Consciousness and Cognition 38 (2015) 107–117

Contents lists available at ScienceDirect

Consciousness and Cognition

journal homepage: www.elsevier.com/locate/concog

Perceptual confidence demonstrates trial-by-trial insight into the precision of audio–visual timing encoding

Brendan Keane [a,*], Morgan Spence [a], Kielan Yarrow [b], Derek Arnold [a]

[a] School of Psychology, University of Queensland, Australia
[b] Department of Psychology, City University London, United Kingdom

ARTICLE INFO

Article history:
Received 23 July 2015
Revised 28 October 2015
Accepted 31 October 2015

Keywords:
Confidence
Metacognition
Temporal
Perception
Insight

ABSTRACT

Peoples' subjective feelings of confidence typically correlate positively with objective measures of task performance, even when no performance feedback is provided. This relationship has seldom been investigated in the field of human time perception. Here we find a positive relationship between the precision of human timing perception and decisional confidence. We first demonstrate that subjective audio–visual timing judgements are more precise when people report a high, as opposed to a low, level of confidence. We then find that this relationship is more likely to result from variance in sensory timing estimates than the application of variable decision criteria, as the relationship held when we adopted a measure of timing sensitivity designed to limit the influence of subjective criteria. Our results suggest analyses of timing perception and associated decisional confidence reflect the trial-by-trial variability with which timing has been encoded.

© 2015 Elsevier Inc. All rights reserved.

1. Introduction

It has repeatedly been shown that humans can successfully report when their perceptual judgements have been accurate, even in the absence of explicit feedback regarding task performance (for reviews see Fleming, Dolan, & Frith, 2012; Yeung & Summerfield, 2012). This insight has been demonstrated in a number of contexts, including the differentiation of motion direction (Zylberberg, Barttfeld, & Sigman, 2012), spatial frequency, orientation (de Gardelle & Mamassian, 2014), and when judging luminance-contrast (Song et al., 2011). This suggests that, in each case, humans have access to an accurate reportable estimate concerning the strength of evidence underlying their perceptual decisions. Therefore, confidence has been classified as a form of metacognition (Fleming et al., 2012).

Despite its demonstration in many types of perceptual judgements, metacognitive insight into human time perception has seldom been investigated. One study, however, has provided suggestive evidence. Allan (1975) had participants make audio–visual temporal order judgements, followed by a confidence categorisation (high or low) concerning their timing judgement. Visual appraisal of distributions of high-confidence order judgements showed a discrepancy relative to overall distributions (which comprised both high and low confidence order judgements). The different sets of distributions seemed non-parallel, suggesting different computational processes had been involved in judgements of the two temporal orders.

* Corresponding author at: McElwain Building, University of Queensland, Brisbane, QLD 4072, Australia.
E-mail address: brendan.keane1@uqconnect.edu.au (B. Keane).

http://dx.doi.org/10.1016/j.concog.2015.10.010
1053-8100/© 2015 Elsevier Inc. All rights reserved.

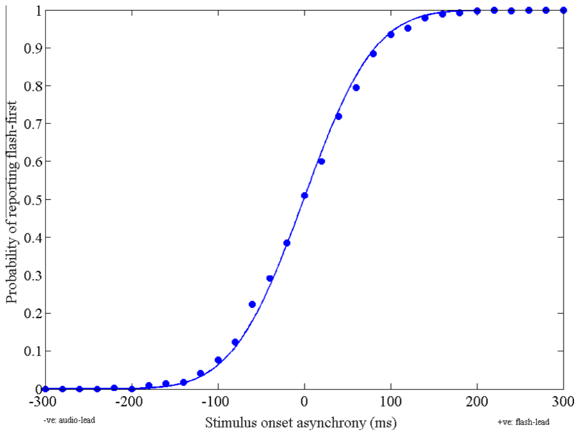

Fig. 1. Simulated temporal order judgement data. The arrival times of audio and visual signals at a central comparator are assumed to be variable from trial-to-trial, generating a Gaussian distribution of encoded arrival time differences for any given physical offset. Here mean arrival time differences are assumed to be equal to the physical timing difference, with standard deviations of 50 ms. These differences in encoded arrival time are compared against a fixed order criterion of 0 ms. Presentations yielding a negative encoded difference are reported as an audio-lead presentation, whereas presentations yielding a positive encoded difference are reporting as a visual lead. The prediction is well established analytically, but for comparison with subsequent simulations, here each physical stimulus onset asynchrony was sampled 1000 times.

Allan's (1975) observations have interesting implications, as they would contradict a prominent class of timing perception models. These assume that encoded signals must propagate to a common neural site, with relative subjective timing scaling with differences in arrival times at the central comparator (e.g., Sternberg & Knoll, 1973). Further, these models assume that neural propagation times vary from trial-to-trial, obeying a Gaussian distribution, and that encoded timing differences are referenced against fixed timing criteria (for instance, to denote when a given signal has preceded or lagged another). Predicted discriminant functions can take the form of a cumulative Gaussian, with a slope determined by trial-to-trial variance in encoded signal arrival time differences (Baron, 1969, see Fig. 1). For simplicity, we can assume an unbiased observer, such that the criterion used for categorising timing differences as denoting a lead or lag is physical synchrony (0 ms), with negative and positive encoded values prompting audio lead and lag categorisations respectively.

The above generic class of human timing perception models can be extended to capture confidence by assuming 2 additional criteria, one denoting the extent by which an encoded signal must fall under the fixed timing criterion to prompt a high-confidence *lead* response, and another denoting the extent by which an encoded signal must exceed the fixed criterion to prompt a high-confidence *lag* response. Low confidence is reported when encoded differences fall between these two confidence criteria. The important point of difference between predictions of this class of model, and Allan's suggestive report, is that they predict *parallel* high-confidence discriminant functions (see Fig. 2, left panel).

Response simulations, based on a generic model of human timing perception, reveal that this scheme also predicts a difference in low and high-confidence discriminant function slopes (see Fig. 2, right panel), something that Allan (1975) did not directly investigate. For audio–visual judgements, this implies one should transition from predominantly responding sound first, to predominantly responding light first, over a smaller expanse of test offsets when confident than when unconfident. This prediction depends on all combinations of order and confidence relying on the same source of information, in this case encoded signal arrival time differences, which vary from trial-to-trial (e.g., Sternberg & Knoll, 1973). This contrasts with Allan's suggestion, that discrepant processes underlie confident sound and light first reports.

