On Confidence in Individual and Group Decision-Making

Dan Bang

A thesis submitted for the degree of
Doctor of Philosophy of the University of Oxford

Department of Experimental Psychology
& Magdalen College

Trinity Term 2015

Table of contents

Abstract

On Confidence in Individual and Group Decision-Making

Dan Bang
Department of Experimental Psychology
& Magdalen College
DPhil, Experimental Psychology
Trinity Term 2015

This thesis is about the human ability to share and combine representations of the uncertainty associated with individual beliefs – an ability which is called *metacognition* and facilitates effective cooperation. We distinguish between two metacognitive representations: an implicit *confidence variable* for oneself and an explicit *confidence report* for sharing with others. Using visual psychophysics and computational modelling, we address the issues of *optimality* and *flexibility* in the formation and the utilisation of these representations. We show that people can compute the confidence variable in an optimal manner (the probability that a given belief is correct as per Bayesian inference). Further, we show that the mapping of this variable onto a confidence report can vary flexibly – with people adjusting their reports according to the history of reports given and feedback obtained. This optimality and flexibility is important for effective cooperation. Being a probability, the optimal confidence variable can be compared across people. However, to facilitate this comparison, people must *adapt* their confidence reports to each other and develop a common metric for reporting the probability that their belief is correct. We show that people solve this *communication problem* sub-optimally; they *match* each other's mean confidence and confidence distributions, regardless of whether they are equally likely to be correct or not. In addition, we show that, while people can take into account differences in underlying competence to some extent, they fail to do so adequately; they exhibit an *equality bias*, weighting their partner's beliefs as if they were as good or as bad as their own, regardless of true differences in their underlying competence. More generally, our results pose a problem for our current understanding of metacognition which assumes that confidence reports are stable over time. In addition, our results show that confidence reports are socially malleable, and thus raise the possibility that well-known biases, such as *overconfidence*, might reflect particular norms for social interaction.

This work was funded by a studentship from the Calleva Research Centre (Magdalen College), and was conducted under the supervision of Jennifer Lau and Christopher Summerfield.

Acknowledgements

Firstly, I would like to thank my supervisors. Jennifer Lau – for seeing my potential and taking me on as her student, for showing trust in my abilities, for selfless support and giving me the freedom to develop as an independent thinker. Christopher Summerfield – for taking me into his lab, for introducing me to decision-making in all its nuances, for always pushing me to refine my arguments, for providing support in times of change.

I would like to thank the Department of Experimental Psychology for providing an outstanding environment in which to work and learn. I would like to thank Magdalen College for support – especially in difficult times – and Stephen Butt for the generous donation which led to the Calleva Research Centre and which funded most of the work for this thesis.

This thesis would never have been possible without the help of many eminent scientists who have been kind with their time, advice and coding. Karsten Olsen stimulated many thoughts about social learning. Andreas Roepstorff always reminded me of the larger picture. Laurence Aitchison and Peter Latham taught me a great deal of tricks. Santiago Herce-Castañón helped me learn how to argue, through our literally endless discussions, from coffee stamps, through "broadly" and "more generally", to priors and likelihoods. Alejo Nevado-Holgado taught me how to use the supercomputer. Konstantinos Tsetsos provided me with a voice of reason. At a workshop in Bristol, Rani Moran noted that one of my analyses could be done analytically instead of through simulation; this turned into a great friendship, and a great collaboration. One person recently noted that "Rani would be successful even on Mars"; I'd tend to agree.

I would like to thank Bahador Bahrami and Chris Frith. I would never have been at this stage without them. Bahador showed so much trust in me as a young student that I always aspired to be a little bit better. We've had so much fun studying the peculiarities of social behaviour; first in Denmark, and then on a trip to China (luckily I was there to look after him), and now in the UK. Chris has been so kind with his time; always taking time to discuss, never failing to reply to an email, even if I was rambling. Chris is by far the largest intellectual influence in my life; I hope he knows how many people owe their interests and thoughts to him.

I would like to thank my friends – especially Salam, Tom, Nick, Thomas, Kevin, Angus, Mark, Per Martin, James, Hamed, Janto, Will, Karsten, Peer, Olm, Rasmus – for all the fun times and perspectives on my work. I'm very fortunate to have so many smart friends.

