Non-sensory information affects confidence during perceptual decision-making

Nicolás Sánchez-Fuenzalida[12], Simon van Gaal[12] , Stephen M. Fleming[345], Julia M. Haaf1 & Johannes J. Fahrenfort[126]

Abstract

In noisy environments, people frequently make decisions based on non-sensory information to maximize rewards. Therefore, a central problem in perceptual decision-making and consciousness research is to distinguish between decisions resulting from changes in sensory experience and those arising from non-sensory information. Sensory experience is often gauged using a decision measure combined with a confidence rating. It has recently been proposed that such metacognitive reports can be used to discern between perceptual and non-perceptual effects. Here we show across two experiments and three bias manipulations that confidence during perceptual decision-making is not uniquely affected by sensory information. Instead, non-sensory manipulations affecting response bias 'leak' into perceptual confidence reports. This occurs for biases resulting from changes in the base-rate of stimuli ('cognitive' priors), but also when biasing information does not inform decision correctness (asymmetric payoff matrix). These results provide compelling evidence that confidence reports are affected by non-perceptual influences during perceptual decision-making.

Keywords: consciousness, perceptual decision-making, decision bias, rewards, predictions, visual illusion, confidence

1 Department of Psychology, University of Amsterdam, The Netherlands
2 Amsterdam Brain & Cognition, University of Amsterdam, The Netherlands.
3 Wellcome Centre for Human Neuroimaging, Institute of Neurology, University College London, UK.
4 Department of Experimental Psychology, University College London, UK
5 Max Planck Centre for Computational Psychiatry and Ageing Research, University College London, UK.
6 Department of Applied and Experimental Psychology, Vrije Universiteit Amsterdam, The Netherlands

Introduction

Humans combine both sensory and non-sensory information when making perceptual decisions. For example, when facing uncertain sensory information, an airport security guard may be inclined to report certain shapes on an x-ray as a potential gun given the costly consequences of allowing such an object to go into a plane, even if the perceptual evidence is meagre. Alternatively, you may be pondering whether the person sitting next to you in the train is that famous actor from TV but decide against it because the odds of such an event are slim, despite the perceptual evidence being ample. What these two examples have in common is that it is not immediately clear how to assess whether these decisions stem from changes in observers' subjective experience, or from a sensory-independent decisional process. Indeed, changes in categorization behaviour may occur even if observers' perceptual experience remains the same (Morgan et al., 2012). A central challenge in research on perceptual decision-making is therefore to separate decisions reflecting the sensory experience of observers from decisions reflecting response bias. Traditionally, this problem has been addressed through the development of signal detection theoretic (SDT) models that dissociate sensitivity from bias in decision making (Green & Swets, 1966). However, the bias parameter in these models has proven unable to dissociate perceptual from non-perceptual response shifts (see Witt et al., 2015 for a detailed account). We have recently solved this problem by asking observers to reproduce their experience in a controlled fashion, thereby distinguishing between experimental manipulations that affect sensory experience (visual illusions) from those that only affect response bias (base rate and payoff manipulations) (Sánchez-Fuenzalida et al., 2022).

However, there is also a long tradition of using confidence judgements to assess changes in perceptual experience. Aside from its use to build Receiver Operating Characteristic curves in SDT, confidence has come to be widely used across consciousness science as a marker of the strength of perceptual experience (Kunimoto et al., 2001; Morales & Lau, 2021; Sandberg et al., 2010). The idea that observers have privileged access to their own experience, and that this experience can be gauged through confidence judgments is intuitively appealing. Put simply, decisions based on strong sensory evidence are more likely to be correct and elicit a greater sense of confidence, while decisions based on weak sensory evidence are likely to be incorrect and elicit greater uncertainty. Indeed, a large body of research has pointed to the ability of observers to accurately track their performance on perceptual discrimination tasks using confidence ratings – known as metacognitive sensitivity (Fleming et al., 2012; Fleming & Lau, 2014; Maniscalco & Lau, 2010; Michel, 2023; Nelson, 1984; Yeung & Summerfield, 2012). However, while it is true that both first order judgements (decisions about the stimuli themselves) and second order judgements (confidence ratings in one's own performance) heavily rely on sensory input, both can potentially be affected by non-sensory information (Evans & Azzopardi, 2007; Jachs et al., 2015; Martinez-Saito, 2022; Morales & Lau, 2021). For example, if one knows that one event is more likely to occur, one can report the prevalent event while also reporting high confidence, even if the true identity of the stimuli is completely unknown. While this behaviour can be considered normatively adaptive (as the most prevalent option is more likely to be correct), it also highlights how non-sensory information can affect confidence judgements. Further, there is also evidence to suggest that changes in the reward value associated with a decision can bias confidence reports (Lebreton et al., 2018; Locke et al., 2020), even when the reward does not provide a relevant cue to decision accuracy. In light of this range of possible non-sensory contributions to confidence, it remains underdetermined how and under what circumstances subjective experience is faithfully tracked by confidence reports.

Nevertheless, it has recently been proposed that metacognitive reports can be used as a diagnostic tool to discern between perceptual and non-perceptual effects. Gallagher and colleagues (2019) showed that when a decision bias results from changes in sensory input (motion aftereffect), it is accompanied by a concomitant metacognitive shift, whereas for biases that arise from the decision process alone (default-option strategy), confidence reports remained unaffected. However, the non-perceptual bias manipulation used in their study did not provide any information about decision accuracy. Rather, participants were offered a default answer they could choose whenever they were unsure. Therefore, it remains unclear whether a non-perceptual manipulation that provides information about the likelihood of the occurrence of certain stimuli would result in biased confidence reports. Furthermore, other non-sensory sources of decision bias were not considered. For instance, it is possible that asymmetric stimulus-response reward contingencies might also influence confidence reports, even in the context of a perceptual task. Finally (and critically), it is impossible to unequivocally assess the relationship between confidence and perceptual experience without an independent behavioural benchmark for perceptual experience. Our lab has recently introduced a novel measure (‘controlled reproduction’) to capture whether a given manipulation affects sensory experience, independently from effects of decision bias (Sánchez-Fuenzalida et al., 2022). Here, we use this measure as a benchmark to determine the perceptual or non-perceptual nature of experimental manipulations and ask how such manipulations affect confidence.

To test whether confidence reports are susceptible to decision bias, we used a perceptual decision-making task in which we could bias decision making using both perceptual and non-perceptual manipulations. Across two experiments, we employed three manipulations, each reflecting a different bias source: a sensory manipulation (Müller-Lyer illusion), a punishment scheme (asymmetric payoff matrix), and changes in the ratio of relevant stimuli (base rate manipulation), see figure 1A for a graphical depiction of the manipulations. We replicated previous results from our lab showing that all these bias manipulations result in strong decision biases, while only the Müller-Lyer illusion affects observers’ perceptual experience. Crucially, however, we show that all bias manipulations -- whether perceptual or non-perceptual -- also result in systematic shifts in confidence. Finally, we show that these effects occur both when confidence is reported simultaneously with the decision, and when confidence is delayed until after the decision has taken place.

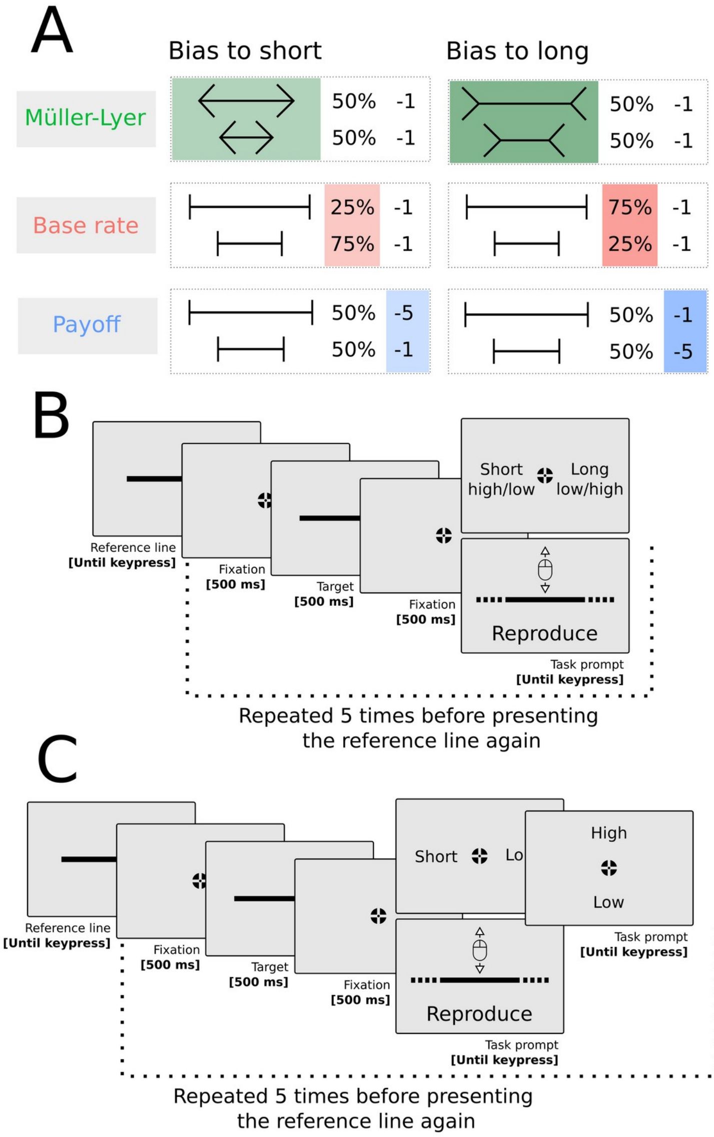


Figure 1: Trial layout and bias manipulations summary. A) Manipulation summary: Target lines presented in the Müller-Lyer condition were flanked by inward-pointing arrowheads when the bias direction was long and by outward-pointing arrowheads when the bias direction was short. In the base rate and payoff condition, vertical lines flanked the target lines. In the base rate condition there were three times more long lines than short lines when the bias direction was long and vice versa when the bias direction was short. In the Müller-Lyer and payoff conditions, there were an equal number of long and short trials. In the payoff condition participants lost 5 points for incorrectly answering long and 1 point for incorrectly answering short when the bias direction was short and vice versa when the bias direction was long. **B) Experiment 1 trial layout:** A typical sequence of trials, from here onward referred to as a mini-block, consisted of the presentation of a reference screen (until keypress) followed by five trials. Each trial consisted of a fixation period (500 ms), followed by a target screen (500 ms), followed by a second fixation period (500 ms). At the end of the trial observers were prompted to either categorise the target line they just saw as being shorter or longer than the reference line (short/long), and to concurrently provide their confidence on their response (high/low), or to reproduce the length of the target line by increasing/decreasing a line by moving the mouse (the prompt was shown until an answer was registered). **C) Experiment 2 trial layout:** Same as A) but observers reported their confidence on their response after the categorisation response.