An alternate possibility is that, instead of trial-by-trial variability in encoded timing differences, there is little or no such variability. Instead, discrepant low and high-confidence discriminant functions could result from people adopting variable decision criteria from trial-to-trial (Ulrich, 1987; Yarrow, Jahn, Durant, & Arnold, 2011; Yarrow, Sverdrup-Stueland, Roseboom, & Arnold, 2013). For instance, people might adopt more variable criteria when low in confidence. This would predict that confidence-based differences in the precision of timing perception should be minimised, or eliminated, by limiting the influence of subjective decisional criteria.

In this study we aimed to determine whether subjective confidence predicts the precision of human timing perception, and to assess whether high-confidence categorical discriminant functions for timing are parallel. We present two experiments. In Experiment 1 we show that high-confidence temporal order judgements are more precise than low-confidence order judgements, and that high-confidence light-first and sound-first discriminant functions are, on average, parallel. In Experiment 2 we show that the greater timing precision suggested in Experiment 1 for high-confidence trials generalises to objective performance in a task designed to minimise the influence of subjective decisional criteria. In combination, our data are consistent with models of human timing perception that assume Gaussian trial-by-trial distributions of encoded timing differences, which are referenced against fixed decisional criteria.

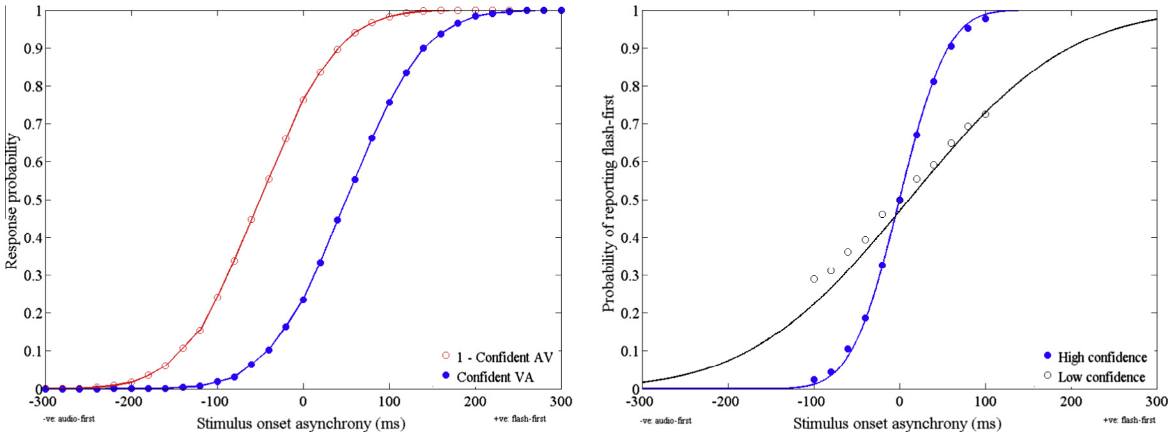

Fig. 2. Simulated temporal order judgements. Order judgements are made according to the process described in main text, and simulated in Fig. 1. Here, two additional fixed confidence criteria are placed (at −50 and +50 ms). (Left) Distributions of high-confidence order judgements, for reporting that sound had led visual signals (plotted in reverse to aid comparison), and for reporting that visual signals had led sound. Note that these predicted discriminant functions are parallel. (Right) Order judgements are plotted for high and low-confidence order judgements. Data are plotted for onset asynchronies yielding between 20% and 80% high confidence. This avoids fitting discriminant functions for the two confidence conditions to different ranges of physical test timings, and prevents under-sampled asynchronies distorting function-fitting procedures (e.g., large asynchronies are likely to be under-sampled in the low-confidence distribution, while small asynchronies are likely to be under-sampled in the high-confidence distribution). Cumulative Gaussian functions were fit to simulated distributions of encoded temporal order using the psignifit toolbox for MATLAB (see http://www.bootstrap-software.org/psignifit/).

2. Experiment 1

In Experiment 1 we made use of a classic temporal perception task—a binary temporal order judgement (Hirsh & Sherrick, 1961)—followed by a binary classification of low or high confidence in the preceding timing decision. In line with previous literature, we predict that psychological experiences of temporal order will be more precise when people report high confidence.

2.1. Methods

Eight adults (five male; $M = 25.38$ years, $SD = 6.59$ years) volunteered to participate in three blocks of 240 trials each (detailed below). Three of the four authors participated, in addition to five volunteers who were naïve as to the purpose of the experiment. All participants had normal or corrected-to-normal vision and reported no history of hearing loss.

At the start of each trial a red dot, with a diameter subtending 0.70 degrees of visual angle (dva) at the retina, was presented in the middle of the display. After a variable delay (1–1.5 s) an auditory and a visual event were presented. The visual event consisted of a white Gaussian blob (diameter = 6.78 dva, $\sigma = 0.88$ dva) with a peak luminance intensity of approximately 120 cd m^{-2}. Visual stimuli were presented at fixation for 8.33 ms (1 frame at 120 Hz) against a black background on a gamma-corrected 21-in. Samsung SyncMaster 1100p+ monitor with a resolution of 1024 × 768 pixels. Visual stimuli were generated using a Cambridge Research Systems ViSaGe stimulus generator and were viewed from a distance of 57 cm with the participants' head restrained by a chin rest. The auditory event was a 1000 Hz pure tone presented for 10 ms with 5 ms linear onset/offset ramps. These were presented binaurally via Sennheiser HDA 200 headphones at an intensity of ~68 dB SPL. Auditory stimuli were generated via a Tucker Davis Technology (TDT) psychoacoustic processor. Auditory triggers, timed to coincide with a monitor refresh, ensured precise timing of stimulus presentation.

On each trial (schematised in Fig. 3) the presentation of visual and audio events was offset by one of 12 pre-defined stimulus onset asynchronies (SOAs; ±300 ms, ±200 ms, ±100 ms, ±75 ms, ±50 ms, & ±25 ms, with negative values indicating an audio lead). These offsets were doubled for one participant. Participants were required to report which stimulus had been presented first by pressing one of two mouse buttons, or to press a third button if they had experienced a lapse in attention/concentration. When this third button was pressed test stimuli were re-presented, before moving on to the next trial. Following each order judgement, participants were prompted to report their level of decision confidence (high or low) by pressing one of two mouse buttons.

During each block of trials, each of 12 audio–visual timing relationships was sampled (without replacement) 20 times in a random order, yielding 240 trials per block (with some trials containing additional presentations, if the participant reported lapses in attention/concentration). Each participant completed three blocks of trials, resulting in each individual successfully completing 720 individual trials. For the purposes of analysis, data were collated across the three blocks of trials completed by each participant.

