From local to global estimations of confidence in perceptual decisions

Quentin Cavalan, Jean-Christophe Vergnaud, Vincent de Gardelle

Title: From local to global estimations of confidence in perceptual decisions

Quentin Cavalan[1,2], Jean-Christophe Vergnaud[1], Vincent de Gardelle[1,2]

[1] Centre d'Economie de la Sorbonne, CNRS & Université Paris 1, France
[2] Paris School of Economics, Paris, France

Corresponding author: Quentin Cavalan, 57 rue de la commune de Paris, 93300 Aubervilliers, FRANCE. 0630593955. quentin.cavalan@hotmail.fr
Dataset link: https://osf.io/btsed/

Summary

Perceptual confidence has been an important topic recently. However, one key limitation in current approaches is that most studies have focused on confidence judgments made for single decisions. In three experiments, we investigate how these local confidence judgments relate and contribute to global confidence judgments, by which observers summarize their performance over a series of perceptual decisions. We report two main results. First, we find that participants exhibit more overconfidence in their local than in their global judgments of performance, an observation mirroring the aggregation effect in knowledge-based decisions. We further show that this effect is specific to confidence judgments and does not reflect a calculation bias. Second, we document a novel effect by which participants' global confidence is larger for sets which are more heterogeneous in terms of difficulty, even when actual performance is controlled for. Surprisingly, we find that this effect of variability also occurs at the level of local confidence judgments, in a manner that fully explains the effect at the global level. Overall, our results indicate that global confidence is based on local confidence, although these two processes can be partially dissociated. We discuss possible theoretical accounts to relate and empirical investigations of how observers develop and use a global sense of perceptual confidence.

Keywords: overconfidence, confidence-frequency effect, aggregation effect

JEL: D91

Introduction

Metacognition is broadly defined as the set of cognitive operations by which individuals monitor and control their own cognitive processes (Nelson & Narens, 1990). One instance of a metacognitive evaluation is the confidence that accompanies our decisions: on some occasions we are certain that we have the correct answer to a question, whereas on others we know that our decision was as good as a guess. The question of how observers rate their confidence is not new at all (Peirce & Jastrow, 1884), yet it has received a lot of interest recently especially in the domain of perceptual decision making (Mamassian, 2016), within the framework of signal detection theory (Green & Swets, 1966; Maniscalco & Lau, 2012; Fleming & Daw, 2017; Rausch & Zehetleitner, 2019; Mamassian & de Gardelle, 2021) or evidence-accumulation processes (Pleskac & Busemeyer, 2010; Balsdon et al., 2020; Desender et al, 2021; Pereira et al, 2022).

Understanding how observers monitor the quality of their own perceptual decisions by expressing confidence is important, not only because confidence provides a tool to study perceptual decisions, but also because such evaluations of confidence can impact behavior in several ways. For instance, confidence can be used to guide changes of mind (van den Berg et al., 2016) and to regulate the acquisition and use of external information (e.g. Desender et al., 2018; Balsdon et al., 2020). In the absence of external feedback, individuals can also use the confidence in their current decision as a teaching signal to drive learning (e.g. Guggenmos et al., 2016; Hainguerlot et al., 2018). Furthermore, confidence, typically defined as the subjective probability of being correct, may constitute a common currency with which distinct perceptual decisions or tasks can be compared (de Gardelle & Mamassian, 2014; de Gardelle et al, 2016; Baer & Odic, 2020) and prioritized within a single individual (e.g. Aguilar-Lleyda et al., 2020; Aguilar-Lleyda & de Gardelle, 2021; Carlebach &

Yeung, 2020), and even combined across individuals (e.g. Bahrami et al., 2010; Koriat, 2015; Massoni & Roux, 2017).

Yet, one issue that is not well understood regarding perceptual confidence is how observers may evaluate it not only over a single decision, but more globally over a given task. As much as the perceptual system seems able to entertain summary representations over sets of stimuli, for sensory features such as orientation, color or shape, or for more abstract features such as emotional content or value (see e.g. Whitney & Yamanashi-Lleib, 2018 for a review), visual confidence could also be aggregated across multiple items. In fact, to describe our performance or to predict our future performance in a task, relying on a global confidence about an ensemble of decisions may be more relevant than relying on a single past decision. The idea that our metacognitive system is capable of aggregating perceptual confidence over multiple decisions is consistent with recent work showing a confidence leak between consecutive decisions (Rahnev et al., 2015; Kantner et al., 2018; Aguilar-Lleyda et al., 2021). However, aside from a few exceptions (Lee et al., 2021; Rouault et al., 2019 and Rouault & Fleming, 2020), to date most studies on perceptual confidence have considered confidence judgments about a single perceptual decision at a time.

Although it has been overlooked in the domain of perceptual decisions, the idea of a global confidence has been already investigated in other domains. For knowledge-based decisions in particular, it has been established that global confidence over a series of decisions tends to be lower than the average of the local confidence ratings made for each decision (Gigerenzer, 1991; Sniezek & Buckley, 1991; Griffin & Tversky, 1992; Treadwell & Nelson, 1996). To explain this phenomenon, referred to as the “confidence-frequency” effect or the “aggregation effect” in the literature, some authors have suggested that global and local confidence judgments for quiz questions are based on distinct cues. Local judgments would be based on the perceived evidence on the current question whereas global judgments would reflect the estimation of performance in previous tests encountered by the participant (Gigerenzer, 1991; Griffin & Tversky, 1992). We note that this lower confidence

for global evaluations of performance has also been found in other types of tasks, e.g. when individuals must estimate their performance in a dexterity task (Stone et al., 2010) or their ability to infer gender from handwriting (Schneider, 1995). However, it has not been documented yet in the domain of perceptual decisions.

The present work had two main objectives. First, we wished to compare local and global confidence judgments made about perceptual decisions, to evaluate a possible aggregation effect in the perceptual domain. Second, we sought to assess more precisely how global confidence judgments are formed, and whether they can be dissociated from local confidence judgments. In this respect, as we would ask individuals to form a global judgment over a set of trials, the question arises of how they deal with the heterogeneity across trials within that set. To the best of our knowledge, this question has never been addressed at the level of global confidence judgments. To investigate this issue, we compared judgments of global confidence over sets of trials that are homogeneous or heterogeneous in terms of task difficulty.

With respect to those objectives, here we describe three experiments in which we engaged participants in a perceptual decision task, in which they rated their confidence after each decision (local judgments), or overall after a series of decisions (global judgments). Trials within a series were either homogeneous (low variability) or heterogenous (high variability) in terms of difficulty (Figure 1).

To anticipate our results, in experiment 1, we replicate the aggregation effect in the perceptual domain by showing that individuals are less confident for global than for local judgments. In experiment 2, we document a novel effect, the variability effect according to which global confidence is higher for sets with greater variability in terms of trial difficulty. Finally, in experiment 3, we provide evidence that the variability effect on global judgments comes from an initial effect on local judgments. Furthermore, we show that the aggregation

effect cannot be reduced to a bias in the way individuals compute the average over a sequence of numbers.

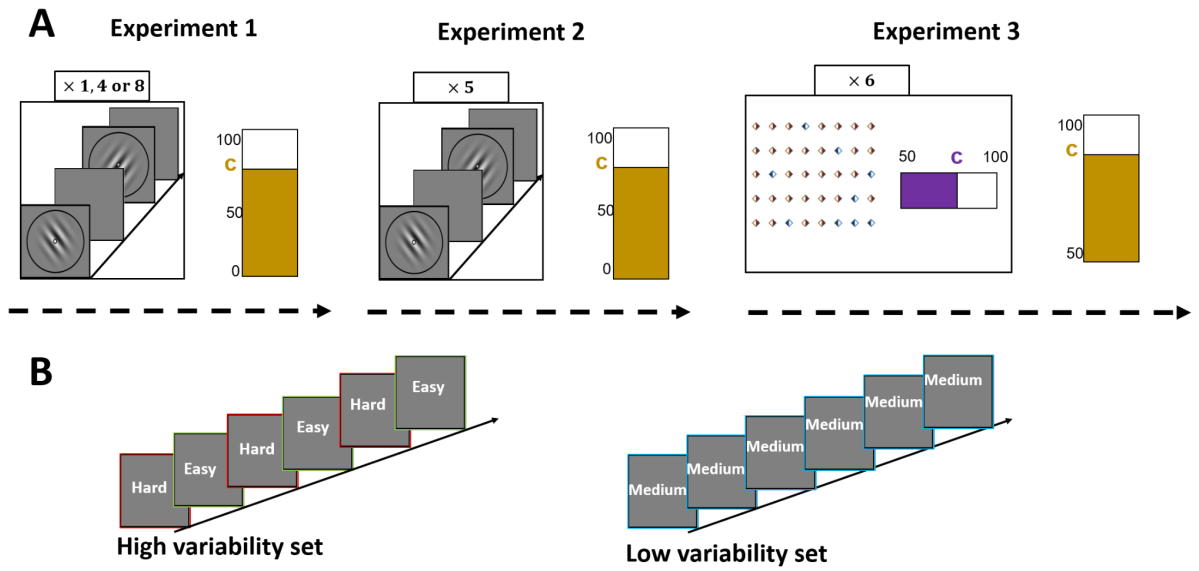

Figure 1. Manipulating set size and variability within a set of trials. (A) Perceptual task and confidence scale in our 3 experiments. In Experiment 1, participants reported their confidence over a set of size 1 (local confidence), 4 or 8 (global confidence). In Experiment 2, participants always reported their global confidence over 5 perceptual trials. In Experiment 3, they reported both local confidence after each trial, and global confidence over the last 6 trials. The perceptual task was an orientation discrimination task in Experiment 1 and 2, and a color numerosity task in Experiment 3. (B) Schematic representation of the variability manipulation featured in each experiment. Within each set, the difficulty of trials exhibits high variability (trials are either hard or easy) or low variability (all trials are of medium difficulty).

