Manuscript accepted for publication in *Psychological Science*

Subjective confidence predicts information seeking in decision making

Kobe Desender[1,2,3], Annika Boldt[4,5], & Nick Yeung[6]

1. Department of Neurophysiology and Pathophysiology, University Medical Center Hamburg-Eppendorf, Germany
2. Department of Experimental Psychology, Ghent University, Belgium
3. Department of Psychology, Vrije Universiteit Brussel, Brussels, Belgium
4. Department of Psychology, University of Cambridge, United Kingdom
5. Institute of Cognitive Neuroscience, University College London, United Kingdom
6. Department of Experimental Psychology, University of Oxford, United Kingdom

Corresponding author:

Dr. Kobe Desender
Department of neurophysiology and pathophysiology
University Medical Center Hamburg-Eppendorf
Martinistrasse 52, 20251 Hamburg
Germany
E-mail: Kobe.Desender@gmail.com

1993/2000 words below (Abstract, Methods, Results, Author contributions, Acknowledgments and References excluded)

Abstract (149/150)

There is currently little direct evidence regarding the function of subjective confidence in decision making: The tight correlation between objective accuracy and subjective confidence makes it difficult to distinguish each variable's unique contribution. Here, we created conditions of a perceptual decision task that were matched in accuracy but differed in subjective evaluation of accuracy, by orthogonally varying the strength versus variability of evidence. Confidence was reduced with variable (vs. weak) evidence, even across conditions matched for difficulty. Building on this dissociation, participants (N = 20) could choose to seek further information before making their decision. The data provided clear support for the hypothesis that subjective confidence predicts information seeking in decision making: Participants were more likely to sample additional information before giving a response in the condition with low confidence, despite matched accuracy. In a preregistered replication (N = 50), these findings were replicated with increased task difficulty levels.

Introduction

People make finely calibrated evaluations of their own performance. In perceptual decision tasks, for example, subjective confidence correlates closely with objective accuracy (Boldt & Yeung, 2015; Fleming, Weil, Nagy, Dolan, & Rees, 2010), indicating that confidence reflects a direct (albeit imperfect) readout of decision processes (De Martino, Fleming, Garrett, & Dolan, 2013; Maniscalco & Lau, 2012; Pasquali, Timmermans, & Cleeremans, 2010). However, the role of decision confidence in adaptive behavior remains unclear. In the memory literature, confidence is thought to serve as a teaching signal during self-regulated learning (Bjork, Dunlosky, & Kornell, 2013; Butler & Winne, 1995; Metcalfe & Finn, 2008). Recent theoretical accounts indeed stress the importance of decision confidence in making predictions, learning from mistakes, and planning subsequent actions in the absence of feedback (Meyniel, Sigman, & Mainen, 2015; Yeung & Summerfield, 2012). However, there is little direct empirical evidence for this hypothesized role of confidence in adaptive decision making.

Animal studies suggest that low confidence motivates exploration before committing to a decision (Call & Carpenter, 2000) or opting-out of difficult decisions with low expected payoff (Foote & Crystal, 2007; Hampton, 2001; Smith et al., 1995). However, this interpretation remains controversial because decision accuracy and confidence are typically highly correlated: When evidence clearly favors one option, performance and confidence will both be high; when evidence is equivocal, accuracy will fall and confidence will be low. Thus, confidence and performance might be separate—but correlated—expressions of evidence quality (Kiani & Shadlen, 2009). For example, a recent study suggested that confidence in a decision can affect trade-offs between speed and accuracy in a subsequent decision (van den Berg et al., 2016a), but the authors inferred confidence from a combination of accuracy, reaction time and evidence, thus confounding confidence with evidence quality. In studies like these, it is therefore difficult to determine whether confidence is explicitly represented and causally influences decision making, or whether apparent adaptive behaviors are driven by lower-level processes such as associative learning to avoid contexts with low evidence quality (Le Pelley, 2012). Crucially, to provide evidence for the former, one needs to demonstrate that confidence predicts strategic decision making while controlling for objective performance (as a proxy for evidence quality).

Importantly, variations in confidence within a single condition are not informative in this respect, because these can still be driven by variations in evidence quality (e.g., due to fluctuations in attention paid across trials; Macdonald, Mathan, & Yeung, 2011). What is needed, therefore, is to induce differences in confidence between conditions that are matched for accuracy. To accomplish this, the current study builds on recent work dissociating accuracy and confidence using perceptual decision tasks that orthogonally manipulate strength and reliability of evidence (de Gardelle & Summerfield, 2011). In these tasks, participants perform categorical judgments on multi-item displays (e.g., judging the average color of 8 different-hued items as red or blue). The task is difficult when mean evidence is close to the decision boundary (i.e., average color is purplish rather than clear red or blue) and when evidence is highly variable

(i.e., the items comprise various red and blue hues). Crucially, evidence variability influences subjective confidence more strongly than does evidence mean (Boldt, de Gardelle, & Yeung, 2017; Spence et al., 2015): Even when performance is matched across conditions with weak versus variable evidence, participants are systematically less confident when variability is high. Here we leverage this dissociation to test the hypothesis that strategic information-seeking behavior—specifically, requesting additional evidence before committing to a decision—is predicted by subjective confidence rather than simple evidence quality as reflected in objective accuracy.

Experiment 1

Method

Participants

Twenty participants (seven men, mean age: 23.7 years, SD = 3.1, range 20 - 31) took part at the University of Oxford for monetary compensation (£8 plus up to £4.92 dependent on performance, range of the rounded actual payments: £10 - £12). Participants were tested individually, and all provided written informed consent. All participants reported normal or corrected-to-normal vision and were naive with respect to the hypothesis. The sample size was chosen because in a previous study this number proved effective in revealing the effect of variance on confidence (Boldt et al., 2017). In that study, the difference in confidence between the two medium difficulty conditions computed on all trials (see below) produced a large effect size ($N = 20$, Cohen's $d = 1.19$), which gives the current study, given $N = 20$, strong power of .99 to detect this. All procedures were approved by the local ethics committee.

Stimuli and apparatus

Stimuli were presented on a gray background on a 20-inch CRT monitor with a 75 Hz refresh rate, using the MATLAB toolbox Psychtoolbox3. Each stimulus comprised eight colored shapes spaced regularly around a fixation point (radius 2.8° visual arc). The mean color of the eight shapes was determined by the variable *C*; the variance of *C* across the eight shapes by the variable *V*. Mean color varied between red ([1, 0, 0]) and blue ([0, 0, 1]) by following a linear transition in RGB space ([*C*,0, 1 −*C*]). At the start of the experiment, *C* could take four different values: 0.450, 0.474, 0.526 and 0.550 (from blue to red, with 0.5 being the category boundary), and *V* could take two different values: 0.0333 and 0.1000 (low and high variability, respectively). On each trial, the color of each individual element was pseudo-randomly selected with the constraint that the mean and variability of the eight elements closely matched the values of *C* and *V*. Each combination of the *C* and *V* values occurred equally often. Four conditions were thus created with different levels of difficulty (see Figure 1B): an easy condition (*high mean – low variance*), two medium conditions (*low mean – low variance* and *high mean – high variance*) and a hard condition (*low mean – high variance*). The individual elements did not vary in shape. All responses were made using a USB mouse.

Procedure

Figure 1A shows an example experimental trial during the main part of the experiment. The stimulus was flashed for 200ms, followed by a fixation point for 200ms. There followed a choice phase consisting of two equally frequent conditions: free choice or forced choice. On free choice trials, the letters R and S (or S and R; counterbalanced across participants) appeared flanking the fixation cross. Participants could either choose to request additional evidence by seeing the stimulus again in an easier version (S), or to give their response (R). They indicated their choice by clicking the corresponding mouse button. On forced choice trials, only an R appeared, and participants were forced to select the option to give their

response. When participants chose to see the stimulus again, the values of the stimulus were slightly altered so that the mean was higher ($C' = C$ +/- .01) and the variability lower ($V' = V$ -.0167). They were then presented with a fixation point for 400ms, the easier stimulus for 200ms and a fixation point for 400ms. When participants opted (or were forced) to give their response without seeing the stimulus again, they simply viewed the fixation point for the same total amount of time (1000ms). Afterwards, a vertical response scale appeared (9.0° high and 0.4° wide) with a slider (0.1° high and 0.4° wide) in the center. The top of the bar was labelled as 'sure blue', the bottom as 'sure red' (counterbalanced across participants). Participants moved the cursor with their mouse to indicate jointly their response and their level of confidence, and confirmed by pressing the space bar. The location of the slider on the scale was translated into a numerical score, ranging from -50 (sure blue) to +50 (sure red), with every three screen-pixel (0.09°) increment resulting in a difference of one confidence point. No response could be given when the cursor was exactly in the middle (0 on the scale), so participants were forced to make the categorical judgment between red or blue. They were instructed to make this judgment at their own pace, and accuracy was stressed. Accuracy for each trial was scored as a binary variable. Confidence was scored as the absolute value on the response scale. To account for between-participant variation in use of the confidence scale and drift in confidence judgments over the course of the experiment, ratings were z-scored separately for each participant and each block.

Participants gained 5 points for correct answers and lost 5 points for errors. They could win up to an additional £4.92 by scoring points (650 points = £1). Crucially, choosing to see the stimulus again in an easier version costs 1 point, giving participants an incentive to ask for the hint only when the benefit of doing so (in terms of increasing the probability of making a correct choice) would outweigh this cost. Participants were explicitly instructed that they could score more points by strategically using the see again option. Participants were not rewarded for the accuracy of their confidence ratings.

The main part comprised 9 blocks of 64 trials, with balanced numbers of trials for each combination of mean, variance and trial type (free choice vs. forced choice), in pseudo-randomized order. Each block started with 8 additional practice trials in which the choice phase was omitted and participants received auditory feedback on the accuracy of their responses. This was done to maintain a stable color criterion over the course of the experiment. Before the main part of the experiment, several practice blocks were administered. In the first block (64 trials), participants practiced the color judgement task with no choice phase and a binary judgment (i.e., the slider could only take three positions, namely centered, up or down, both 2.3°), with auditory feedback to signal decision accuracy. In blocks 2-6 (64 trials each), participants jointly indicated their response and their level of confidence, using the continuous response scale. No feedback was delivered during these blocks, which served to familiarize participants with the confidence rating scale and to allow staircase-based matching of the difficulty of the two medium difficulty conditions, as described below. Finally, practice block 7 was identical to the main part of the experiment.