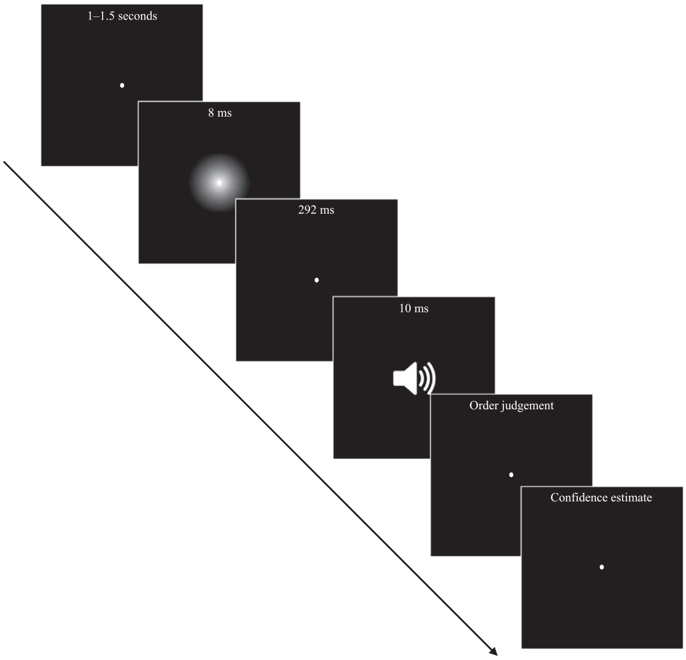

Fig. 3. Stimulus presentation sequence in a 300 ms audio-lag trial.

We fit cumulative Gaussian functions (using the psignifit toolbox for MATLAB[1]) to each participant's order judgements (expressed as a proportion of trials on which the participant reported an audio-lag) separately for high- and low-confidence trials. Participants' just noticeable timing differences (JNDs) were estimated from the difference between the SOAs associated with the 50% and 80% points on the fitted functions. For each participant we identified and analysed order judgements from SOAs associated with between 20% and 80% reports of high confidence (SOAs associated with extremely high or low confidence were not included in these analyses). This was done to ensure that low and high confidence data related to a common range of test SOAs.

2.2. Results

A paired-samples *t*-test compared participants' JNDs associated with high- and low-confidence judgements of temporal order (see Fig. 4). This analysis revealed that participants reliably differentiated temporal order at smaller SOAs when reporting high (M = 33.98 ms, SD = 20.09 ms) as opposed to low (M = 80.68 ms, SD = 46.09 ms) confidence ($t_7 = 3.89$, $p = .006$). This pattern was consistent across the majority of observers; see individual plots in Fig. 5.

Low-confidence responses might disproportionately represent trials associated with a lapse in attention/concentration, resulting in guessing behaviour and thus solely explaining apparent sensitivity differences. Attentional lapses should, however, be uniformly distributed across sampled SOAs. To assess the possibility that low confidence is driven by lapses, we fit Gaussian distributions to individual data describing proportions of low-confidence trials as a function of test SOA. These provided a good description of individual data, and were centred approximately about physical synchrony (see Fig. 6). These analyses show that low-confidence distributions were highly non-uniform relative to test SOA, with participants more likely to report low confidence for small SOAs.

To further test whether lapse rate made a substantial contribution to apparent timing sensitivity differences, we estimated individual lapse rates from overall numbers of incorrect responses at large SOAs (±300 ms and ±200 ms), guided by the assumption that erroneous order judgements at these large SOAs resulted from lapses in attention. Lapse rates were estimated as double the individual error rate averaged across these SOAs. We then calculated the increase in individual guessing rates required to explain discrepancies between high and low confidence data. This was estimated from residual differences (high confidence–low confidence) in proportions of light/sound first judgements from chance (50%). Individual data were averaged across SOAs associated with high confidence on between 20% and 80% of trials. A paired-samples *t*-test revealed that the requisite increases in guessing rates (M = 22.71%, SD = 13.85%) were greater than estimated lapse rates (M = 4.17%, SD = 6.04%; $t_7 = 3.24$, $p = .014$).

Allan (1975) claimed (without inferential statistical support) that participants in her temporal order judgement task had produced skewed responses to the two temporal orders under conditions of high confidence, with light-first responses

[1] See http://www.bootstrap-software.org/psignifit/.

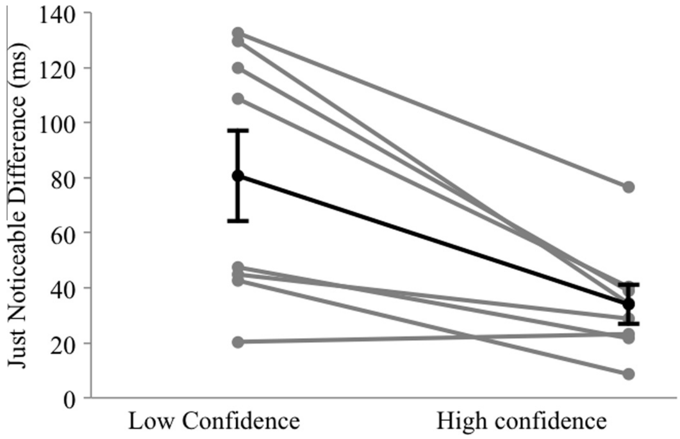

Fig. 4. Just noticeable difference (JND) magnitudes for low- and high-confidence trials for each participant. Grey lines depict individual participants, while the black line indicates group averages. Error bars depict standard error of the mean. Smaller JNDs indicate greater sensitivity.

