Full title:

Two sides of the same coin: monetary incentives concurrently improve and bias confidence judgments.

Running title:

How motivation impacts confidence accuracy.

Authors:

Maël Lebreton[1,2+] Shari Langdon[3,4], Matthijs J. Slieker[3,4], Jip S. Nooitgedacht[3,4], Anna E. Goudriaan[3,4], Damiaan Denys[4,5], Judy Luigjes[3,4*], Ruth J. van Holst[3,4*]

Affiliations:

[1]Amsterdam Brain and Cognition (ABC),

[2]CREED, Amsterdam School of Economics (ASE), Universiteit van Amsterdam, 1018 WB Amsterdam, the Netherlands.

[3]Amsterdam Institute for Addiction Research,

[4]Department of Psychiatry, Academic Medical Centre, 1100 DD Amsterdam, the Netherlands

[5]Netherlands Institute for Neuroscience, Institute of the Royal Netherlands Academy of Arts and Sciences, 1105 BA Amsterdam, the Netherlands

[+] To whom correspondence should be addressed (m.p.lebreton@uva.nl)

*shared last authorship.

Abstract:

Most decisions are accompanied by a feeling of confidence, i.e., a belief about the decision being correct. Confidence accuracy is critical, notably in high-stakes situations such as medical or financial decision-making. Here, we experimentally investigated how incentive motivation influences confidence accuracy by combining a perceptual task with a confidence incentivization mechanism. Importantly, by varying the magnitude and valence (gains or losses) of the monetary incentives, we orthogonalized their motivational and affective components. Corroborating theories of rational decision-making and motivation, our results first reveal that the motivational value of incentives improves aspects of confidence accuracy. However, in line with a value-confidence interaction hypothesis we further show that the affective value of incentives robustly biases confidence reports, thus degrading confidence accuracy. Finally, we demonstrate that the motivational and affective effects of incentives differentially impact how confidence builds on perceptual evidence. Altogether, these findings may provide new hints about confidence miscalibration in healthy or pathological contexts.

Introduction:

Imagine you have to cross a road. It's dark and raining, the visibility is low. At some point, you estimate that there does not seem to be any danger, and decide to cross. However, because you feel quite unsure about crossing with such low visibility, you decide to check one last time, and spot a car coming at high speed to your direction. Luckily, you have enough time to withdraw from the street and avoid being hit by the car.

Just like in this example, most decisions in everyday life are accompanied by a subjective feeling of confidence emerging from the constant monitoring of our own thoughts and actions by metacognitive processes[1,2]. Formally, confidence is a decision maker's estimate of the probability –or belief- that her action, answer or statement is correct, based on the available evidence[3,4]. Although high confidence accuracy seems critical to monitor and re-evaluate previous decisions[5] or to arbitrate between different strategies[6,7], converging evidence suggests that confidence judgments often significantly differ from the actual probability of being correct. Notably, we often overestimate the probability of being correct, a phenomenon called overconfidence[8]. This bias, potentially detrimental for the decision-maker or society, has been consistently reported in numerous domains and situations, from simple sensory psychophysics[9] or knowledge[10] tasks in the laboratory, to medical[11], financial, and managerial[12,13] decision-making.

Altogether, the importance of confidence as a cognitive variable mitigating decision-making and the societal relevance of its miscalibration have considerably stimulated the search for factors which modulate or bias confidence estimation. An influential line of research, encapsulated under the term "*motivated cognition*" or "*motivated reasoning*", has suggested that beliefs are influenced by individuals' desires[14–16]. In other terms, individuals tend to estimate desirable events to be more likely than undesirable ones, potentially leading to overconfidence[17]. In a different line of research, studies have established links between incidental psychological states such as elevated mood[18] or emotional arousal[19,20] and (over-)confidence. Recently, functional neuroimaging studies reported neural correlates of confidence in the ventromedial prefrontal cortex[21,22], a brain region associated with the encoding of economic, motivational and affective values[23]. Such an overlap in the neural correlates of confidence and values suggests that these variables also interact at the behavioral level. In practice, this hypothesis entails that a decision-maker reports higher confidence because she is in a high expected- or experienced- value context. This values-confidence interaction could both explain associations between positive affective-states and overconfidence[17–20,24], and underpin new biases in confidence judgments.

In this report, we methodically investigated the interactions between incentive motivation and confidence, in an attempt to explain features of human confidence (mis)judgment. To do so, we designed an original task where participants had to first make a difficult perceptual decision, and then judge the probability of their answer being correct, i.e., their confidence in their decision (**Fig. 1.A**). Critically, we incentivized the truthful and accurate reporting of confidence using the Matching Probability (MP) mechanism, a well-validated method from behavioral economics adapted from the Becker-DeGroot-Marschak auction[25,26]. This incentivization was implemented after the perceptual choice, which made it possible to separately motivate the accuracy of confidence judgments without directly influencing the performance on the perceptual decision (**Fig. 1.A-C** and **Methods**). Briefly, the MP mechanism considers participants' confidence reports as bets on the correctness of their answers, and implements trial-by-trial comparisons between these bets and random lotteries. Under utility maximization assumptions, this guarantees that participants maximize their earnings by reporting their most precise and truthful confidence estimation[27,28]. In order to identify the critical features of the interactions between incentive motivation and confidence accuracy, we varied the magnitude and valence of the monetary incentives. This allowed us to independently examine the effect of the net value of incentives (i.e., the affective component of the incentive, which can take both positive and negative values, and thereafter indexed as V) and the absolute value of incentives (ie., the motivational value of incentives, regardless of their valence, thereafter indexed as |V|) (**Fig. 1.D**). Importantly, this simple manipulation enabled us to test three opposing predictions from three different theories (**Fig. 1.E**). First, standard theories of rational decision-making and motivation from behavioral economics[29–31] and cognitive psychology[32] predict that higher stakes increase participants tendency to conform to rational model predictions, which should improve confidence accuracy to maximize the payoff. In this case, if participants are generally overconfident, an increase in absolute value should decrease confidence judgments (and thereby overconfidence). Second, motivated cognition theories[14] –i.e. in the form of the desirability bias[17]- predict that participants should be more motivated to believe that they are correct when more money is at stake, irrespective of the valence (gain or loss). In this case, an increase in absolute value should increase confidence judgments (and exaggerate overconfidence if participants are generally overconfident). Finally, our values-confidence interaction hypothesis predicts higher monetary incentives should bias confidence judgments upwards in a gain frame, and downward in a loss frame, despite the potentially detrimental consequences on the final payoff. In this case, the net value should bias confidence judgments.

In four experiments, we systematically varied the magnitude and valence (gains or losses) of the monetary stakes used to incentivize confidence accuracy, and repeatedly found behavioral patterns which seem to favor the value-confidence interaction hypothesis.

Results:

We collected data in four experiments in which participants performed different versions of a confidence elicitation task (**Fig. 1, Table 1** and **Methods**); in each trial, participants briefly saw a pair of Gabor patches first, then had to indicate which one had the highest contrast, and finally had to indicate how confident they were in their answer (from 50 to 100%). Critically, the confidence judgment was incentivized: after the binary choice and before the confidence judgment, a monetary stake was displayed, which could be neutral (no incentive) or indicate the possibility to gain or lose a certain payoff (e.g. 10 cents, 1 euro, 2 euros), which differed between the experiments. Participants could maximize their chance to gain (respectively not lose) the stake by reporting their confidence as accurately and truthfully as possible, because a Matching Probability (MP) mechanism[25,26] determined the outcome of the trial. Prior to the task, participants performed a calibration session, which was used to generate the main task stimuli, such that the decision situations spanned a pre-defined range of individual subjective difficulties (see **Methods**).