Experiment 1

The goal of experiment 1 was to compare participants' local and global evaluations of confidence, in the context of a perceptual decision task. On each trial, participants had to make a decision on a visual stimulus. In addition, trials were grouped in sets (of size 1, 4 or 8) and after each set, participants had to give their overall confidence regarding their decisions in the set. To evaluate whether the aggregation effect would occur for perceptual decisions, we then calculate overconfidence for each setsize, and compare it across setsizes. Moreover, in sets of 4 or 8 trials, the variability of the difficulty of stimuli within the set was manipulated.

Methods

Participants

Twenty adults (age range 18-24) were recruited. They were naïve with respect to the goal of the study. Three participants were removed from all the analyses, two because they performed the entire perceptual task at random (less than 55% of correct answers) and one because he performed the second half of the task at random.

A power analysis revealed that our sample size was above the 16 participants that were needed to detect (one tailed t-test, with power = .8 and error probability =.05) a similar aggregation effect size (d= $.66$) than Treadwell & Nelson (1996).

Written informed consent was obtained from all participants before the experiment. The research was non-invasive; it involved negligible risks and no collection of nominative/identifying information or health information. The study was approved by our local ethics committee at Paris School of Economics (IRB00010601, decision 2017-009).

Session

The session took place at the Parisian Experimental Economics Laboratory (LEEP) in 2019. First, participants received a first training with the perceptual task alone and no confidence evaluation (100 trials). They then received a second training that also included the metacognitive task (104 perceptual trials, divided in 24 sets). Finally, they completed the main part of the experiment (216 sets, divided in 36 blocks).

Perceptual task

For each trial, two Gabor stimuli were presented in succession (inter-stimulus interval: 500ms) at the center of the screen. Stimulus width was about 5° of visual angle.

Stimuli were presented at 40% contrast on a gray background, for 100ms. Observers had to judge whether the second stimulus was rotated “clockwise” or “counterclockwise” relative to the first stimulus, which was determined randomly at each trial with equal probability. The orientation of the first stimulus was tilted (+/- 4° around vertical), and the orientation of the second stimulus with respect to the first was controlled to achieve 75% of correct answers in the task (see Calibration). Observers reported their responses on a keyboard. Visual symbols were present on the screen to indicate the mapping between the stimulus categories and the response keys. Responses were not speeded and no feedback on performance was given.

Metacognitive task

After the initial training, the metacognitive task was introduced to participants. After a set of 1, 4 or 8 perceptual trials, participants had to give their global confidence in their answers for that set. This was done using a 0 to 100 scale where 100% confidence meant “sure to be correct to all answers” and 0% confidence meant “sure to be incorrect to all answers”. These confidence ratings were not speeded. The division of perceptual trials into sets was emphasized by a frame whose color changed from one set to the other, and a gauge located at the bottom and at the top of the screen indicating the number of total and remaining trials in the current set. This was done to ensure that participants recognized which set of trials their global confidence was supposed to be about.

This confidence rating was incentivized using a probability matching rule (Massoni et al, 2014). For each trial, the participant is offered an exchange between his response and a lottery ticket with an unknown probability P of success. The number P is generated randomly on each confidence judgment, and compared to the confidence response. If P (the success probability) is greater than the confidence, then the participant’s reward is determined by the lottery. If not, it is determined by the accuracy of the response. In our experiment, the reward was equal to 4 euros. For sets of 4 and 8 trials, the same mechanism was applied to each

trial, using the confidence rating given for the set, and participant's reward for the set was the mean reward over all the trials within that set. The mechanism was presented to participants as a way to maximize their earnings by providing accurate confidence ratings.

Instructions, examples, and a training phase with feedback (with 8 sets for each set size) were included to make sure that participants understood the mechanism. Participants then completed 936 trials (216 confidence judgments) in the main task. At the end of the experiment, 2 rounds were chosen randomly for payment and their earnings for those rounds were added to a fixed participation fee of 6 euros.

Calibration

During the first training, at each trial the orientation of the second stimulus was determined by randomly interleaved staircases aiming at 75% accuracy for clockwise and counterclockwise rotations. Specifically, we used adaptive staircases (accelerated stochastic approximation from Kesten, 1958). Four staircases were interleaved, with 2 staircases converging at 75% of "counterclockwise" responses and 2 at 75% of "clockwise" responses.

In the main part of the experiment, the difficulty levels of the stimuli were adjusted every 26 trials, to keep track of potential fluctuation in perceptual performance. To do so, we estimated a psychophysical curve describing the proportion of clockwise responses as a function of the difference in degrees between the two visual orientations, over the last 104 perceptual decisions. This curve was fitted with a cumulated Gaussian, using maximum likelihood. From this curve, we estimated two thresholds (x_{25} and x_{75}) corresponding to the stimulus levels for which the participant would have 75% of correct answers when the rotation was clockwise (x_{75}) or counterclockwise (x_{25}). These values were used in the next 26 trials, and so on and so forth.

Design

Our design involves a manipulation of setsize (the size of the set over which confidence was probed: 1, 4 and 8 trials) and a manipulation of variability (the variance of stimulus difficulty within the set: high vs. low) for sets larger than one. In 'low-variability sets and for the setsize 1 condition, the second orientation was either x_{25} or x_{75}. For 'high-variability sets, it was drawn from a normal distribution centered on x_{25} or x_{75} and with standard deviation $\sigma^{stim}=(x_{75}-x_{25})/4$. This design results in 6 conditions, which were presented in a random order within each block of the main experiment.

A proxy for local confidence based on ideal confidence

Local confidence was only measured for sets of 1 trial in experiment 1. To investigate how global confidence evaluations would depend on local confidence for setsizes larger than 1, we relied on a proxy for local confidence. Our proxy was based on the expected value of ideal local confidence, in a statistical model of perceptual decisions, where ideal local confidence is the subjective probability that a correct response was made on a given trial. According to Signal Detection Theory (Green & Swets, 1966), this subjective probability can be expressed mathematically as a random variable, whose distribution depends on the participant's sensitivity for the current trial and on the subjective prior probability over the different stimuli. Then, once the probability distribution of ideal local confidence is known, it is possible to evaluate its expected value given the stimulus strength and the accuracy of the perceptual decision in the current trial.

For each participant, we fitted this expected ideal confidence to the confidence ratings for setsize 1, with 2 free parameters to account for a linear mapping between expected ideal confidence and participant's reported confidence. We then used this fitted model to predict participant's local confidence in each trial for sets of size 4 and 8. This prediction corresponds to our proxy for local confidence. Finally, for each set, we averaged this local confidence proxy over the 4 or 8 trials of the set, to obtain a proxy of the average

local confidence for this set. More details on this local confidence proxy (including mathematical derivations for expected ideal confidence and goodness of fit) are presented in the Supplementary Material.

Analyses

When comparing means between two conditions, we perform t-tests and report systematically the mean and standard deviation of the difference, the t-statistic associated with its degrees of freedom, the p-value and Cohen's d as a measure of effect size. When comparing means between more than two conditions, we perform ANOVAs and report systematically the F-statistics associated with its degrees of freedom, the p-value and partial eta squared as a measure of effect size for each effect (main effects and interactions when applicable).

Transparency and openness

Data and code needed to reproduce all analyses in the paper are available and can be downloaded at https://osf.io/btsed/. Experiments were 1 and 2 were not pre-registered. Experiment 3 was pre-registered (more details in the interim discussion of Experiment 2).

Results

Perceptual performance

Before examining confidence evaluations, we first analyzed performance in the perceptual task, as performance would be an important determinant of confidence. Although performance was controlled in our study, it appeared that it differed between setsizes ($F(2,32)=6.91$, $p=0.0032$, $\eta_p^2=.357$). This was driven by a position effect, by which response accuracy was better at the beginning of the series (Figure 2A), irrespectively of setsize. A two-way analysis of variance (ANOVA) on performance, with setsize and trial position (first

vs. second half of the set) as within-participant factors indicated indeed a significant main effect of position ($F(1,16)=23.46$, $p<.001$, $\eta_p^2=.594$), with no significant main effect of setsize ($F(1,16)=1.43$, $p=.250$, $\eta_p^2=.082$), and no significant interaction ($F(1,16)=0.57$, $p=.463$, $\eta_p^2=.034$).

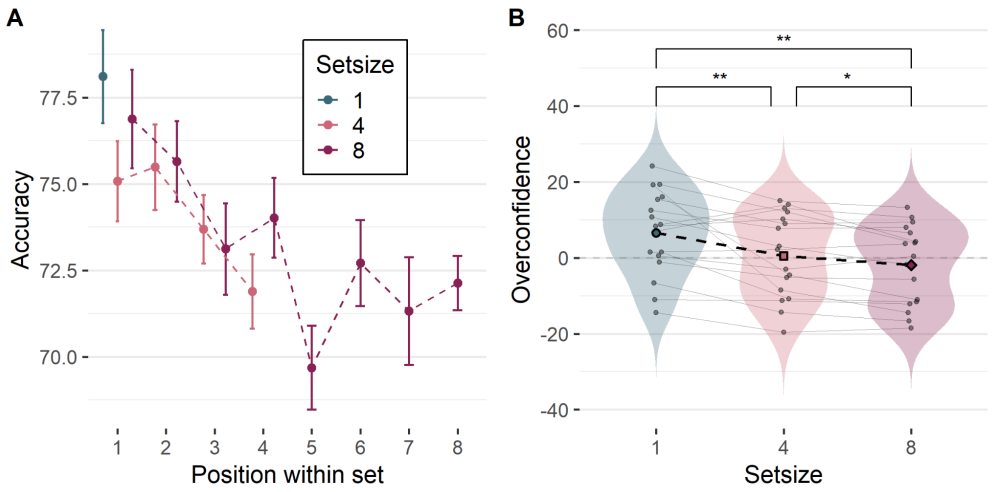

Figure 2. (A) Accuracy in the perceptual task against the position of the trial within the set, separately for each setsize. Error bars represent mean and s.e.m. across participants. (B) Distribution of overconfidence across participants in each setsize. Large dots and thick lines represent the mean overconfidence across participants. Small dots and thin lines represent individual data. Pairwise comparisons are performed using t-tests (***: p<.001, **: p<.01 *: p<.05: n.s.:p>.05). See Figure S2 in Supplementary Material to see overconfidence by setsize and variability conditions.