Methods

Method and analysis are the same for experiment 1 and 2 unless stated otherwise.

Ethics

All experimental procedures were approved by the University of Amsterdam Ethics Review Board. Informed consent was obtained in accordance with the ethics board approved procedures.

Participants

122 participants (29 males, 21.3 years old on average, SD=3.3) and 123 participants (19 males, 20.5 years old on average, SD=2.8) took part in experiment 1 and 2 respectively. All participants were recruited through the University of Amsterdam lab pool and were guaranteed to be rewarded 1.5 research credits (or 15 euros). On top of this, participants could earn an extra reward (up to .5 research credits or 5 euros) that depended on their performance during the experiment. On average the total payment was similar across conditions. Each participant completed one session of roughly 90 minutes in the Müller-Lyer and payoff condition, or 120 minutes in the base rate condition, including instructions, practice and breaks.

Since all the confirmatory analyses in this experiment were conducted in a Bayesian framework, in experiment 1, we collected the data of 30 participants on each of the three bias manipulation conditions, removed outliers, and then ran a Bayesian t-test between the biased to long and short conditions in the decision, confidence and in the reproduction task. If there was moderate evidence for the effect of our manipulation in all tasks (for either the null or the alternative hypothesis), we stopped data collection ($BF^{10} > 3$ or $BF^{10} < 0.3$), otherwise we collected five more subjects and repeated the process. In experiment 2 we aimed to collect the same number of participants as in experiment 1, given that both experiments were identical aside from the moment at which participants provided their confidence rating. In this framework optional stopping or data peaking is not considered problematic (Rouder, 2014). Participants were considered outliers if their signal detection theory criterion, d' or reproduction error fell outside four standard deviations from the sample mean, this is, around the grand average across bias manipulation conditions. Participants with a signal detection theory *d'* below zero were also filtered out. In experiment 1 two participants were removed (one reproduction error outlier in the base rate condition and one participant with *d'* < 0 in the Müller-Lyer condition). In experiment 2 four participants were removed (one reproduction error outlier in the base rate condition and three participants with *d'* < 0, two in the payoff condition and one in the Müller-Lyer condition).

Tasks

On each trial, participants had to either categorise a line as being shorter or longer than a reference line and report their confidence on this decision (decision/confidence task) or they had to reproduce the length of the lines presented to them (controlled reproduction task). Crucially, observers did not know which task they would be performing while they viewed the target line, thus preventing specific task demands from affecting stimulus processing. Confidence was either provided concurrently with their categorization decision (experiment 1) or after the categorization decision was submitted (experiment 2). See figure 1B and C for the trial layout of each experiment. The instructions participants received regarding how to report confidence were similar to those in the experiment by Gallagher et al. (2019) (see supplementary figure S1 and S2 for the full confidence report instructions).

Stimuli

Stimuli consisted of black horizontal lines presented at the centre of the screen over a grey background. The reference line was 400 pixels long and there were seven possible target lengths that ranged between 370 and 430 in steps of 10 pixels. All stimuli were presented on a 23” (58.4 cm) monitor with a resolution of 1920x1080 pixels, at a distance of approximately 75 cm. At this distance the size of each pixel was 0.265mm, or 0.02 visual angle degrees. The monitor refresh rate was 120 Hz. On length reproduction trials, the reproduction line was initially presented with a length of 40 pixels, bounded between 0 and 800 pixels. In the Müller-Lyer condition, each diagonal line that formed the arrowheads on the target line was 60 pixels long, and subtended a 45- or 135-degree angle with the horizontal line. The vertical flanking lines in the payoff and base rate condition were 70 pixels long. All lines had a width of 4 pixels. On each trial a fixation period (500 ms) was followed by the target line (500 ms), further followed by a second fixation period (500 ms). The experiment was programmed on Python 3.6 (Rossum et al., 2010) and Psychopy 2 (Peirce et al., 2019).

Design

We employed a 3 between-participant bias source (Müller-Lyer illusion, payoff and base rate) x 2 nested within-participants bias direction (biased to long and biased to short) design. The three bias conditions were identical except for the following details: In the Müller-Lyer condition target lines were flanked by outward-pointing arrowheads when the bias direction was short, or inward-pointing arrowheads when the bias direction was long. In the payoff and base rate condition target lines were flanked by vertical lines. In the base rate condition, the ratio of target lines that were longer or shorter than the reference line was uneven, so one category was three times more likely to be presented (see Supplementary figure S3 for a histogram describing the frequency of each target length value). In the payoff and Müller-Lyer condition the ratio between target lines that were longer or shorter than the reference was even. In all conditions participants could earn an extra reward that depended on the number of mistakes made during the experiment, but in the payoff condition participants were differentially punished for incorrect decision responses. If the bias direction was short, incorrect ‘long’ responses cost them five times more than incorrect ‘short’ responses, and vice versa when the bias direction was long. In the base rate and Müller-Lyer condition there was a flat penalty of one point for every type of mistake, and in all conditions (including payoff) reproduction mistakes were also punished with one point regardless of the direction of the error. See figure 1A for a graphical depiction of the manipulations.

General procedures

For each task, participants received extensive instructions and extensive practice (see Supplementary text T1 for a detailed description). During the experiment the reference line was presented every five trials (see Figure 1B and C). In the base rate condition, the relative frequency of short and long lines was presented next to the reference line, while in the payoff condition the cost for incorrectly answering ‘short’ and ‘long’ was presented next to the reference line. In the Müller-Lyer condition no extra information was presented with the reference line (see supplementary figure S4 for an example of the reference screen and target screen of each condition). On every trial the presentation of the categorization or reproduction prompt was counterbalanced. After every 50 trials, participants would receive block-level feedback about their performance in the decision and reproduction task during the previous 50 trials. The block-level feedback they received about the decision task was different for the different conditions. In the payoff and base-rate condition, they would be shown how often they had incorrectly answered ‘short’ or ‘long’. In the Müller-Lyer condition they were informed about the overall number of mistakes in the previous 50 trials. The block-level feedback regarding their performance in the reproduction task was the same for all conditions and consisted of the overall number of reproduction errors. A reproduction ‘error’ was defined as a deviation of more than 22 pixels from the actual length of the target line, regardless of the direction of the error (over or underestimation), and it was merely provided to motivate participants to reproduce as accurately as they could. We decided on 22 pixels based on the data from our previous study (Sánchez-Fuenzalida et al., 2022) as this threshold would yield on average above chance performance. Additionally, participants received feedback on their relative use of low and high confidence reports. Participants received an extra message reminding them to use the low and high confidence option evenly if they answered ‘low’ or ‘high’ in less than 25% of the trials of a 50-trial block. An example of a block-level feedback screen is provided in Supplementary figure S5.

The experiment was divided into two bias direction blocks (the order was counterbalanced so roughly half of the participants started with the biased to long condition). In the Müller-Lyer and payoff condition observers completed 280 decision trials and 140 reproduction trials per bias direction, summing up to 840 in total (420 per bias direction). Each of the seven target lines was presented 40 times in the decision/confidence task and 20 in the reproduction task. In the base rate condition observers completed 1170 trials, where the decision/confidence and reproduction task was respectively presented 390 and 195 times. In the decision/confidence task, depending on the bias direction, the targets in the more prevalent category were presented 120, 90 and 60 times going from the farthest (longest/shortest) to the closest length to the reference line (see Supplementary figure S3 for a histogram describing the frequency of each target length value). To increase the effect of the base rate manipulation we made the shortest or longest target lines more prevalent within the prevalent category (short/long). The target lines of the non-frequent category and the target that had the same length of the reference line were presented 30 times each. In the reproduction task each target was presented half of the times as described for the decision/confidence task. Every 50 trials participants received block-level feedback and could choose to have a break or to continue immediately. In addition, they were forced to take a 3 minute break every 200 trials.

Analysis

Signal detection analysis. To determine performance and bias on the tasks we computed signal detection sensitivity (*d’*) and criterion (*c*) based on hit rate and false alarms as follows:

$d' = Z(HR) - Z(FAR)$

and

$c = 1/2 \times (Z(HR) + Z(FAR))$

Where *Z()* denotes the inverse of the standard normal cumulative distribution (often denoted as the Z-transform, as it has a mean of 0 and a standard deviation of 1). The formula can be easily translated to R code by replacing the *Z()* with the *qnorm()* function from the R stats package. HR denotes hit rate, FAR denotes false alarm rate. In this setting correct ‘long’ responses are considered hits and correct ‘short’ responses correct rejections.

