and discrimination of spatial offsets. *Vision Research, 35*(1), 51–58.

Hirsch, J., Hylton, R., & Graham, N. (1982). Simultaneous recognition of two spatial-frequency components. *Vision Research, 22*(3), 365–375.

Huang, P.-C., Kingdom, F., & Hess, R. (2006). Only two phase mechanisms, ±cosine, in human vision. *Vision Research, 46*(13), 2069–2081.

Jachs, B., Blanco, M. J., Grantham-Hill, S., & Soto, D. (2015). On the independence of visual awareness and metacognition: A signal detection theoretic analysis. *Journal of Experimental Psychology: Human Perception and Performance, 41*(2), 269.

Jenkins, J., & Langton, S. (2003). Configural processing in the perception of eye-gaze direction. *Perception, 32*(10), 1181–1188.

Jun, Y. Y., Mareschal, I., Clifford, C. W., & Dadds, M. R. (2013). Cone of direct gaze as a marker of social anxiety in males. *Psychiatry Research, 210*(1), 193–198.

Klein, S. A. (1985). Double-judgment psychophysics: Problems and solutions. *Journal of the Optical Society of America. A, 2*(9), 1560–1585.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3. *Perception, 36*(14), 1.

Lagarias, J. C., Reeds, J. A., Wright, M. H., & Wright, P. E. (1998). Convergence properties of the Nelder–Mead simplex method in low dimensions. *SIAM Journal on optimization, 9*(1), 112–147.

Legge, G. E., & Campbell, F. W. (1981). Displacement detection in human vision. *Vision Research, 21*(2), 205–213.

Lobmaier, J. S., Fischer, M. H., & Schwaninger, A. (2006). Objects capture perceived gaze direction. *Experimental Psychology, 53*(2), 117–122.

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), 422–430.

Maniscalco, B., & Lau, H. (2014). Signal Detection Theory analysis of type 1 and type 2 data: Meta-d′, response-specific Meta-d′, and the unequal variance SDT model. *The Cognitive Neuroscience of Metacognition* (pp. 25-66): Springer Berlin Heidelberg.

Marcel, A. J. (1980). Conscious and preconscious recognition of polysemous words: Locating the selective effects of prior verbal context. *Attention and Performance VIII*, 435-457.

Marcel, A. J. (1983). Conscious and unconscious perception: An approach to the relations between phenomenal experience and perceptual processes. *Cognitive Psychology, 15*(2), 238–300.

Mareschal, I., Calder, A. J., & Clifford, C. W. (2013). Humans have an expectation that gaze is directed toward them. *Current Biology, 23*(8), 717–721.

Mareschal, I., Otsuka, Y. & Clifford, C. W. G. (2014). A generalized tendency towards direct gaze with uncertainty. *Journal of Vision, 14*(12). doi:10.1167/14.12.27.

McCauley, C., Parmelee, C., Sperber, R. D., & Carr, T. H. (1980). Early extraction of meaning from pictures and its relation to conscious identification. *Journal of Experimental Psychology: Human Perception and Performance, 6*(2), 265.

Otsuka, Y., Mareschal, I., Calder, A. J., & Clifford, C. W. (2014). Dual-route model of the effect of head orientation on perceived gaze direction. *Journal of Experimental Psychology: Human Perception and Performance, 40*(4), 1425.

Otsuka, Y., Mareschal, I., & Clifford, C. W. (2015). Gaze constancy in upright and inverted faces. *Journal of Vision, 15*(1), 21.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*(4), 437–442.

Quick, R. F. (1974). A vector-magnitude model of contrast detection. *Biological Cybernetics, 16*(2), 65–67.

Ricciardelli, P., Baylis, G., & Driver, J. (2000). The positive and negative of human expertise in gaze perception. *Cognition, 77*(1), B1–B14.

Rollman, G. B., & Nachmias, J. (1972). Simultaneous detection and recognition of chromatic flashes. *Perception & Psychophysics, 12*(3), 309–314.

Sinha, P. (2000). Last but not least. *Perception, 29*(8), 1005–1008.

Thomas, J. P. (1985). Detection and identification: How are they related? *Journal of the Optical Society of America. A, 2*(9), 1457–1467.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*(4), 273.

Timmermans, B., & Cleeremans, A. (2015). How can we measure awareness? An overview of current methods. *Behavioural Methods in Consciousness Research*, 21-46.

Tolhurst, D., & Dealy, R. (1975). The detection and identification of lines and edges. *Vision Research, 15*(12), 1367–1372.

Weiskrantz, L., Barbur, J., & Sahraie, A. (1995). Parameters affecting conscious versus unconscious visual discrimination without V1. *Proceedings of the National Academy of Sciences, 92*(13), 6122–6126.

Weiskrantz, L., Warrington, E. K., Sanders, M., & Marshall, J. (1974). Visual capacity in the hemianopic field following a restricted occipital ablation. *Brain, 97*(1), 709–728.

Wollaston, W. H. (1824). On the apparent direction of eyes in a portrait. *Philosophical Transactions of the Royal Society of London, 114,* 247–256.