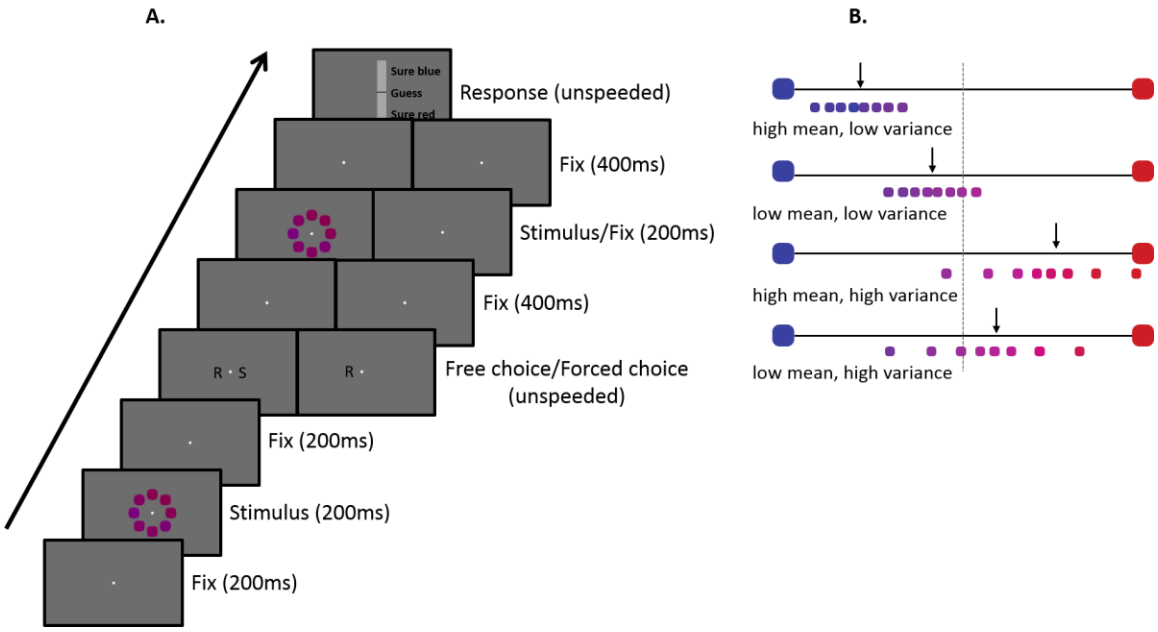

Figure 1. ***A.*** *Example of an experimental trial during the main experiment. After being presented with the stimulus, on half of the trials participants chose either to see the stimulus again in an easier version (by clicking S) or to give their response (by clicking R). These constitute the free choice trials. One the other half of the trials, participants could only choose to give their response (i.e., forced choice trials). Then, participants jointly indicated their response and level of confidence on a vertical continuous response scale.* ***B.*** *The four conditions that were created by crossing mean and variance, in order of increasing difficulty. Note that only four out of eight possible trial types are shown, because each condition equally often appeared with red and blue as the correct answer.*

To match performance between the two medium conditions (i.e., *high mean – high variance* and *low mean – low variance*), *C* levels in the *low mean* condition were adaptively changed. At the end of blocks 2 – 6, *C* was adjusted whenever the two medium conditions (16 trials per condition) were not matched in accuracy. In blocks 8 – 16, the change in *C* was based on the accuracies of the preceding two blocks of the forced choice trials only (16 trials per condition). The magnitude of the change depended on the size of the difference. If there was a difference in error rate of at least 6%, 10% or 15%, *C* values of the *low mean* condition were adjusted by 0.0005, 0.0012 or 0.0025, respectively. Inspection of the data revealed that overall the *C* values in the low mean condition at the end of the adaptive staircase procedure (M = .526 for 'red' stimuli, range: .518 to .534) did not differ from the value that was chosen at the start of the experiment (.526), $|t| < 1$, $BF = .24$[1].

[1] Because the default frequentist approach to statistics does not allow to interpret null effects, we additionally calculated a Bayes Factor (*BF*) in R using the BayesFactor package (Morey & Rouder, 2014) and the BayesMed package for correlations (Nuijten, Wetzels, Matzke, Dolan, & Wagenmakers, 2015), using the default priors. BFs for ANOVA's were calculated using a model comparison approach. Compared to classical *p*-values, a *BF* has the advantage that it can dissociate between data in favor of the null hypothesis ($BF < 1/3$), data in favor of the alternative hypothesis ($BF > 3$) and data that is uninformative ($BF \approx 1$).

Results

Decision accuracy

Decision accuracy was computed based on the data of the forced choice trials (ignoring the level of subjective confidence), and submitted to a 2 (mean: low or high) by 2 (variance: low or high) repeated measures ANOVA. The 95% confidence intervals were computed after logit transforming the data, and were then transformed back. Replicating earlier reports (Boldt et al., 2017; de Gardelle & Summerfield, 2011), both mean, $F(1,19) = 125.06$, $p < .001$, $\eta_p^2 = 0.87$, $BF = 7.65e{+}09$, and variance, $F(1,19) = 77.15$, $p < .001$, $\eta_p^2 = 0.80$, $BF = 3.03e{+}10$, affected accuracy, and there was a reliable interaction between the two factors, $F(1,19) = 37.73$, $p < .001$, $\eta_p^2 = 0.66$, $BF = 674$. Accuracy was highest in the easy condition (M = 97.7%, $CI_{95\%}$ [98.7, 99.8]) and lowest in the difficult condition (M = 75.2%, $CI_{95\%}$ [71.0, 82.1]). Crucially, follow-up contrasts focusing on the two medium difficulty conditions revealed that accuracy was effectively matched across the *low mean – low variance* and *high mean – high variance* conditions: M = 91.6%, $CI_{95\%}$ [90.3, 95.3], versus M = 91.1%, , $CI_{95\%}$ [90.0, 95.7], $|t| < 1$, $BF = .26$. Thus, our staircase procedure was successful in creating two conditions that were matched in terms of objective accuracy (Figure 2A). See Table 1 for a summary of all the dependent variables in this experiment, split over the different conditions. Given that the task was unspeeded, we did not observe any reliable differences in reaction time, all Fs < 1, all BFs < .33.

Confidence judgments

Next, the data of forced choice trials were used to examine how subjective confidence related to accuracy, how mean and variance influenced subjective confidence, and whether subjective confidence was as predicted reduced in the *high mean – high variance* condition compared to the *low mean – low variance* condition, despite their matched performance.

Confidence resolution. Participants tended to be more accurate when they expressed higher levels of confidence. Logistic regression models were fitted for each participant separately predicting accuracy based on the level of confidence, and positive significant slopes were observed for all fits, all ps < .017. To examine whether these results hold when controlling for difficulty, the factors evidence mean (high or low) and variance (high or low) were additionally entered into the regression. Again, positive slopes for confidence were observed for all fits, and these were significantly different from zero for 18 out of 20 participants. Correspondingly, after dividing trials into quintile bins according to confidence (for each participant separately) we observed a monotonic increase in accuracy with level of confidence (Figure 2B).

Average Confidence. The effect of mean and variance on z-scored confidence ratings was assessed on the data of the forced choice trials using a 2 (mean: low or high) by 2 (variance: low or high) repeated measures ANOVA. Here we report the results of confidence on all trials. The results for correct trials only, which did not differ materially from those reported here, can be found in the Supplementary Materials. Both evidence mean, $F(1,19) = 152.60$, $p < .001$, $\eta_p^2 = 0.89$, $BF = 8.44e{+}08$, and variance, $F(1,19) = 53.64$,

$p < .001$, $\eta_p^2 = 0.74$, *BF* = 1.25e+17, influenced confidence ratings. There was also a significant interaction between these factors, but the Bayes factor indicated weak ability to reject the null, $F(1,19) = 9.89$, $p = .005$, $\eta_p^2 = 0.34$, *BF* = 0.75. Confidence was highest in the easy condition ($M = 0.52$, $CI_{95\%}$ [0.39, 0.65]) and lowest in the hard condition ($M = -0.63$, $CI_{95\%}$ [-0.77, -0.50]). Crucially, as predicted, and replicating findings from Boldt et al. (2017), confidence ratings were significantly lower in the *high mean – high variance* condition ($M = -0.27$, $CI_{95\%}$ [-0.37, -0.17]), than the *low mean – low variance* condition ($M = -0.02$, $CI_{95\%}$ [-0.11, 0.08]), $t(19) = 3.22$, $p = .004$, *BF* = 10 (Figure 2C). Despite matching for difficulty, individual differences in confidence between the two conditions did not reliably correlate with individual differences in accuracy between them, $r(18) = .20$, $p = .400$, *BF* = .24.