As a pre-requisite, we assessed the quality of our experimental design and the validity of our experimental variables, irrespective of the effects of monetary incentives. We notably show that, in all four experiments, *ex ante* choice predictions from our psychophysical model closely match participants' actual choice behavior (**SI results**). Additionally, we show that in all four experiments, participants' confidence judgments exhibit three fundamental properties[33]: (1) confidence ratings correlate with the probability of being correct; (2) the link between confidence ratings and evidence is positive for correct and negative for incorrect responses; (3) the link between evidence and performance differs between high and low confidence trials (**SI results**). Overall, these preliminary results suggest that the confidence measure elicited in our task actually corresponds to subjects' estimated posterior probability of being correct[33]. They also address potential concerns about the validity of confidence elicitation in general[34,35], and additionally demonstrate that our MP incentivization mechanism did not bias or distort confidence.

Effects of incentives on confidence judgments. Twenty-four subjects participated in our first experiment, where the combination of their choice and confidence ratings could lead, depending on the trial, to a gain or no-gain of 1 euro, a loss or no-loss of 1 euro, or to a neural outcome (**Fig. 1.C**). In order to investigate the interaction between incentive motivation and confidence, and compare the predictions of the different theories (**Fig 2.A**), we implemented linear mixed-effect models, with incentive net values and incentive absolute values as independent variables (see **Methods**). In line with our value-confidence interaction hypothesis, our results first show that participants' confidence judgments are specifically modulated by the incentive's net value ($\beta_V = 2.06\pm0.42$, $p<0.001$; $\beta_{|V|} = -0.97\pm1.03$, $p=0.38$). Critically, and as expected from our task design, this effect of incentives on confidence is not driven by an effect on performance, given that neither the net nor absolute value of incentive have any effect on performance (both $p>0.20$).

Effects of incentives on confidence accuracy. In order to explore how incentives impact confidence *accuracy*, we then focused on two distinctive metacognitive metrics: discrimination and calibration (see **Methods** for formal definitions). Discrimination (or resolution) measures how confidence distinguishes between correct and incorrect responses, and is computed as the difference between the average confidence for correct answers and the average confidence for incorrect answers; the higher the discrimination, the more accurate the confidence judgments. On the other hand, calibration measures how averaged confidence matches averaged objective performance, and is computed as the difference between the averaged confidence and the averaged performance. Therefore, a calibration of zero signals high confidence accuracy, whereas a positive (respectively negative) calibration signals overconfidence (respectively underconfidence). Our results first show that discrimination is specifically modulated by the absolute value of incentive ($\beta_V = -0.06\pm0.36$, $p = 0.88$; $\beta_{|V|} = 2.72\pm0.90$, $p < 0.01$; **Fig. 2.B**); this means that positive and negative incentivization symmetrically improve participants' discrimination compared to the no-incentives condition. We refer to this first effect as the *motivational* effect of incentives on confidence accuracy. Second, mirroring the effects on confidence judgements, calibration monotonically increases with the net value of incentive ($\beta_V = 2.78\pm0.67$, $p < 0.001$; $\beta_{|V|} = -0.71\pm1.42$, $p = 0.62$; **Fig 2.B**). Since participants are overconfident on average, calibration is thereby improved by loss prospects, but paradoxically deteriorated by gain prospects. We refer to this second effect as the *biasing* effect of incentives on confidence.

Effects of incentives on confidence formation.

Confidence builds on noisy perceptual evidence[33], resulting in a positive correlation for correct choices and a negative correlation for incorrect choices (**SI results**). Incentive effects can affect confidence per se (suggesting a simple bias of confidence) or influence the relationship between confidence and evidence (suggesting that incentives affect the integration of evidence in the formation of the confidence signal). To investigate this question, we next explore how incentives modulate the relationship between confidence and evidence, for correct and incorrect answers (**Fig. 2.C**). For each individual and each incentive level, we built a multiple linear regression modelling confidence ratings as a combination of a confidence baseline, and of two terms capturing the linear integration of perceptual evidence for correct and incorrect answers. For each incentive level, regression coefficients were estimated at the individual level and the effect of incentives on the different regression coefficients were subsequently tested in our linear mixed-effect model. Our results show a clear dissociation between the motivational and biasing effects of incentives on confidence formation. On the one hand, the motivational effect of incentives can be found in the slopes of those regressions: in both cases gains and losses increase the linear relationship between confidence and evidence, compared to no incentives (correct answers: $\beta_{|V|} = 0.04 \pm 0.02$, $p < 0.05$; incorrect answers: $\beta_{|V|} = -0.24 \pm 0.08$, $p < 0.01$). On the other hand, the biasing effect of incentives is found in the intercept of those regressions ($\beta_V = 2.18 \pm 0.47$, $p < 0.001$). Therefore, while the motivational effect of incentives actually influences the way confidence is built from evidence by increasing the weight of evidence in the ratings in opposite direction for correct and incorrect answers, the biasing effect of incentives appears to be a purely additive effect of incentives on confidence, unrelated to the amount of evidence.

To further investigate the effects of incentive effects confidence, and to control for alternative explanations, we next conducted three additional experiments.

The effects of incentives are incidental.

To rule out that participants deliberately and strategically increase their confidence with incentive's net value, due to some misconceptions induced by the incentivization (i.e., MP) mechanisms, we collected data from twenty-one new participants who performed a second *performance* task in addition to our standard *confidence* task. In the performance task, confidence is simply elicited with ratings after choice; confidence accuracy is not incentivized with the MP mechanisms, and subjects are only rewarded according to their choice performance -correct/incorrect (see **Methods** and **Fig. Sup. 2**). Still, in line with the value-confidence hypothesis, confidence is found to be specifically modulated by the net values of incentives in both the confidence and the performance task (confidence task: $\beta_V = 2.47 \pm 0.68$, $p < 0.001$; performance task: $\beta_V = 1.88 \pm 0.59$, $p < 0.01$ **Fig. 3.A**). To the contrary, and as expected from the task design, incentives have no effects on performance in either task (all ps > 0.38). Importantly, the biasing effect of net value is also seen as a trend in calibration in both tasks (confidence task: $\beta_V = 2.01 \pm 1.24$, $p = 0.11$; performance task: $\beta_V = 1.75 \pm 0.89$, $p = 0.05$; **Fig. 3.B**). Interestingly, the motivational effect of the absolute value on discrimination is replicated when confidence is incentivized, but not when performance is incentivized (confidence task: $\beta_{|V|} = 2.41 \pm 0.87$, $p < 0.01$; performance task: $\beta_{|V|} = -0.07 \pm 1.32$, $p = 0.96$; **Fig. 3.B**).

We then replicate and extend the findings of the first experiment on the confidence formation model (**Figure 3.C**); when biasing effects of the net value of incentive are present (in both tasks), they impact the intercept of the confidence formation model (both ps < .001). On the other hand, the motivational effects of incentives are only found on the slope of incorrect trials in the confidence task (incorrect answers; confidence task: $\beta_{|V|} = -0.317 \pm 0.13$, $p < 0.05$; performance task: $\beta_{|V|} = -0.03 \pm 0.10$, $p = 0.81$). Again, this means that the motivational effect of the incentivization of confidence accuracy is underpinned by a better integration of perceptual evidence in the confidence rating when stakes increase, whereas this effect is absent in the task where confidence accuracy is not incentivized. In sum, these results indicate that the biasing effects of incentives on confidence judgments are not induced by the incentivization mechanism, and that the motivational effects of incentives are only found when confidence is incentivized.