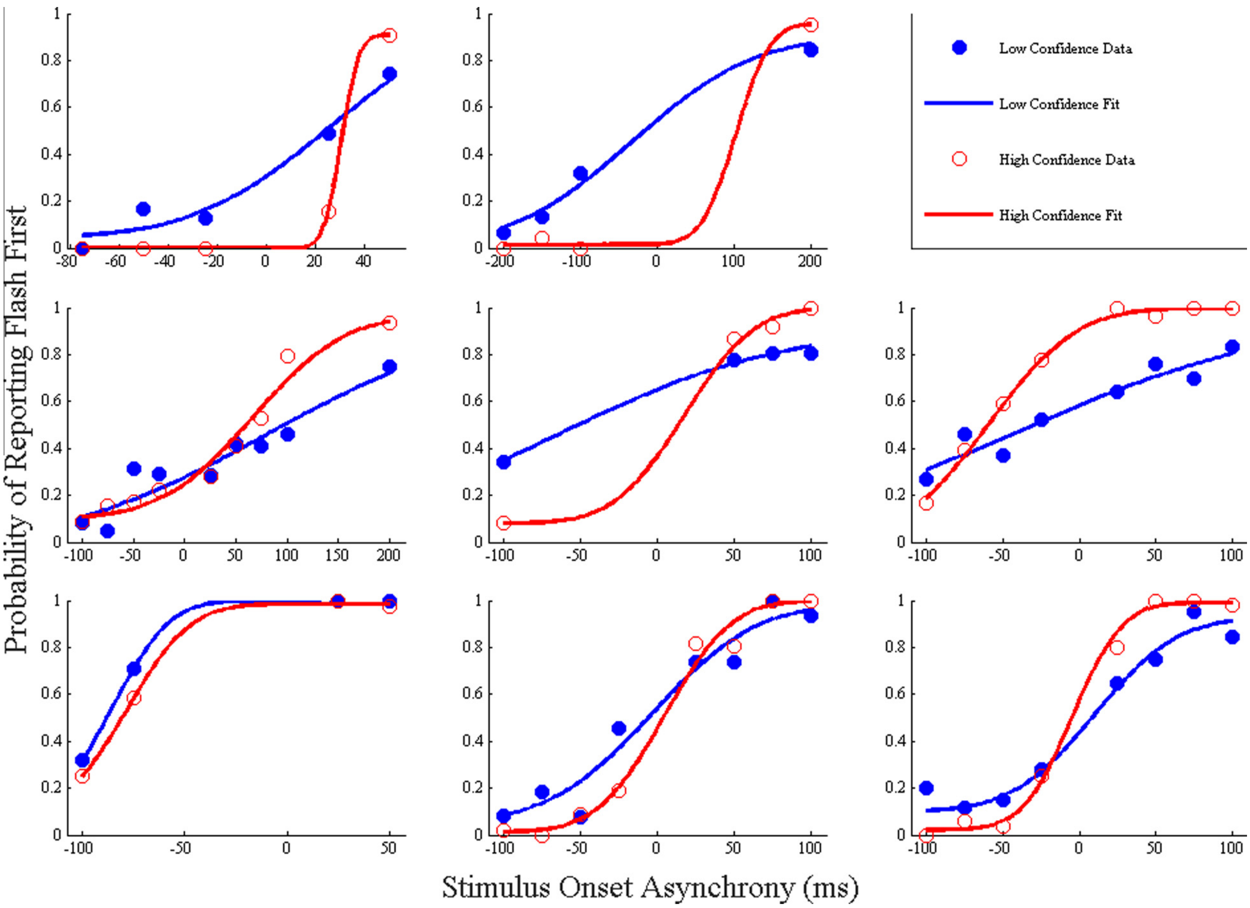

Fig. 5. Cumulative Gaussian distributions fitted to participants' low- and high-confidence reports of temporal order. Negative stimulus onset asynchronies (SOAs) indicate audio-first presentations. These distributions were fit to SOAs associated with high confidence on between 20% and 80% of trials. Asterisks indicate author responses.

negatively skewed and tone-first responses positively skewed. This would call some models of time perception into question, as they predict parallel discriminant functions for high confidence judgements. Here (Fig. 7) we plot individual discriminant functions for high confidence order reports, along with data averaged across 7 participants (inclusive of data from all participants, except one, who had required that physical offsets be doubled). We see no evidence of a systematic difference

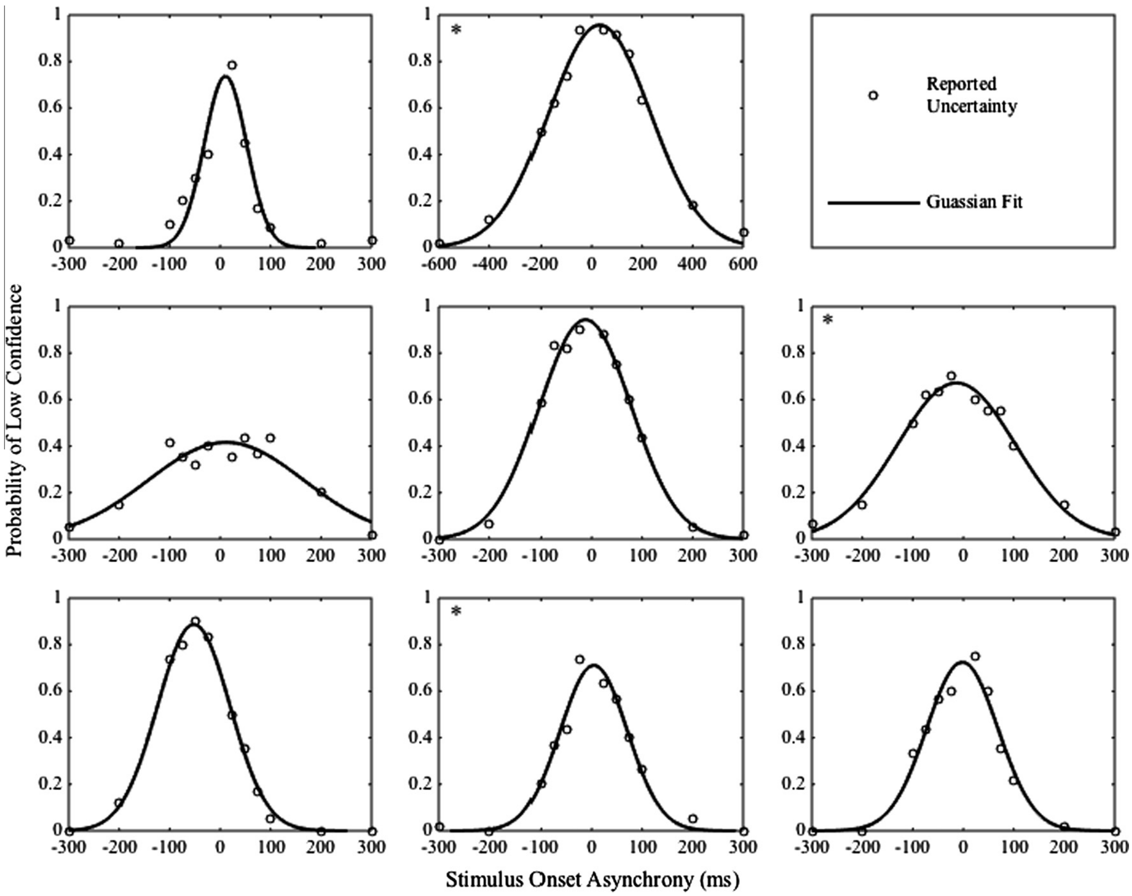

Fig. 6. Gaussian distributions fitted to participants' reports of low confidence at each stimulus onset asynchrony (SOA). Negative SOAs indicate audio-first presentations. Each panel in this figure corresponds to the panel in the same relative location of Fig. 5. Asterisks indicate author responses.

in the shape of discriminant functions. We fit cumulative Gaussian functions to individual discriminant functions. A within-participant *t*-test found that the slope of functions fitted to audio-lag data ($M = 70.76$, $SD = 14.4$) was not reliably different from audio-lead data ($M = 74.19$, $SD = 12.88$), suggesting the two sets of functions are best described as parallel across participants ($t_7 = 0.27$, $p = .797$).