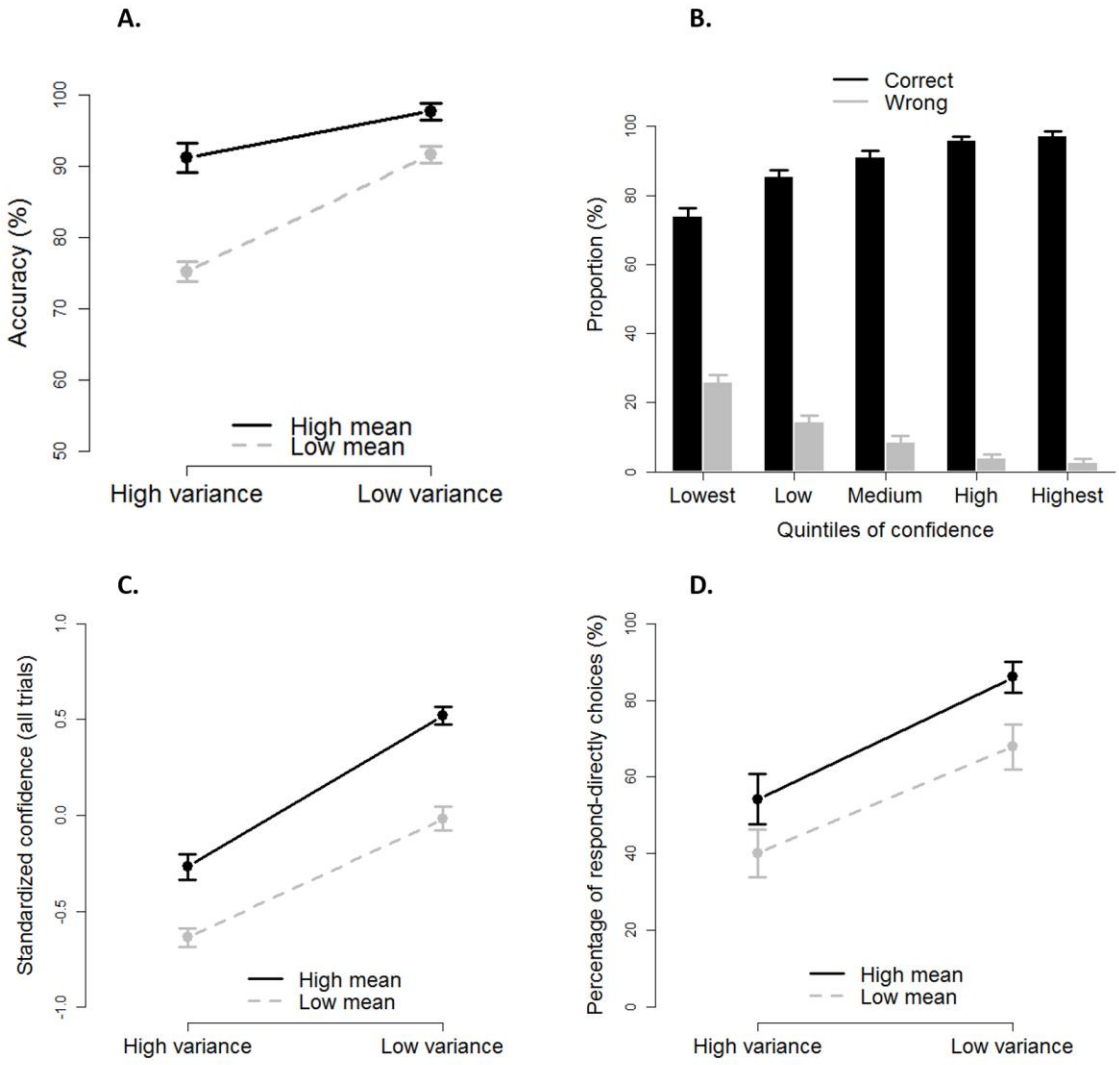

Figure 2. A. *First-order task performance, irrespective of the level of confidence, calculated from forced choice trials. The staircase procedure was successful in matching objective accuracy in the* high mean – high variance *and the* low mean – low variance *conditions.* ***B.*** *Distribution of objective accuracy as a function of confidence, calculated from forced choice trials.* ***C.*** *Mean of standardized confidence, calculated from forced choice trials.* ***D.*** *Percentage of trials where participants chose to give their response rather than to see the stimulus again, separately for each category, calculated from free choice trials. Error bars indicate standard errors of the mean.*

Information seeking

On free choice trials, participants on average chose to the see the stimulus again in an easier version on 37.9% of the trials. There was considerable variability between participants in this percentage (range 9% - 80%), but this did not correlate reliably with individual differences in overall mean confidence on forced choice trials, $r(18) = -.26$, $p = .268$, $BF = .31$, or accuracy on forced choice trials, $r(18) = .24$, $p = .299$, $BF = .29$.

To examine how evidence mean and variance influenced strategic information-seeking behavior, the percentage of see again choices in free choice trials was calculated separately for each trial type and then subjected to a 2 (mean: low or high) by 2 (variance: low or high) repeated measures ANOVA. Mirroring the results of the confidence judgments, effects of both mean, $F(1,19) = 71.46$, $p < .001$, $\eta_p^2 = 0.79$, $BF = 2.25e+05$, and variance, $F(1,19) = 82.23$, $p < .001$, $\eta_p^2 = 0.81$, $BF = 4.78e+13$, were significant, but with no interaction, $F < 1$, $\eta_p^2 = 0.04$, $BF = 0.39$. As expected, in the difficult condition participants asked to see the stimulus again on a majority of these trials ($M = 59.9\%$, $CI_{95\%}$ [47.0, 72.8]) whereas in the easy condition this occurred rarely ($M = 14.0\%$, $CI_{95\%}$ [5.7, 22.3]). Crucially, and as predicted, participants more often chose to see the stimulus again in the *high mean – high variance* condition ($M = 45.8\%$, $CI_{95\%}$ [32.1, 59.5]), that was associated with lower confidence ratings, compared to the *low mean – low variance* condition ($M = 32.1\%$, $CI_{95\%}$ [19.8, 44.5]), $t(19) = 6.23$, $p < .001$, $BF = 3749$ (see Figure 2D). This finding proved to be highly robust, with 19 out of 20 participants showing the pattern. Moreover, the one participant who requested to see the stimulus again more often in the *low mean – low variance* condition also expressed lower confidence in this condition compared to the *high mean – high variance* condition. Indeed, as can be seen in Figure 3, there was a trend correlation across participants between differences in confidence on forced choice trials and differences in see again choices across the two medium difficulty conditions, $r(18) = .42$, $p = .063$, $CI_{95\%}$ [-.02, .73], $BF = 0.95$. However, this relation should be cautiously interpreted given that the Bayes Factor indicates that these data are uninformative.

Importantly, it should be ruled out that participants more often chose to view the stimulus again in the *high mean – high variance* condition than in the *low mean – low variance* condition, simply because it was more helpful to do so for these trials. To this end, we compared these two conditions for see-again trials only. Due to a lack of at least 10 trials in one of these cells, only 13 participants were included in this analysis. Crucially, after participants asked to see the stimulus again, accuracy still did not differ reliably

between the two medium difficulty conditions, $t(12) = -1.45$, $p = .17$, $BF = .65$. If anything, accuracy was numerically lower in the *high mean – high variance* condition ($M = 96.2\%$, $CI_{95\%}$ [93.4, 98.8]) than the *low mean – low variance* condition ($M = 97.5\%$, $CI_{95\%}$ [95.1, 99.9]). This result remained unchanged when including the seven participants who had fewer than 10 trials in one of the cells, $|t| < 1$, $BF = .25$. These findings argue strongly against the possibility that participants' chose to see the stimulus again more often in the *high mean – high variance* condition because it was more useful to do so than in the *low mean – low variance* condition.

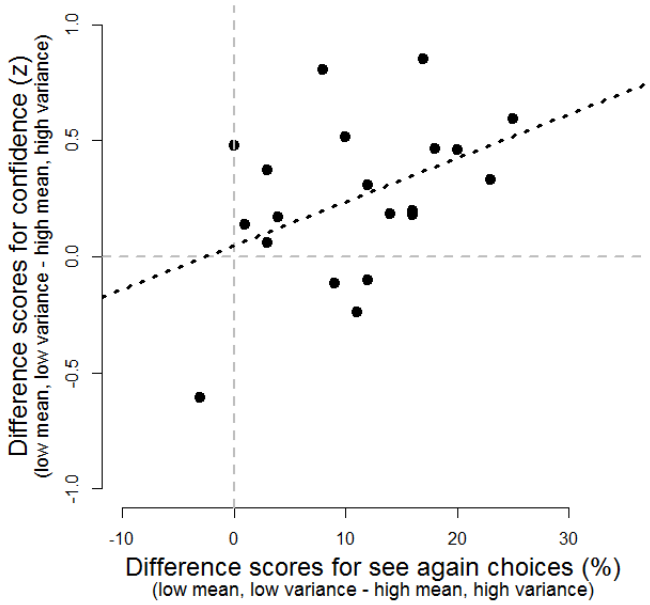

Figure 3. *Scatterplot showing the relationship between the difference scores for the two medium conditions for confidence judgments and see again choices.*

Metacognitive ability

We next performed exploratory analyses to investigate whether the impact of evidence variability on confidence judgments and see again choices relates to individual differences in metacognitive ability. Based on the data from forced choice trials we computed *M*-ratios (Fleming & Lau, 2014) to quantify the efficiency with which each participant's confidence ratings discriminated between their correct and incorrect responses (so-called meta-*d'*) while controlling for their first-order performance (*d'*). When this ratio is 1, all available first-order information is used in the confidence judgment. When the ratio is smaller than 1, metacognitive sensitivity is suboptimal, meaning that not all available information from the first-order response is used in the metacognitive judgment (Fleming & Lau, 2014). Averaged across conditions, mean *d'* was 2.58 (*SD* = .65), and mean meta-*d'* was 1.76 (*SD* = .63), giving a ratio meta-*d'*/*d'* of 0.69 (*SD* = .26). Interestingly, individual differences in meta-*d'*/*d'* ratio were positively correlated with differences in

confidence between the two medium conditions, $r(18) = .72$, $p < .001$, $CI_{95\%}$ [.42, .88], $BF = 109$, as well as with differences in see again choices between these conditions, $r(18) = .49$, $p = .028$, $CI_{95\%}$ [.06, .77], $BF = 1.88$ (Figure 4). Thus, the influence of evidence variability on subjective confidence and strategic information seeking was more profound in participants with *good* metacognitive ability. However, our sample size was not chosen to test this specific question and these analyses were performed in an exploratory rather than hypothesis-driven fashion. Therefore, a high-powered replication of these correlations is necessary to evaluate their robustness.