Dissociating incentive value effects from (mere) valence effects. In order to demonstrate that the motivational and biasing effects of incentives are due to incentive *values*, rather than to simple valence (gain/loss) effects, we next invited thirty-five subjects to participate in a third experiment, where incentives for confidence accuracy varied in both valence (gains and losses) and magnitude (1€ vs 10¢) (see **Table 1**). We modified our linear mixed-effect models to include a valence variable (=1 if incentives are positive and 0 if negative, thereafter indexed by +/-), in addition to the incentive's net and absolute value independent variables. Results show that both the net value of incentive and the valence variable impact confidence judgments ($\beta_V = 1.02 \pm 0.38$, $p < 0.01$; $\beta_{+/-} = 4.01 \pm 1.11$, $p < 0.001$; **Fig. 4.A**). This means that the biasing effects of incentive previously reported are not simply due to an effect of valence, but are truly underpinned by the incentive net value. Again, no effects of incentives are found on performance (all ps > 0.22). Importantly, the linear effect of incentive's net value on confidence judgments percolates confidence accuracy, as a similar effect is found on calibration ($\beta_V = 2.15 \pm 0.97$, $p < 0.05$; **Fig. 4.B**). Note that in the case of discrimination, we do not find significant effects of incentive's absolute value on confidence judgement (all ps > 0.14). Although this seemingly contradicts to the results of the two previous experiments, this difference can be explained by the lack of a neutral incentive condition in the present experiments. This means that motivational effects previously reported are primarily due to the mere presence of incentives. Replicating our previous finding, the biasing effect of incentive's net value is found to be independent from the amount of evidence, impacting the intercepts of the linear relationship between evidence and confidence (intercept: $\beta_V = 1.34 \pm 0.50$, $p < 0.01$; **Fig. 4.C**). No effects of incentives could be found on the slopes characterizing the integration of evidence in confidence judgments (all ps > 0.11). In sum, the results from this third experiment replicate the biasing effect of incentive's net value on confidence, and furthermore demonstrate that these effects depend on the magnitude of incentives.

Accounting for valence difference.

While the biasing effect of incentives on confidence and calibration revealed in our first three experiments appeared robust and replicable, it seemed to be driven by the loss frame. This could mean that this biasing effect is purely restricted to the loss frame. However, an alternative hypothesis is that subjects are simply less sensitive to gains, as suggested by prospect theory[36]. In order to dissociate between those two hypotheses, we invited twenty-four subjects to participate in a final study which included higher stakes (10c, 1€, 2€) in both gain and loss frames (**Table 1**). In this case, our linear mixed-effect model included three independent variables: two variables accounted for the signed magnitude of incentives in the gain (V+), and in the loss frame (V-); in addition, and in line with the previous experiment, the third variable captured the effect of the valence framing (+/-). Our results reveal a significant effect of the magnitude of incentives on confidence, in both the gain and loss frames ($\beta_{V+} = 20.79\pm0.25$, $p < 0.001$; $\beta_{V-} = 1.22\pm0.38$, $p < 0.001$; **Fig. 5.A**). This result confirms our initial hypothesis: following expected values, higher incentives bias confidence judgments upwards in a gain frame, and downward in a loss frame. Similar to our third experiment, the motivational effect on discrimination is not found (all ps > 0.11; **Fig. 5.B**), suggesting that it is mostly driven by the incentive versus no-incentive contrast. In this experiment, however, the effects of incentives on calibration are mostly driven by the valence variable ($\beta_{+/-} = 4.26\pm1.72$, $p < 0.05$; **Fig. 5.B**). Given that calibration combines the noise from confidence and performance, and that the presence of six incentive-levels decreases the number of trials used to estimate those measure, we interpret the absence of effects of the magnitude of incentives on calibration as a lack of power. Supporting this interpretation, those effects can be found on the intercept of our confidence formation model, a more sensitive measure of our bias ($\beta_{V+} = 0.72\pm0.40$, $p = 0.08$; $\beta_{V-} = 1.26\pm0.58$, $p < 0.05$; **Fig. 5.C**). This last set of results replicates, for the fourth time, the biasing effects of incentives on confidence, and confirms that monetary gains and losses both contribute to biasing confidence in perceptual decisions.

Discussion:

In this article, we combined a perceptual decision task and an auction procedure inspired from behavioral economics[25,26] to investigate how monetary incentives influence confidence. In addition to replicating important features of a recent model of confidence formation[33], we reveal and dissociate two effects of monetary incentives on confidence accuracy.

The first effect is a motivational effect of incentives: in line with theories of rational decision making and motivation, incentivizing confidence judgments improves discrimination. This means that high (respectively low) confidence is more closely associated with correct (respectively incorrect) decisions when confidence reports are incentivized, regardless of the valence or magnitude of the incentive. This nicely extends a recent study reporting a similar effect of incentivization on discrimination, although limited to the gain domain[37]. Here, we further show that this motivational effect of incentives is underpinned by a better integration of perceptual evidence in the confidence judgment when stakes increase.

The second effect, the biasing effect of incentives on confidence accuracy, is entirely novel: confidence judgments are parametrically biased by the net value of the incentive. The prospect of gains increases confidence, while the prospect of losses decreases confidence. Because people generally exhibited overconfidence in our experiment, the gain prospects detrimentally increased overconfidence (i.e. deteriorated calibration), while prospects of losses improved calibration. As opposed to the motivational effect, the biasing effect of incentive was purely additive, i.e., independent of the amount of evidence on which decisions and confidence judgments are based. The biasing effect was also found to be incidental, i.e., also present when the confidence judgments are not incentivized. Importantly, we show that this bias is unpredicted by motivated cognition theories such as the desirability bias[17], which predicts increased overconfidence also with increased negative values because avoiding a loss is desirable. This bias is also unpredicted by the theories of rational decision making and motivation, which predicts decreased overconfidence with increased positive values because it would lead to a higher reward. Yet, the biasing effect of incentives is in line with the value-confidence hypothesis. One plausible interpretation for this effect is an affect-as-information effect: people use their momentary affective states as information in decision-making[38] which, in our case, means that they integrate the trial expected value into their confidence judgment.

In order to incentivize confidence reports, we used a mechanism inspired from Becker-DeGroot-Marschak auction procedures[25,26], referred to as Reservation or Matching Probability(MP), which conveniently allowed us to manipulate the monetary stakes on a trial-by-trial basis. Contrary to other incentivization methods such as the quadratic scoring rule (QSR), the MP mechanism is valid under simple utility maximization assumptions, i.e., remains incentive-compatible when subjects are not risk neural[28,37]. The MP mechanism is even incentive-compatible when considering probability distortions, on the assumption that both subjective (confidence) and objective (lotteries) probabilities are transformed identically[39,40]. Several studies have investigated the impact of different incentivization mechanisms on subjective probability judgments (confidence or belief), and report that MP is among the best methods available, both at the theoretical and experimental levels[28,37,39]. Our findings nonetheless demonstrate that regardless of the type of incentivization mechanism, the mere presence of incentives can induce distortions in probability estimations such as confidence judgments.

In this collection of experiments, we only used relatively small monetary amounts as incentives; how the motivational and biasing effect of incentive scales when monetary stakes increase significantly remains an open question. Critically, higher stakes may also impact physiological arousal, which influence confidence and interoceptive abilities[20,41]. In general, the effects of incentives on confidence accuracy could also be mediated by inter-individual differences in metacognitive or interoceptive abilities[41,42] and by incentive motivation sensitivity[43]. Because our subject sample was mostly composed of university students, the generalization of those findings in the general population will have to be assessed in further studies.

Our results confirm that confidence judgments do not just represent rational estimates of the probability of being correct[4], but also integrate information and potential biases processed after a decision is made[44]. The mere notion of confidence biases, notably overconfidence, and the actual conditions under which they can be observed sparked an intense debate in psychophysics[9,45,46] and evolutionary theories[47,48]. Critically, here, confidence accuracy was properly incentivized, hence deviations from perfect calibration can be appropriately interpreted as cognitive biases[48]. The striking effects of incentive net values on confidence seem to make sense when considering evolutionary perspective: in natural settings, whereas overconfidence might pay off when prospects are potential gains (e.g., when claiming resources[47]), a better calibration might be more appropriate when facing prospects of losses (e.g., death or severe injuries), given their potential dramatic consequences on reproductive chances. Interestingly, the observed valence difference in the effect of incentives magnitude –higher in the loss than in the gain domain- seem to mimic valence asymmetries observed in economic decision-making theories such as prospect theory[36].

How confidence is formed in the human brain and how neurophysiological constraints explain biases in confidence judgments remain an open question[4,49]. Although functional and structural neuroimaging studies initially linked confidence and metacognitive abilities to dorsal prefrontal regions[1], confidence activations were also recently reported in the ventro-medial prefrontal cortex[21,22,50], a region which has been consistently involved in motivation and value-based decision making[23]. It is therefore possible that this region plays a role in the motivational and biasing effects of incentives on confidence, and constitutes the neurophysiological basis for the affect-as-information theory[38]. However, this remains highly speculative and should be investigated in future neuroimaging studies.