2.3. Discussion

In Experiment 1 we aimed to determine whether humans have metacognitive insight into the strength of sensory evidence supporting their classifications of temporal order. We did so by measuring the magnitude of the temporal offset required for participants to reliably discern order when they had expressed either a high or low level of confidence in their order judgements. We found participants were able to discern temporal order when presented with *smaller* temporal offsets on trials in which they had expressed high as opposed to low confidence.

A common criticism of analyses of data split according to confidence is that low-confidence responses might disproportionately capture trials on which participants had experienced a lapse in attention/concentration, resulting in guessing behaviour. In the extreme, it could be argued that there is no difference in sensitivity between high and low confidence data after accounting for this influence. Two features of our data speak against this possibility. First, low-confidence reports were non-uniformly distributed as a function test SOA, concentrated at the smallest SOAs. This is inconsistent with low confidence data resulting from lapses in attention, as such lapses should not systematically coincide with a specific range of test SOAs. Second, we were able to estimate what increase in guessing rate would be required to explain discrepancies between the precision of low and high confidence timing data, and compare this with lapse rates estimated from performance on large SOA trials. We found that guessing rates required to explain confidence-based differences in task performance precision (~23%) were considerably greater than estimates of the observed lapse rate (~4%).

Overall, the results of Experiment 1 suggest that humans have insight into the strength of evidence underlying their decision when making temporal order judgements. This was evidenced by participants needing a smaller temporal offset

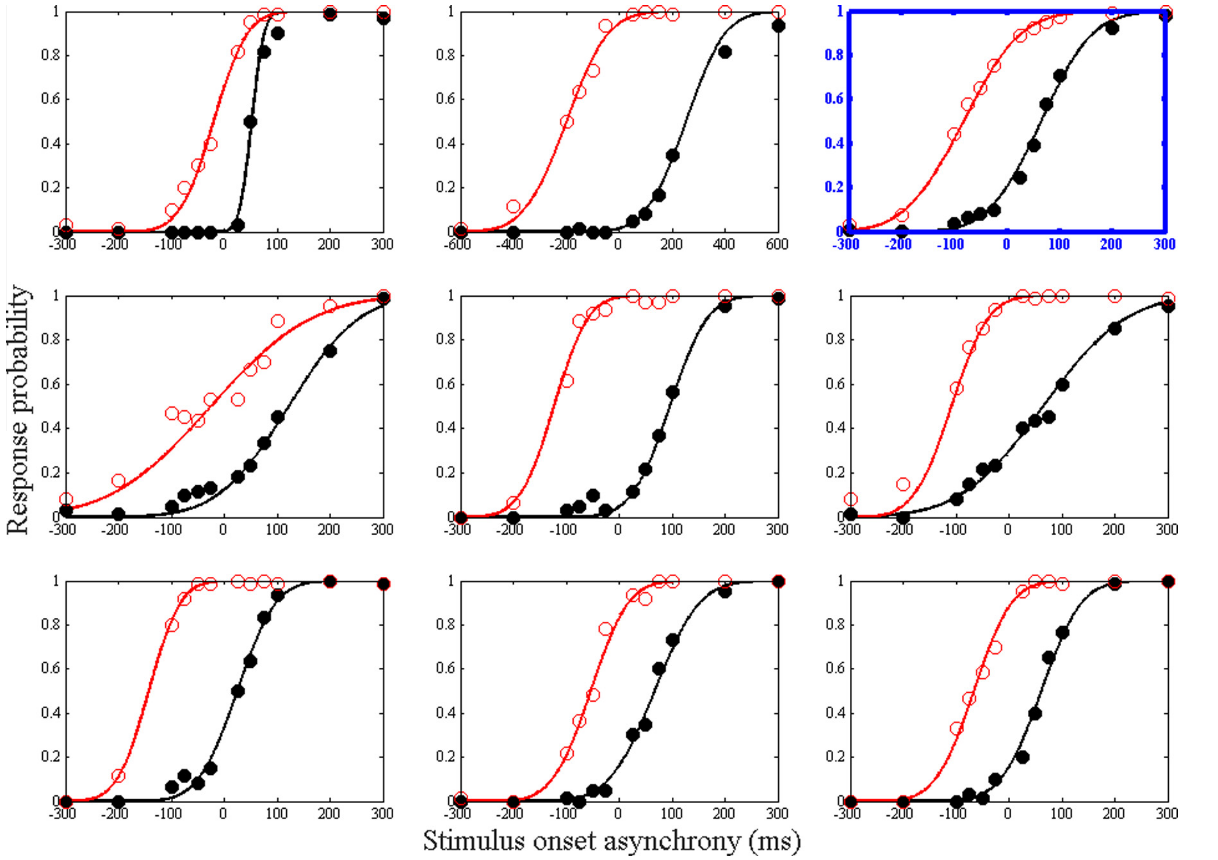

Fig. 7. High confidence reports of flash-first (black), and 1 – audio-first (red) for all participants, along with data average across 7 participants (top-right). Negative stimulus onset asynchronies (SOAs) indicate audio-first presentations. Audio-first response probabilities have been subtracted from 1 to facilitate comparison. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

to transition between perceived orders (auditory event before visual, or vice versa) when reporting a high, as opposed to a low level of confidence. However, it remains unclear whether the effect observed here is reflective of variable timing encodings from trial-to-trial, or to the application of variable subjective decisional criteria. In Experiment 2 we aim to determine which of these explanations is more plausible by taking steps to minimise the influence of subjective timing criteria.