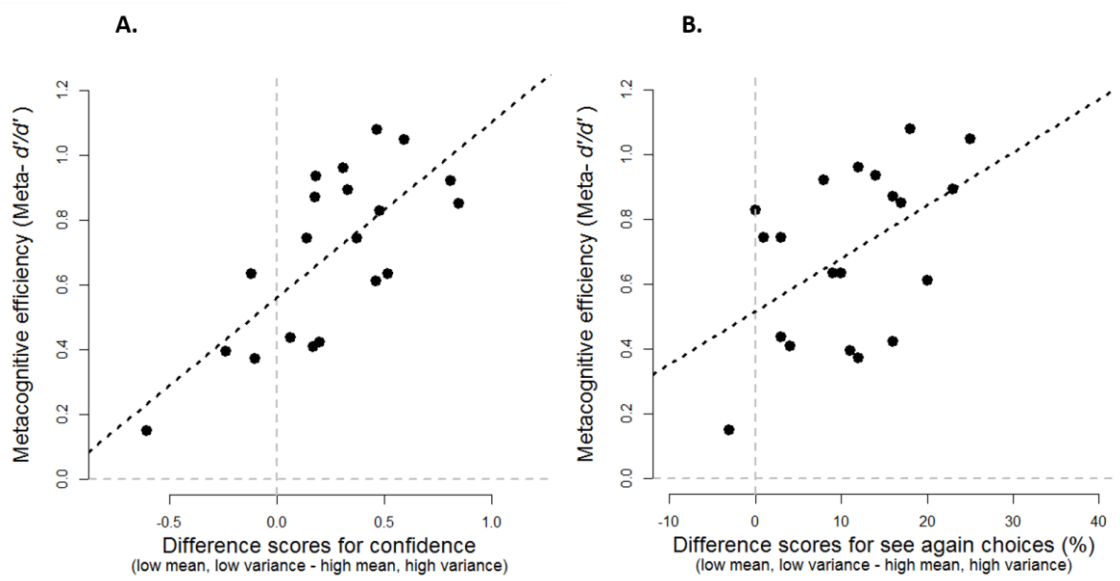

Figure 4. *Relationship between metacognitive ability, quantified as meta-*d'/d', *which was calculated from forced choice trials, and the difference between the two medium condition in both confidence judgments (A) and see again choices (B), which were both calculated from the free choice trials.*

A unique contribution of confidence

The previous results are consistent with the hypothesis that choices to sample more information are driven by subjective confidence, not by performance. To provide further support for this interpretation, we fitted mixed regression models to demonstrate that confidence predicts information seeking over and above measurable factors, and we performed a causal mediation analysis to test whether the influence of evidence variability on see again choices is mediated by subjective confidence. For both analyses, for each participant we computed (i) mean confidence based on the forced-choice data, (ii) the proportion of see again choices based on the free choice data, and iii) mean accuracy based on the forced-choice data, separately for the factors evidence variability (high or low), evidence mean (high or low), and color (red or blue). Values were calculated separately for each color to partition the data in a more fine-grained manner, but this variable was not taken into account in the analyses. For ease of interpretation, low variability and

high mean were dummy coded as reference categories so that a positive effect of each factor corresponds to an increase in difficulty.

To demonstrate that confidence has a unique influence on see again choices over and above the other experimental factors, we fitted a mixed regression model predicting see again choices by confidence, evidence variability, mean evidence and mean accuracy (using the lme4 package; Bates, Maechler, Bolker, & Walker, 2015). Random slopes were added for confidence, variance and their interaction, because this significantly increased the fit compared to a model without random slopes. Degrees of freedom were estimated using Satterthwaite's approximation. As predicted, confidence (β = -.16), t = -4.13, p < .001, variance (β = .20), t = 9.02, p < .001, and mean (β = .09), t = 4.97, p < .001, all affected see again choices, whereas mean accuracy did not, p > .67. This regression analysis clearly demonstrates that confidence predicts information seeking over and above performance.

Subsequently, we used causal mediation analyses to test whether the effect of variance on see again choices is mediated by confidence. First, a mediator mixed model was fit predicting mean confidence by variance (2 levels: high or low), mean (2 levels: high or low) and mean accuracy, with random slopes for variance. Second, an outcome mixed model was fit predicting the proportion of see again choices by mean confidence, variance (2 levels: high or low), mean (2 levels: high or low) and mean accuracy, with random slopes for mean confidence and variance. A mediation analyses was then performed based on these two models (using the mediation package; Tingley, Yamamoto, Hirose, Keele, & Imai, 2014), testing whether the influence of variance on see again choices is mediated by confidence, conditional on evidence mean and mean accuracy (a more detailed explanation can be found in the Supplementary Materials). The results showed that from the total effect of variance on see again choices (β = .277, $CI_{95\%}$ [.207, .350]), p < .001, there was 44.6% of total variance that was mediated by confidence (β = .125, $CI_{95\%}$ [.066, .200]), p < .001 (see Figure 5). This finding suggests that confidence is a crucial mediator between evidence variability and see again choices, although we acknowledge that mediation analysis is a correlational technique and therefore cannot provide ultimate evidence for causality (but see Supplementary Materials for further analysis that supports our interpretation).

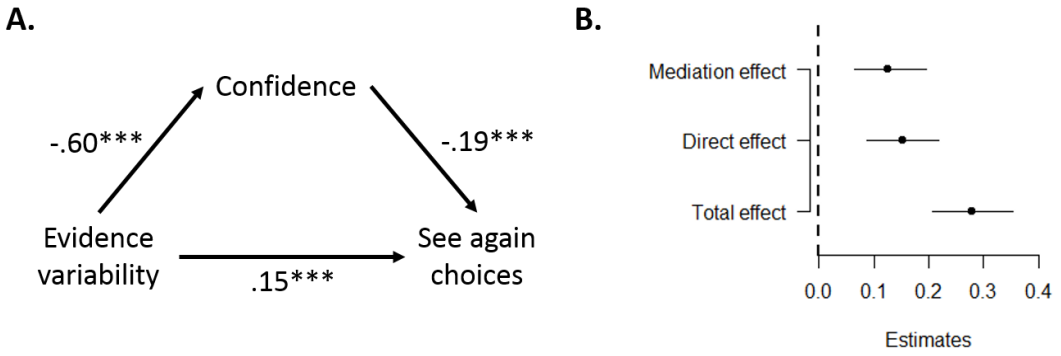

Figure 5. A. *Regression coefficients obtained from the mediation and the outcome mixed regressions models fit to the data of Experiment 1. Note: *** = p < .001. Degrees of freedom for*

calculating significance are based on Satterthwaite's approximation. ***B.*** *Causal mediation analysis. Error bars reflect quasi-Bayesian 95% confidence intervals.*

Comparing forced choice and free choice trials

A final set of analyses focused on possible performance benefits (or costs) of information seeking. As a manipulation check, we first confirmed that the opportunity to see the stimulus again led to improved performance and increased confidence. Simple *t*-tests revealed that participants were on average more accurate ($M_{\textit{free choice – forced choice}}$ = 3.5%, $CI_{95\%}$ [1.9, 5.2], $t(19) = 4.59$, $p < .001$, $BF = 151$) and more confident ($M_{\textit{free choice – forced choice}}$ = 0.23, $CI_{95\%}$ [0.14,0.33], $t(19) = 5.30$, $p < .001$, $BF = 616$) in the free choice condition (in which they could choose to see the stimulus again) compared to the forced choice condition. In relation to these findings, we note that free choice trials are more demanding than forced choice trials. Free choice trials constitute a dual task situation (two choices have to be made) whereas in forced choice trials only one decision has to be made. Although a dual task would typically reduce performance, and thus cannot account for the increase in performance and confidence in the free choice condition, it remains possible that confidence and accuracy might be affected in other ways by this difference in demand.

Focusing on the medium difficulty conditions, the data further showed that the increase in accuracy in free choice trials relative to forced choice trials was larger for the *high mean – high variance* condition ($M_{\textit{free choice – forced choice}}$ = 4.4%, $CI_{95\%}$ [2.2, 6.6]) than for the *low mean – low variance* condition ($M_{\textit{free choice – forced choice}}$ = 2.1%, $CI_{95\%}$ [-0.2, 4.3], $F(1,19) = 5.07$, $p = .036$, $BF = 1.78$, although the Bayes Factor indicated weak evidence. This observation reflects the fact that participants more often requested to see the stimulus again in the *high mean – high variance* condition, and as reported above, seeing the stimulus again increases performance. If this conjecture is true, the performance difference between trials in which participants waived the see again option versus forced choice trials (which are identical in terms of visual input) should be larger for the *high mean – high variance* condition than for the *low mean - low variance* condition. This is because participants were less confident in the *high mean – high variance* condition, so they needed to be more accurate on average before turning down the option of seeing the stimulus again. First, we confirmed that indeed overall performance was better on trials were participants waived the option of seeing the stimulus again (M = 94.0%, $CI_{95\%}$ [91.2, 96.9]), compared to forced decision trials (M = 88.9%, $CI_{95\%}$ [86.4, 91.4]), $t(19) = 5.07$, $p < .001$, $BF = 394$ (for a similar finding in monkeys, see Hampton, 2001). Importantly, as predicted, this difference was larger for the *high mean – high variance* condition ($M_{\textit{choose to respond – forced}\ \text{decision}}$ = 4.3%, $CI_{95\%}$ [1.6, 7.0]), than for the *low mean – low variance* condition ($M_{\textit{choose to respond – forced}\ \text{decision}}$ = 1.6%, $CI_{95\%}$ [-0.8, 4.0]), $F(1,19) = 6.22$, $p = .022$, $\eta_p^2 = 0.25$, BF = 2.67.

The above analyses suggest that participants indeed strategically used the option to receive additional evidence to increase their performance. Additional evidence increased performance more in the *high mean – high variance* condition than in the *low mean – low variance* condition, because in the former

condition participants were less confident and thus more willing to ask for additional evidence. This pattern was not observed for confidence judgments: The increase in confidence on free choice trials over forced choice trials was not different between the two medium difficulty conditions, $F < 1$. Likewise, participants were more confident when they turned down additional evidence in the free choice condition ($M = 0.29$, $CI_{95\%}$ [0.16, 0.42]) compared to being forced to respond ($M = -0.10$, $CI_{95\%}$ [-0.14, -0.06]), $t(19) = 5.21$, $p < .001$, $BF = 524$, but this increase did not differ reliably between the two medium difficulty conditions, $F(1,19) = 1.22$, $p = .283$, $\eta_p^2 = 0.06$, $BF = 0.39$. In sum, these additional analyses demonstrate that participants indeed strategically used the see again option to increase their performance.

Table 1.

Summary of the dependent variables separated by condition. Numbers between brackets represent standard deviations.