Overall, our results suggest that investigating the interactions between incentive motivation and confidence judgments might provide valuable insights on the cause of confidence miscalibration in healthy and pathological settings. For instance, high monetary incentives in financial or managerial domains may in fact create or exaggerate overconfidence, leading to overly risky and suboptimal decisions. In the clinical context, inflated levels of overconfidence in pathological gamblers[51] could be amplified by high monetary incentives, contributing to compulsive gambling in the face of great loss. Moreover, if indeed value-induced affective states modulate confidence judgements, other disorders with abnormal incentive processing such as addictions, mood-disorders, obsessive-compulsive disorder and schizophrenia could be at particular risk for confidence miscalibration[52–54]. Field experiments and clinical research will be needed to further explore the individual and societal consequences of the interactions between incentive motivation and confidence accuracy.

Experimental Procedures

Subjects. All studies were approved by the local Ethics Committee of the University of Amsterdam Psychology Department. All subjects gave informed consent prior to partaking in the study. The subjects were recruited from the laboratory's participant database (https://www.lab.uva.nl/spt/). A total of 83 subjects took part in this study (see Table 1). They were compensated with a combination of a base amount (10€), and additional gains and/or losses from randomly selected trials (one per incentive condition per session for experiment 1, and one per incentive condition from one randomly selected session for experiments 2 and 3).

Tasks. All tasks were implemented using MATLAB® (MathWorks) and the COGENT toolbox (http://www.vislab.ucl.ac.uk/cogent.php). In all four experiments, trials of the confidence incentivization task shared the same basic steps (**Fig. 1.A**): after a brief fixation cross (750 ms), participants viewed a pair of Gabor patches displayed on both sides of a computer screen (150 ms), and judged which had the highest contrast (self-paced) by using the left or right arrow. They were thereafter presented with a monetary stake (1000 ms, accompanied by the sentence "You can win[/lose] X euros") and asked to report their confidence C in their answer on a scale from 50 to 100% by moving a cursor with the left and right arrows, and selecting their desired answer by pressing the spacebar (self-paced). The steps following the confidence rating, and the relation between the monetary stake, the confidence and the correctness of the answer were manipulated in two main versions of this task. In the ***Extended Version,*** at the trial level, the *lottery draw* step was separated in two smaller steps. First, a lottery number L was drawn in a uniform distribution between 50 and 100% and displayed as a scale under the confidence scale. After 1200 ms, the scale with the highest number was highlighted for 1200 ms. Then, during the *resolution step*, if C happened to be higher than L, a clock was displayed for 750 ms together with the message "*Please wait*". Then, a *feedback* was displayed which depended on the correctness of the initial choice. Back at the *resolution step*, if the L happened to be higher than C, the lottery was implemented. A wheel of fortune, with a L% chance of losing was displayed, and played; the lottery arm spun for ~750 ms, and would end up in the winning (green) area with L% probability or in the losing (red) area with 1-L% probability. Then, a *feedback* informed whether the lottery was winning or losing.

Subject would win (gain frame) or not lose (loss frame) the incentive in case of a "winning" trial, and they would not win (gain frame) or lose (loss frame) the incentive in case of a "losing" trial. Thanks to this MP procedure, the strategy to maximize one's earnings is to always report on the confidence scale one's subjective probability of being correct as truthfully and accurately as possible (**SI Text**). Subjects were explicitly instructed so. In addition to extensive instructions explaining the MP procedure, participants gained direct experience with this procedure through a series of 24 training trials that did not count towards final payment.

In the ***Short version***, the incentivization scheme was the same as in the ***Extended Version***, but part of it was run in the background. Basically, the lottery scale appeared, and the scale with the highest number was highlighted concomitantly (1200ms).Additionnaly, the *resolution step* was omitted. Still, the complete feedback relative to the lottery and or the correctness of the answer was given to subjects in the *feedback* step. there was no difference in our participants' behavior when using the extended or short version of our task.

In the ***Performance Version***, the layout was similar to the ***Short Version*** (see **Sup. Fig. 2**). The monetary stake screen was accompanied by a different sentence (You may have won[/lost] X euros). The *lottery draw* /comparison step was replaced with a screen of similar duration (1200ms) simply displaying the confidence scale and the chosen rating. A *feedback*, on the correctness of the answer and the trial outcome was finally given at every trial (1000ms).

Stimuli & design: Participants initially performed a 144 trials calibration session (~5min), where they only performed the Gabor contrast discrimination task, without incentive or confidence measure (**Fig. 1.A**). During this calibration, the distribution of contrast difference (i.e. difficulty) was adapted every 12 trials following a staircase procedure, such that performance reached approximatively 70% correct.

The calibration data was used to estimate individual psychometric function:

$$p(\mathbf{ch_L}) = 1 + \exp(-\mu - \sigma \times (\mathbf{C_L} - \mathbf{C_R}))^{-1}$$

where $p(\mathbf{ch_L})$ is the probability of subjects choosing the left Gabor, and $\mathbf{C_L}$ and $\mathbf{C_R}$ are the contrast intensities of the left and right Gabors. In this formalization, μ quantifies subjects' bias toward choosing the left Gabor in the absence of evidence, and σ quantifies subjects' sensitivity to contrast difference. The estimated parameters (μ and σ) were used to generate stimuli for the confidence task, spanning defined difficulty levels (i.e. known $p(\mathbf{ch_L})$) for all incentives levels. After the first session of the confidence task, μ and σ were re-estimated for each session from the data of the preceding session (experiments 1, 2 and 4), or from a new calibration session (experiment 3).

Metacognitive metrics. Calibration was computed as

$$\mathbf{C} = \frac{1}{n}\sum_{k=1}^{n} C_k - \frac{1}{n}\sum_{k=1}^{n} P_k$$

where n is the total number of trials, Ck is the reported confidence at trial k, and Pk is the performance at trial k (1 for a correct answer and 0 for an incorrect answer);

Discrimination was computed as

$$\mathbf{D} = \frac{1}{n_c}\sum_{k_c=1}^{n_c} C_{k_c} - \frac{1}{n_i}\sum_{k_i=1}^{n_i} C_{k_i},$$

where n_c (respectively n_i) is the number of correct (respectively incorrect answer).

Statistics. All statistical analyses were performed with Matlab R2015a. All statistical analyses reported in the main text result from linear mixed-models (estimated with the glmefit function). For each behavioral measure Y (e.g. confidence, performance, calibration, discrimination), we computed the average of Y per incentive level per individual. For the confidence formation model, we used the regression coefficient from the individual linear regressions linking confidence and evidence for correct and incorrect choices, estimated per individual and incentive level. We then used as predictor variables the incentive absolute value (|V|), the incentive net value (V) and the incentive valence (+/-, only for Experiments 3 and 4). All mixed models included random intercepts and random slopes. As an example, in Wilkinson-Rogers notation, the linear mixed-effect models for Experiment 1 can be written: Y ~ 1 + |V| + V + (1 + |V| + V |Subject). Detailed results on all linear mixed-effect models used in the manuscript can be found in the **SI results**.

Acknowledgments. ML, JL and RJvH were supported by Amsterdam Brain and Cognition (ABC) Talent Grants (Universiteit van Amsterdam). ML was additionally supported by an NWO Veni Fellowship, and the Bettencourt Schueller Fondation. AEG was supported by an NWO VIDI grant. The authors thank Stefano Palminteri, Ivan Soraperra, Frans van Winden, Jan B. Engelmann and Joel van der Weele for helpful discussions and comments on the manuscript, and Thomas A. Davison for checking the English.