Curve fitting. For the decision data (categorization task) we fitted a Cumulative Gaussian function to each participant's distribution of ‘short’ and ‘long’ responses as a function of the length of the target lines separately for each bias source and bias direction condition. The point of subjective equality (PSE), that is, where the probability of answering ‘short’ was 50%, was interpolated from each fitted curve. For the confidence data we fitted a quadratic function to each participant’s distribution of ‘low’ and ‘high’ responses as a function of the length of the target lines separately for each bias source and bias direction condition. The point of peak uncertainty was then interpolated by identifying the lowest confidence point for each fitted curve.

Results

Experiment 1: Simultaneous decision and confidence rating

Across all conditions observers were able to distinguish between short and long lines (see Supplementary Figure S6 for sensitivity data for each condition). To assess decision bias, we fitted a cumulative Gaussian function to each participant’s distribution of ‘short’ and ‘long’ responses as a function of the length of the target lines, separately for each bias source and bias direction condition (see figure 2 top row). We then approximated the point of subjective equality (PSE), that is, the target length at which observers are equally likely to answer ‘short’ or ‘long’. Overall, all manipulations resulted in large effects such that observers preferred the biased choice (‘short’ or ‘long’ depending on the bias direction), resulting in effect sizes (Cohen’s d) of 0.9 (Müller-Lyer), 1.2 (base-rate) and 1.72 (payoff). A simple paired Bayesian t-test showed extreme evidence for a difference between the biased to long and biased to short conditions in all bias manipulations (BF^{10} > 100 for all conditions; all t-tests reported are one-sided and have a default Cauchy prior of $\sqrt{2} \div 2$ unless stated otherwise). These effects replicate our previous findings that all these manipulations result in a response shift towards the biased option.

Next, to assess whether these decision shifts resulted from changes in perception we fitted a straight line to observers’ distribution of length reproductions as a function of the length of the target lines, separately for each bias source and bias direction condition (see figure 2 middle row). We then estimated the target line (x-axis) for which observers reproduced the length of the reference line (x-axis). While the Müller-Lyer condition showed a large effect (*d* = 1.03) reflecting that biased to long lines were reproduced as longer and biased to short were reproduced as shorter, the reproduction error magnitudes in the base rate and payoff conditions were nearly identical for the biased to short and biased to long conditions (*d* = -0.13 and *d* = 0.26). A paired Bayesian t-test revealed extreme evidence for an effect in the Müller-Lyer condition (BF^{10} = 230), and substantial evidence for a null effect in the base rate and payoff conditions (BF^{10} = 0.1 and BF^{10} = 0.08, respectively; see Supplementary Figure S7 for the decision and reproduction results plotted as in (Sánchez-Fuenzalida et al., 2022). Thus, as in (Sánchez-Fuenzalida et al., 2022) we show that despite the fact that decision bias for payoff and base-rate are largest compared to Müller-Lyer, reproduction effects uniquely occur in Müller-Lyer and not in payoff or base-rate, confirming that only the Müller-Lyer manipulation causes a change in perception.

Finally, we assessed whether confidence reports were able to distinguish between perceptual and non-perceptual manipulations as identified by the reproduction task. We reasoned that if confidence uniquely captures perceptual shifts, the bias manipulations towards ‘long’ and ‘short’ should translate into confidence shifts only for the Müller-Lyer, but not for the other bias manipulations. To assess shifts in confidence, we fitted a quadratic function to observers’ distribution of ‘low’ and ‘high’ confidence responses as a function of the length of the target lines, separately for each bias source and bias direction condition (see figure 2 bottom row). We followed Gallagher et al. (2019) and approximated the bottom point of each fitted curve to estimate the target length associated with the point of maximal uncertainty. All manipulations resulted in medium to large sized effects and a paired Bayesian t-test revealed very strong evidence for a difference in confidence between bias to long and bias to short in all bias manipulations ($BF^{10} > 60$ for all conditions). That is, when the bias direction was ‘long’ observers were most uncertain when the length of the target line was slightly shorter than the reference line, and vice versa when the bias direction was ‘short’.

Taken together, these data replicate our previous results by showing that only the Müller-Lyer illusion had an effect both on observers’ decisions (decision task) and subjective experience (controlled reproduction task), whereas the payoff and base rate manipulations affected observers’ decisions without changing perception. Crucially, however, we further show that all three manipulations resulted in biased confidence reports, regardless of whether they affected perceptual experience.

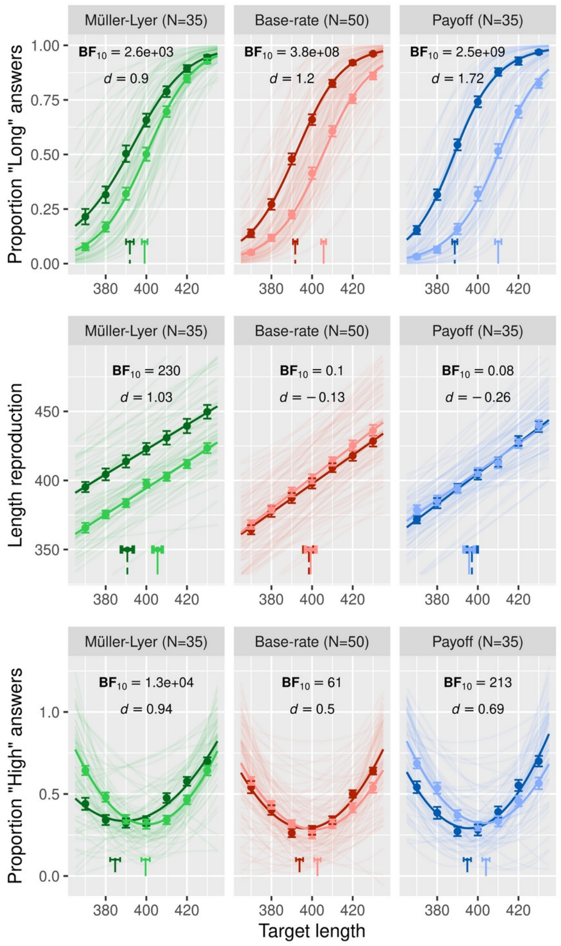


Figure 2: Decision concurrent confidence results. Top row: Decision bias. A cumulative Gaussian function was fitted to each observer’s distribution of ‘short’ and ‘long’ answers as a function of the length of the target lines separately for each bias source and bias direction condition. The group-level point of subjective equality (PSE) is plotted at the bottom of each panel. **Middle row: Length reproduction.** Same as described for the decision bias except that a straight line was fitted to the distribution of length reproductions. At the bottom of each panel is plotted the target length associated with reproductions equal to the reference line (400 pixels) at the group-level. **Bottom row: Confidence bias.** Same as described for the decision bias except that a quadratic function was fitted to the distribution of ‘high’ and ‘low’ confidence

responses. The group-level point of peak uncertainty (lowest confidence rating) is plotted at the bottom of each panel. Across all plots dots and error bars correspond to empirical data along with fitted functions plotted as lines. All error bars indicate one standard error from the mean. Conditions biased to long are plotted as dark red, green or blue, while biased to short conditions are plotted as light red, green or blue. BF values correspond to one-sided t-tests with a default Cauchy prior of $\sqrt{2} \div 2$. *d* values correspond to Cohen's d effect size coefficients.

Experiment 2: Post-decision confidence ratings

One possible explanation for the results of the previous experiment is that observers may have conflated their preference for the biased-choice with their confidence reports such that the act of choosing the biased option contaminated simultaneous ratings of confidence. To control for such effects we ran a second experiment with the same experimental design with the exception that observers first answered whether the target lines were shorter or longer than the reference and then only afterwards had to report their confidence on the decision they just made. For all three tasks we followed the same analysis procedure as described in the previous experiment.

Across all conditions observers were able to distinguish between short and long lines (see Supplementary Figure S8 for each condition's sensitivity data) and all manipulations resulted in large effects and extreme evidence for a difference between bias direction conditions ($BF^{10} > 100$ for all conditions; see Figure 3, top row). In the reproduction task we again found extreme evidence for an effect in the Müller-Lyer condition ($BF^{10} = 1.1e+03$), moderate evidence for no effect in the base rate ($BF^{10} = 0.08$), and anecdotal evidence for a null-effect in the payoff ($BF^{10} = 0.49$) condition (see Figure 3, middle row; see Supplementary Figure S9 for the decision and reproduction results plotted as in (Sánchez-Fuenzalida et al., 2022)). When combining the reproduction data from experiment 1 and 2 (for which the stimuli and the task were identical), the evidence is even more compelling with a BF^{10} of 9.8e+05 for Müller-Lyer (extreme evidence for an effect), a BF^{10} of 0.05 for base-rate and a BF^{10} of 0.1 for payoff (both strong evidence for the null) (see Supplementary figure S10 for combined reproduction data). Crucially however, when testing for a difference in the point of peak uncertainty based on the confidence responses, we find compelling evidence for effects in all conditions: extreme evidence in the Müller-Lyer condition ($BF^{10} = 5.0e+04$), very strong evidence in the base rate condition ($BF^{10} = 92$) and substantial evidence in the payoff condition ($BF^{10} = 9$) (see Figure 3, bottom row). This confirms the findings from experiment 1, that unlike reproduction, confidence responses do not uniquely shift as a result of changes in perception but also as a result of non-sensory information, even if such information does not aid decision accuracy.