Experiment 1

	high mean – low variance	low mean – low variance	high mean – high variance	low mean – high variance
Accuracy (%) $_{\text{forced choice}}$	97.69 (5.3)	91.56 (5.2)	91.12 (6.2)	75.19 (9.3)
Accuracy (%) $_{\text{free choice}}$	98.5 (4.4)	93.6 (7.5)	95.5 (5.0)	82.1 (10.7)
Confidence (z) $_{\text{forced choice (all trials)}}$	0.52 (0.28)	-0.02 (0.21)	-0.27 (0.22)	-0.64 (0.29)
Confidence (z) $_{\text{forced choice (corrects)}}$	0.53 (0.27)	0.07 (0.21)	-0.21 (0.22)	-0.56 (0.27)
Confidence (z) $_{\text{free choice (all trials)}}$	0.65 (0.30)	0.23 (0.25)	0.03 (0.22)	-0.38 (0.28)
Confidence (z) $_{\text{free choice (corrects)}}$	0.67 (0.27)	0.28 (0.24)	0.05 (0.21)	-0.26 (0.29)
See again (%) $_{\text{free choice (all trials)}}$	14.0 (17.8)	32.1 (26.4)	45.8 (29.3)	59.9 (27.6)
See again (%) $_{\text{free choice (corrects)}}$	14.0 (22.2)	32.4 (33.9)	45.9 (34.3)	59.9 (35.9)

Experiment 2

	high mean – low variance	low mean – low variance	high mean – high variance	low mean – high variance
Accuracy (%) $_{\text{forced choice}}$	87.0 (9.2)	72.9 (7.6)	73.0 (8.4)	62.2 (7.6)
Accuracy (%) $_{\text{free choice}}$	90.3 (6.7)	79.8 (11.3)	79.8 (10.8)	68.9 (10.9)
Confidence (z) $_{\text{forced choice (all trials)}}$	0.04 (0.30)	-0.11 (0.32)	-0.17 (0.27)	-0.21 (0.29)
Confidence (z) $_{\text{forced choice (corrects)}}$	0.08 (0.30)	-0.07 (0.32)	-0.09 (0.27)	-0.16 (0.32)
Confidence (z) $_{\text{free choice (all trials)}}$	0.17 (0.35)	0.06 (0.33)	-0.04 (0.26)	-0.11 (0.25)
Confidence (z) $_{\text{free choice (corrects)}}$	0.20 (0.35)	0.09 (0.33)	-0.00 (0.26)	-0.06 (0.27)
See again (%) $_{\text{free choice (all trials)}}$	32.5 (24.0)	38.9 (24.7)	45.4 (26.7)	47.3 (27.1)
See again (%) $_{\text{free choice (corrects)}}$	33.0 (24.7)	40.8 (24.9)	47.7 (25.5)	50.1 (27.8)

Experiment 2: Preregistered replication

Experiment 2 aimed to replicate the results of Experiment 1, with an increased sample size and with increased task difficulty to address possible concerns about ceiling effects in accuracy (accuracy on see again choices in the critical medium difficulty conditions exceeded 96% in Experiment 1). The experimental protocols and all analyses were preregistered on the Open Science Framework. The replication specifically enabled us to evaluate the robustness of individual difference correlations observed in Experiment 1, now using a sample size appropriately powered for these analyses.

Method

Participants

Following Simonsohn (2015), we increased our sample size by a factor of 2.5 (equaling $N = 50$). This ensured 80% power to reject the null hypothesis that effects observed in Experiment 1 were too small to detect with that sample. This new sample size provides strong power (.96) to detect the correlation reported in Figure 4B. After replacing the data from one participant who stopped halfway through the experiment, fifty participants from the Free University of Brussels (seventeen men, mean age = 20.1 years, $SD = 2.7$, range 18 – 29, seven left-handed) took part in return for course credit and additional monetary compensation depending on performance (range €1 – €3.7). Participants were tested in groups of five people at most, seated in individual cubicles. They provided written informed consent. All participants reported normal or corrected-to-normal vision and were naive with respect to the hypothesis. All procedures were approved by the local ethics committee.

Stimuli, apparatus and procedure

Stimuli, apparatus and procedure were identical to Experiment 1, except that new mean color (C) values were chosen at the start of the experiment to increase task difficulty. Specifically, values were chosen so that the two medium difficulty conditions were slightly more difficult than the hard condition of Experiment 1, aiming for an accuracy level in these conditions of 70-75%. Thus C could now take four different values: 0.470, 0.485, 0.515 and 0.530 (from blue to red, with 0.5 being the category boundary).

Results

Decision accuracy

Decision accuracy was computed based on the data from forced choice trials (ignoring the level of subjective confidence), and submitted to a 2 (mean: low or high) by 2 (variance: low or high) repeated measures ANOVA. Both mean, $F(1,49) = 432.52$, $p < .001$, $\eta_p^2 = 0.90$, $BF = 6.96e+33$, and variance, $F(1,49) = 214.92$, $p < .001$, $\eta_p^2 = 0.81$, $BF = 2.34e+33$, affected accuracy. Although there was a significant interaction between the two factors, $F(1,49) = 5.68$, $p = .021$, $\eta_p^2 = 0.10$, $BF = 1.58$, the Bayes Factor indicated that the data were uninformative on this point. Accuracy was highest in the easy condition (M = 87.0%, $CI_{95\%}$ [87.0, 91.7]) and lowest in the difficult condition (M = 62.2%, $CI_{95\%}$ [60.2, 64.7]). Note that

the interaction between mean and variance was much weaker than the interaction reported in Experiment 1 (where $\eta_p^2 = 0.66$), and was over- rather than under-additive. In Figure 2a it can be seen that there was a clear ceiling effect in the *high mean – low variance* condition of Experiment 1, which might be the cause of this interaction. Visual inspection of Figure 6a demonstrates that this ceiling effect was effectively dealt with in Experiment 2 by increasing the difficulty of the task. Crucially, follow-up contrasts focusing on the two medium difficulty conditions revealed that accuracy was effectively matched across the *low mean – low variance* and the *high mean – high variance* conditions: M = 72.9%, $CI_{95\%}$ [71.3, 75.8], versus M = 73.0%, $CI_{95\%}$ [71.4, 76.3], $|t| < 1$, BF = .15. On average, C values in the low mean condition at the end of the experiment (M = .515 for 'red' stimuli, range: .502 to .528) did not differ from the value that was chosen at the start of the experiment (M_{red} = .515), $|t| < 1$, BF = .16. Given that the task was unspeeded, we did not observe any reliable differences in reaction time, all ps > .157, all BFs < .55.

Confidence judgments

Confidence resolution. Logistic regression models were fitted for each participant separately predicting accuracy based on the level of confidence. Positive slopes were observed for 47 out of 50 fits, however, these were only significant for 29 participants ($p < .05$), whereas they did not reach significance ($p > .05$) for the other 21 participants. When the factors mean (high or low) and variance (high or low) were additionally entered into the regression, positive slopes for confidence were observed for 47 out of 50 participants. For seventeen participants these were significant ($p < .05$) in the expected direction. On the group level, after dividing trials into quintile bins according to confidence (for each participant separately), we observed a monotonic increase in accuracy with the level of confidence (Figure 6B).

Average Confidence. The effect of mean and variance on z-scored confidence ratings was assessed on the data of the forced choice trials using a 2 (mean: low or high) by 2 (variance: low or high) repeated measures ANOVA. Here we report the results of confidence on all trials; the results for correct trials only, which did not differ materially from those reported here, can be found in the Supplementary Materials. Both mean, $F(1,49) = 31.17$, $p < .001$, $\eta_p^2 = 0.39$, BF = 322, and variance, $F(1,49) = 22.67$, $p < .001$, $\eta_p^2 = 0.32$, BF = 3.23e+07, influenced confidence ratings. There was also a significant interaction between these factors, $F(1,49) = 14.28$, $p < .001$, $\eta_p^2 = 0.22$, BF = 1.82, although the evidence was weak. Confidence was highest in the easy condition (M = 0.04, $CI_{95\%}$ [-0.04, 0.12]) and lowest in the hard condition (M = -0.21, $CI_{95\%}$ [-0.29, -0.13]). As predicted, confidence ratings were numerically lower in the *high mean – high variance* condition (M = -0.17, $CI_{95\%}$ [-0.25, -0.09]), than the *low mean – low variance* condition (M = -0.11, $CI_{95\%}$ [-0.20, -0.01]), but the effect was only marginally significant in a two-tailed contrast, $t(49) = 1.79$, p = .079, BF = 0.67 (Figure 6C), perhaps due to greater between-subject variability in confidence judgments in this experiment with an overall more difficult version of the task (see Supplementary Materials). When we analyzed median confidence, as suggested by a reviewer, the difference in confidence between the two medium difficulty conditions was clearer: $t(49) = 2.44$, p = .018, BF = 2.27. Finally,

individual differences in confidence between the two conditions did not reliably correlate with individual differences in accuracy between them, $r(48) = .15$, $p = .296$, $BF = .19$.