3. Experiment 2

Experiment 2 aims to determine whether the positive relationship between confidence and the precision of subjective timing judgements generalises to more objective measures of perceptual sensitivity. Further, this experiment aims to determine whether it is more likely this relationship is shaped by insight into the strength of encoded timing information, or the application of variable timing decisional criteria. Here we use a 3-alternative odd-one-out paradigm. This minimises the impact of subjective criteria, by avoiding any requirement for participants to categorise inputs along a nominated dimension (such as sound first, or light first). Instead, sensitivity to any difference is assessed by having people choose which of three stimuli differs from the other two. Accuracy can therefore be interpreted as a measure of perceptual sensitivity (Green & Swets, 1974; Stanislaw & Todorov, 1999).

3.1. Methods

Eight adults (six male; $M = 26.14$ years, $SD = 6.39$ years) volunteered to participate in four blocks of 120 trials each (detailed below). Three of the four authors participated, in addition to five volunteers who were naïve as to the purpose of the experiment. Seven of eight participants had participated in Experiment 1. All participants reported normal or corrected-to-normal vision and no history of hearing loss.

This task involved sequential presentations of three pairs of audio–visual signals. Visual signals consisted of a single 10 ms pulse of a green light-emitting diode, with a peak luminance intensity of ~125 cd m^{-2} and 5 ms linear onset and offset ramps. Auditory events consisted of a 1000 Hz pure tone presented for 10 ms, with 5 ms linear onset/offset ramps. Tones had a peak auditory intensity of ~63 dB SPL at 57 cm and were presented via a single PC speaker. Both visual and auditory signals

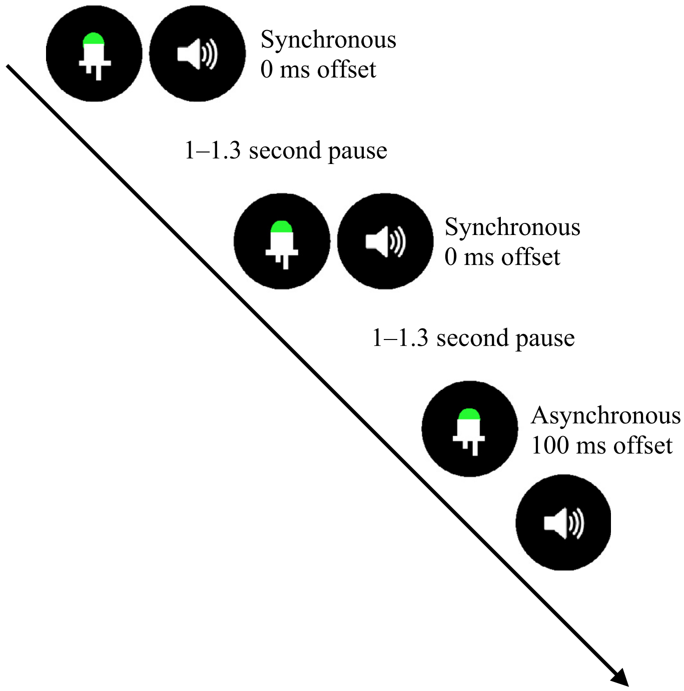

Fig. 8. Stimulus sequence in an audio-lag-deviant block, on a trial where the target pairing was presented third.

were generated and presented with precision timing by a TDT psychoacoustic workstation. The diode was positioned directly in front of the speaker. Participants viewed the apparatus in a darkened room from a distance of 57 cm, with their head restrained by a chin rest.

On each trial, participants were required to report which of the three test presentations was temporally distinct. Audio and visual events during two of the three presentations were physically synchronous, whereas audio and visual signals in the third pairing were asynchronous (see Fig. 8, below). Participants indicated which presentation they thought was temporally distinct by pressing one of three buttons on a computer mouse, or they could report a lapse in attention/concentration by holding down any mouse button until a distinct tone occurred. If a participant reported a lapse, the trial was repeated with presentation order re-randomised. Participants then reported their confidence (low or high) in their judgement. Inter-stimulus intervals varied randomly between 1 and 1.3 s on a presentation-by-presentation basis, to avoid participants using presentation rhythmicity as a reliable timing cue. No feedback was provided regarding task performance.

In two blocks of trials the target pairing was an audio-lead, whereas in the other two the target was an audio-lag. Participants first completed block one of the audio-lead targets, and then all remaining blocks in randomised order. Participants were not told what type of target they were attempting to identify. The first 40 trials of the first block (for each target type) were used to estimate a 75% accuracy threshold, with sampling controlled via a one-up/one-down staircase procedure. This process involved the presentation of trials containing an asynchronous target with an SOA determined by the participant's accuracy on the previous trial. Stimuli in the target pairing were initially offset by 400 ms. Each correct/incorrect identification of the target decreased/increased the SOA of the target pairing by 25 ms. Upon the completion of these trials, a logistic function was fit to participants' accuracy as a function of SOA using the psignifit toolbox for MATLAB. We used this logistic function to estimate the SOA each participant would be able to accurately identify on 75% of trials. These SOAs are presented in below with their corresponding overall and conditional accuracies, as well as the proportion of trials on which participants reported high confidence.

In the final 80 individual trials of the first block of trials, and for all 120 trials in the second, the target audio–visual offset was set to the participants' 75% threshold for that experimental condition. The order in which the deviant target was presented (i.e., 1st, 2nd, or 3rd) was randomised (without replacement) for each trial within each block (including the threshold-assessment trials in the first block). Each participant's accuracy was calculated for high- and low-confidence trials separately (based only on non-staircase trials) and then directly compared within participants.

3.2. Results

In general, our calibration measures equated accuracy across target-type and between participants at ~75%. Tables 1 and 2 present each participant's SOA in each target condition, with their overall and conditional accuracy, and the proportion of trials responded to with high confidence.

Table 1
Summary of data collected from participants completing audio-lead target blocks.

SOA (ms)	High confidence trials (%)	Accuracy (%)		
		Overall	Low confidence	High confidence
−94	57	81.5	65.12	93.86
−189	60	92.5	81.25	100
−352	68.5	82.5	58.73	93.43
−166	52.5	72.5	65.26	79.05
−97	37	63	50.79	83.78
−181	61.5	53	31.17	66.67
−229	67	79	66.67	85.07
−106	72	71	41.07	82.64
Mean	59.44	74.38	57.51	85.56

Table 2
Summary of data collected from participants completing audio-lag target blocks.