References

1. Fleming, S. M. & Dolan, R. J. The neural basis of metacognitive ability. *Phil Trans R Soc B* **367,** 1338–1349 (2012).
2. Yeung, N. & Summerfield, C. Metacognition in human decision-making: confidence and error monitoring. *Philos. Trans. R. Soc. B Biol. Sci.* **367,** 1310–1321 (2012).
3. Adams, J. K. A Confidence Scale Defined in Terms of Expected Percentages. *Am. J. Psychol.* **70,** 432–436 (1957).
4. Pouget, A., Drugowitsch, J. & Kepecs, A. Confidence and certainty: distinct probabilistic quantities for different goals. *Nat. Neurosci.* **19,** 366–374 (2016).
5. Folke, T., Jacobsen, C., Fleming, S. M. & Martino, B. D. Explicit representation of confidence informs future value-based decisions. *Nat. Hum. Behav.* **1,** 0002 (2016).
6. Donoso, M., Collins, A. G. E. & Koechlin, E. Foundations of human reasoning in the prefrontal cortex. *Science* **344,** 1481–1486 (2014).
7. Vinckier, F. *et al.* Confidence and psychosis: a neuro-computational account of contingency learning disruption by NMDA blockade. *Mol. Psychiatry* **21,** 946–955 (2016).
8. Lichtenstein, S., Fischhoff, B. & Phillips, L. D. Calibration of probabilities: the state of the art to 1980. in *Judgment Under Uncertainty: Heuristics and Biases* (eds. Kahneman, D., Slovic, P. & Tversky, A.) 306–334 (Cambridge University Press, 1982).
9. Baranski, J. V. & Petrusic, W. M. The calibration and resolution of confidence in perceptual judgments. *Percept. Psychophys.* **55,** 412–428 (1994).
10. West, R. F. & Stanovich, K. E. The domain specificity and generality of overconfidence: Individual differences in performance estimation bias. *Psychon. Bull. Rev.* **4,** 387–392 (1997).
11. Berner, E. S. & Graber, M. L. Overconfidence as a Cause of Diagnostic Error in Medicine. *Am. J. Med.* **121,** S2–S23 (2008).
12. Camerer, C. & Lovallo, D. Overconfidence and Excess Entry: An Experimental Approach. *Am. Econ. Rev.* **89,** 306–318 (1999).

13. Malmendier, U. & Tate, G. CEO Overconfidence and Corporate Investment. *J. Finance* **60,** 2661–2700 (2005).
14. Kunda, Z. The case for motivated reasoning. *Psychol. Bull.* **108,** 480 (1990).
15. Bénabou, R. & Tirole, J. Mindful Economics: The Production, Consumption, and Value of Beliefs. *J. Econ. Perspect.* **30,** 141–164 (2016).
16. Epley, N. & Gilovich, T. The Mechanics of Motivated Reasoning. *J. Econ. Perspect.* **30,** 133–140 (2016).
17. Giardini, F., Coricelli, G., Joffily, M. & Sirigu, A. Overconfidence in Predictions as an Effect of Desirability Bias. in *Advances in Decision Making Under Risk and Uncertainty* (eds. Abdellaoui, P. M. & Hey, P. D. J. D.) 163–180 (Springer Berlin Heidelberg, 2008). doi:10.1007/978-3-540-68437-4_11
18. Koellinger, P. & Treffers, T. Joy Leads to Overconfidence, and a Simple Countermeasure. *PLOS ONE* **10,** e0143263 (2015).
19. Jönsson, F. U., Olsson, H. & Olsson, M. J. Odor emotionality affects the confidence in odor naming. *Chem. Senses* **30,** 29–35 (2005).
20. Allen, M. *et al.* Unexpected arousal modulates the influence of sensory noise on confidence. *eLife* **5,** e18103 (2016).
21. De Martino, B., Fleming, S. M., Garrett, N. & Dolan, R. J. Confidence in value-based choice. *Nat. Neurosci.* **16,** 105–110 (2013).
22. Lebreton, M., Abitbol, R., Daunizeau, J. & Pessiglione, M. Automatic integration of confidence in the brain valuation signal. *Nat. Neurosci.* **18,** 1159–1167 (2015).
23. Levy, D. J. & Glimcher, P. W. The root of all value: a neural common currency for choice. *Curr. Opin. Neurobiol.* **22,** 1027–1038 (2012).
24. Massoni, S. Emotion as a boost to metacognition: How worry enhances the quality of confidence. *Conscious. Cogn.* **29,** 189–198 (2014).

25. Becker, G. M., DeGroot, M. H. & Marschak, J. Measuring Utility by a Single-Response Sequential Method. *Behav. Sci.* **9,** 226–232 (1964).
26. Ducharme, W. M. & Donnell, M. L. Intrasubject comparison of four response modes for "subjective probability" assessment. *Organ. Behav. Hum. Perform.* **10,** 108–117 (1973).
27. Schotter, A. & Trevino, I. Belief Elicitation in the Laboratory. *Annu. Rev. Econ.* **6,** 103–128 (2014).
28. Schlag, K. H., Tremewan, J. & Weele, J. J. van der. A penny for your thoughts: a survey of methods for eliciting beliefs. *Exp. Econ.* **18,** 457–490 (2015).
29. Smith, V. L. & Walker, J. M. Monetary Rewards and Decision Cost in Experimental Economics. *Econ. Inq.* **31,** 245–261 (1993).
30. Bonner, S. E. & Sprinkle, G. B. The effects of monetary incentives on effort and task performance: theories, evidence, and a framework for research. *Account. Organ. Soc.* **27,** 303–345 (2002).
31. Wilcox, N. T. Lottery Choice: Incentives, Complexity and Decision Time. *Econ. J.* **103,** 1397–1417 (1993).
32. Botvinick, M. & Braver, T. Motivation and Cognitive Control: From Behavior to Neural Mechanism. *Annu. Rev. Psychol.* **66,** 83–113 (2015).
33. Sanders, J. I., Hangya, B. & Kepecs, A. Signatures of a Statistical Computation in the Human Sense of Confidence. *Neuron* **90,** 499–506 (2016).
34. Drugowitsch, J., Moreno-Bote, R. & Pouget, A. Relation between Belief and Performance in Perceptual Decision Making. *PLOS ONE* **9,** e96511 (2014).
35. Drugowitsch, J. Becoming Confident in the Statistical Nature of Human Confidence Judgments. *Neuron* **90,** 425–427 (2016).
36. Tversky, A. & Kahneman, D. Loss Aversion in Riskless Choice: A Reference-Dependent Model. *Q. J. Econ.* **106,** 1039–1061 (1991).
37. Hollard, G., Massoni, S. & Vergnaud, J.-C. In search of good probability assessors: an experimental comparison of elicitation rules for confidence judgments. *Theory Decis.* **80,** 363–387 (2015).

38. Schwarz, N. & Clore, G. L. Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *J. Pers.* **45,** 513–523 (1983).

39. Trautmann, S. T. & van de Kuilen, G. Belief Elicitation: A Horse Race among Truth Serums. *Econ. J.* **125,** 2116–2135 (2015).

40. Wakker, P. P. On the composition of risk preference and belief. *Psychol. Rev.* **111,** 236 (2004).

41. Kandasamy, N. *et al.* Interoceptive Ability Predicts Survival on a London Trading Floor. *Sci. Rep.* **6,** 32986 (2016).

42. Song, C. *et al.* Relating inter-individual differences in metacognitive performance on different perceptual tasks. *Conscious. Cogn.* **20,** 1787–1792 (2011).

43. Harsay, H. A., Cohen, M. X., Reneman, L. & Ridderinkhof, K. R. How the aging brain translates motivational incentive into action: The role of individual differences in striato-cortical white matter pathways. *Dev. Cogn. Neurosci.* **1,** 530–539 (2011).

44. Navajas, J., Bahrami, B. & Latham, P. E. Post-decisional accounts of biases in confidence. *Curr. Opin. Behav. Sci.* **11,** 55–60 (2016).

45. Olsson, H. & Winman, A. Underconfidence in sensory discrimination: The interaction between experimental setting and response strategies. *Percept. Psychophys.* **58,** 374–382 (1996).

46. Juslin, P., Winman, A. & Olsson, H. Naive empiricism and dogmatism in confidence research: A critical examination of the hard-easy effect. *Psychol. Rev.* **107,** 384–396 (2000).