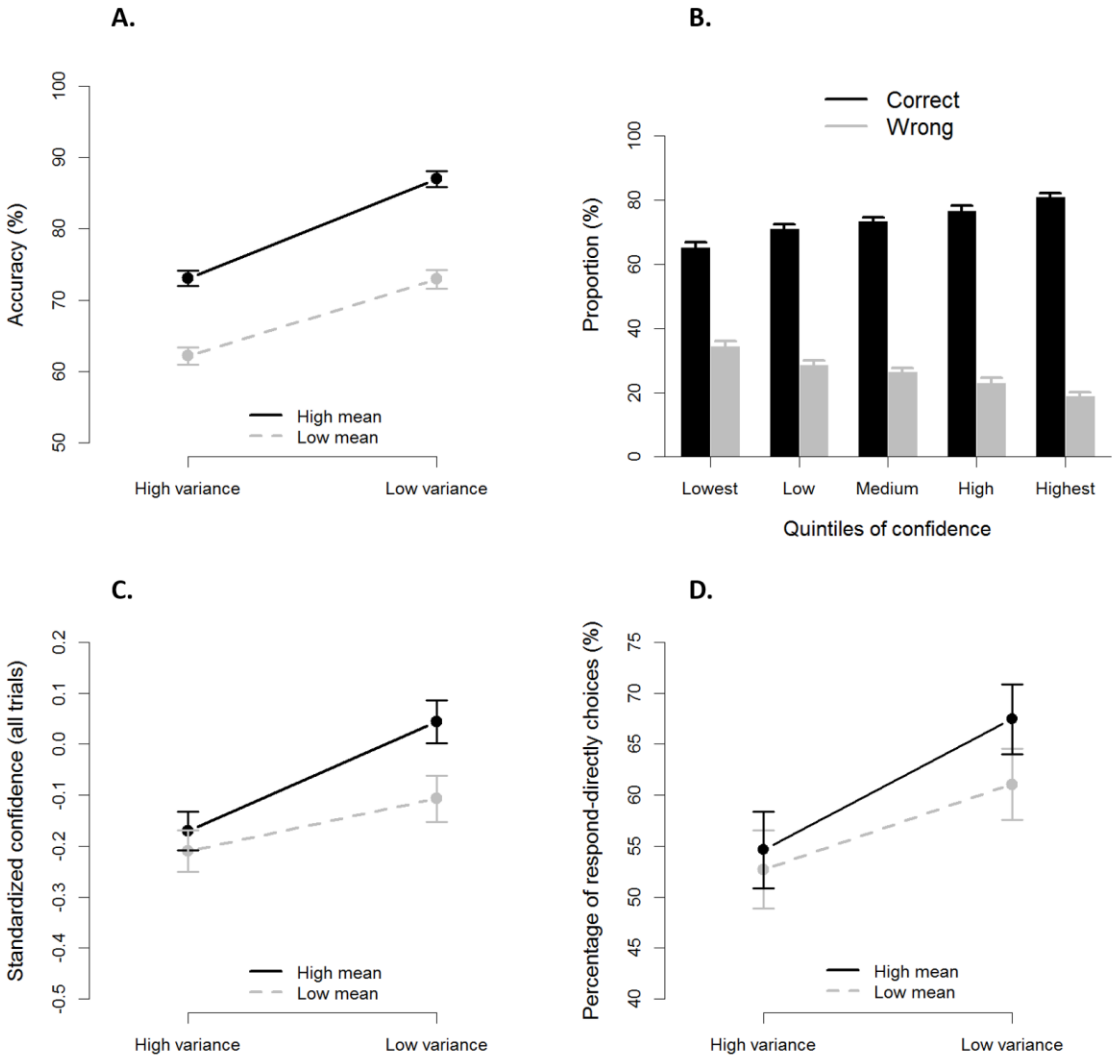

Figure 6. A. *First-order task performance, irrespective of the level of confidence, calculated from forced choice trials. The staircase procedure was successful in matching performance across the* high mean – high variance *and the* low mean – low variance *conditions.* ***B.*** *Distribution of objective accuracy as a function of confidence, calculated from forced choice trials.* ***C.*** *Mean of standardized confidence, calculated from forced choice trials.* ***D.*** *Percentage of trials where participants chose to give their response rather than to see the stimulus again, separately for each category, calculated from free choice trials. Error bars indicate standard errors of the mean.*

Information seeking

On free choice trials, participants on average chose to the see the stimulus again in an easier version on 41.0% of the trials. This rate is similar to that observed in Experiment 1 (37.9%), and once again there was considerable variability between participants in this percentage (range 1.5% - 96.5%), that did not correlate reliably with individual differences in overall mean confidence on forced choice trials, $r(48) = -.12$, $p = .388$, $BF = .16$, or accuracy on forced choice trials, $r(48) = -.03$, $p = .816$, $BF = .11$.

To examine how evidence mean and variance influenced strategic information-seeking behavior, the percentage of see again choices in free choice trials was calculated separately for each trial type and then subjected to a 2 (mean: low or high) by 2 (variance: low or high) repeated measures ANOVA. Effects of both mean, $F(1,49) = 28.830$, $p < .001$, $\eta_p^2 = 0.37$, $BF = 95$, and variance, $F(1,49) = 46.65$, $p < .001$, $\eta_p^2 = 0.49$, $BF = 2.40e+14$, were significant, as well as the interaction between the two factors, $F(1,49) = 9.99$, $p = .003$, $\eta_p^2 = 0.17$, $BF = 1.47$, although the evidence for the latter was weak. As expected, in the difficult condition participants asked to see the stimulus again most frequently ($M = 47.3\%$, $CI_{95\%}$ [39.6, 55.0]), whereas in the easy condition this occurred less often ($M = 32.5\%$, $CI_{95\%}$ [25.7, 39.4]). This difference was consistently observed, though numerically it was much smaller than in Experiment 1, where the task was easier overall and where see again choices were less evenly distributed across conditions. Crucially, replicating Experiment 1, participants more often chose to see the stimulus again in the *high mean – high variance* condition ($M = 45.4\%$, $CI_{95\%}$ [37.8, 52.9]), compared to the *low mean – low variance* condition ($M = 38.9\%$, $CI_{95\%}$ [31.9, 46.0]), $t(49) = 4.47$, $p < .001$, $BF = 468$ (see Figure 6D). Moreover, there was a reliable correlation across participants between differences in confidence and differences in see again choices across the two medium difficulty conditions, $r(48) = .36$, $t(48) = 2.65$, $p = .011$, $CI_{95\%}$ [.08, .58], $BF = 2.78$ (Figure 7). Interestingly, although not designed with this contrast in mind, the data of Experiment 2 provide further support for a coupling between confidence and information seeking in relation to the contrast between the two high variance conditions. In particular, whereas the *high mean – high variance* and the *low mean – high variance* conditions show marked and highly consistent differences in accuracy, $t(49) = 11.60$, $p < .001$, $BF > 5.64e+12$ (Figure 6a), there was only inconclusive evidence for a difference in see again choices between the two conditions, $t(49) = 1.868$, $p = .067$, $BF = 0.76$ (Figure 6d). Consistent with our proposed coupling between confidence and information sampling, confidence also differed only modestly between these conditions, $t(49) = 2.58$, $p = .013$, $BF = 3.04$, (Figure 6c).

To rule out that participants more often choose to view the stimulus again in the *high mean – high variance* condition than in the *low mean – low variance* condition simply because it was more helpful to do so for these trials, we compared these two conditions for see-again trials only. Due to a lack of at least 10 trials in one of these cells, only 42 participants were included in this analysis. Crucially, after participants asked to see the stimulus again, accuracy still did not differ reliably between the two medium difficulty conditions, $|t| < 1$, $BF = .25$ (*high mean – high variance* condition: $M = 86.0\%$, $CI_{95\%}$ [82.4, 89.6]; *low mean – low variance* condition: $M = 84.1\%$, $CI_{95\%}$ [79.9, 88.4]).

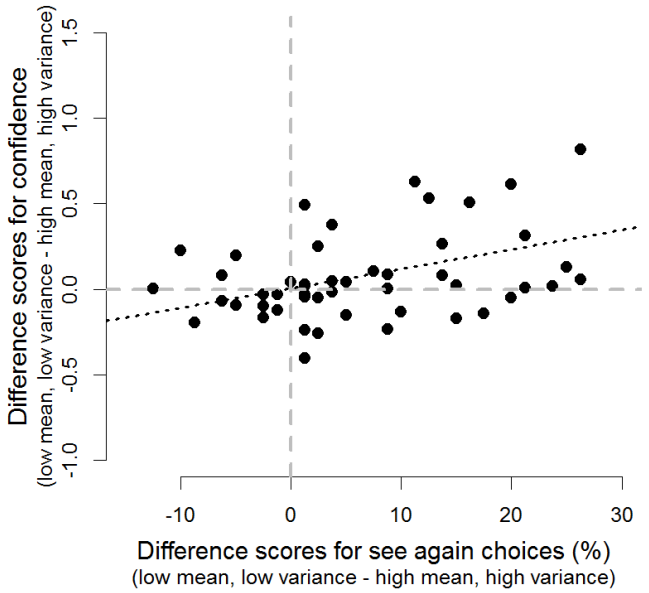

Figure 7. *Scatterplot showing the relationship between the difference scores for the two medium conditions for confidence judgments and see again choices.*

Metacognitive ability

We next investigated whether the impact of evidence variability on confidence judgments and see again choices relates to individual differences in metacognitive ability. Averaged across conditions, mean *d'* was 1.30 , $CI_{95\%}$ [1.17, 1.43], and mean meta-*d'* was 0.50, $CI_{95\%}$ [0.39, 0.62], resulting in a ratio meta-*d'*/*d'* of 0.41, $CI_{95\%}$ [0.32, 0.49]. Contrary to the results of exploratory analyses in Experiment 1, individual differences in meta-*d'*/*d'* ratio did not correlate reliably with differences in confidence between the two medium conditions, $r(48) = .08$, $p = .585$, $CI_{95\%}$ [-.20, .35], $BF = .13$, nor with differences in see again choices between these conditions, $r(48) = .14$, $p = .323$, $CI_{95\%}$ [-.14, .40], $BF = .18$ (Figure 8). This lack of correlation did not result from the reduced correspondence between confidence and accuracy compared to Experiment 1 (see *confidence resolution*): both correlations were far from significant in the subset of 29 participants for whom confidence significantly predicted accuracy (correlation involving the difference scores for confidence: $r(27) = -.16$, $p = .41$, $BF = .20$, and the correlation involving the difference scores for see again choices: $r(27) = -.15$, $p = .43$, $BF = .20$).

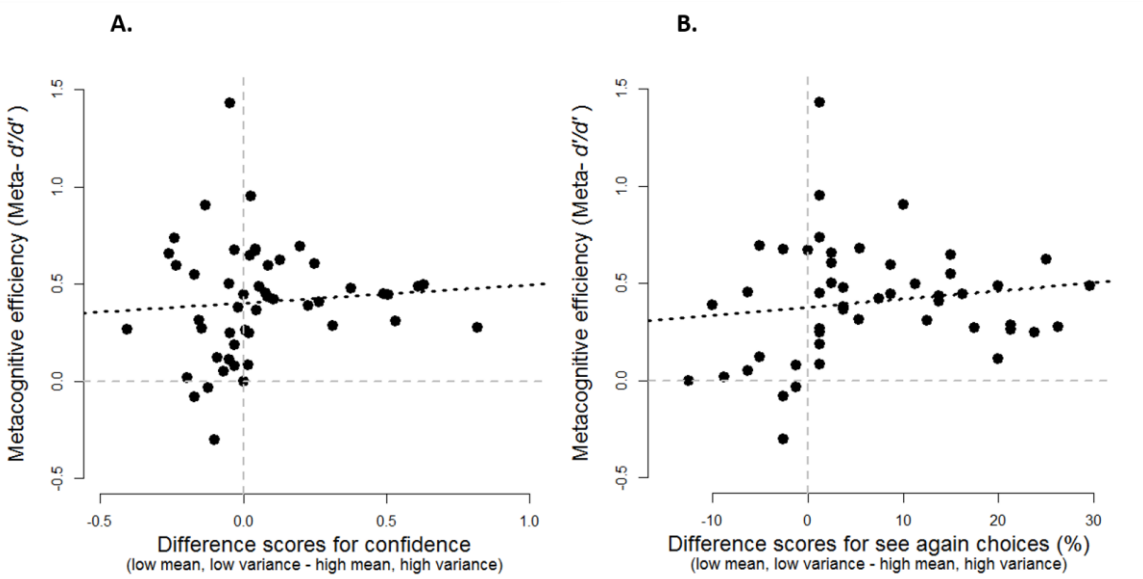

Figure 8. *Relationship between metacognitive ability, quantified as meta-d'/d', which was calculated from forced choice trials, and the difference between the two medium condition in both confidence judgments (A) and see again choices (B), which were both calculated from the free choice trials.*

A unique contribution of confidence

The same mixed regression model as in Experiment 1 was fit to the data, with random slopes for confidence, variance and accuracy. Similar to Experiment 1, confidence (β = -.10), t = -4.47, p < .001, variance (β = .07), t = 5.24, p < .001, and accuracy (β = .12), t = 3.91, p < .001, all affected see again choices, whereas evidence mean did so less consistently, p = .084. This finding demonstrates that confidence predicts see again choices over and above the other experimental variables.