This higher performance for early trials in the set may reflect the increased arousal or attention level of participants at the beginning of the series. Based on these results, we explicitly included position as a modulator of performance in our model of ideal confidence (see Supplementary Material).

Effect of setsize and variability on overconfidence

We then looked at overconfidence, defined as confidence minus accuracy. To evaluate the variability manipulation on overconfidence we first focused on sets larger than 1, and conducted a two-way ANOVA with variability (low vs. high) and setsize (4 vs. 8) as within-participant factors. We found a significant main effect of setsize on overconfidence ($F(1,16)=7.40$, $p=.015$, $\eta_p^2=.316$), with no significant effect of variability ($F(1,16)=4.35$,

p=.053, η_p^2=.214) and no significant interaction between setsize and variability (F(1,16)=3.34,p=.086, η_p^2=.173). In other words, our manipulation of variability did not seem to have a strong impact on confidence, possibly because the experimental manipulation was not strong enough. In subsequent analyses, we thus pooled high and low variability conditions together, which enabled us to analyze all setsizes together (See Supplementary Material for a plot of overconfidence by variability conditions).

When analyzing overconfidence across all setsizes using an ANOVA, we found again a significant main effect of setsize on overconfidence bias (F(2,32)=10.47, p<.001, η_p^2=.396). Specifically, overconfidence decreased when set size increased (Figure 2B). All pairwise comparisons between setsizes were significant (setsize 1 vs. setsize 4: M_Dif=6.05, SD_Dif=8.52, t(16)=2.93, p=.010, d=.71; setsize 4 vs. setsize 8: M_Dif=2.38, SD_Dif=3.61, t(16)=2.72, p=.015, d=.66; setsize 1 vs. setsize 8: M_Dif=8.43, SD_Dif=9.93, t(16)=3.50, p=.003, d=.85). In other words, we found an aggregation effect for perceptual decisions: overconfidence was greater for local than for global confidence judgments. Note that since performance is lower for larger sets (because of the position effect described in section “Perceptual performance”), and that overconfidence is known to decrease with performance (“hard-easy” effect, Lichtenstein and Fischhoff, 1977), the aggregation effect that we found is likely to have been slightly underestimated.

When examining each setsize separately, we found that overconfidence was significantly positive when the setsize was 1 (Mean=6.57, t(16)=2.50, p=.024), but not when it was 4 (Mean=.51, t(16)=0.20, p=.842) or 8 (Mean=-1.87, t(16)=-0.73, p=.476). In other words, overconfidence was present only for local confidence judgments in this dataset.

Local confidence contributes to global confidence

Next, our goal was to assess whether local confidence contributed to global confidence, above and beyond accuracy. However, because we did not ask for local confidence ratings after each trial in sets of size 4 and 8, we relied on a proxy for local confidence in those trials, which was based on a SDT model of confidence and on participant's local confidence in setsize 1 (see Methods and Supplementary Material for details).

Global confidence was significantly correlated with our proxy (Figure 3), both across participants (Cor_4=.75, p<.001; Cor_8=.65, p=.005) and within each participant (Mean_Cor4=.22, t(16)=4.46, p<.001; Mean_Cor8=.26, t(16)=8.11, p<.001). These correlations suggest that local confidence contributes to global confidence ratings. Moreover, the aggregation effect was present in all participants and setsizes, as indicated by the fact that data points in Figure 3A lie below the diagonal. Besides, in this analysis the aggregation effect could not be explained by confounding differences (in accuracy and stimulus difficulty) between setsize conditions. To confirm that the relation between global and local confidence was not simply the result of fluctuations in accuracy, we regressed global confidence against accuracy for each participant and setsize, and we examined the residuals of this regression. We found that these residuals were still significantly correlated with our proxy of average local confidence, on average across participants (Mean_Cor4=.094, t(16)=3.24, p<.001; Mean_Cor8=.098, t(16)=4.91, p<.001). Note that these individual correlations can only be modest numerically, given that our proxy is only the expected value of confidence, given the stimulus strength and the accuracy of the perceptual decision in the current trial (see Supplementary Material for details).

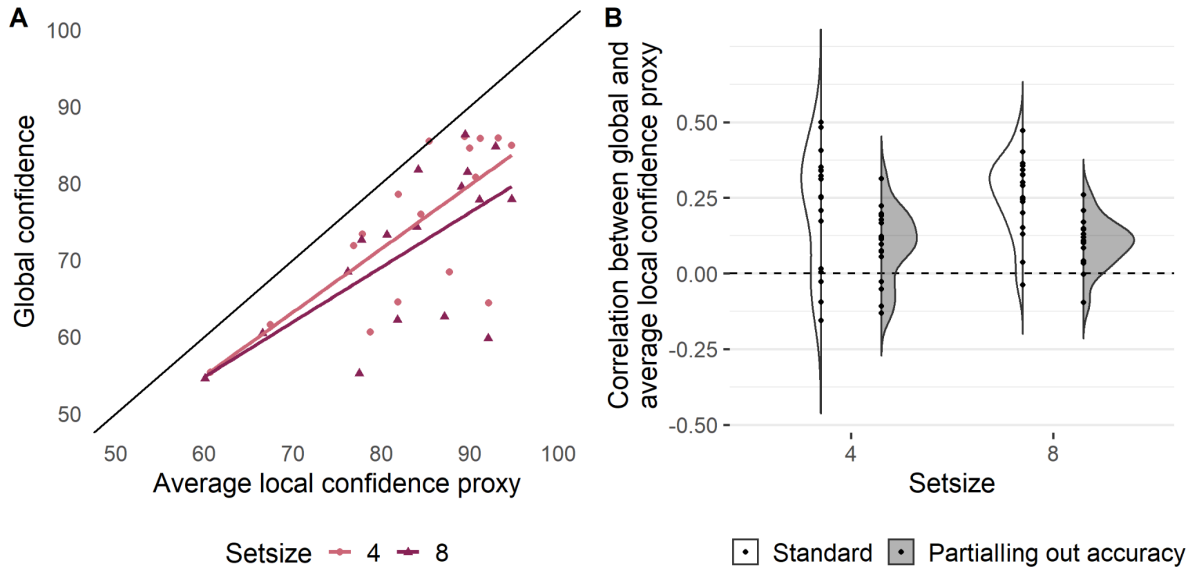

Figure 3. (A) Global confidence against our average local confidence proxy for setsize 4 and 8. Each dot represents a participant in a given setsize. Lines represent linear regressions across participants for each setsize. See main text for details on the computation of average local confidence proxy. (B) Distribution of the within-participant correlations between global confidence and average local confidence proxy, separately for setsize 4 and 8. The white and gray distributions correspond to the standard Pearson correlation and to the correlation partialling out accuracy, respectively. Each dot represents a single participant. See main text for details on how we partial out accuracy in the correlation.

Recency effect on global confidence

Additionally, we examined how global confidence was based on the different trials in the series. First, for each participant and setsize, we estimated how much global confidence ratings depended on our proxy for local confidence, in a multivariate regression. We then compared regression weights between the first vs. second half of the set, and between setsizes, using a two-way ANOVA across participants. We found a significant main effect of position (first vs. second half, $F(1,16)=13.57$, $p=.002$, $\eta_p^2=.459$) with no significant main effect of setsize ($F(1,16)=4.26$, $p=.056$, $\eta_p^2=.210$) and no significant interaction ($F(1,16)=0.04$, $p=.836$, $\eta_p^2=.003$). As can be seen in Figure 4, global confidence showed a recency effect, with a greater influence of local confidence in the more recent trials. This was observed both for sets of size 4 and 8.

Figure 4. Influence of the ideal local confidence onto global confidence. Weights of the different trial positions were estimated in a multivariate regression of global confidence against ideal local confidence, separately for each participant and setsize (4 and 8). Error bars represent mean and s.e.m. across participants.

Stability of overconfidence in time and across setsizes

Finally, we asked whether overconfidence was a stable feature of each individual, in time and across setsizes. To investigate this, we estimated it separately for the first and the second half of the experiment, and for each setsize. We found significant correlations across time for each setsize (Cor_1=.66, p=.004; Cor_4=.88, p<.001; Cor_8=.85, p<.001). Our measures of overconfidence were thus stable over the time course of the experiment (1h30), both for local and global evaluations of confidence.

In terms of stability across setsizes, we also found significant correlations across setsize ($r_{1,4}$=.70, p=.002; $r_{1,8}$=.57, p=.016; $r_{4,8}$=0.95, p<.001), suggesting a common source of overconfidence at the local and global level. However, when we compared these correlations with one another (Zou, 2007), we also found clear evidence that the correlation between the two global measures of overconfidence ($r_{4,8}$) was higher than the correlations between a local and a global measure of overconfidence ($r_{4,8} > r_{1,8}$ and $r_{4,8} > r_{1,4}$, with both p<.01 for

both comparisons). The latter two correlations were not statistically different (Figure 5). This pattern suggests that across-participants heterogeneity in global overconfidence also involves unique variance, and that global overconfidence can be partially isolated from local confidence.