Thus, on the one hand, the results across both experiments confirm that the reproduction task consistently tracks the effect of the Müller-Lyer manipulation while being unaffected by the payoff and base rate manipulations. On the other hand, confidence reports were biased across all conditions – both when confidence reports were elicited simultaneously with a decision, and also when confidence was rated after the decision.

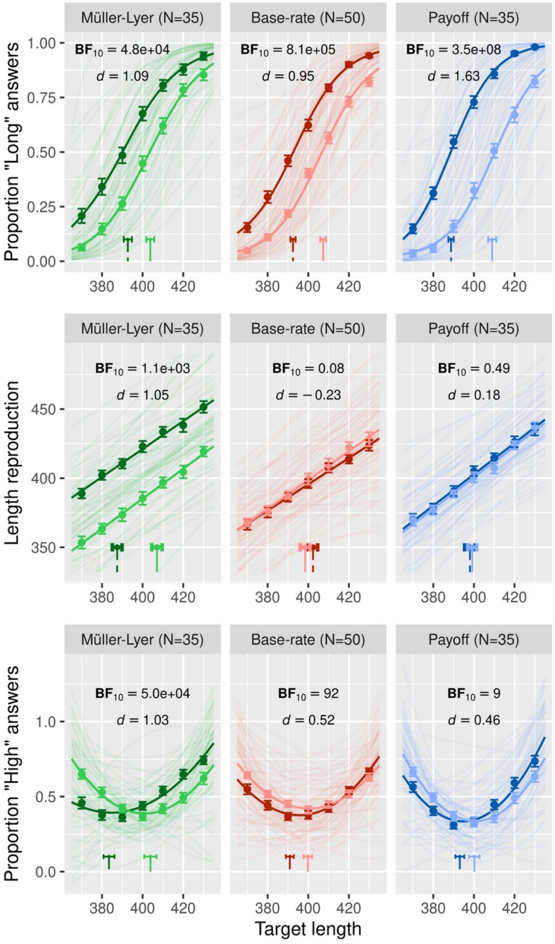


Figure 3: Post-decision confidence results. Decision bias (top row), **length reproduction (middle row)** and **confidence bias (bottom row)** are depicted in the same way as described in figure 2

Discussion

Does confidence track perception? Yes -- but not uniquely. Across two experiments using three well-known bias manipulations (the Müller-Lyer illusion, a base rate manipulation and a payoff manipulation) we showed that confidence reports in a perceptual decision-making context are susceptible to decision bias regardless of the perceptual or non-perceptual nature of the manipulation. First, we replicated our previous results (Sánchez-Fuenzalida et al., 2022) showing that the Müller-Lyer illusion biases both decisions as well as perceptual experience, whereas base rate and payoff manipulations selectively biased decisions without affecting perception. Having confirmed the perceptual or non-perceptual nature of each of our manipulations, we then showed that all bias manipulations resulted in biased confidence reports. This suggests that, although confidence may well be tracking perception as argued by Gallagher et al. (2019), it is also influenced by non-sensory information such as cognitive priors and expected rewards.

Confidence has been conceptualised as emerging during the decision-making process (Dotan et al., 2018; Fetsch et al., 2014; Geurts et al., 2022; Kiani & Shadlen, 2009; Meyniel et al., 2015), with additional processing occurring post-decisionally, after a judgment has been made (Desender et al., 2022; Pleskac & Busemeyer, 2010). Under inferential models of confidence formation, a number of cues -- such as response time, or motor fluency -- may be leveraged to inform a summary judgement of the likelihood of a decision being correct (Kiani et al., 2014). To investigate how decision biases influence these different aspects of confidence formation we asked observers to either report their confidence concurrently with their decision (experiment 1) or report confidence after the decision had been made (experiment 2). Although post-decisional confidence reports are usually considered to be more affected by non-sensory information (Balakrishnan & Ratcliff, 1996; Kahneman & Tversky, 1982; Pleskac & Busemeyer, 2010), we found that both concurrent and post-decision confidence reports were affected by decision bias across all bias manipulations.

Both perceptual (Müller-Lyer illusion) and non-perceptual (base rate and payoff) manipulations, as identified by the reproduction task, resulted in biased confidence reports regardless of whether the confidence report was concurrent or delayed. As the Müller-Lyer illusion affects sensory experience, the effect of this manipulation in confidence reports is in line with Gallagher and colleagues' (2019) findings. However, despite not affecting perception, the base rate and payoff manipulations also resulted in biased confidence reports. Indeed, it has been argued before (Morales & Lau, 2021), and shown experimentally (Constant et al., 2023; Dunning et al., 1990; Lebreton et al., 2018, p. 201; Manis et al., 1980; Otten et al., 2023, p. 202; Rahnev et al., 2015; Sherman et al., 2015), that observers can use prior information to inform their confidence judgments. And while this behaviour can be considered normatively optimal, as prior information can be used to influence decision performance, and therefore confidence, it has been also shown that rewards associated with a decision can confidence even if they have no bearing on decision accuracy (Lebreton et al., 2018, p. 201; Locke et al., 2020). Unlike previous studies, here we not only show that confidence reports are susceptible to stimulus-relevant (base rate) and stimulus-irrelevant (expected rewards) information, but we also firmly establish that these types of manipulations do not affect sensory experience. In contrast, we establish that these non-perceptual manipulations systematically affect confidence judgments.

Overall, our data show that confidence tracks both sensory experience and other non-sensory information in a similar fashion to first-order judgements. Such results are more consistent with computational models that propose confidence tracks choice consistency rather than sensory experience (Boundy-Singer et al., 2022; Mamassian & de Gardelle, 2021). Our findings also echo previous literature that have warned about the dangers of equating confidence level with subjective experience (Fleming & Lau, 2014; Maniscalco & Lau, 2010; Morales et al., 2019). An alternative proposal is that consciousness researchers should focus on metacognitive sensitivity – the mapping between confidence and performance computed over many trials (Michel, 2023).The logic here is that when we are conscious of a stimulus, we are usually also sensitive to the correctness of our responses to that stimulus, and metacognitive sensitivity should be high. However, this methodology is difficult to apply to cases, such as those studied here, which seek to quantify a qualitative change in subjective experience (the perceived length of a line), rather than whether a stimulus is consciously perceived or not. Thus, while we do not argue for discarding the use of confidence reports in studies of perceptual experience, our study urges caution in using confidence ratings as direct assay of changes in perception. More broadly, our results shed light on the combination of factors that affect subjective confidence in perceptual decision-making.

Bibliography

Balakrishnan, J. D., & Ratcliff, R. (1996). Testing models of decision making using confidence ratings in classification. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(3), 615–633. https://doi.org/10.1037/0096-1523.22.3.615

Boundy-Singer, Z. M., Ziemba, C. M., & Goris, R. L. T. (2022). Confidence reflects a noisy decision reliability estimate. *Nature Human Behaviour*, 1–13. https://doi.org/10.1038/s41562-022-01464-x

Constant, M., Pereira, M., Faivre, N., & Filevich, E. (2023). Prior information differentially affects discrimination decisions and subjective confidence reports. *Nature Communications*, *14*(1), Article 1. https://doi.org/10.1038/s41467-023-41112-0

Desender, K., Vermeylen, L., & Verguts, T. (2022). Dynamic influences on static measures of metacognition. *Nature Communications*, *13*(1), Article 1. https://doi.org/10.1038/s41467-022-31727-0

Dotan, D., Meyniel, F., & Dehaene, S. (2018). On-line confidence monitoring during decision making. *Cognition*, *171*, 112–121. https://doi.org/10.1016/j.cognition.2017.11.001

Dunning, D., Griffin, D. W., Milojkovic, J. D., & Ross, L. (1990). The overconfidence effect in social prediction. *Journal of Personality and Social Psychology*, *58*, 568–581. https://doi.org/10.1037/0022-3514.58.4.568

Evans, S., & Azzopardi, P. (2007). Evaluation of a “bias-free” measure of awareness. *Spatial Vision*, *20*(1–2), 61–77. https://doi.org/10.1163/156856807779369742

Fetsch, C. R., Kiani, R., & Shadlen, M. N. (2014). Predicting the Accuracy of a Decision: A Neural Mechanism of Confidence. *Cold Spring Harbor Symposia on Quantitative Biology*, *79*, 185–197. https://doi.org/10.1101/sqb.2014.79.024893

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

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*, 443. https://doi.org/10.3389/fnhum.2014.00443

Gallagher, R. M., Suddendorf, T., & Arnold, D. H. (2019). Confidence as a diagnostic tool for perceptual aftereffects. *Scientific Reports*, *9*(1), 7124. https://doi.org/10.1038/s41598-019-43170-1

Geurts, L. S., Cooke, J. R. H., van Bergen, R. S., & Jehee, J. F. M. (2022). Subjective confidence reflects representation of Bayesian probability in cortex. *Nature Human Behaviour*, *6*(2), 294–305. https://doi.org/10.1038/s41562-021-01247-w

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (pp. xi, 455). John Wiley.