To perform the same causal mediation analysis as in Experiment 1, the same two mixed models were fit to the data, with random slopes for the factor variance in the mediator model, and random slopes for confidence, variance and accuracy in the outcome model. The results showed that from the total effect of variance on see again choices (β = .084, $CI_{95\%}$ [.055, .110]), p < .001, there was 17.7% of total variance that was mediated by confidence (β = .015, $CI_{95\%}$ [.007, .023]), p < .001, conditional on mean accuracy and evidence mean (see Figure 9). Thus, similar to Experiment 1, this finding is in line with our hypothesis that confidence mediates between evidence variability and see again choices.

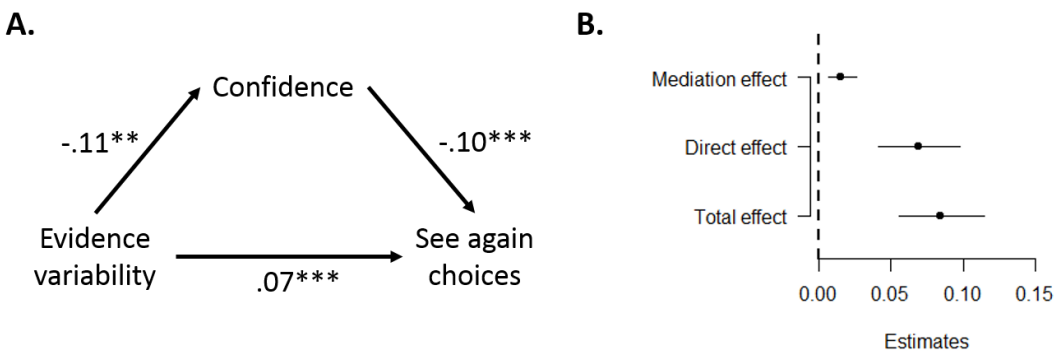

Figure 9. A. *Regression coefficients obtained from the mediation and the outcome mixed regressions models fit to the data of Experiment 2. Note: *** = p < .001. Degrees of freedom for calculating significance are based on Satterthwaite's approximation.* ***B.*** *Causal mediation analysis. Error bars reflect quasi-Bayesian 95% confidence intervals.*

Comparing forced choice and free choice trials

As a manipulation check, we again confirmed that the opportunity to see the stimulus again led to improved performance and increased confidence. Simple *t*-tests revealed that participants were on average more accurate ($M_{\textit{free choice – forced choice}}$ = 5.92%, $CI_{95\%}$ [4.58, 7.26], $t(49) = 8.90$, $p < .001$, $BF = 1.09e+09$) and more confident ($M_{\textit{free choice – forced choice}}$ = 0.13, $CI_{95\%}$ [0.08,0.17], $t(49) = 6.02$, $p < .001$, $BF = 67299$) in the free choice condition (in which they could choose to see the stimulus again) compared to the forced choice condition. Contrary to Experiment 1, however, the increase in accuracy in free choice trials relative to forced choice trials was not different between the *high mean – high variance* ($M_{\textit{free choice – forced choice}}$ = 6.73%, $CI_{95\%}$ [4.57, 8.89]) and the *low mean – low variance* ($M_{\textit{free choice – forced choice}}$ = 6.74%, $CI_{95\%}$ [4.23, 9.25]) conditions, $|t| < 1$, $BF = .15$. This observation is noteworthy, because even though participants more often requested to see the stimulus again in the *high mean – high variance* condition, and seeing the stimulus again improved performance, this did not improve accuracy more than it did so in the *low mean – low variance* condition. Also the increase in confidence in free choice relative to forced choice trials was not different between the two medium difficult conditions, $t(49) = -1.34$, $p = .18$, $BF = .36$.

We further found that performance was better on trials in which participants waived the option of seeing the stimulus again (M = 77.7%, $CI_{95\%}$ [74.8, 80.7]), compared to forced decision trials (M = 73.8%, $CI_{95\%}$ [71.8, 75.8]), $t(49) = 4.57$, $p < .001$, $BF = 629$. However, again unlike the results of Experiment 1, this was not different between the two medium difficulty conditions, $F < 1$, $\eta_p^2 < .01$, $BF = 0.15$ (*high mean – high variance* condition: $M_{\textit{choose to respond – forced}\text{ decision}}$ = 4.4%, $CI_{95\%}$ [1.3, 7.4], $t(49) = 2.91$, $p = .005$, BF = 6.38; *low mean – low variance* condition: $M_{\textit{see again – forced}\text{ decision}}$ = 4.1%, $CI_{95\%}$ [1.4, 6.8], $t(49) = 3.02$, p = .004, $BF = 8.27$), perhaps reflecting the more even distribution of see again choices across conditions in this experiment with a more difficult task overall. The same pattern was observed for confidence judgments. Participants were more confident when they waived additional evidence in the free choice condition (M = 0.03, $CI_{95\%}$ [-0.06, 0.12]) compared to being forced to respond (M = -0.11, $CI_{95\%}$ [-0.18,-

0.04]), $t(49) = 4.14$, $p < .001$, $BF = 173$, but this increase did not differ reliably between the two medium difficulty conditions, $F(1,49) = 1.80$, $p = .185$, $\eta_p^2 = 0.03$, $BF = .17$.

Discussion

This study provides novel evidence that subjective confidence influences strategic decision making, such that people seek further information when they lack confidence in an initial choice. Our critical manipulation was to contrast conditions matched for objective accuracy but differing in confidence, thus breaking the confound between confidence and objective evidence quality that complicates most existing paradigms. Crucially, participants solicited additional information before committing to a decision more often in a condition with high evidence variance (associated with lower confidence) than a condition with low mean (associated with higher confidence), despite matched objective accuracy.

Models of decision making and confidence

The dominant characterization of decision making is in terms of continuous accumulation of evidence (Gold & Shadlen, 2007; Ratcliff & McKoon, 2008). Notably, in this framework, deciding whether to sample more information from the environment before committing to a decision depends primarily on whether the level of accumulated evidence has reached a predefined threshold. It has been argued that confidence can also be understood as reflecting the accumulated evidence for a decision (Kiani & Shadlen, 2009; Pleskac & Busemeyer, 2010), with dissociations between confidence and accuracy thought to arise because confidence judgments are typically collected only after a decision has been made (Van Den Berg et al., 2016b) allowing time for further evidence accumulation after the initial choice (Resulaj, Kiani, Wolpert, & Shadlen, 2009; Moran, Teodorescu, & Usher, 2015). This explanation, however, cannot explain the dissociation between confidence and accuracy reported here, because responses were unspeeded and participants provided their decision and confidence judgments in a single response.

In the present study, two conditions were created that were matched for accuracy but differed in subjective confidence: Confidence was lower in a condition with high evidence variability relative to a condition with low evidence mean. This observation is already difficult to interpret within evidence accumulation models, which predict a close link between accuracy and confidence (see Yeung & Summerfield, 2012, for discussion). However, our crucial observation was that the decision to seek more information tracked subjective confidence, not objective accuracy: Participants consistently chose to see the stimulus again more often when evidence variability was high than when evidence mean was low. This effect was observed in two experiments differing markedly in overall task difficulty, in datasets collected in different experimental settings, attesting to the robustness of the effect. The effect varied systematically across individuals according to the sensitivity of their confidence judgments to evidence variability.

Thus, the decision whether or not to sample more information related to subjective evaluations of accuracy, rather than objective performance. As such, our findings indicate a distinction between systems representing the decision variable that underpins choice (e.g., Donner, Siegel, Fries, & Engel, 2009), and systems that compute confidence based on a noisy read-out of these decision processes (De Martino et al., 2013). Indeed, our findings are naturally explained by theories assuming that confidence depends on

computations that are at least partly distinct from those guiding the initial choice, with potential input from other sources of information (Pasquali et al., 2010). Our results indicate a role for these metacognitive computations in adaptive behavior (Meyniel et al., 2015), in line with findings that choices made with low confidence are less likely to be repeated (Folke, Jacobsen, Fleming, & De Martino, 2016) and have less influence on decisions whether or not to switch environments (Purcell & Kiani, 2016).

The present study exploited a previously observed effect of evidence variability to dissociate objective accuracy from subjective confidence, without specifically aiming to understand the cause of this dissociation. Nonetheless, in replicating previous findings using the same paradigm (Boldt et al., 2017) and related manipulations (Spence et al., 2015), our results stand in apparent contrast to recent findings that evidence mean more strongly influences choice and confidence than evidence weight (Kvam & Pleskac, 2016). In that study, however, evidence weight was operationalized as the amount of evidence, as distinct from the variability of presented evidence studied here. Given this difference, it is difficult to compare Kvam & Pleskac's findings to ours. Future work could usefully attempt to unravel the interplay between evidence mean, weight and variability in shaping decisions, confidence and information seeking.

The importance of replication

In Experiment 1, we observed a correlation between individual differences in metacognitive ability and our key effects. However, the sample size was not chosen to examine correlations, and caution is warranted in drawing conclusions from such exploratory analyses (Lindsay, 2015). Moreover, whereas Experiment 1 had high power to detect confidence differences between the two medium difficulty conditions (.99), it had much lower power to detect the observed correlations (e.g., power of .62 for the correlation in Figure 4B). In light of concerns that exploratory studies with low experimental power are prone to false positive findings, it is notable that the correlations observed in Experiment 1 were not replicated in our high-powered preregistered replication, despite its increased statistical power. It is therefore unlikely that the significant correlations reported in Experiment 1 are meaningful (Simonsohn, 2015). Overall, these findings highlight how good research practice can help to shed better light on the robustness of effects, using high-powered preregistered replications to rigorously examine key findings (here: the influence of confidence on decision making) and examine the validity of effects observed in exploratory analyses.