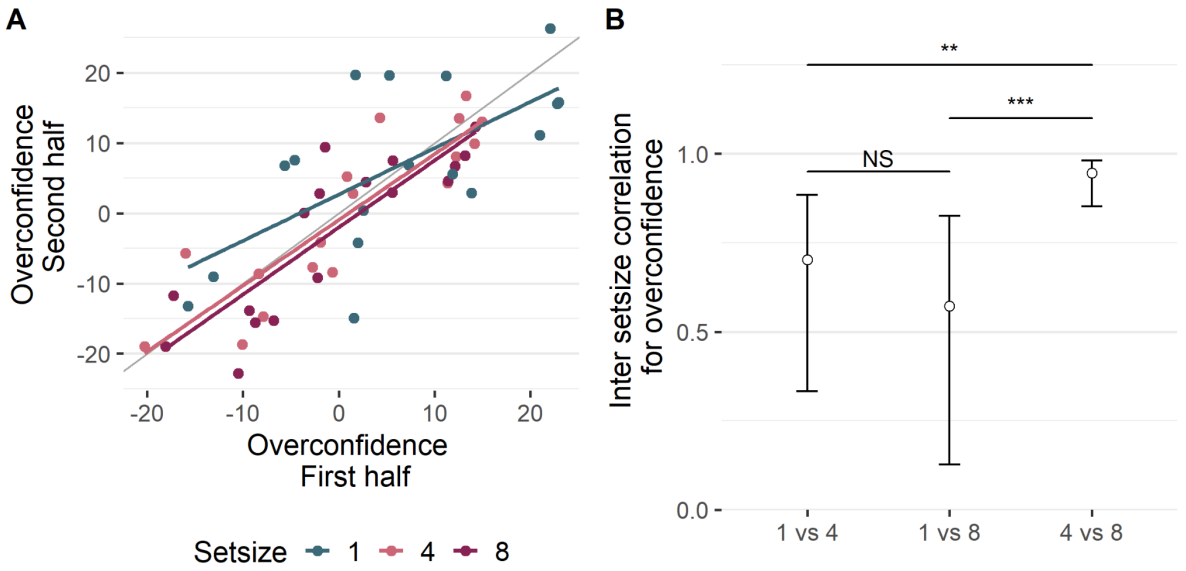

Figure 5. (A) Overconfidence in the second half of the experiment against overconfidence in the first half, separately for each setsize. Each dot represents a participant in a given setsize. Lines represent linear regressions for each setsize. (B) Correlations of overconfidence across setsizes. Dots represent the Pearson correlation estimate and error bars refer to 95% confidence intervals around those estimates. Comparison of correlations were performed using Zou's test (2007).

Discussion

In Experiment 1, we generalize in the perceptual domain the aggregation effect already documented for knowledge-based decisions: participants' overconfidence was larger when assessing their performance on a single trial than when they considered it over a set of trials. Further evidence for a partial dissociation between global and local confidence judgments was found when looking at how confidence biases were correlated across setsizes: confidence biases appeared to be more similar when compared between the two global levels (sets of size 4 and 8) than when compared between the local (sets of size 1) and the global level. Yet, we also provided evidence that local and global confidence are not

completely distinct processes since using our ideal confidence model, we showed that global confidence was, at least partially, based on local confidence.

Additionally, we did manipulate the within-set variability of difficulty, but our manipulation failed to produce a significant effect on global confidence in our data. However, we realized ex-post that our experimental manipulation may have been weaker than expected. Indeed, performance unexpectedly changed according to the position of the trial within the set which created some unexpected variability in the low variability condition and consequently reduced the magnitude of the manipulation. Although we were aiming at 0 variability in the low variability condition, ex-post analyses taking into account the impact of position on performance showed that standard deviation of expected performance in this condition was actually .047 (against .100 in the high variability condition). Our aim in experiment 2 was thus to strengthen our manipulation of variability.

Experiment 2

Experiment 2 was thus focused on the variability manipulation, with three conditions (low, medium and high variability). Set size was no longer manipulated: all sets had 5 trials. Thus, only global confidence was measured in Experiment 2.

Method

Participants, session and tasks

32 adults (age range 18-24) participated in experiment 2. 3 additional participants were tested but removed from all the analyses due to the failure of the difficulty procedure for those participants (accuracy < 55% of correct answers whereas performance was controlled at 75%).

The session took place at the LEEP as follows. First, participants trained for the perceptual task (100 trials). Then they had a second training that also included the metacognitive task

(60 perceptual trials, divided in 12 sets). Finally, they completed the main part of the experiment (168 sets, divided in 56 blocks).

The perceptual and metacognitive tasks in the experiment were identical to experiment 1. Set size was however no longer manipulated, and fixed to 5 for all sets. Thus, in experiment 2, we only measured confidence at the global level, for sets of 5 trials. In the main experiment, each block contained our 3 variability conditions presented in a random order.

Variability manipulation

In Experiment 2, we aimed at keeping expected performance constant (instead of mean orientation, as in Experiment 1) across variability conditions. To do so, for each set, we generated 5 stimuli such that the probability of making a correct response, measured via the psychophysical curve for each participant, was equal to 75% on average over the set. In practice, those probabilities of being correct were generated from 5 values drawn from a $Beta(\alpha, \beta)$ distribution with parameters $\alpha = \beta = 2$ for the medium variability and $\alpha = \beta = 4.5$ for the high variability conditions, and then rescaled to the interval [0.5, 1]. In the low variability condition, the expected difficulty was .75 for all 5 trials within the set. The resulting mean of expected difficulty is 75% in every condition, with standard deviation of either 0 (low variability), 0.158 (medium) or 0.224 (high). Thus, the manipulation of standard deviation is expected to be numerically 4 times larger than in Experiment 1.

Results

Perceptual performance

We first ran a two-way ANOVA on performance, with variability and trial position (between 1 and 5) as within participant factors. This analysis showed no main effect of variability ($F(2,62)=0.08$, $p=.928$, $\eta_p^2 =.002$), no main effect of position ($F(4,124)=1.04$, $p=.391$,

η_p^2=.032) and no interaction (F(8,248)=0.97, p=.461, η_p^2=.030). Thus, perceptual performance was comparable between experimental conditions and, contrarily to experiment 1, did not depend on the position of the trial in the set. This absence of a position effect might be attributed to set size being constant in Experiment 2, unlike in Experiment 1.

Effect of variability on overconfidence

We then turned to the main goal of Experiment 2, that is to evaluate how global confidence would be affected by a stronger manipulation of variability in trial difficulty across the set. To better control for fluctuations in performance across individuals and conditions, we calculated global overconfidence (i.e. global confidence minus performance) separately for each participant and variability conditions (see Figure 6). Critically, a one-way ANOVA with variability as a within-participant factor indicated a significant main effect of variability on global overconfidence (F(2,62)=6.49, p=.003, η_p^2=.173). Pairwise comparisons further indicated that participants were more underconfident in the low variability condition, compared to the medium and high variability conditions, which were not significantly different from one another (low vs. medium: M_Dif=-1.21, SD_Dif=3.11, t(31)=-2.20, p=.036, d=.39; low vs. high: M_Dif=-2.10, SD_Dif=3.10, t(31)=-3.84, p<.001, d=.68; medium vs. high: M_Dif=-0.90, SD_Dif=3.71, t(31)=-1.37, p=.18, d=.24). Finally, when examining each variability conditions independently, we found significant underconfidence when variability was low (Mean=-4.58, t(31)=-2.43, p=.021), but not when it was medium (Mean=-3.38, t(31)=-1.62, p=.11) or high (Mean=-2.48, t(31)=-1.33, p=.193).

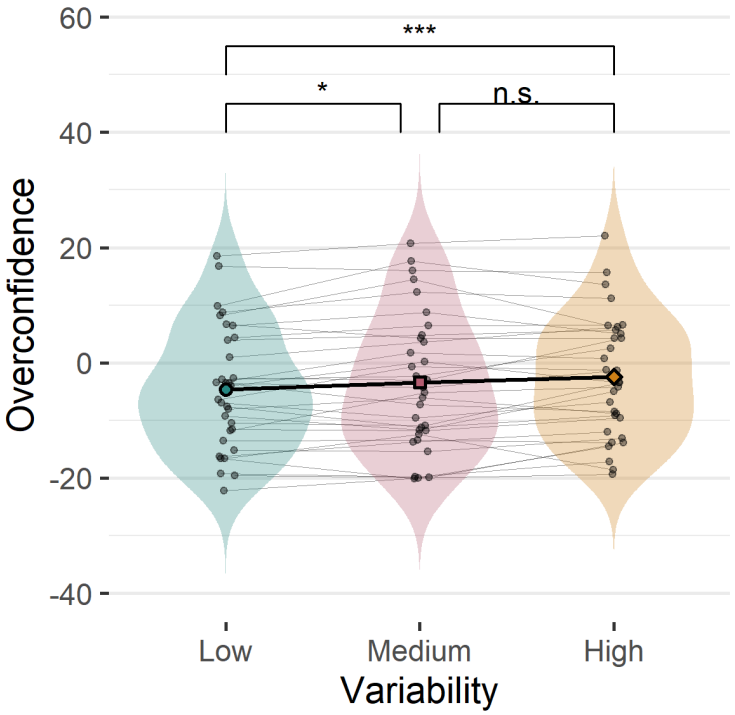

Figure 6. Distribution of global overconfidence across participants in each variability condition. Large dots and thick lines illustrate the mean overconfidence across participants. Small dots and thin lines represent individual data. Pairwise comparisons are performed using t-tests (***: p<.001, **: p<.01 *: p<.05: n.s.:p>.05).

Discussion

By increasing the strength of the variability manipulation compared to Experiment 1, Experiment 2 revealed a significant effect of variability on global overconfidence: the higher the variability of the difficulty within a set, the higher participants' global confidence when assessing their overall performance on this set. In Experiment 3, our goal was to replicate this effect and investigate its sources, by measuring not only global confidence over a set of trials but also local confidence for these trials. More specifically, we tested two hypotheses. First, we examined whether the variability effect on global confidence could be attributed to local confidence or whether it occurred during the aggregation process. Second, we evaluated the possibility that the aggregation effect may be due to biases in how participants aggregate numerical values over a set. To do so, in a second session we asked participants

to compute a numerical average over a set of numbers, which were in fact taken from their local confidence ratings in the first session. The design of this experiment as well as those two main hypotheses were pre-registered (the pre-registration can be accessed at https://aspredicted.org/QTS_4WR).