SOA (ms)	High confidence trials (%)	Accuracy (%)		
		Overall	Low confidence	High confidence
89	67.5	88.5	67.69	98.52
20	24	52.5	49.34	62.5
282	47	64	48.11	81.91
108	71	79.5	65.52	85.21
70	28.5	46	41.26	57.89
385	53.5	39	35.48	42.06
320	81	80.5	55.26	86.42
70	66.5	79.5	52.24	93.23
Mean	54.88	66.19	51.86	75.97

A 2 (target-type: auditory-lead, auditory-lag) × 2 (confidence: high, low) repeated-measures ANOVA on ternary detection accuracy revealed a main effect of confidence on accuracy, with participants demonstrating greater accuracy when reporting high ($M = 81\%$, $SD = 16\%$) as opposed to low ($M = 55\%$, $SD = 14\%$) confidence ($F_{1,7} = 71.87$, $p = <.001$; see Fig. 9). There was, however, no evidence for a main effect of target-type on accuracy, with participants similarly able to identify audio-lead ($M = 71\%$, $SD = 19\%$) and audio-lag ($M = 64\%$, $SD = 20\%$; $F_{1,7} = 2.25$, $p = .177$) targets. This was expected, since performance was equated at ~75% for each target type. Finally, there was no evidence of an interaction between target-type and confidence ($F_{1,7} = 0.72$, $p = .424$).

Accuracy can be artefactually depressed by a response bias, for example by a participant preferentially choosing one of three intervals when in doubt as to which interval might contain the target (McNicol, 1972). To determine whether response biases changed as a function of confidence, confounding our finding, we conducted a 3 (interval: first, second, third) × 2 (confidence: high, low) × 2 (target-type: auditory-lead, auditory-lag) repeated-measures ANOVA on interval choice frequency. This yielded no systematic biases in interval choice frequency. There was no main effect of interval ($F_{2,14} = 0.27$, $p = .746$, $\varepsilon = .90$), nor an interaction between interval and confidence ($F_{2,14} = 0.82$, $p = .433$, $\varepsilon = .75$). Hence we can conclude that the observed difference in accuracy between high- and low-confidence trials was not a result of distinct biases.

3.3. Discussion

The results of Experiment 2 again demonstrate that human participants have insight into the accuracy of their timing judgements. By design, in this experiment we limited the influence of subjective timing criteria by asking participants to identify which of three presentations was distinct from the other two. This provides a measure of objective timing sensitivity, as opposed to the more common subjective timing categorisation task (TOJ). Performance on the latter task can be strongly affected by any preference to categorise trials as having been sound first, or light first, whereas the former task disregards this subjective impression. Instead, the present task measures participants' ability to differentiate stimuli on the basis of physical differences, regardless of how those differences are experienced. Performance in *m*AFC tasks can be affected by any preference to select one presentation over the other two, but an analysis of our data revealed no such tendency. We can therefore conclude that participants displayed insight into the degree to which they had encoded timing correctly.

As in Experiment 1, participants had the option of indicating that they had suffered a lapse in attention/concentration. When this occurred, the trial was repeated with presentation order re-randomised. This meant that the oddball target was randomly re-assigned to any one of the three presentation intervals. Assuming participants are aware of their own lapses, this measure achieved two ends. First, it indicates low confidence responses are not simply lapses in concentration.

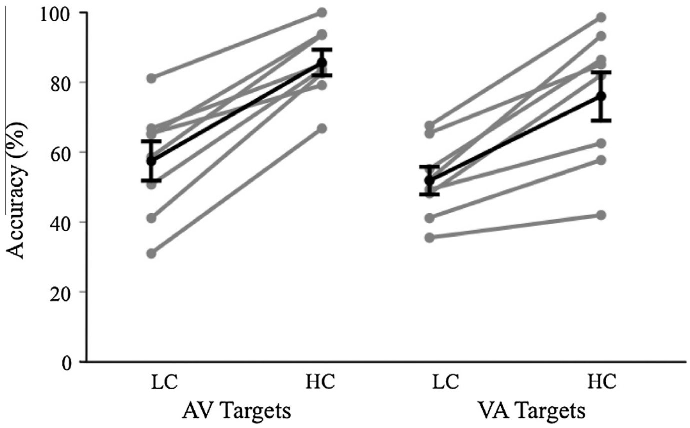

Fig. 9. Participants' accuracy in a 3AFC asynchrony detection task across low (LC) and high confidence (HC). Targets were (in separate blocks) an audio lead (AV; left) and lag (VA; right). As in Fig. 4, grey lines indicate performance of individual participants, while black bars indicate group means. Error bars represent standard error of the mean.

Moreover, the re-randomisation of presentation order dictated that participants could not benefit from a repeated trial. If presentation order had not had not been re-randomised, the repetition of a given trial order could have resulted in improved performance and high confidence.

4. General discussion

Our results demonstrate that humans have insight into the strength of encoded information underlying audio–visual timing judgements. This was apparent as, without task performance feedback, confidence predicted the precision of order judgements. More precise order judgements on high, as opposed to low confidence trials cannot be attributed solely to people *guessing* on a disproportionate number of low confidence trials, as lapse rates suggested by errors on simple trials were too few to account for confidence-based timing sensitivity differences. Nor can our data be attributed to people adopting more rigid subjective criteria on high, as opposed to low, confidence trials, as we also found evidence for heighted timing sensitivity on high confidence trials when people completed a three-alternative odd-one-out protocol, which minimised the influence of subjective response criteria.

Overall, our data are consistent with a generic class of timing perception models, which assume that temporal order judgements, and confidence in those decisions, is informed by a common source of variable encoded information (e.g., Allan, 1975; Sternberg & Knoll, 1973). Here the pertinent information refers to variable encodings of audio–visual timing differences. Importantly, in Experiment 2, our data related to a constant set of physical timing relationships, repeatedly sampled on different trials (two synchronous audio–visual presentations and one constant audio–visual asynchrony). Yet, despite the constant nature of stimulation, confidence predicted the precision with which people could identify the discrepant asynchronous stimulus presentation on a trial-by-trial basis. Moreover, as analyses revealed no evidence for a bias to preferentially select one of the three test intervals on each trial, our data suggest this relationship was due solely to trial-by-trial differences in how the pertinent timing relationships were encoded.