I would like to thank my family – especially Mor, Far, David, Camilla, Mormor, Morfar. You have always encouraged and supported me; when I moved to China, to Hong Kong, to the Netherlands, and then to the UK; when I switched from Chinese, to linguistics, to anthropology and then to psychology. Living apart for so many years, more than a decade, has been hard, but you've always been there for me.

And I would like to thank Jennifer, who hopefully knows how much this work depends upon her efforts, patience and support. I couldn't ask for more.

Contributions

My thesis builds on a series of articles that I have (co-)authored as part of my research. I have taken care to include work for which I was responsible. However, in a few instances, the work of collaborators has been necessary to include directly for clarity and cohesion:

+ In **Chapter 2**, the Bayesian method for performing model comparison (**Section 2.2.4: Model comparison**) was mainly developed by Laurence Aitchison (University College London) and Peter Latham (University College London).

+ In **Chapter 3**, the dataset for Experiment 3 was provided by Rani Moran (Tel Aviv University).

+ In **Chapter 4**, I collected the data with the assistance of Banafshe Rafiee (University of Tehran; Experiment 1) and Santiago Herce-Castañón (University of Oxford; Experiment 4).

+ In **Chapter 4**, the analytical derivation of individual and joint accuracy (**Section 4.3.1.2: Deriving accuracy**) was mainly developed by Rani Moran (Tel Aviv University).

+ In **Chapter 4**, the numerical derivation of maximum entropy distributions (**4.3.1.5: Maximum entropy distributions**) was mainly developed by Rani Moran (Tel Aviv University).

+ In **Chapter 6**, I collected the data with the assistance of Ali Mahmoodi (University of Tehran; Experiments 1 to 3), Karsten Olsen (University of Aarhus; Experiment 1), Yuanyuan Aimee Zhao (Peking University; Experiment 1) and Christos Sideras (UCL; Experiment 4).

+ In **Chapter 6**, I implemented the Bayesian learning model together with Ali Mahmoodi (University of Tehran) – the source code was provided by Tim Behrens (University of Oxford).

List of figures

The items are labelled so that the first number indicates the chapter and the second number indicates the position within the chapter (e.g., "Figure 5–1" is the first figure in chapter five).

List of tables

The items are labelled so that the first number indicates the chapter and the second number indicates the position within the chapter (e.g., “Table 5–1” is the first table in chapter five).

List of equations

The items are labelled so that the first number indicates the chapter and the second number indicates the position within the chapter (e.g., “Eq. 5–1” is the first equation in chapter five).

List of abbreviations

2HBT1	Two-Heads-Better-Than-One
BDT	Bayesian Decision Theory
MCS	Maximum Confidence Slating
MRTS	Minimum Reaction Time Slating
SDT	Signal Detection Theory

Chapter 1: Sharing and combining representations of uncertainty

This chapter will set the stage for our empirical work. We outline a component account of the cognitive processes involved in sharing and combining representations of uncertainty. Broadly, two cognitive processes play a key role: decision-making and metacognition. We use Bayesian decision theory to formalise how an optimal agent would take into account the sources of uncertainty in a given decision task. We then relate this treatment of decision-making to metacognition and use Bayesian decision theory to formalise how an optimal agent would compute meta-level representations of uncertainty: an implicit "confidence variable" for oneself and an explicit "confidence report" for sharing with others. We discuss the sparse evidence that humans compute such meta-level representation in an optimal manner and identify the outstanding questions that we will address in this thesis. Broadly, we will focus on the issues of metacognitive optimality and flexibility – both factors are important for effective cooperation. Lastly, we briefly outline our experimental and computational framework, involving visual psychophysics and signal detection theory.

1.1 Introduction

Humans must often make decisions on the basis of imperfect information. Consider a football referee who has to decide whether the ball briefly crossed the goal line. The referee only had a brief glimpse of the situation; so they must make their decision on the basis of an imprecise estimate of the ball position. In this case, the referee might rely on prior experience; maybe goal keepers tend to act suspiciously when it was a goal. If it still seems uncertain that the ball crossed the goal line, then awarding the goal might not be worth the risk. If the referee turns out to be wrong, they could be dropped for future matches. Research has shown that we are generally good at making decisions under these different sources of uncertainty – from noise inside the nervous system to noise in the outside world (Ernst & Bülthoff, 2004; Ma & Jazayeri, 2014; Trommershäuser, Maloney, & Landy, 2008; Wolpert & Flanagan, 2010).[1]

[1] We use *noise* to refer to objective random or unpredictable variability (e.g., in sensory responses to a constant stimulus) and *uncertainty* to refer to an agent's subjective representation of such variability.