Limitations

We interpret our findings as showing that confidence influences strategic information seeking. However, an alternative interpretation is that confidence and information seeking are only indirectly linked via their relationship with a third, correlated variable. Specifically, as suggested by a reviewer, it could be that high variance stimuli have two separate effects: reducing confidence and cueing the need to seek further information (cf. Le Pelley, 2012). As described above, our data favor the former interpretation because confidence strongly mediates the relationship between evidence variability and information

seeking, and confidence explains differences in information seeking over and above effects of evidence variability. This mediation effect is far from trivial, given that the values for confidence and see again choices are based on different parts of the data (forced choice versus free choice), and that see again choices are sampled earlier in time than confidence judgments. Nevertheless, we acknowledge that our results cannot conclusively rule out the latter account. However, converging conceptual considerations strongly favor our interpretation. In particular, our account is consistent with the foundational hypothesis that confidence (like other metacognitive representations) supports adaptive behaviors that are flexible and, crucially, generalizable: Any factor that decreases confidence can produce the same adaptive behavior of information seeking. In contrast, the alternative account provides no account of why confidence should be explicitly represented and, crucially, suggests an account of adaptive behavior that is implausibly rigid: It implies that people would need to learn separately to seek information in an indefinitely large number of conditions: not only when evidence variability is high, but also when evidence mean is low, stimulus presentation is brief, lighting conditions are poor, attention has waned, etc. Thus, for example, this hypothesis cannot explain the finding in our data that subjective confidence likewise mediates the relationship between mean evidence and information seeking (see Supplemental Materials), a finding that falls out naturally from our interpretation.

Although we argue that our interpretation in terms of confidence provides the most parsimonious explanation of our data, we acknowledge that deeper understanding of the mechanisms by which evidence variability affects confidence will be important to further our understanding of this effect, for example by ruling out that the dissociation between confidence and accuracy emerges from differences in scale-dependent interactions (Wagenmakers, Krypotos, Criss, & Iverson, 2012). Similarly, future work might usefully investigate how seeing additional evidence affects performance. Although we found that additional evidence was equally informative in both medium difficulty conditions, this could only be examined when participants actively chose to sample more information. By introducing a condition where participants are forced to sample more evidence on some trials, it could be examined whether the value of additional evidence depends on the degree of variability. More generally, this might help to shed light on the question whether our proposed link between evidence variability, confidence and information seeking is adaptive, or a suboptimal bias.

Conclusion

In the current study, subjective confidence was dissociated from decision accuracy. This enabled us to show that the strategic choice to seek information before committing to a decision is determined by the subjective evaluation of accuracy, as reflected in confidence ratings, rather than by objective accuracy.

Acknowledgments

The authors would like to thank Tom Verguts and Cristian Buc Calderon for helpful comments on an earlier draft.

Author contributions

K.D., A.B., and N.Y. designed the study. K.D. performed testing and data analysis. K.D. drafted the paper and A.B. and N.Y. provided critical revisions. All authors approved the final version of the manuscript for submission.

Funding

This work was supported by grants of the Research Foundation Flanders, Belgium (FWO-Vlaanderen) awarded to K.D. (grant numbers 11H3415N and V447115N), and by an Economic and Social Research Council UK PhD studentship to A.B.

Author note

The preregistration protocol and the raw data of both Experiments are publicly available on the OSF platform (https://osf.io/sh3wy/).

References (40/40)

Bates, D. M., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-Regulated Learning: Beliefs, Techniques, and Illusions. *Annual Review of Psychology*, *64*(1), 417–444. https://doi.org/10.1146/annurev-psych-113011-143823

Boldt, A., de Gardelle, V., & Yeung, N. (2017). The Impact of Evidence Reliability on Sensitivity and Bias in Decision Confidence. *Journal of Experimental Psychology: Human Perception and Performance*, *43(8), 1520-1531*. https://doi.org/10.1037/xhp0000404

Boldt, A., & Yeung, N. (2015). Shared Neural Markers of Decision Confidence and Error Detection. *Journal of Neuroscience*, *35*(8), 3478–3484. https://doi.org/10.1523/JNEUROSCI.0797-14.2015

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, *65*(3), 245–81.

Call, J., & Carpenter, M. (2000). Do apes and children know what they have seen? *Animal Cognition*, *3*(4), 207–220. https://doi.org/10.1007/s100710100078

de Gardelle, V., & Summerfield, C. (2011). Robust averaging during perceptual judgment. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(32), 13341–6. https://doi.org/10.1073/pnas.1104517108

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

Donner, T. H., Siegel, M., Fries, P., & Engel, A. K. (2009). Buildup of Choice-Predictive Activity in Human Motor Cortex during Perceptual Decision Making. *Current Biology*, *19*(18), 1581–1585. https://doi.org/10.1016/j.cub.2009.07.066

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

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science*, *329*(5998), 1541–3. https://doi.org/10.1126/science.1191883

Folke, T., Jacobsen, C., Fleming, S. M., & De Martino, B. (2016). Explicit representation of confidence informs future value-based decisions. *Nature Human Behaviour*, *2*(November), 17–19. https://doi.org/10.1038/s41562-016-0002

Foote, A. L., & Crystal, J. D. (2007). Metacognition in the Rat. *Current Biology*, *17*(6), 551–555. https://doi.org/10.1016/j.cub.2007.01.061

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, *30*, 535–561. https://doi.org/10.1146/annurev.neuro.29.051605.113038

Hampton, R. R. (2001). Rhesus monkeys know when they remember. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(9), 5359–5362. https://doi.org/10.1073/pnas.071600998

Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science (New York, N.Y.)*, *324*(5928), 759–764. https://doi.org/10.1126/science.1169405

Kvam, P., & Pleskac, T. (2016). Strength and weight : The determinants of choice and confidence. *Cognition*, *152*(April), 170–180. https://doi.org/10.1016/j.cognition.2016.04.008

Le Pelley, M. E. (2012). Metacognitive monkeys or associative animals? Simple reinforcement learning explains uncertainty in nonhuman animals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(3), 686–708. https://doi.org/10.1037/a0026478

Lindsay, D. S. (2015). Replication in Psychological Science. *Psychological Science*, *26*(12), 1827–1832. https://doi.org/10.1177/0956797615616374

Macdonald, J. S. P., Mathan, S., & Yeung, N. (2011). Trial-by-trial variations in subjective attentional state are reflected in ongoing prestimulus EEG alpha oscillations. *Frontiers in Psychology*, *2*(MAY), 1–16. https://doi.org/10.3389/fpsyg.2011.00082

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive

sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422–30. https://doi.org/10.1016/j.concog.2011.09.021
Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review*, *15*(1), 174–179. https://doi.org/10.3758/PBR.15.1.174
Meyniel, F., Sigman, M., & Mainen, Z. F. (2015). Confidence as Bayesian Probability: From Neural Origins to Behavior. *Neuron*, *88*(1), 78–92. https://doi.org/10.1016/j.neuron.2015.09.039
Moran, R., Teodorescu, A. R., & Usher, M. (2015). Post choice information integration as a causal determinant of confidence: Novel data and a computational account. *Cognitive Psychology*, *78*, 99–147. https://doi.org/10.1016/j.cogpsych.2015.01.002
Morey, R. D., & Rouder, J. N. (2014). BayesFactor: Computation of Bayes factors for common design.
Nuijten, M. B., Wetzels, R., Matzke, D., Dolan, C. V, & Wagenmakers, E.-J. (2015). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review*, *47*(1), 85–97. https://doi.org/10.3758/s13423-012-0295-x
Pasquali, A., Timmermans, B., & Cleeremans, A. (2010). Know thyself: metacognitive networks and measures of consciousness. *Cognition*, *117*(2), 182–90. https://doi.org/10.1016/j.cognition.2010.08.010
Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review*, *117*(3), 864-901. https://doi.org/10.1037/a0022399
Purcell, B. a., & Kiani, R. (2016). Hierarchical decision processes that operate over distinct timescales underlie choice and changes in strategy. *Proceedings of the National Academy of Sciences*, 201524685. https://doi.org/10.1073/pnas.1524685113
Ratcliff, R., & McKoon, G. (2008). The Diffusion Decision Model : Theory and Data for Two-Choice Decision Tasks. *Neural Computation*, *20*, 873–922.
Resulaj, A., Kiani, R., Wolpert, D. M., & Shadlen, M. N. (2009). Changes of mind in decision-making. *Nature*, *461*(7261), 263–266. https://doi.org/10.1038/nature08275
Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*(January 2015), 1–11. https://doi.org/http://dx.doi.org/10.2139/ssrn.2259879
Smith, J. D., Strote, J., Egnor, R., Schull, J., McGee, K., & Erb, L. (1995). The uncertain response in the bottlenosed dolphin (Tursiops truncatus). *Journal of Experimental Psychology: General*, *124*(4), 391–408. https://doi.org/10.1037/0096-3445.124.4.391
Spence, M. L., Dux, P. E., & Arnold, D. H. (2016). Computations Underlying Confidence in Visual Perception Computations Underlying Confidence in Visual Perception. *Journal of Experimental Psychology : Human Perception and Performance*, *42*(5), 671-82.
Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). mediation: R Package for Causal Mediation Analysis. *Journal of Statistical Software*, *59*(5), 1–38. https://doi.org/10.18637/jss.v059.i05
van den Berg, R., Zylberberg, A., Kiani, R., Shadlen, M. N., & Wolpert, D. M. (2016a). Confidence Is the Bridge between Multi-stage Decisions. *Current Biology*, *26*(23), 3157–3168. https://doi.org/10.1016/j.cub.2016.10.021
Van Den Berg, R., Anandalingam, K., Zylberberg, A., Kiani, R., Shadlen, M. N., & Wolpert, D. M. (2016b). A common mechanism underlies changes of mind about decisions and confidence. *eLife*, 1–21. https://doi.org/10.7554/eLife.12192
Wagenmakers, E.-J., Krypotos, A.-M., Criss, A. H., & Iverson, G. (2012). On the interpretation of removable interactions: A survey of the field 33 years after Loftus. *Memory & Cognition*, *40*(2), 145–160. https://doi.org/10.3758/s13421-011-0158-0
Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. *Philosophical Transactions of the Royal Society of London*, *367*(1594), 1310–21. https://doi.org/10.1098/rstb.2011.0416