Jachs, B., Blanco, M. J., Grantham-Hill, S., & Soto, D. (2015). On the independence of visual awareness and metacognition: A signal detection theoretic analysis. *Journal of Experimental Psychology: Human Perception and Performance*, *41*, 269–276. https://doi.org/10.1037/xhp0000026

Kahneman, D., & Tversky, A. (1982). Variants of uncertainty. *Cognition*, *11*(2), 143–157. https://doi.org/10.1016/0010-0277(82)90023-3

Kiani, R., Corthell, L., & Shadlen, M. N. (2014). Choice Certainty Is Informed by Both Evidence and Decision Time. *Neuron*, *84*(6), 1329–1342. https://doi.org/10.1016/j.neuron.2014.12.015

Kiani, R., & Shadlen, M. N. (2009). Representation of Confidence Associated with a Decision by Neurons in the Parietal Cortex. *Science*, *324*(5928), 759–764. https://doi.org/10.1126/science.1169405

Kunimoto, C., Miller, J., & Pashler, H. (2001). Confidence and Accuracy of Near-Threshold Discrimination Responses. *Consciousness and Cognition*, *10*(3), 294–340. https://doi.org/10.1006/ccog.2000.0494

Lebreton, M., Langdon, S., Slieker, M. J., Nooitgedacht, J. S., Goudriaan, A. E., Denys, D., van Holst, R. J., & Luigjes, J. (2018). Two sides of the same coin: Monetary incentives concurrently improve and bias confidence judgments. *Science Advances*, *4*(5), eaaq0668. https://doi.org/10.1126/sciadv.aaq0668

Locke, S. M., Gaffin-Cahn, E., Hosseinizaveh, N., Mamassian, P., & Landy, M. S. (2020). Priors and payoffs in confidence judgments. *Attention, Perception, & Psychophysics*, *82*(6), 3158–3175. https://doi.org/10.3758/s13414-020-02018-x

Mamassian, P., & de Gardelle, V. (2021). Modeling perceptual confidence and the confidence forced-choice paradigm. *Psychological Review*. https://doi.org/10.1037/rev0000312

Manis, M., Dovalina, I., Avis, N. E., & Cardoze, S. (1980). Base rates can affect individual predictions. *Journal of Personality and Social Psychology*, *38*, 231–248. https://doi.org/10.1037/0022-3514.38.2.231

Maniscalco, B., & Lau, H. (2010). Comparing signal detection models of perceptual decision confidence. *Journal of Vision*, *10*(7), 213–213. https://doi.org/10.1167/10.7.213

Martinez-Saito, M. (2022). Probing doors to visual awareness: Choice set, visibility, and confidence. *Visual Cognition*, *30*(6), 393–424. https://doi.org/10.1080/13506285.2022.2086333

Meyniel, F., Sigman, M., & Mainen, Z. F. (2015). Confidence as Bayesian Probability: From Neural Origins to Behavior. *Neuron*, *88*(1), 78–92. https://doi.org/10.1016/j.neuron.2015.09.039

Michel, M. (2023). Confidence in consciousness research. *Wiley Interdisciplinary Reviews. Cognitive Science*, *14*(2), e1628. https://doi.org/10.1002/wcs.1628

Morales, J., & Lau, H. (2021). *Confidence tracks consciousness*.

Morales, J., Odegaard, B., & Maniscalco, B. (2019). *The Neural Substrates of Conscious Perception without Performance Confounds*. PsyArXiv. https://doi.org/10.31234/osf.io/8zhy3

Morgan, M., Dillenburger, B., Raphael, S., & Solomon, J. A. (2012). Observers can voluntarily shift their psychometric functions without losing sensitivity. *Attention, Perception, & Psychophysics*, *74*(1), 185–193. https://doi.org/10.3758/s13414-011-0222-7

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, *95*, 109–133. https://doi.org/10.1037/0033-2909.95.1.109

Otten, M., Seth, A. K., & Pinto, Y. (2023). Seeing Ɔ, remembering C: Illusions in short-term memory. *PLOS ONE*, *18*(4), e0283257. https://doi.org/10.1371/journal.pone.0283257

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review*, *117*, 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

Rossum, G. van, Drake, F. L., & Van Rossum, G. (2010). *The Python language reference* (Release 3.0.1 [Repr.]). Python Software Foundation.

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301–308. https://doi.org/10.3758/s13423-014-0595-4

Sánchez-Fuenzalida, N., Gaal, S. van, Fleming, S., Haaf, J. M., & Fahrenfort, J. J. (2022). *Predictions and rewards affect decision making but not subjective experience.* PsyArXiv. https://doi.org/10.31234/osf.io/5v8dh

Sandberg, K., Timmermans, B., Overgaard, M., & Cleeremans, A. (2010). Measuring consciousness: Is one measure better than the other? *Consciousness and Cognition*, *19*(4), 1069–1078. https://doi.org/10.1016/j.concog.2009.12.013

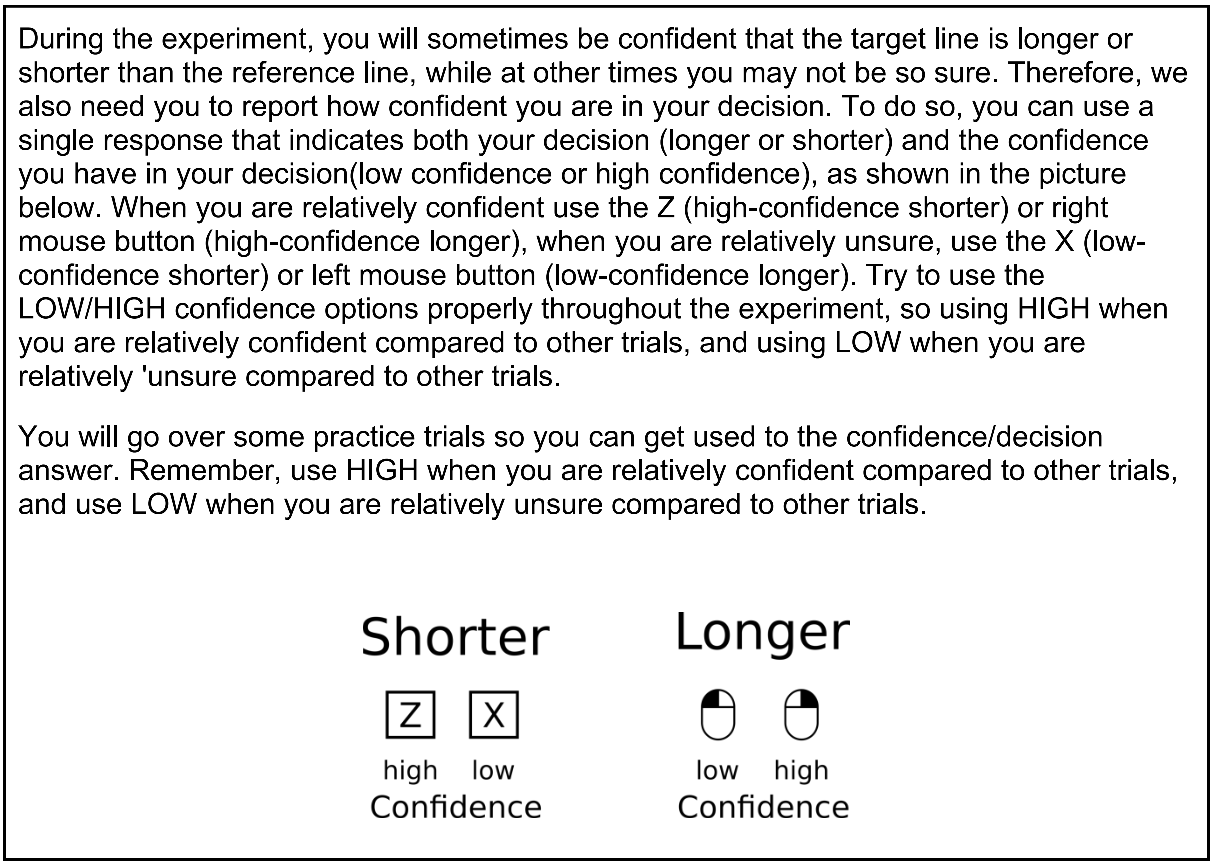
Sherman, M. T., Seth, A. K., Barrett, A. B., & Kanai, R. (2015). Prior expectations facilitate metacognition for perceptual decision. *Consciousness and Cognition*, *35*, 53–65. https://doi.org/10.1016/j.concog.2015.04.015

Witt, J. K., Taylor, J. E. T., Sugovic, M., & Wixted, J. T. (2015). Signal Detection Measures Cannot Distinguish Perceptual Biases from Response Biases. *Perception*, *44*(3), 289–300. https://doi.org/10.1068/p7908