Could we make better decisions under uncertainty? Yes – if there is more than one of us. In this case, we could often make better decisions by sharing and combining our representations of task-relevant variables and their associated uncertainty. This might, for example, be achieved by communicating explicit point estimates with different levels of confidence (e.g., "I'm pretty sure that the ball landed one meter behind the goal line"). In general, the combination of multiple noisy estimates improves the signal-to-noise ratio and thus decision accuracy; under Gaussian noise, for example, the signal scales with the number of estimates, whereas the noise only scales with the square root of the number of estimates. In addition, the signal-to-noise ratio can be further improved if the individual estimates are weighted according to their associated uncertainty. Indeed, this promise of pooling multiple sources of information has sparked numerous attempts to harness the "wisdom of crowds" (Surowiecki, 2004), including expert panels, committees and prediction markets, and the development of methods for eliciting and combining subjective judgements in situations where the user of a method has no means of evaluating the respondents' honesty or competence (Prelec, 2004).

Unfortunately, despite seemingly using accurate representations of uncertainty to guide our own behaviour, the sharing of this information might not be as straightforward. For example, the information might not be available for higher-order cognition.[2] As an analogy, while most of us are good at catching balls, we seem to have no insight into how this is done (Reed, McLeod, & Dienes, 2010). Instead, we may come up with heuristic approximations. However, even if the information is available for higher-order cognition, it might be partially or even fully lost when transformed into a format which can be shared with others. Again, as an analogy, while we can perceive subtle differences in colour hue, our colour terms are coarse and often group together visibly different hues (Berlin & Kay, 1969).

[2] We will use *cognition* in a very general sense to encompass all information processing within the brain. We take *lower-order* and *higher-order* to imply unconscious and conscious, respectively.

Further, even if accurate representations of uncertainty are shared, the combination of this information – a process often studied in the context of group decision-making – may easily be jeopardised by social biases. For example, group members may be subject to *groupthink* (Janis, 1982) and suppress valid but unpopular viewpoints (Packer, 2009) or isolate themselves from outside influences (Minson & Mueller, 2012). Further, group members may be subject to the *hidden-profile paradigm* (Stasser & Titus, 1985, 2003) and discard information that is highly relevant but not shared by everyone. Similarly, group members may be subject to *conformity* (Asch, 1951) and adopt each other's attitudes and beliefs even if clearly wrong. It remains debated why these social biases exist; for example, group members might seek to minimise conflict and maximise cohesion. However, social biases may sometimes be beneficial; for example, when uncertain, integrating other's beliefs into one's own judgement can improve accuracy (Soll & Larrick, 2009).

This thesis is concerned with our ability to *share* and *combine* representations of uncertainty. We will focus on tasks in which sensory noise is the limiting factor on decision accuracy. First, sensory tasks have an objectively true answer, allowing us to specify how an optimal agent or a pair of such agents should solve a given decision problem. Second, we can fully control the task-relevant information (cf. participants can rely on complex memories in general-knowledge tasks). Lastly, sensory noise in two brains is uncorrelated, allowing us to study cooperative situations in which a pair of agents can gain a significant advantage by sharing their respective representations of uncertainty. Briefly, we will study how people compute an estimate of the uncertainty associated with a decision for report; whether the report of this estimate is flexible; whether people can learn to report this estimate in a manner that facilities effective cooperation; and lastly, how people evaluate the reports made by others.[3] Our goal is not to provide a simple recipe for eliciting and combing representations of uncertainty.

[3] We use *to estimate* / *an estimate* and *to represent* / *a representation* interchangeably.

Instead, our goal is to provide a framework for studying and modelling how this process unfolds in social interaction.

1.2 General foundation for thesis

In this section, we will lay a general foundation for the thesis. In particular, we will identify and specify which cognitive processes are involved in sharing and combining representations of uncertainty. The purpose of this section is to develop a more precise language for the thesis and identify the questions that we will test empirically in this thesis. We will use the general foundation laid in this section to guide the choice of the experimental and computational framework for our empirical work.