Experiment 3

In experiment 3 we thus kept a similar design than experiment 2 except that on top of global confidence judgments after each set of trials, we also asked for local confidence judgments after each trial. We also incorporated another task, the numerical averaging task, which participants did one day after the main session.

Method

Initially, we had planned to run this last experiment in-lab. However, due to the Covid-19 pandemic, we could not access the LEEP during the fall of 2020 and had to bring the experiment online. All changes between Experiment 2 and Experiment 3 are explicitly mentioned below.

Participants and sessions

The experiment took place online and 31 adults were recruited using the Prolific platform. They were naïve with respect to the goal of the study. Four additional participants were tested but removed from all the analyses, because they were not able to successfully answer an attention check (during the experiment, participants were explicitly asked to report the number 57 on a scale).

Our sample size was above the 14 participants that were needed to detect (one tailed t-test, power=.8 and error probability = .05) a similar aggregation effect size (d=.71) than in Experiment 1 (between setsize 1 and 4). It was also above the 15 participants needed to

detect a similar variability effect size (d=.68) than in Experiment 2 between high and low variability. We selected a bigger sample than required as we expected that the size of the aggregation effect would be reduced in Experiment 3 (where local and global confidence were elicited on the same sets of trials) compared to Experiment 1 (where they were elicited on different sets of trials). In practice, our sample of 31 participants allows us to detect an aggregation effect size d=.46.

The experiment was made of two sessions. During the first session, participants started with a training to the perceptual task of 100 trials. Then they had a second training that also included the metacognitive tasks (36 trials, divided in 6 sets). Finally, the main part of the session was delivered (48 series of 6 trials). A second session took place the next day. During this session, participants had a quick training for the averaging task (10 trials) followed by the main part (48 trials).

Perceptual task

To provide a shorter and slightly more engaging task for online testing, we replaced the previous task with a color numerosity task. On each perceptual trial, an array of 10x20 diamonds, each being blue or brown, was presented to participants during 1 second. They had to judge whether the array contained more blue or more brown diamonds, which was determined randomly at each trial with equal probability. The difficulty of the task was determined by the proportion of the dominant color relative to the other color, and calibrated for each participant (see below).

Metacognitive tasks

After the initial training, the metacognitive tasks were introduced to participants. After making their decision on a perceptual stimulus, participants had to give their local confidence in their answer (local confidence task) on an horizontal scale ranging from 50% to 100%. After each set of 6 trials, they had to give their global confidence in their answers (global

confidence task) on a vertical scale ranging from 50% to 100%. Participants were informed that 50% was the chance level. Both local and global confidence ratings were incentivized using the same probability matching rule as in Experiment 1 and 2.

We can thus define two measures of overconfidence. Local overconfidence refers to the mean difference between local confidence and performance, whereas global overconfidence refers to the mean difference between global confidence and performance.

Online calibration

Stimulus strength, that is the difference in proportion between brown and blue diamonds, was calibrated for each participant. Every 24 trials, a psychophysical curve describing the proportion of brown responses as a function of stimulus strength was estimated based on the last 72 perceptual decisions, by fitting a cumulated Gaussian with maximum likelihood. Using this curve, we estimated 6 thresholds (x_{05}, x_{25}, $x_{45,}$ x_{55}, x_{75}, x_{95}) corresponding to the stimulus levels for 55% (hard trials), 75% (intermediate trials) or 95% (easy trials) of correct answers when the dominant color was brown (respectively x_{55}, x_{75}, or x_{95}) and when it was blue (respectively $x_{45,}$ x_{25}, x_{05}). These values were used in the next 24 trials.

Design

For the perceptual task in the first session, we manipulated the variability of the difficulty within each set of trials. In the low variability condition, sets consisted of 6 intermediate trials. In the high variability condition, sets consisted of 3 hard trials and 3 easy trials presented in random orders. In the end, this induced a mean perceived difficulty equal to 75% in every condition, with standard deviation of either 0 (low variability), or 0.219 (high variability). Our variability manipulation was thus comparable to the one in experiment 2. The

main part of the session involved 24 blocks, and each block contained our 2 variability conditions presented in a random order.

Numerical averaging task

In the second session, the numerical averaging task was introduced to participants. In the numerical averaging task, each trial consisted in averaging 6 numbers which appeared sequentially on their screen for 1 second each (each number was preceded by a fixation cross which appeared for 500ms). Participants had to average 48 sequences of numbers in a row. The numbers that were presented were in fact the local confidences reported by participants the day before. Participants had 5 seconds to submit their answer which ensured they did not have time to explicitly compute the average. Additionally, payment in this task did not depend on performance to prevent participants from being tempted to compute the average using an online calculator. They received a fixed fee of 4 euros for this task.

Results

Perceptual performance and local confidence

We first ran a two-way ANOVA on performance, with variability and trial position (factor between 1 and 6) as within-participant factors. This analysis showed a significant main effect of variability ($F(1,30)=5.30$, $p=.029$, $\eta_p^2 =.150$), no significant main effect of position ($F(5,150)=1.24$, $p=.293$, $\eta_p^2=.040$) and no significant interaction ($F(5,130)=1.33$, $p=.253$, $\eta_p^2=.043$). Performance was higher under high variability than under low variability ($M\_Dif=1.95$, $SD\_Dif=4.71$, $t(30)=2.30$, $p=.028$, $d=.41$). This performance difference, between the two variability conditions, presumably due to an imperfect calibration, would be taken into account in our subsequent analysis.

In addition, we ran a two-way ANOVA on local confidence, with variability and trial position (between 1 and 6) as within-participant factors. We found a significant main effect of variability ($F(1,30)=43.01$, $p<.001$, $\eta_p^2 =.589$), no significant main effect of position ($F(5,150)=1.07$, $p=.380$, $\eta_p^2=.035$) and no significant interaction ($F(5,130)=0.45$, $p=.814$, $\eta_p^2=.001$). In line with what was found for performance, local confidence was higher under high variability conditions compared to low variability conditions (M_Dif=2.58, SD_Dif=2.19, $t(30)=6.56$, $p<.001$, $d=1.18$).

Effect of variability on global and local overconfidence

We first sought to confirm whether global overconfidence was higher for high variability sets, as was found in Experiment 2. To address this question, we had to control for the difference in performance between the two conditions. Indeed, the higher performance in the high variability condition might lead to a reduced overconfidence, given the “hard-easy effect” (Lichtenstein and Fischhoff, 1977). Thus, we used a linear regression across participants to evaluate the variability effect of global overconfidence, while controlling for performance. This regression is shown in purple on Figure 7A. Critically, we found a positive and significant intercept in this regression (Mean=2.30, Std=.478, t=4.818, $p<.001$), indicating that when the two variability conditions are equated in performance, global overconfidence is 2.30 points higher for high variability than for low variability sets. This effect is comparable to what was found in experiment 2, where global overconfidence was 2.1 points higher for high variability.

From then, we examined this regression for local overconfidence (blue dots in Figure 7A). Surprisingly, we found that the variability effect was present also for local confidence (Mean=2.20, Std=.394, t=5.588, $p<.001$), and similar in magnitude to that observed for global confidence (in Figure 7A, the two regression lines are perfectly overlapping). To compare global and local conditions, we conducted the same regression analysis over both

conditions which indicated no significant main effect of condition (i.e. no difference between the two intercepts, ΔMean=.10, Std=.62, t=.164, p=.871) and nos significant interaction between condition and the effect of performance (i.e. no difference between the two slopes, Δ Mean=.004, Std=.123, t=.034, p=.973). In other words, variability affected overconfidence in the same manner at the global and at the local level.

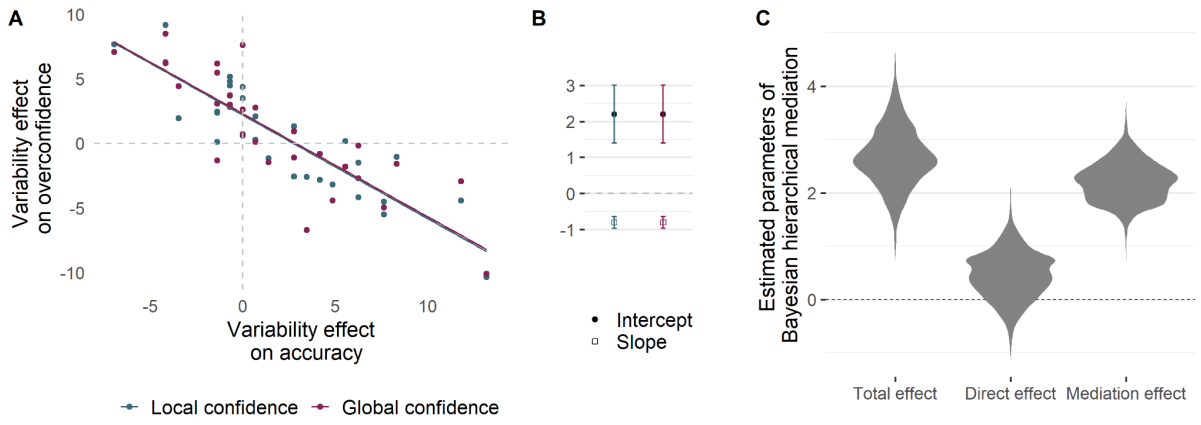

Figure 7. (A) Variability effect on overconfidence against variability effect on accuracy by the type of report (local vs global). Each dot represents a participant in a given condition. Lines represent linear regressions in each condition. (B) Intercept and slope estimates with 95% confidence intervals for the regressions in panel A. (C) Distribution of the main estimated parameters of the Bayesian hierarchical mediation model. Total effect refers to the effect of variability on global confidence without taking into account the mediating impact of local confidence. This total effect is decomposed into a direct effect which does not go through local confidence, and a mediation effect, which does.