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

Supplementary text T1: General procedure

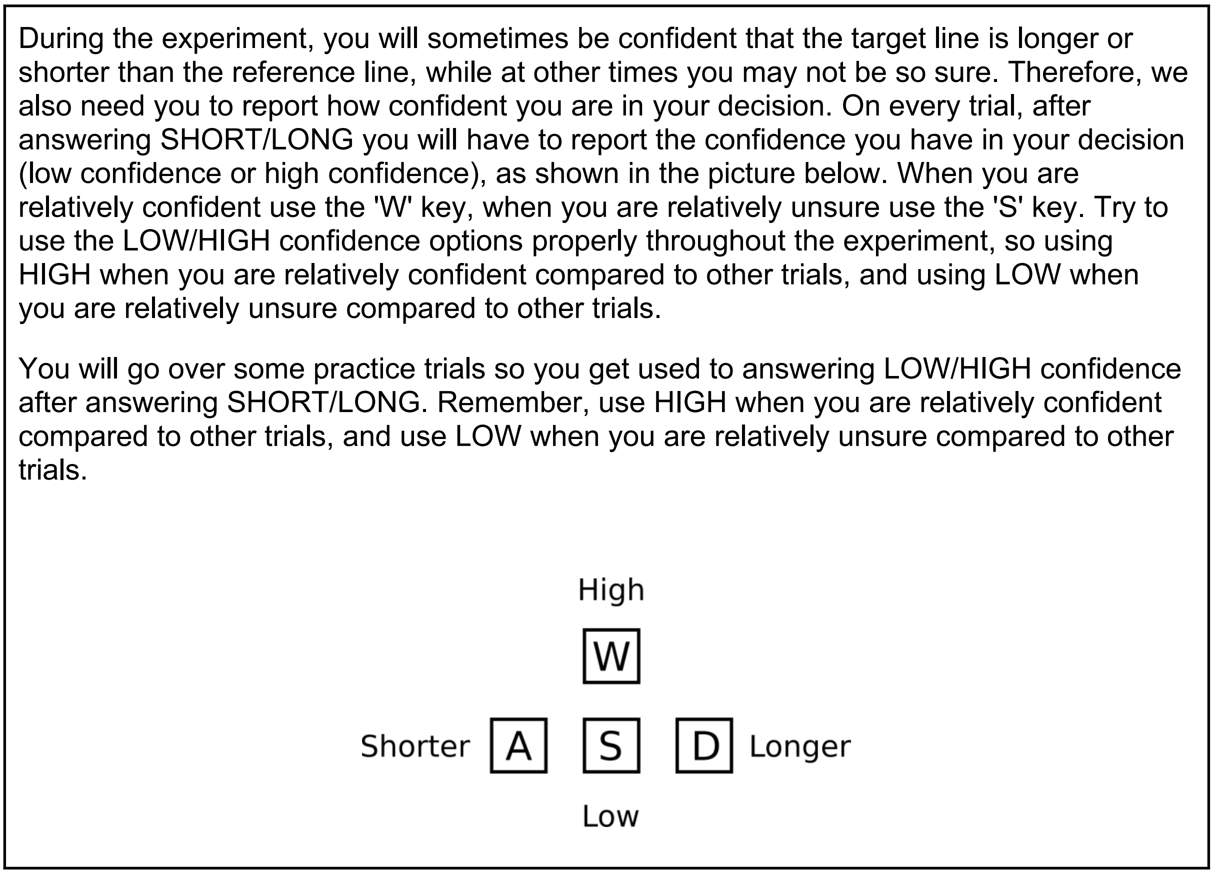
For the line length categorization task, participants first completed 10 trials with feedback with no performance demands, then 10 correct practice trials in a row with feedback, 10 correct practice trials without feedback and finally a longer, more difficult block of 25 trials without feedback with at least 80% correct responses. After the categorization instructions, participants received instructions about how to provide confidence reports about their decision (low and high) (see supplementary figure S1 and S2 for the full confidence report instructions). Participants completed 20 confidence practice trials where half of the trials were very difficult compared with the other half of the trials. Participants were required to evenly use high and low confidence reports, so that difficult trials were more often labelled as low confidence trials, and easy trials were more often labelled as high confidence trials. Participants then received instructions for the reproduction task in the same way as described for the length categorization (decision) task. In the reproduction task a deviation greater than 40 pixels from the length of the target line was considered an error (regardless of whether it was above or below the target line length). Afterwards, participants completed 25 practice trials where both tasks (categorization and reproduction) were intermixed, just as in the actual experiment (see Supplementary figure S11 for a graphical depiction of the procedure). After the tasks' instructions and practice, participants in the payoff and base rate conditions were instructed about the asymmetrical punishment and stim-prevalence scheme just before the experimental trials started. In both the payoff and base rate conditions, participants completed an extra practice block that consisted of 25 trials with trial-by-trial feedback where either the payoff or base rate manipulation was in place to confirm they understood the instructions. In the Müller-Lyer condition participants were explicitly instructed to ignore the flanking arrowheads and to solely judge the length of the horizontal target lines. In the payoff and base rate condition a similar instruction was given about the flanking vertical lines.



During the experiment, you will sometimes be confident that the target line is longer or shorter than the reference line, while at other times you may not be so sure. Therefore, we also need you to report how confident you are in your decision. To do so, you can use a single response that indicates both your decision (longer or shorter) and the confidence you have in your decision(low confidence or high confidence), as shown in the picture below. When you are relatively confident use the Z (high-confidence shorter) or right mouse button (high-confidence longer), when you are relatively unsure, use the X (low-confidence shorter) or left mouse button (low-confidence longer). Try to use the LOW/HIGH confidence options properly throughout the experiment, so using HIGH when you are relatively confident compared to other trials, and using LOW when you are relatively 'unsure compared to other trials.

You will go over some practice trials so you can get used to the confidence/decision answer. Remember, use HIGH when you are relatively confident compared to other trials, and use LOW when you are relatively unsure compared to other trials.

Fig. S1: Concurrent decision confidence instructions. During the discrimination instructions participants were prompted to concurrently report a confidence rating. After these instructions participants went over a practice block where they were required to evenly use “high” and “low” confidence responses.



During the experiment, you will sometimes be confident that the target line is longer or shorter than the reference line, while at other times you may not be so sure. Therefore, we also need you to report how confident you are in your decision. On every trial, after answering SHORT/LONG you will have to report the confidence you have in your decision (low confidence or high confidence), as shown in the picture below. When you are relatively confident use the 'W' key, when you are relatively unsure use the 'S' key. Try to use the LOW/HIGH confidence options properly throughout the experiment, so using HIGH when you are relatively confident compared to other trials, and using LOW when you are relatively unsure compared to other trials.

You will go over some practice trials so you get used to answering LOW/HIGH confidence after answering SHORT/LONG. Remember, use HIGH when you are relatively confident compared to other trials, and use LOW when you are relatively unsure compared to other trials.

Fig. S2: Post-decision confidence instructions. During the discrimination instructions participants were prompted to concurrently report a confidence rating. After these instructions participants went over a practice block where they were required to evenly use "high" and "low" confidence responses.

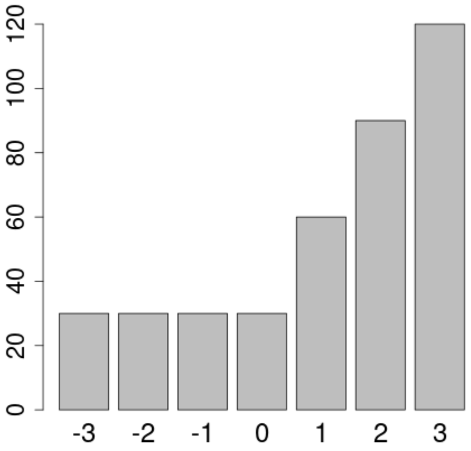
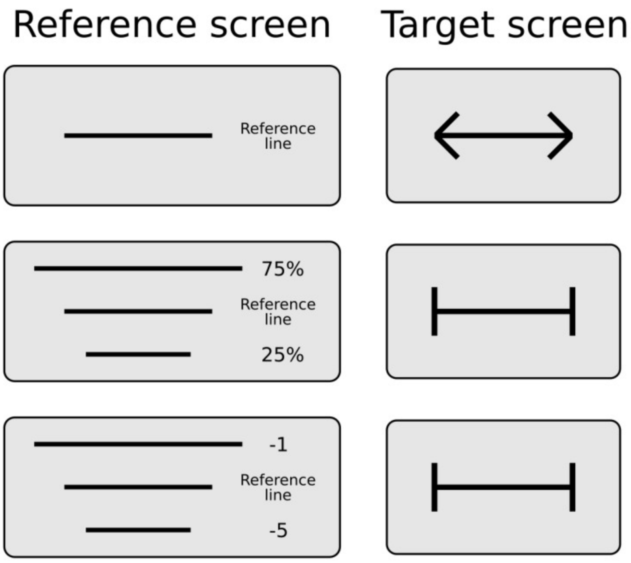
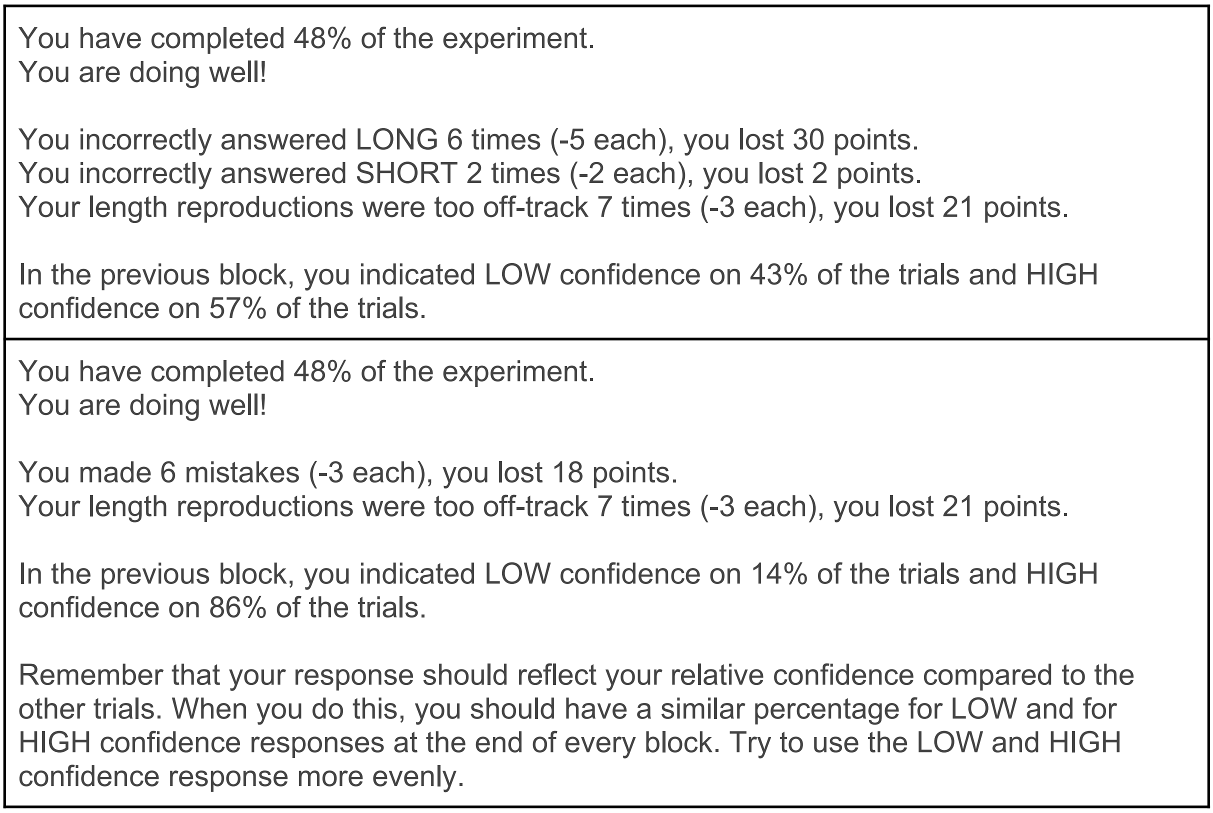