47. Johnson, D. D. P. & Fowler, J. H. The evolution of overconfidence. *Nature* **477,** 317–320 (2011).

48. Marshall, J. A. R., Trimmer, P. C., Houston, A. I. & McNamara, J. M. On evolutionary explanations of cognitive biases. *Trends Ecol. Evol.* **28,** 469–473 (2013).

49. Meyniel, F., Sigman, M. & Mainen, Z. F. Confidence as Bayesian Probability: From Neural Origins to Behavior. *Neuron* **88,** 78–92 (2015).

50. Hebart, M. N., Schriever, Y., Donner, T. H. & Haynes, J.-D. The Relationship between Perceptual Decision Variables and Confidence in the Human Brain. *Cereb. Cortex* bhu181 (2014). doi:10.1093/cercor/bhu181

51. Goodie, A. S. The role of perceived control and overconfidence in pathological gambling. *J. Gambl. Stud.* **21,** 481–502 (2005).

52. Goldstein, R. Z. *et al.* Is decreased prefrontal cortical sensitivity to monetary reward associated with impaired motivation and self-control in cocaine addiction? *Am. J. Psychiatry* **164,** 43–51 (2007).

53. Figee, M. *et al.* Dysfunctional reward circuitry in obsessive-compulsive disorder. *Biol. Psychiatry* **69,** 867–874 (2011).

54. Whitton, A. E., Treadway, M. T. & Pizzagalli, D. A. Reward processing dysfunction in major depression, bipolar disorder and schizophrenia. *Curr. Opin. Psychiatry* **28,** 7 (2015).

	Exp. 1	Exp. 2	Exp. 3	Exp. 4
M/F (total)	6/18 (24)	7/17 (24)	9/26 (35)	7/17 (24)
Age	23.8±6.00	23.0±5.67	24.8±5.43	24.3±7.79
Stakes (€)	0; ±1;	0; ±1;	±0.1; ±1;	±0.1; ±1;±2
P(c=L)	50 ± [10:10:40] (×4)	50±[15:10:35] (×2)	50±[0:05:15] (×2)	50±[0:10:40] (×2)
Tasks	Short (×4)	Short (×2) + Perf (×2)	Ext. (×1) + Short(×1)	Short (×3)

Table.1: Demographics and experimental design. P(c=L) indicates the level of difficulty (i.e probability of choosing the left Gabor) used to generate the stimuli. The number of times all levels of difficulty were repeated per incentive and per session is indicated between brackets. Tasks indicate which task version (short: confidence incentivization Short Version, extended: confidence incentivization Extended Version, perf: Performance incentivization Version) was offered to participants, and the number of session per task is indicated between brackets.

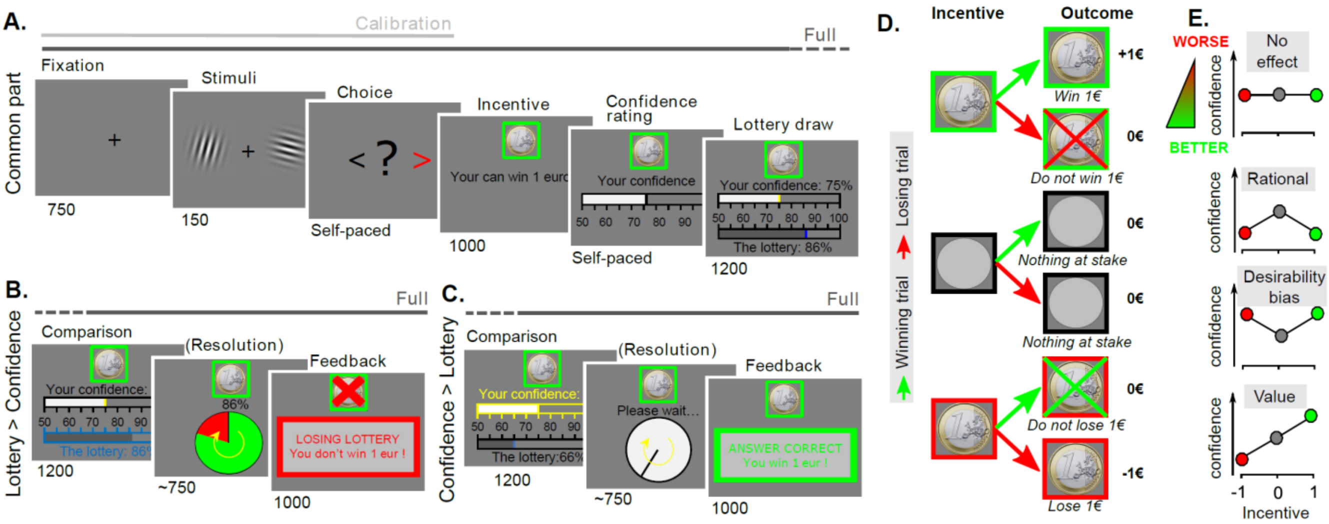

Fig. 1. Behavioral task and hypotheses.

Successive screens displayed in one trial are shown from left to right with durations in ms.
A. Behavioral task - Common part. Participants viewed a couple of Gabor patches displayed on both sides of a computer screen, and judged which had the highest contrast. They were thereafter presented with a monetary stake (in a green frame for gain, grey for neutral and red for losses), and asked to report their confidence C in their answer on a scale from 50 to 100%. Then, a lottery number L was drawn in a uniform distribution between 50 and 100%, displayed as a scale under the confidence scale and the scale with the highest number was highlighted.
B. Behavioral task - Lottery > Confidence. If the L > C, the lottery was implemented. A wheel of fortune, with a L% chance of losing was displayed, and played. Then, a *feedback* informed whether the lottery was winning or losing.
C. Behavioral task - Confidence > Lottery. If C > L, a clock was displayed together with the message "*Please wait*", followed by a *feedback* which depended on the correctness of the initial choice.
Subject would win (gain frame) or not lose (loss frame) the incentive in case of a "winning" trial, and they would not win (gain frame) or lose (loss frame) the incentive in case of a "losing" trial.
D. Behavioral task – Payoff matrix. Depending on the combination of a trial's offered incentive and the trial's final win or loss (regardless of whether the lottery or the correctness of the answer determined it), participants could receive various outcomes, from winning the proposed incentive to losing the proposed incentive.
E. Hypotheses. The expected modulation of confidence by incentives (-1€; 0€ or +1€)) is presented under different theoretical hypotheses. (1st line) H0: No effects of incentives. Participants are similarly overconfident in the 3 incentive conditions. (2nd line) H1: rational decision making. Under higher incentives, participant are more rational, i.e. better calibrated. The absolute value of incentives therefore decreases confidence, if participants are generally overconfident. (3rd line) H2: desirability bias. Participants are more inclined to believe that they are correct when higher incentives are at stake. The absolute value of incentives increases confidence. (4th line) H3: value-confidence interaction. The confidence judgment of participants is affected by the affective component of incentives. The incentive net value impacts confidence.