Given this result, and the link from local to global confidence, it is possible that variability over trial difficulty within a set had an initial effect on local judgments, which then spread on global judgments. A Bayesian hierarchical mediation analysis, using the "bmlm" package in R (Vuorre, 2017) confirmed this hypothesis (Figure 7C). Variability affected local confidence (Mean coefficient=2.57, 95% CI =[1.80,3.38]), which in turn affected global confidence (Mean coefficient=.86, 95% CI =[0.80,0.93]). Critically when taking into account this mediating role of local confidence on global confidence, the effect of variability on global confidence decreased from 2.65 (average total effect, 95% CI = [1.59,3.78]) to .42 points (average direct effect, 95% CI = [-0.43, 1.32])), which was no longer significantly different from zero. By contrast, the mediation effect of local confidence on global confidence was

significantly positive (Mean 2.23, 95% CI = [1.56, 2.96]). In these estimations, 86% of the effect of variability on global confidence was mediated by a change in local confidence.

The formation of global confidence judgments

The other main goal of Experiment 3 was to evaluate how global confidence evaluations are derived from local confidence, by collecting both types of judgments over the same set of trials. First, we confirmed that confidence was lower in global than in local judgments (i.e. the aggregation effect), as was found in Experiment 1. Indeed, an ANOVA on overconfidence with judgment type (local vs global) and variability as factors indicated a significant main effect of the judgment type ($F(1,30)=10.16$, $p=.003$, $\eta_p^2=.253$), no significant main effect of variability ($F(1,30)=10.16$, $p=.373$, $\eta_p^2=.027$) and no significant interaction ($F(1,30)=0.08$, $p=.773$, $\eta_p^2=.003$). As illustrated on Figure 8A, overconfidence was significantly higher for local judgments ($M\_Dif=2.23$, $SD\_Dif=3.90$, $t(30)=3.19$, $p=.003$, $d=.57$), and in fact it was only significant for local judgments ($Mean\_local=3.90$, $t(30)=2.34$, $p=.026$; $Mean\_global=1.67$, $t(30)=1.023$, $p=.315$). This result confirms the aggregation effect for perceptual decisions, as found in experiment 1, while additionally controlling for variations in performance, since the same trials were used for global and local confidence judgments.

Second, we also sought to confirm the recency effect found in Experiment 1. We thus estimated the weight of each trial's local confidence on global confidence ratings, and in particular we evaluated how these weights differed between positions within the set (between 1 and 6) and between variability conditions. An ANOVA on the weights indicated a significant main effect of position ($F(5,150)=8.63$, $p<.001$, $\eta_p^2=.223$), with no significant effect of variability ($F(1,30)=0.88$, $p=.355$, $\eta_p^2=.029$) and no significant interaction ($F(5,150)=1.46$, $p=.208$, $\eta_p^2=.046$). This position effect corresponds to a recency effect, as illustrated by Figure 8B (purple dots), which shows larger weights for the last trials of the set, that are the most recent at the time of the global confidence evaluation. Furthermore, we confirmed that this recency effect was not simply the result of fluctuations in performance, by regressing

global confidence against accuracy at each position, and regressing the residuals of this regression against local confidence at each position, for each participant and variability condition. An ANOVA on those weights for which the impact of accuracy had been partialled out indicated again a significant main effect of position ($F(5,150)=7.76$, $p<.001$, $\eta_p^2=.206$), with no significant effect of variability ($F(1,30)=0.58$, $p=.751$, $\eta_p^2=.022$) and no significant interaction ($F(5,150)=1.15$, $p=.230$, $\eta_p^2=.039$). As illustrated by Figure 8B (black dots), weights were larger for the last trials of the set.

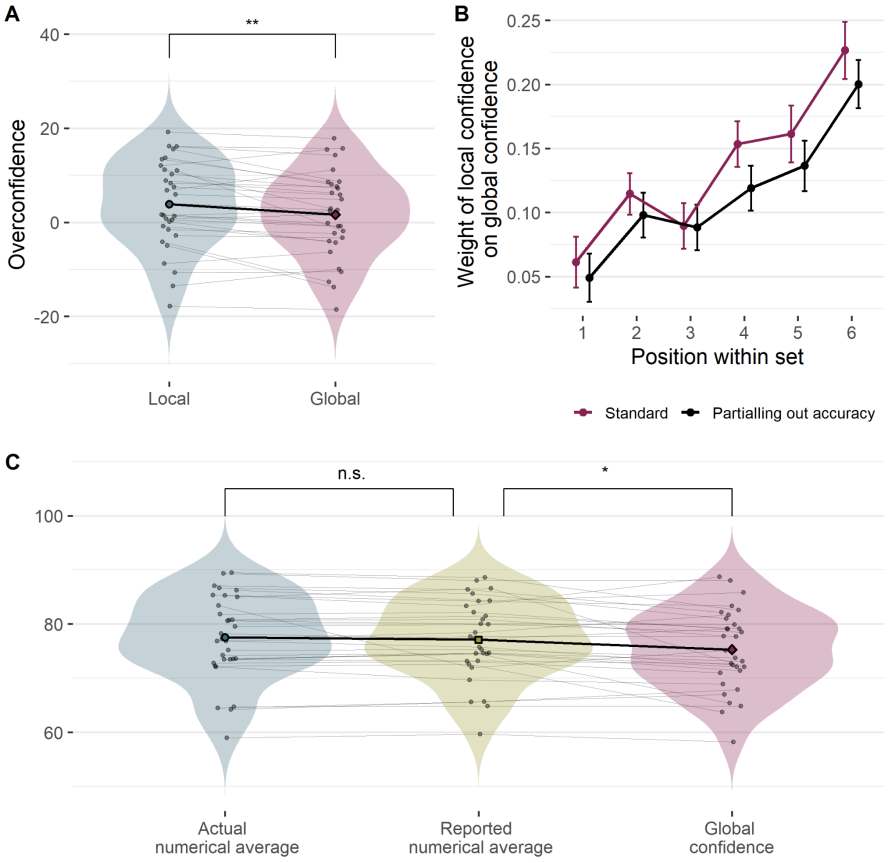

Figure 8. (A) Distribution of overconfidence across participants by type of report (local vs global). Big dots represent mean overconfidence across participants and Heavy lines connect those means. Small dots represent each participant in a given condition and thin lines connect participants across conditions. Comparison is performed using a t-test. (B) In purple, weight of local confidence on global confidence in a linear regression by position of the trial in the set. In black, similar weights for which the effect of accuracy on global confidence has been partialled out. Error bars represent mean and s.e.m. across participants. See main text for details on how we partial out accuracy. (C) Distribution of actual numerical average, reported numerical average and global confidence across participants. Big dots represent mean overconfidence across participants and heavy lines connect those means. Small dots represent each participant in a given condition and thin lines connect participants across conditions. Pairwise comparisons are performed using t-tests (***: p<.001, **: p<.01 *: p<.05: n.s.:p>.05).

Finally, in Experiment 3, we evaluated the possibility that participants would base their global confidence on the numerical average of their reported local confidence. To do so, we relied on a second session where participants had to perform such a numerical averaging task, on numbers which were in fact the local confidence ratings from the first session (see section Design). Participants performed relatively well in this task, as assessed by the correlation between the stimuli and responses within each participant (Mean_Cor=.83, $t(30)=48.91$, $p<.001$). If participants had underestimated the average in this numerical averaging task, this could have suggested that the aggregation effect was simply a computational bias, unspecific to the formation of confidence judgments. However, as shown in Figure 8C, participants did not exhibit any bias in the task: the actual and reported averages in the numerical averaging task were not significantly different (M_Dif=.41, SD_Dif=2.31, $t(30)=-0.98$, $p=.333$, $d=.18$) while the reported averages were significantly higher than the global confidences in the first session (M_Dif=2.38, SD_Dif=3.97, $t(30)=2.55$, $p=.016$, $d=.46$).

Discussion

First, this experiment showed that the variability effect on global confidence actually came from an initial effect of variability on local confidence. Second, we confirmed the presence of an aggregation effect, with larger overconfidence in local confidence than in global confidence judgments, even when participants reported local and global judgments on the very same trials. This shows the robustness of this effect since one could have expected its disappearance in this context, which facilitates the adoption of a numerical averaging strategy to produce global ratings as the mean of the local confidence ratings produced a few seconds ago. Third, and relatedly, by comparing global confidence judgments and numerical averages within the same participants, we showed that the aggregation effect is not simply due to biases in numerical averaging.

General discussion

In this study, we have investigated the characteristics of local and global confidence as well as their relationship in a perceptual task. We provide two main results. First, we show an aggregation effect in the perceptual domain: individuals were overconfident when estimating their accuracy in one trial, but not over larger sets of trials (Experiment 1 and 3). In addition, this effect was not simply due to computational biases when estimating averages over numerical values (Experiment 3). Second, we document a new phenomenon, the variability effect, by which individuals were more confident overall for sets of trials which were more heterogeneous in terms of difficulty (Experiment 2 and 3). We also show that this effect was mediated by local confidence (Experiment 3). In addition to these two main issues, we also replicate recency effects (Lee et al., 2021) in the formation of global confidence judgments (Experiments 1 and 3).