Fig. S3: Base rate frequency of each target length. Frequency of each target length value in the base rate condition biased to long (histogram was mirrored when the bias direction was ‘short’). The y-axis indicates the absolute frequency of each target length in the x-axis. In the x-axis zero represents the length of the reference line and each step away from the centre indicates an decrease/increase of 10 pixels.



Supplementary figure S4: Reference and target screen examples. On the left column there is an example of the reference screen of the Müller-Lyer (top row), base rate biased to long (middle row) and payoff biased to long (bottom row) conditions. In the Müller-Lyer condition the reference screen looked the same for biased to long and short conditions. In the base rate and payoff conditions the numbers indicating the relative frequency of short and long lines, or the cost for incorrectly answering 'short' or 'long' was inverted depending on the bias direction. On the right column there is an example of the target screen for the same conditions in the same order as for the left column. In the Müller-Lyer condition the arrowheads pointed outwards when the bias direction was short or inward when the bias direction was long. In the base rate and payoff condition the target lines were always flanked by vertical lines, regardless of the bias direction.



You have completed 48% of the experiment.
You are doing well!

You incorrectly answered LONG 6 times (-5 each), you lost 30 points.
You incorrectly answered SHORT 2 times (-2 each), you lost 2 points.
Your length reproductions were too off-track 7 times (-3 each), you lost 21 points.

In the previous block, you indicated LOW confidence on 43% of the trials and HIGH confidence on 57% of the trials.

You have completed 48% of the experiment.
You are doing well!

You made 6 mistakes (-3 each), you lost 18 points.
Your length reproductions were too off-track 7 times (-3 each), you lost 21 points.

In the previous block, you indicated LOW confidence on 14% of the trials and HIGH confidence on 86% of the trials.

Remember that your response should reflect your relative confidence compared to the other trials. When you do this, you should have a similar percentage for LOW and for HIGH confidence responses at the end of every block. Try to use the LOW and HIGH confidence response more evenly.

Fig. S5: Block-level feedback example. Examples of the feedback participants received at the end of each experimental block in the payoff condition (top panel) and in the Müller-Lyer condition (bottom panel). In the base rate condition participants received the same message as in the payoff condition with the exception that the cost for incorrect “long” and “short” responses was the same (-3 points). In all conditions participants also received feedback on the relative use of “high” and “low” confidence responses. When participants reported “high” or “low” confidence in less than 25% of the trials of a block they were prompted to try to use both “high” or “low” as evenly as possible (see bottom panel).

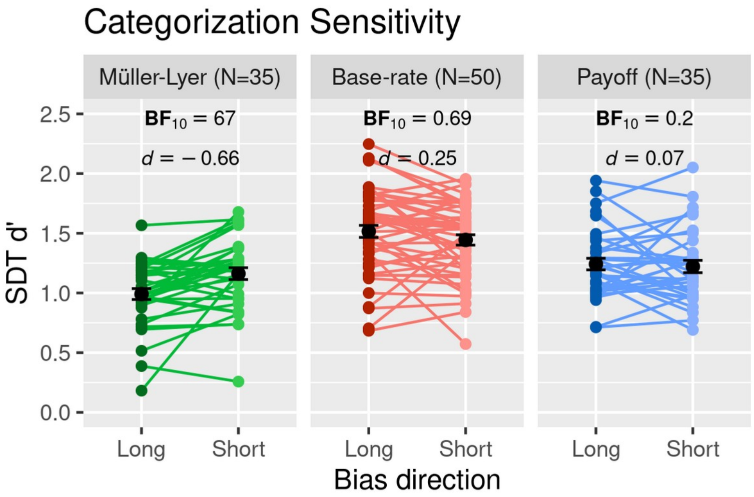


Fig. S6: Experiment 1: Categorization sensitivity. SDT d' values for each subject are plotted along with the group average for each bias source and bias direction condition. All error bars are standard errors of the mean. Higher values indicate a better performance at the task. BF values correspond to a two-sided Bayesian t-test with a default Cauchy prior of $\sqrt{2} \div 2$. *d* values correspond to Cohen's d effect size coefficients.

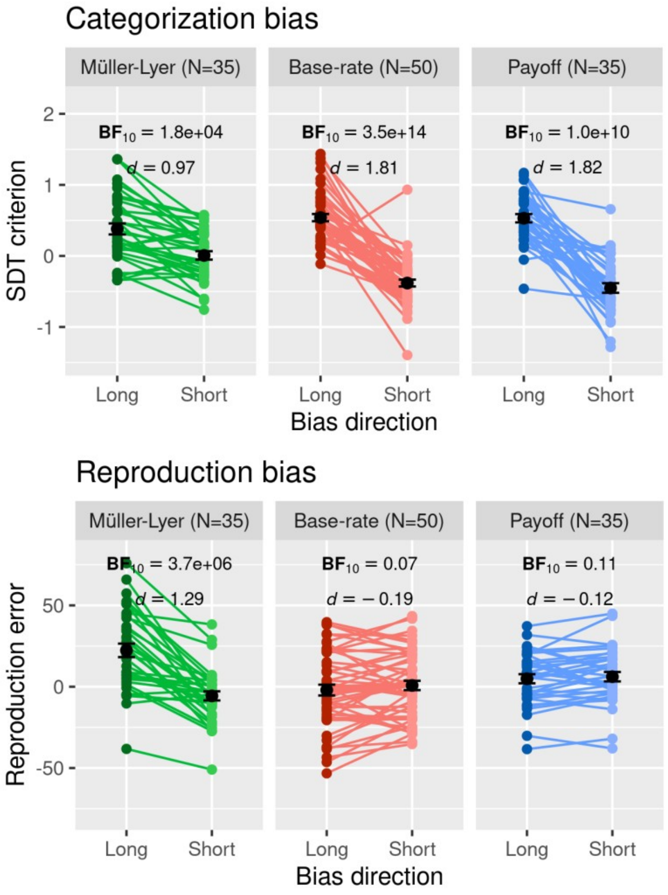


Fig. S7: Experiment 1. Decision bias and reproduction results. A) Categorization bias. The SDT criterion value for each subject along with the group average for each bias source and bias direction condition. Higher values indicate a stronger bias towards answering 'long' while lower values indicate a stronger bias towards answering 'short'. **B) Reproduction error.** The average reproduction error (*length reproduction - target length*) for each subject is displayed for each bias source and bias direction condition. Higher values indicate lines reproduction longer than the target line while lower values indicate lines reproduction shorter than the target line. All error bars are standard errors of the mean. BF values correspond to a one-sided Bayesian t-test with a default Cauchy prior of $\sqrt{2} \div 2$. *d* values correspond to Cohen's d effect size coefficients.

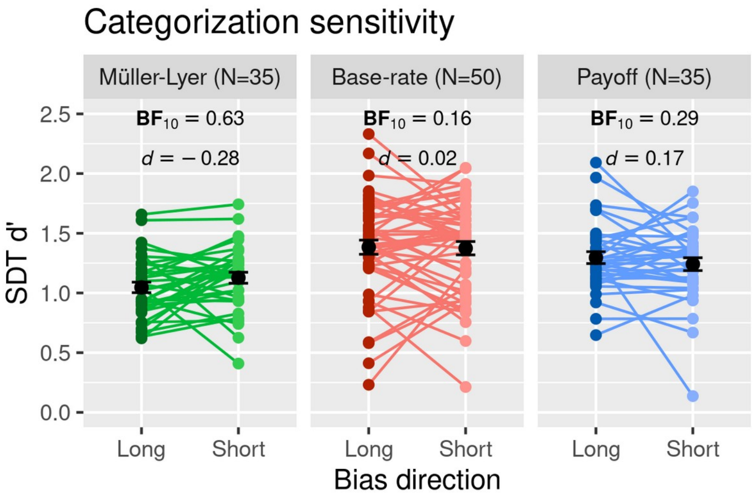


Fig. S8: Experiment 2: Categorization sensitivity. SDT d' values for each subject are plotted along with the group average for each bias source and bias direction condition. All error bars are standard errors of the mean. Higher values indicate a better performance at the task. BF values correspond to a two-sided Bayesian t-test with a default Cauchy prior of $\sqrt{2} \div 2$. d values correspond to Cohen's d effect size coefficients.