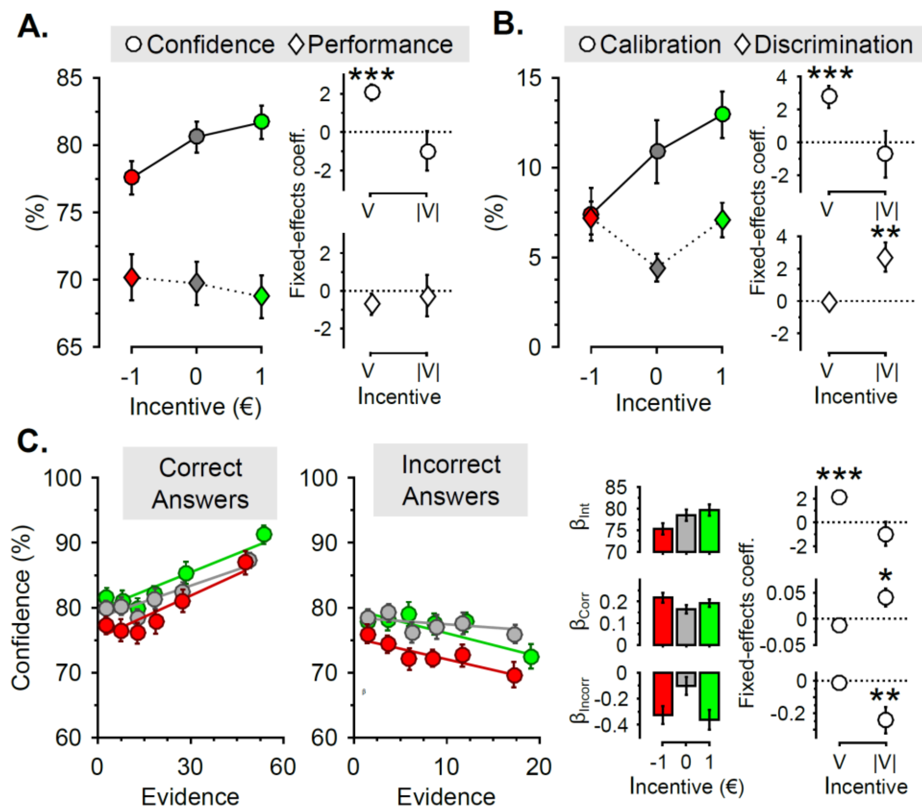

Fig. 2. Experiment 1.

A. Incentive effects on behavior. Reported confidence (dots) and performance (diamonds) –i.e. % correct- as a function of incentives. The insets presented on the right-hand side of the graph depict the results of the linear mixed-effect model, estimated for each behavioral measure.

B. Incentive effects on confidence accuracy. Computed calibration (dots) and discrimination (diamonds) as a function of incentives. The insets presented on the right-hand side of the graph depict the results of the linear mixed-effect model, estimated for each behavioral measure.

C. Incentive effects on confidence formation

Linking incentives, evidence and confidence for correct (left) and incorrect (right) answers. In those two panels, the scatter plots display reported confidence as a function of evidence, for the different incentive levels. The solid line represents the best linear regression fit at the population level.

The histograms represent the intercepts (top) and slope for correct (middle) and incorrect answers (bottom) of this relationship, estimated at the individual level and averaged at the population level.

The insets presented on the right-hand side of the graph depict the results of the linear mixed-effect model, estimated for each parameter of this regression (i.e. intercept (top), slope for correct answers (middle) and for incorrect answers (bottom).

V: incentive net value. |V|: incentive absolute value. Error bars indicate inter-subject standard errors of the mean. *: P<0.05; ** P<0.01; ***P<0.001;

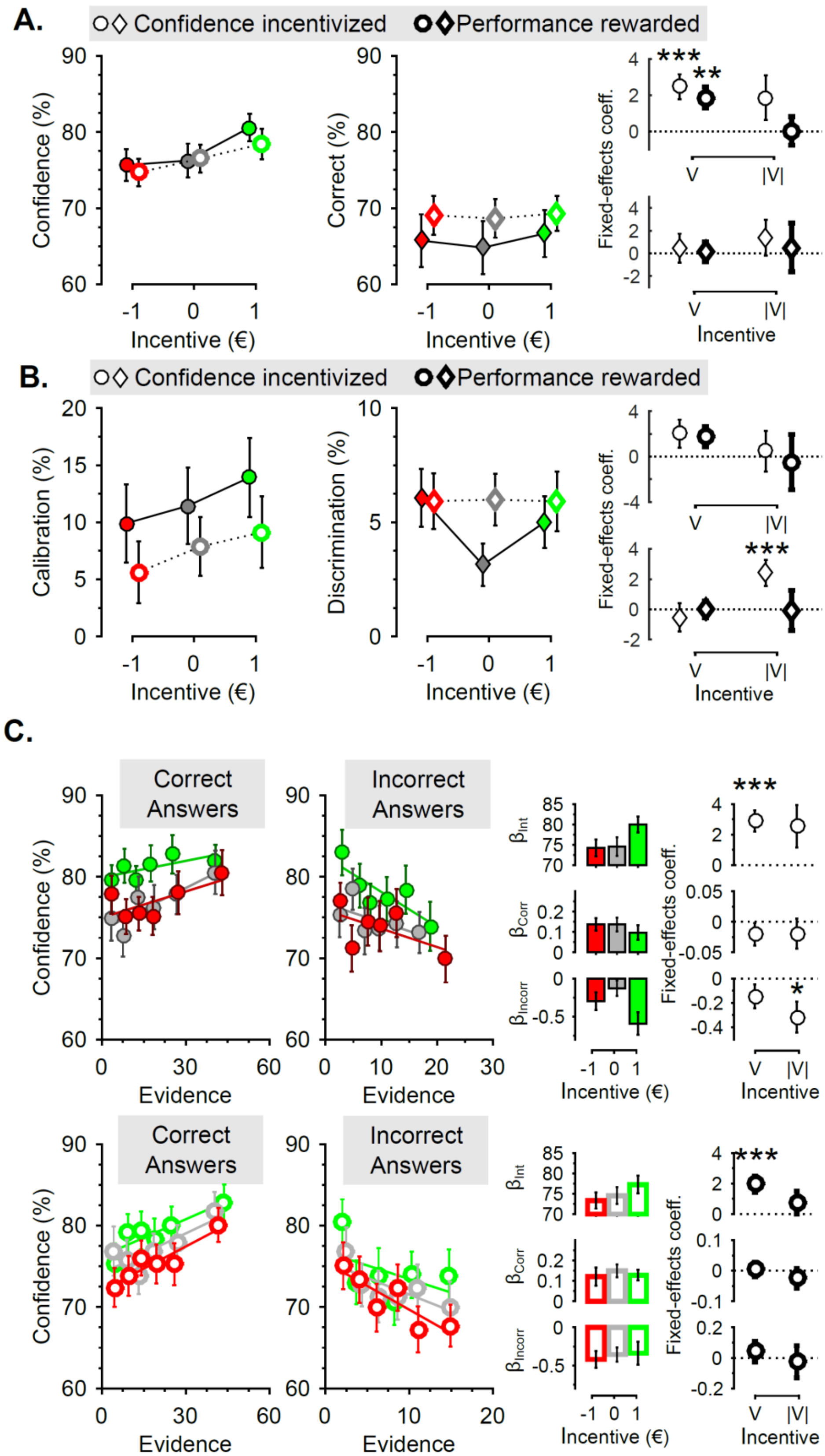

Fig. 3. Experiment 2.

A. Incentive effects on behavior. Reported confidence (dots – leftmost scatter plot) and performance (diamonds – rightmost scatter plot) –i.e. % correct- as a function of incentives. The insets presented on the right-hand side of the graph depict the results of the linear mixed-effect model, estimated for each behavioral measure. Empty markers with thick edges indicate the Performance Rewarded task.

B. Incentive effects on confidence accuracy. Computed calibration (dots – leftmost scatter plot) and discrimination (diamonds – rightmost scatter plot) as a function of incentives. The insets presented on the right-hand side of the graph depict the results of the linear mixed-effect model, estimated for each behavioral measure. Empty markers with thick edges indicate the Performance Rewarded task.

C. Incentive effects on confidence formation

Linking incentives, evidence and confidence for the Confidence incentivized (top row) and the Performance Rewarded (bottom row) tasks, for both correct (left) and incorrect (right) answers. In those two panels, the scatter plots display reported confidence as a function of evidence, for the different incentive levels. The solid line represents the best linear regression fit at the population level.

The histograms represent the intercepts (top) and slope for correct (middle) and incorrect answers (bottom) of this relationship, estimated at the individual level and averaged at the population level.

The insets presented on the right-hand side of the graph depict the results of the linear mixed-effect model, estimated for each parameter of this regression (i.e. intercept (top), slope for correct answers (middle) and for incorrect answers (bottom).