The aggregation effect has been well-documented when individuals give confidence judgments about their performance in a quiz (Gigerenzer, 1991; Sniezek & Buckley, 1991; Griffin & Tversky, 1992; Treadwell & Nelson, 1996). Importantly, here we show that it also applies to the domain of perceptual decision making. Our design has two critical features that allow us to demonstrate the existence and the specificity of the aggregation effect. First, by eliciting both types of confidence on the same set of trials in Experiment 3, we ensure that differences between local and global reports cannot be explained by differences in performance. Additionally, this shows that the aggregation effect is hard to suppress since participants reported local estimates of their performance more than 2 points higher than their global estimates, even though the latter were made less than 15 seconds after the former. Second, we show that the same participants that are subject to this effect are, on the other hand, quite able to produce numerical averages of values that correspond to their local confidences when explicitly asked to do so. We thus isolate the aggregation effect from computational biases suggesting that it is specific to confidence judgments.

In his first demonstration of the aggregation effect, Gigerenzer (1991) suggested that individuals use different probabilistic mental models or, put simply, different sets of probabilistic cues, when forming local and global confidence estimates. In a similar fashion, Griffin & Tversky (1992) suggested that local confidence mainly depends on the strength and weight of evidence, while global confidence also integrates task difficulty or past experience. Reformulating these ideas in our perceptual framework, this means that when assessing their confidence after a single trial, observers would mainly use the perceptual stimulus on the screen, while when assessing their confidence after several trials, they would also use their knowledge about their perceptive abilities. Relatedly, in Experiment 1, we found that local and global confidence biases were less correlated than global confidence biases between themselves, which supports the idea of two distinct systems for global and local confidence judgments.

Importantly however, it is also clear from our data and from prior work that global and local confidence are also partially related, as global confidence judgments are influenced by the local properties of prior decisions (Lee et al, 2021; Rouault & Fleming, 2019). In terms of a biological implementation, a recent study by Rouault and Fleming (2020) shows that whereas the ventral striatum is associated only with global confidence judgments, local and global confidence signals seem to interact in other brain areas such as the ventromedial prefrontal cortex. At the behavioral level, Lee and colleagues (Lee et al, 2021) found that individuals were better at choosing which of two series of trials they were more successful at (a decision similar to a global confidence judgment), as the number of trials in the series increased. Such an increase in the accuracy of the choice would not have been possible if global confidence was based on a single trial. The corollary of this finding is that fluctuations in local confidence should contribute to global confidence. In the present study, we illustrate these contributions of local confidence to global confidence both in Experiment 1, where we estimated local confidence indirectly using a proxy, and in Experiment 3, where we estimated it directly from participants. In these experiments, we show the presence of

recency effects in the formation of global confidence judgments. As previously found in Lee et al. (2021), when integrating local confidences, individuals assign more weight to the most recent trials. Note that those recency effects do not appear to be specific to the formation of global confidence since we also found them in the numerical averaging task (see Supplementary Material for more details). This is not surprising since such recency effects are well known in the memory literature (see e.g. Glanzer and Cunitz, 1966; Cohen; 1970; Broadbent & Broadbent, 1981; Greene, 1986; Davelaar et al, 2005).

The variability effect documented in the present study is another illustration of a perfect transmission from local to global confidence. In this effect, individuals are more confident for sets of trials that are more heterogeneous in terms of difficulty. Importantly, the variability effect on global confidence was fully explained by the level of local confidence. One can discuss here several accounts of this phenomenon. Firstly, we can discard a scenario under which the variability effect is due to participants remembering more the easiest trials when forming a global confidence judgment retrospectively. Indeed, this scenario would not account for the variability effect in local confidence judgments. Secondly, such an effect could emerge if individuals automatically increased their local confidence when presented with heterogeneous trials in terms of difficulty. However, we do not find it plausible since heterogeneity is not known by participants at the beginning of the set and its effect is constant during the set (see Supplementary Material). Instead, a more plausible hypothesis would be that the variability effect is a consequence of individuals' distorting objective probabilities at the level of local confidence. For instance, if objective probabilities are subject to a linear transformation in log odds when converted to confidence, this could generate (under some parameter values) a convexity in the relationship between confidence and objective probability to be correct, which ultimately would produce the variability effect. Interestingly, it could also be derived in a SDT model of confidence assuming individuals' misperceive the variability of stimuli's difficulty over trials (see Supplementary Material for an illustration). Evaluating how well these hypotheses explain the variability effect is left for

further research. In particular, to properly estimate how individuals distort objective probabilities and how it could explain the variability effect, we would need to increase the set of possible difficulty levels faced by participants and to evaluate their confidence at each difficulty level, separately for each variability condition.

To the best of our knowledge, the present study is the first to document this effect of variability in perceptual difficulty on global confidence judgments. As we wanted to examine global judgements of confidence over a set of trials, it seemed natural to us to ask whether the variability in task difficulty or local confidence within this set would impact observers' global judgments. We note that this approach may seem similar to that of prior studies examining the effect of variability when observers have to compute the average of a set of numbers (Spencer; 1963; Anderson; 1964; Beach & Swenson, 1966; Levin, 1975; Brezis et al., 2018), or the average along a specific perceptual dimension of a set of stimuli (for a review on this 'ensemble perception' literature, see e.g. Whitney & Yamanashi Leib, 2018). Such studies have typically found that within-trial variability in the evidence degrades performance, although it can have various effects on confidence (de Gardelle & Mamassian, 2015, Zylberberg et al., 2016; Spence et al., 2016; Boldt et al., 2017). To be clear, however, our study does not address this question at the metacognitive level. To do this, one should ask whether the variability of task difficulty across a set of trials would affect observers' confidence regarding their global evaluation of performance. This judgment would be a form of meta-metacognitive evaluation, which we did not measure in the present study. We note however that this notion of meta-metacognition has been put forward in a recent study demonstrating that participants are able to make such judgments (Recht et al., 2022), and we believe that investigating further this ability might offer exciting perspectives for future research.

To conclude, we may speculate about the functional roles of global confidence. In the literature on metacognition, the initial interest in studying global confidence judgments is that intuitively, such global judgments seem more relevant for predicting future performance

in the task. Along these lines, one interesting topic for future research is the similarity between prospective confidence judgments, by which participants indicate how much they anticipate to succeed in a future task (Siedlecka et al., 2016; Fleming et al., 2016), and retrospective global confidence. A prior study indicated that such prospective judgments aggregate confidence ratings over several trials in the past (Fleming et al., 2016), much like global confidence judgments in the present study. In a broader perspective, one could consider global confidence as a first step in bridging the conceptual gap between confidence in a single decision and higher-level notions of self-confidence and self-efficacy of individuals (Bandura, 1977).

COG statement

Our findings provide evidence of two effects: the aggregation effect and the variability effect. The aggregation effect had already been observed for a diverse range of participants in different contexts, and we generalize it to the perceptual domain. We thus expect it to be very common and to generalize to other domains as well. The variability effect, according to which individuals' confidence is larger for sets which are more heterogeneous in terms of difficulty, has not been documented before, and whether it holds true in other experimental conditions and in other domains remains to be established. In particular, it is not yet clear whether the effect would still be present in case participants were facing the exact same difficulty levels in the heterogeneous and homogeneous condition. To go further, one would thus need to construct sets with the same difficulty levels and manipulate the heterogeneity through the probability of occurrence of extremely hard or easy trials. More generally, although we have no reason to believe that our results should depend on the characteristics of participants, materials, context, or that they would depend on the cognitive domain under consideration (beyond perception), we acknowledge that such issues remain open empirical questions.

Public significance statement

Most research about confidence has looked at people's confidence after making a single decision, neglecting the study of the overall sense of confidence experienced after making multiple decisions. Our study shows that people tend to be less overconfident when they evaluate their confidence after a series of decisions than after a single decision. We also show that this overall sense of confidence is partly based on the confidence expressed after each individual decision, and increases with the heterogeneity of the set of decisions.

Bibliography

Aguilar-Lleyda, D., de Gardelle, V. Confidence guides priority between forthcoming tasks. *Sci Rep* 11, 18320 (2021). https://doi.org/10.1038/s41598-021-97884-2

Aguilar-Lleyda, D., Konishi, M., Sackur, J., & de Gardelle, V. (2021). Confidence can be automatically integrated across two visual decisions. *Journal of Experimental Psychology: Human Perception and Performance*, *47*(2), 161-171. https://doi.org/10.1037/xhp0000884

Aguilar-Lleyda, D., Lemarchand, M., & de Gardelle, V. (2020). Confidence as a priority signal. *Psychological Science*, *31*(9), 1084-1096. https://doi.org/10.1177/0956797620925039

Anderson, N. H. (1964). Test of a model for number-averaging behavior. *Psychonomic Science, 1*(7), 191–192.

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences*, *15*(3), 122-131. https://doi.org/10.1016/j.tics.2011.01.003

Baer, C., & Odic, D. (2020). Children flexibly compare their confidence within and across perceptual domains. Developmental Psychology, 56(11), 2095.

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science (New York, N.Y.)*, *329*(5995), 1081-1085. https://doi.org/10.1126/science.1185718

Balsdon, T., Wyart, V., & Mamassian, P. (2020). Confidence controls perceptual evidence accumulation. *Nature Communications*, *11*(1), 1753. https://doi.org/10.1038/s41467-020-15561-w

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*(2), 191–215.

Beach, L. R., & Swenson, R. G. (1966). Intuitive estimation of means. *Psychonomic Science, 5*(4), 161–162.

Brezis, N., Bronfman, Z. Z., & Usher, M. (2018). A Perceptual-Like Population-Coding Mechanism of Approximate Numerical Averaging. *Neural computation*, *30*(2), 428–446. https://doi.org/10.1162/neco_a_01037

Broadbent, D. E., & Broadbent, M. H. (1981). Recency effects in visual memory. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 33A*(1), 1–15. https://doi.org/10.1080/14640748108400762

Cavalan, Q. (2022, October 18). From local to global estimations of confidence in perceptual decisions. Retrieved from osf.io/btsed

Carlebach N & Yeung N (2020). Subjective confidence acts as an internal cost-benefit factor when choosing between tasks. *Journal of Experimental Psychology: Human Perception and Performance*, 46(2), 131-154.