A plausible interpretation of our data is that, in this context, confidence might be a reportable estimate of the extent by which sensory evidence has exceeded decisional criteria (Ferrell & McGoey, 1980). This presumes that temporal judgements, and confidence in those judgements, are informed by a common source of information: encodings of timing differences that fluctuate from trial-to-trial. This implies that, unless some additional factor exists, confidence and the sensitivity of timing judgements should be invariably related. However, in other contexts, additional factors have been implicated (for reviews, see Fleming et al., 2012; Yeung & Summerfield, 2012). For instance, Spence et al. (in press) recently reported that variable motion-direction signals could be equated in terms of discriminability, but result in different levels of decisional confidence (see also de Gardelle & Mamassian, 2015). Such findings suggest that sensitivity and decisional confidence rely on independent and differentially weighted transformations of encoded signals. It remains to be seen whether similar dissociations of confidence and sensitivity will emerge in timing perception.

A final implication of our data is that encoded timing relationships are likely to be quite variable. Within an individual, encoded timing relationships seem to vary from trial-to-trial, resulting in marked confidence-based differences in the precision of timing judgements from-to-trial. Perusal of Fig. 5 also suggests marked individual differences in the extent of trial-by-trial variability. This is interesting, as relatively poor timing judgements have been linked to such conditions such as schizophrenia, depression and dysthymia (Bonnot et al., 2011; Gil & Droit-Volet, 2009; Rammsayer, 1990), Parkinson's disease (Pastor, Artieda, Jahanshahi, & Obeso, 1992), and attention-deficit hyperactivity disorder (Meaux & Chelonis, 2003; Smith, Taylor, Warner Rogers, Newman, & Rubia, 2002). Our data thus suggest that confidence could be used as a tool to investigate the precision of timing encoding, both on an individual basis and within specific clinical sub-populations.

Acknowledgement

These works were funded by an Australian Research Council Future Fellowship held by Associate Professor Derek H. Arnold.

References

Allan, L. (1975). Temporal order psychometric functions based on confidence-rating data. *Perception & Psychophysics, 18*(5), 369–372. http://dx.doi.org/10.3758/BF03211214.

Baron, J. (1969). Temporal ROC curves and the psychological moment. *Psychonomic Science, 15*(6), 299–300. http://dx.doi.org/10.3758/BF03336309.

Bonnot, O., de Montalembert, M., Kermarrec, S., Botbol, M., Walter, M., & Coulon, N. (2011). Are impairments of time perception in schizophrenia a neglected phenomenon? *Journal of Physiology – Paris, 105*(4), 164–169. http://dx.doi.org/10.1016/j.jphysparis.2011.07.006.

de Gardelle, V., & Mamassian, P. (2014). Does confidence use a common currency across two visual tasks? *Psychological Science, 25*(6), 1286–1288. http://dx.doi.org/10.1177/0953797614528956.

de Gardelle, V., & Mamassian, P. (2015). Weighting mean and variability during confidence judgements. *PLoS ONE, 10*(3), e0120870. http://dx.doi.org/10.1371/journal.pone.0120870.

Ferrell, W., & McGoey, P. (1980). A model of calibration for subjective probabilities. *Organizational Behaviour and Human Performance, 26*(1), 32–53. http://dx.doi.org/10.1016/0030-5073(80)90045-8.

Fleming, S., Dolan, R., & Frith, C. (2012). Metacognition: Computation, biology and function. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*(1594), 1280–1286. http://dx.doi.org/10.1098/rstb.2012.0021.

Gil, S., & Droit-Volet, S. (2009). Time perception, depression and sadness. *Behavioural Processes, 80*(2), 169–176. http://dx.doi.org/10.1016/j.beproc.2008.11.012.

Green, D., & Swets, J. (1974). *Signal detection theory and psychophysics*. Huntington, New York: R.E. Krieger Pub. Co.

Hirsh, I., & Sherrick, C. (1961). Perceived order in different sense modalities. *Journal of Experimental Psychology, 62*(5), 423–432. http://dx.doi.org/10.1037/h0045283.

McNicol, D. (1972). *A primer of signal detection theory*. Sydney, Australia: Australasian Publishing Company.

Meaux, J., & Chelonis, J. (2003). Time perception differences in children with and without ADHD. *Journal of Pediatric Health Care, 17*(2), 64–71. http://dx.doi.org/10.1067/mph.2003.26.

Pastor, M., Artieda, J., Jahanshahi, M., & Obeso, J. (1992). Time estimation and reproduction is abnormal in Parkinson's disease. *Brain, 115*(1), 211–225. http://dx.doi.org/10.1093/brain/115.1.211.

Rammsayer, T. (1990). Temporal discrimination in schizophrenic and affective disorders: Evidence for a dopamine-dependent internal clock. *International Journal of Neuroscience, 53*(2–4), 111–120. http://dx.doi.org/10.3109/00207459008986593.

Smith, A., Taylor, E., Warner Rogers, J., Newman, S., & Rubia, K. (2002). Evidence for a pure time perception deficit in children with ADHD. *Journal of Child Psychology and Psychiatry, 43*(4), 529–542. http://dx.doi.org/10.1111/1469-7610.00043.

Song, C., Kanai, R., Fleming, S., Weil, R., Schwarzkopf, D., & Rees, G. (2011). Relating inter-individual differences in metacognitive performance on different perceptual tasks. *Consciousness and Cognition, 20*(4), 1787–1792. http://dx.doi.org/10.1016/j.concog.2010.12.011.

Spence, M., Dux, P., & Arnold, D. (in press). Computations underlying confidence in visual perception. *Journal of Experimental Psychology: Human Perception and Performance*.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behaviour Research Methods, Instruments, & Computers, 31*(1), 137–149. http://dx.doi.org/10.3758/BF03207704.

Sternberg, S., & Knoll, R. L. (1973). The perception of temporal order: Fundamental issues and a general model. *Attention and Performance, IV*, 629–685.

Ulrich, R. (1987). Threshold models of temporal-order judgments evaluated by a ternary response task. *Perception & Psychophysics, 42*(3), 224–239. http://dx.doi.org/10.3758/BF03203074.

Yarrow, K., Jahn, N., Durant, S., & Arnold, D. (2011). Shifts of criteria or neural timing? The assumptions underlying timing perception studies. *Consciousness and Cognition, 20*(4), 1518–1531. http://dx.doi.org/10.1016/j.concog.2011.07.003.

Yarrow, K., Sverdrup-Stueland, I., Roseboom, W., & Arnold, D. (2013). Sensorimotor temporal recalibration within and across limbs. *Journal of Experimental Psychology: Human Perception and Performance, 39*(6), 1678–1689. http://dx.doi.org/10.1037/a0032534.

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: Confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Science, 367*(1594), 1310–1321. http://dx.doi.org/10.1098/rstb.2011.0416.

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