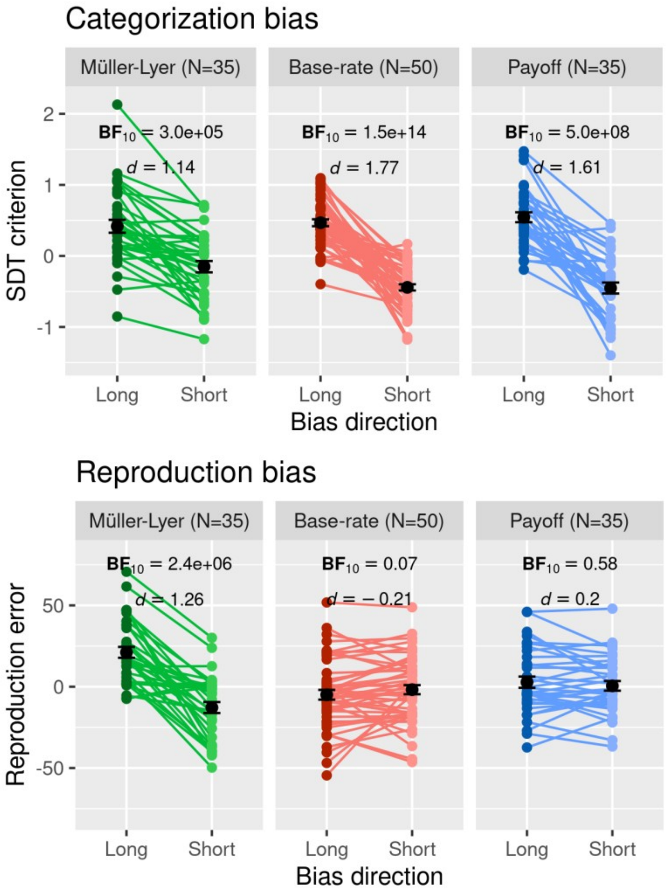
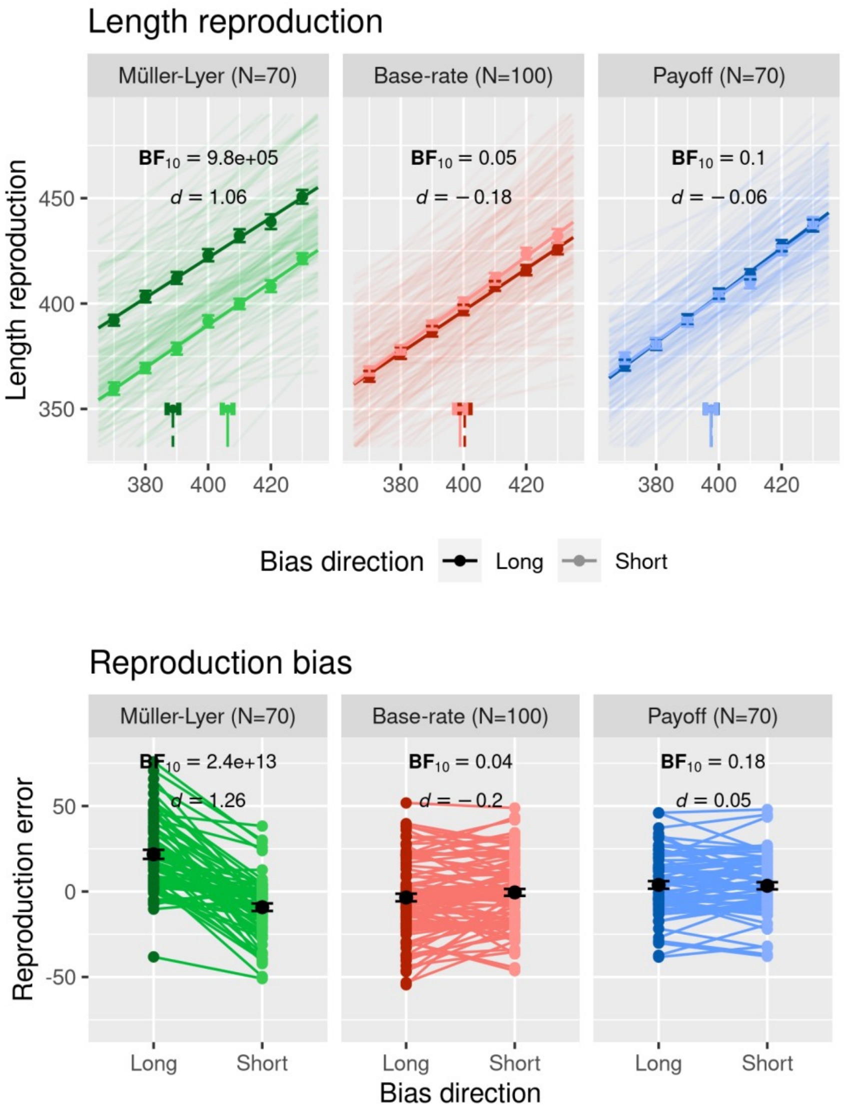
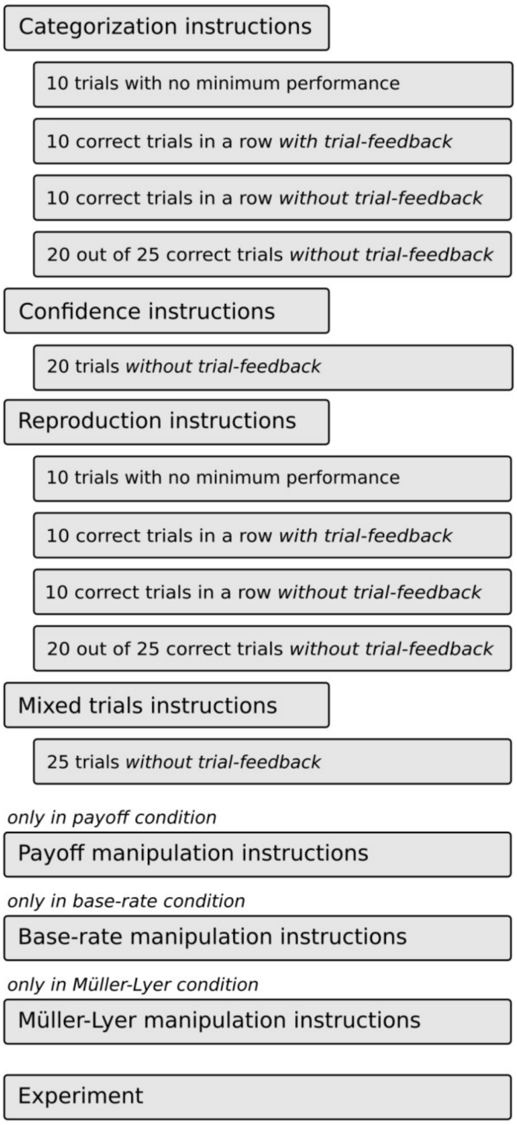


Fig. S9: Experiment 2. Decision bias and reproduction results. A) Categorization bias. The SDT criterion value for each subject along with the group average for each bias source and bias direction condition. Higher values indicate a stronger bias towards answering ‘long’ while lower values indicate a stronger bias towards answering ‘short’. **B) Reproduction bias.** The average reproduction error (*length reproduction - target length*) for each subject is displayed for each bias source and bias direction condition. Higher values indicate lines reproduction longer than the target line while lower values indicate lines reproduction shorter than the target line. All error bars are standard errors of the mean. BF values correspond to a one-sided Bayesian t-test with a default Cauchy prior of $\sqrt{2} \div 2$. d values correspond to Cohen’s d effect size coefficients.



Supplementary figure S10: Experiment 1 and 2 combined reproduction data. Top panel: A straight line was fitted to each observer's distribution of length reproductions of experiment 1 and 2. At the bottom of each panel is plotted the target length associated with reproductions equal to the reference line (400 pixels). Points plotted over the fitted line correspond to the average reproduced length for each target line presented. **Bottom panel:** The average reproduction error (*length reproduction - target length*) for each subject is displayed for each bias source and bias direction condition. Higher values indicate lines reproduction longer than the target line while lower values indicate lines reproduction shorter than the target line. All error bars are standard errors of the mean. BF values correspond to a one-sided Bayesian t-test with a default Cauchy prior of $\sqrt{2} \div 2$. *d* values correspond to Cohen's d effect size coefficients.


Categorization instructions
10 trials with no minimum performance
10 correct trials in a row *with trial-feedback*
10 correct trials in a row *without trial-feedback*
20 out of 25 correct trials *without trial-feedback*
Confidence instructions
20 trials *without trial-feedback*
Reproduction instructions
10 trials with no minimum performance
10 correct trials in a row *with trial-feedback*
10 correct trials in a row *without trial-feedback*
20 out of 25 correct trials *without trial-feedback*
Mixed trials instructions
25 trials *without trial-feedback*
only in payoff condition
Payoff manipulation instructions
only in base-rate condition
Base-rate manipulation instructions
only in Müller-Lyer condition
Müller-Lyer manipulation instructions
Experiment

Supplementary figure S11: Experiment general procedure. For each task participants received extensive instructions and completed multiple practice trials. All participants completed the categorization, confidence and reproduction instructions and mixed trials instructions where all tasks were interleaved. Depending on the bias manipulation condition participants received specific instructions related to the manipulation. From top to bottom the figure depicts the order and stages of each part of the instructions. When there was a minimum performance requirement in any of the practice trials sections, participants repeated the practice block until they achieved the expected performance. See General procedure in Methods and Materials for a more detailed description of the procedure.