Cohen, R. L. (1970). Recency effects in long-term recall and recognition. *Journal of Verbal Learning & Verbal Behavior, 9*(6), 672–678. https://doi.org/10.1016/S0022-5371(70)80031-7

Davelaar, E. J., Goshen-Gottstein, Y., Ashkenazi, A., Haarmann, H. J., & Usher, M. (2005). The Demise of Short-Term Memory Revisited: Empirical and Computational Investigations of Recency Effects. *Psychological Review, 112*(1), 3–42.

de Gardelle, V., Le Corre, F., & Mamassian, P. (2016). Confidence as a common currency between vision and audition. *PLoS ONE*, *11*(1). https://doi.org/10.1371/journal.pone.0147901

de Gardelle, V., & Mamassian, P. (2014). Does Confidence Use a Common Currency Across Two Visual Tasks? *Psychological Science*, *25*(6), 1286-1288. https://doi.org/10.1177/0956797614528956

de Gardelle V, Mamassian P (2015) Weighting Mean and Variability during Confidence Judgments. PLOS ONE 10(3): e0120870. https://doi.org/10.1371/journal.pone.0120870

Desender, K., Boldt, A., & Yeung, N. (2018). Subjective Confidence Predicts Information Seeking in Decision Making. *Psychological Science*, *29*(5), 761-778. https://doi.org/10.1177/0956797617744771

Desender, K., Donner, T. H., & Verguts, T. (2021). Dynamic expressions of confidence within an evidence accumulation framework. *Cognition*, *207*, 104522. https://doi.org/10.1016/j.cognition.2020.104522

Fleming, S. M., & Daw, N. D. (2017). Self-Evaluation of Decision-Making : A General Bayesian Framework for Metacognitive Computation. *Psychological Review*, *124*(1), 91. https://doi.org/10.1037/rev0000045

Fleming, S. M., Massoni, S., Gajdos, T., & Vergnaud, J.-C. (2016). Metacognition about the past and future : Quantifying common and distinct influences on prospective and retrospective judgments of self-performance. *Neuroscience of Consciousness*, *2016*(1), niw018. https://doi.org/10.1093/nc/niw018

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models : A Brunswikian theory of confidence. *Psychological Review*, *98*(4), 506-528. https://doi.org/10.1037/0033-295x.98.4.506

Glanzer, M., & Cunitz, A. R. (1966). Two storage mechanisms in free recall. *Journal of Verbal Learning & Verbal Behavior, 5*(4), 351–360. https://doi.org/10.1016/S0022-5371(66)80044-0

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. John Wiley.

Greene, R. L. (1986). Sources of recency effects in free recall. *Psychological Bulletin*, 99(2), 221.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, *24*(3), 411-435. https://doi.org/10.1016/0010-0285(92)90013-R

Guggenmos, M., Wilbertz, G., Hebart, M. N., & Sterzer, P. (2016). Mesolimbic confidence signals guide perceptual learning in the absence of external feedback. *eLife*, *5*, e13388. https://doi.org/10.7554/eLife.13388

Hainguerlot, M., Vergnaud, J.-C., & de Gardelle, V. (2018). Metacognitive ability predicts learning cue-stimulus associations in the absence of external feedback. *Scientific Reports*, *8*(1), 5602. https://doi.org/10.1038/s41598-018-23936-9

Kantner, J., Solinger, L. A., Grybinas, D., & Dobbins, I. G. (2019). Confidence carryover during interleaved memory and perception judgments. *Memory & Cognition*, *47*(2), 195-211. https://doi.org/10.3758/s13421-018-0859-8

Kesten, H. (1958). Accelerated Stochastic Approximation. *The Annals of Mathematical Statistics*.

Klayman, J., Soll, J. B., González-Vallejo, C., & Barlas, S. (1999). Overconfidence : It Depends on How, What, and Whom You Ask. *Organizational Behavior and Human Decision Processes*, *79*(3), 216-247. https://doi.org/10.1006/obhd.1999.2847

Koriat, A. (2015). When two heads are better than one and when they can be worse : The amplification hypothesis. *Journal of Experimental Psychology: General*, *144*(5), 934-950. https://doi.org/10.1037/xge0000092

Lee, A., de Gardelle, V., & Mamassian, P. (2021). Global visual confidence. *Psychonomic Bulletin and Review*, *28*, 1233-1242. https://doi.org/10.3758/s13423-020-01869-7

Levin, I. P. (1975). Information integration in numerical judgments and decision processes. *Journal of Experimental Psychology: General, 104*(1), 39–53. https://doi.org/10.1037/0096-3445.104.1.39

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, *20*(2), 159-183. https://doi.org/10.1016/0030-5073(77)90001-0

Mamassian, P. (2016). Visual Confidence. *Annual Review of Vision Science*, *2*, 459-481. https://doi.org/10.1146/annurev-vision-111815-114630

Mamassian, P., & de Gardelle, V. (2021). Modeling perceptual confidence and the confidence forced-choice paradigm. *Psychological Review*, No Pagination Specified-No Pagination Specified. https://doi.org/10.1037/rev0000312

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422-430. https://doi.org/10.1016/j.concog.2011.09.021

Massoni, S., Gajdos, T., & Vergnaud, J.-C. (2014). Confidence measurement in the light of signal detection theory. *Frontiers in Psychology*, *5*(1455). https://doi.org/10.3389/fpsyg.2014.01455

Massoni, S., & Roux, N. (2017). Optimal group decision : A matter of confidence calibration. *Journal of Mathematical Psychology*, *79*, 121-130. https://doi.org/10.1016/j.jmp.2017.04.001

Nelson, T. O. (1990). Metamemory : A Theoretical Framework and New Findings. In G. H. Bower (Éd.), *Psychology of Learning and Motivation* (Vol. 26, p. 125-173). Academic Press. https://doi.org/10.1016/S0079-7421(08)60053-5

Peirce, C. S., & Jastrow, J. (1884). On Small Differences in Sensation. *Memoirs of the National Academy of Sciences*, *3*, 75-83.

Pereira, M., Perrin, D., & Faivre, N. (2022). A leaky evidence accumulation process for perceptual experience. *Trends in Cognitive Sciences*, *26*(6), 451-461. https://doi.org/10.1016/j.tics.2022.03.003

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/a0019737

Rahnev, D., Koizumi, A., McCurdy, L. Y., D'Esposito, M., & Lau, H. (2015). Confidence Leak in Perceptual Decision Making. *Psychological Science*, *26*(11), 1664-1680. https://doi.org/10.1177/0956797615595037

Rausch, M., & Zehetleitner, M. (2019). The folded X-pattern is not necessarily a statistical signature of decision confidence. *PLoS Computational Biology*, *15*(10), e1007456. https://doi.org/10.1371/journal.pcbi.1007456

Recht, S. , Jovanovic, L. , Mamassian, P. & Balsdon, T. (2022). Confidence at the limits of human nested cognition. *Neuroscience of consciousness*, 2022(1)

Rouault, M., Dayan, P., & Fleming, S. M. (2019). Forming global estimates of self-performance from local confidence. *Nature Communications*, *10*(1), 1141. https://doi.org/10.1038/s41467-019-09075-3

Rouault, M., & Fleming, S. M. (2020). Formation of global self-beliefs in the human brain. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, *117*(44), 27268-27276. https://doi.org/10.1073/pnas.2003094117

Schneider, S. L. (1995). Item Difficulty, Discrimination, and the Confidence-Frequency Effect in a Categorical Judgment Task. *Organizational Behavior and Human Decision Processes*, *61*(2), 148-167. https://doi.org/10.1006/obhd.1995.1012

Siedlecka, M., Paulewicz, B., & Wierzchoń, M. (2016). But I Was So Sure ! Metacognitive Judgments Are Less Accurate Given Prospectively than Retrospectively. *Frontiers in Psychology*, *7*. https://www.frontiersin.org/articles/10.3389/fpsyg.2016.00218

Sniezek, J. A., & Buckley, T. (1991). Confidence depends on level of aggregation. *Journal of Behavioral Decision Making*, *4*(4), 263-272. https://doi.org/10.1002/bdm.3960040404

Spence, M. L., Dux, P. E., & Arnold, D. H. (2016). Computations underlying confidence in visual perception. *Journal of Experimental Psychology: Human Perception and Performance, 42*(5), 671–682. https://doi.org/10.1037/xhp0000179

Spencer, J. (1963). A further study of estimating averages. *Ergonomics, 6*(3), 255–265. https://doi.org/10.1080/00140136308930705

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

Stone, E. R., Rittmayer, A. D., Murray, N. T., & McNiel, J. M. (2011). Demonstrating and eliminating the aggregation effect in physical skill tasks. *Journal of Behavioral Decision Making*, *24*(1), 1-22. https://doi.org/10.1002/bdm.675

Treadwell, J. R., & Nelson, T. O. (1996). Availability of information and the aggregation of confidence in prior decisions. *Organizational Behavior and Human Decision Processes*, *68*(1), 13-27. https://doi.org/10.1006/obhd.1996.0086

van den Berg, R., Anandalingam, K., Zylberberg, A., Kiani, R., Shadlen, M. N., & Wolpert, D. M. (2016). A common mechanism underlies changes of mind about decisions and confidence. *eLife*, *5*, e12192. https://doi.org/10.7554/eLife.12192

Vuorre, M. (2017). *bmlm : Bayesian Multilevel Mediation*. https://cran.r-project.org/web/packages/bmlm/citation.html

Whitney, D., & Yamanashi Leib, A. (2018). Ensemble Perception. *Annual Review of Psychology*, *69*, 105-129. https://doi.org/10.1146/annurev-psych-010416-044232

Zylberberg A., Fetsch C.R., Shadlen M.N. (2016). The influence of evidence volatility on choice, reaction time and confidence in a perceptual decision. *eLife*