V: incentive net value. |V|: incentive absolute value. Error bars indicate inter-subject standard errors of the mean. *: $P<0.05$; ** $P<0.01$; ***$P<0.001$;

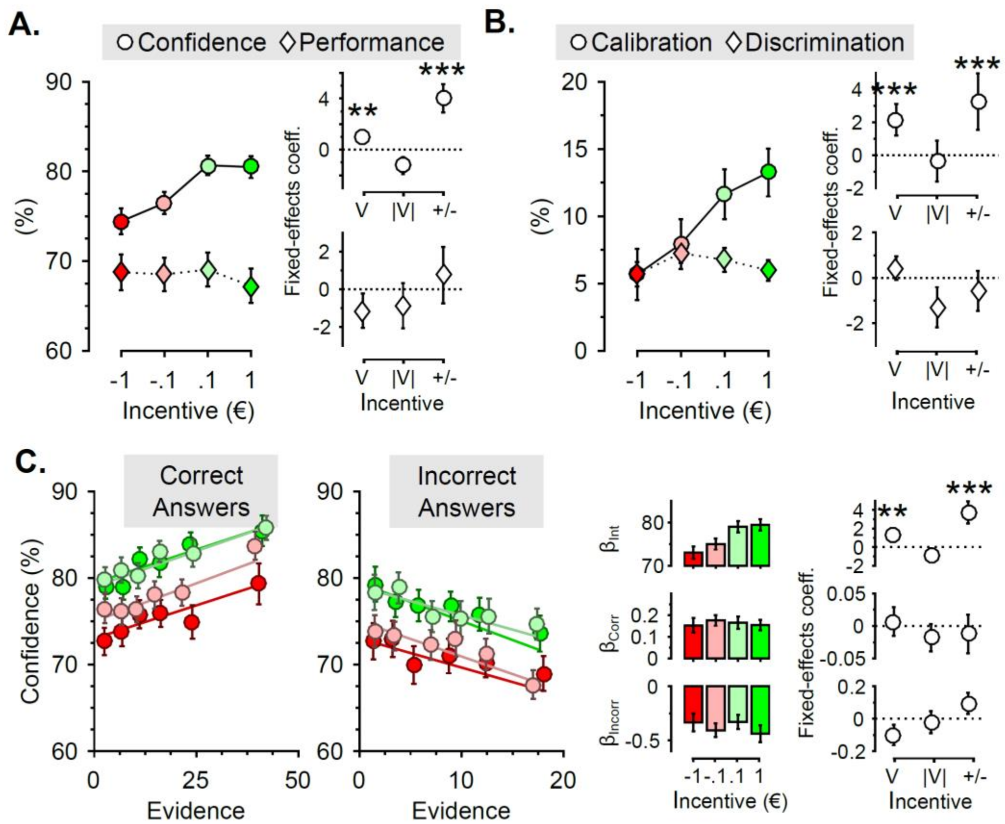

Fig. 4. Experiment 3.

A. Incentive effects on behavior. Reported confidence (dots) and performance (diamonds) –i.e. % correct- as a function of incentives. The insets presented on the right-hand side of the graph depict the results of the linear mixed-effect model, estimated for each behavioral measure.

B. Incentive effects on confidence accuracy. Computed calibration (dots) and discrimination (diamonds) as a function of incentives. The insets presented on the right-hand side of the graph depict the results of the linear mixed-effect model, estimated for each behavioral measure.

C. Incentive effects on confidence formation

Linking incentives, evidence and confidence for correct (left) and incorrect (right) answers. In those two panels, the scatter plots display reported confidence as a function of evidence, for the different incentive levels. The solid line represents the best linear regression fit at the population level.

The histograms represent the intercepts (top) and slope for correct (middle) and incorrect answers (bottom) of this relationship, estimated at the individual level and averaged at the population level.

The insets presented on the right-hand side of the graph depict the results of the linear mixed-effect model, estimated for each parameter of this regression (i.e. intercept (top), slope for correct answers (middle) and for incorrect answers (bottom).

V: incentive net value. |V|: incentive absolute value. +/-: incentive valence. Error bars indicate inter-subject standard errors of the mean. *: P<0.05; ** P<0.01; ***P<0.001;

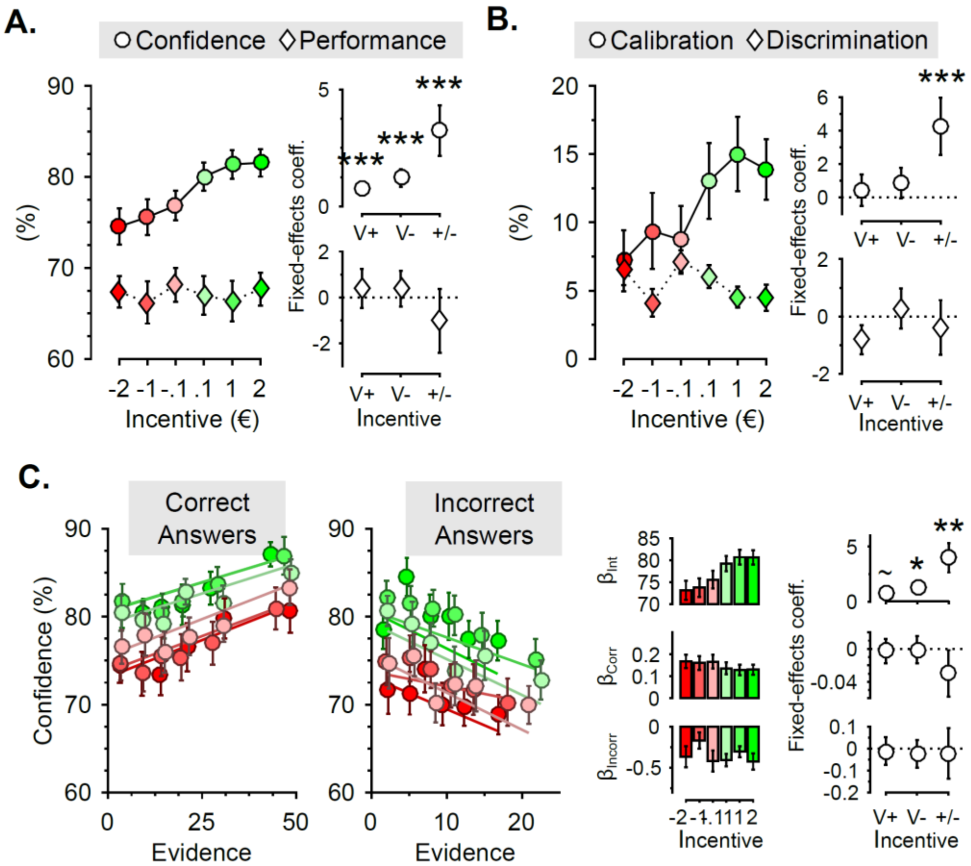

Fig. 5. Experiment 4.

A. Incentive effects on behavior. Reported confidence (dots) and performance (diamonds) –i.e. % correct- as a function of incentives. The insets presented on the right-hand side of the graph depict the results of the linear mixed-effect model, estimated for each behavioral measure.

B. Incentive effects on confidence accuracy. Computed calibration (dots) and discrimination (diamonds) as a function of incentives. The insets presented on the right-hand side of the graph depict the results of the linear mixed-effect model, estimated for each behavioral measure.

C. Incentive effects on confidence formation

Linking incentives, evidence and confidence for correct (left) and incorrect (right) answers. In those two panels, the scatter plots display reported confidence as a function of evidence, for the different incentive levels. The solid line represents the best linear regression fit at the population level.

The histograms represent the intercepts (top) and slope for correct (middle) and incorrect answers (bottom) of this relationship, estimated at the individual level and averaged at the population level.

The insets presented on the right-hand side of the graph depict the results of the linear mixed-effect model, estimated for each parameter of this regression (i.e. intercept (top), slope for correct answers (middle) and for incorrect answers (bottom).

V+: incentive net value for gains. V-: incentive net value for losses. +/-: incentive valence. Error bars indicate inter-subject standard errors of the mean. ~P<0.10; *: P<0.05; ** P<0.01; ***P<0.001;