1.2.1 Component process account

Which cognitive operations must a pair of agents perform in order to share and combine their representations of uncertainty? We can think of the whole process as being made up of a series of component processes (**Figure 1-1**). Consider *two* football referees who have to decide whether the ball briefly crossed the goal line. First, the task determines the input to the whole processes; for example, the visual image of the ball position at the critical moment in time. Second, each referee must decide whether they think the ball crossed the goal line. Third, each referee must evaluate the uncertainty associated with their individual decision. Fourth, they must report their individual decision and its associated evaluation; for example, stating that they are confident that the ball crossed the goal line. Lastly, they must combine their respective reports according to some weighting rule; for example, giving more weight to the report of the more experienced referee. As such, the whole process is underpinned by two levels of decision-making – at the individual level and at the group level.

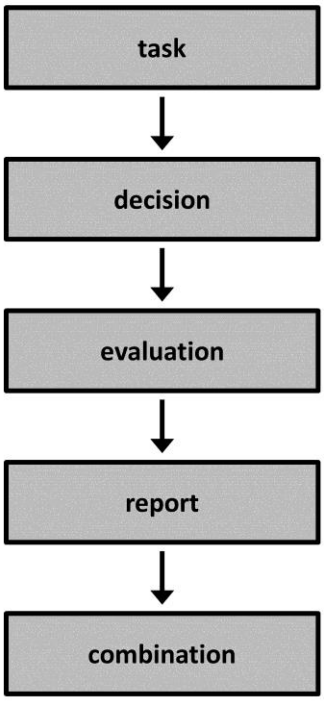


Figure 1-1. Component process account of sharing and combining representations of uncertainty. See text for details.

How do agents perform these cognitive operations? This is the overarching question of this thesis. To address this question, however, we need a better understanding of each of the component processes, and, in particular, how they *should* be performed. In developing this understanding, two cognitive processes will be central: decision-making and metacognition. In broad terms, decision-making is the process of selecting some cognitive state among a set of alternatives, whereas metacognition is a supervisory mechanism that allows us to evaluate the uncertainty associated with the selected cognitive state. While normative theories of decision-making have been developed, theories of metacognition are largely conceptual – we will therefore seek to sketch out a normative theory of metacognition for our empirical work.

The rest of this section will proceed in three steps. First, we will use Bayesian Decision Theory (BDT) to formalise how a decision maker should take into account different types of noise in decision tasks.[4] We will then present evidence that people make decisions as predicted by BDT. We will consider different types of task (sensory, motor and reward-based) so as to build up a strong intuition about what it means to make optimal decisions. Second, we will introduce a conceptual definition of metacognition that builds on the existing literature. We will then use BDT to formalise (albeit in general terms) the computations and the representations involved in metacognition. The notion of confidence will be central – both in the sense of an *implicit* estimate, which can be used to guide individual behaviour, and in the sense of an *explicit* estimate, which can be shared and used to guide group behaviour. We will discuss whether people compute the two types of estimate in an optimal manner. Lastly, having laid a general foundation, we will specify the questions that we will test empirically.

[4] The formalisation of BDT in this thesis builds on Ma & Jazayeri (2014); for similar treatments of decision-making, see Dayan & Daw (2008), Kording (2007) and Maloney & Zhang (2010).

1.2.2 Bayesian decision theory

1.2.2.1 Statistical description of decision tasks

In the BDT framework, a decision problem can be described by a so-called generative model which characterises the statistical relationship between task-relevant variables (**Figure 1-2**). Consider a *single* football referee who has to decide whether the ball briefly crossed the goal line. We will denote the true state of the world, C; for example, that it was a goal. The referee does not have direct access to the true state of the world, but only to a relevant sensory feature, s, such as the visual image of the ball at the critical point in time. In general, each value of C and s have a certain frequency of occurring together in the world, which can be presented by a joint probability distribution, $P(C, s)$. The sensory feature s evokes a neural response, x; for example, a visual response to the ball position. Because of external and internal noise, the relationship between s and x is generally indeterminate and is best described as a probability distribution, $P(x|s)$, with a neural response having a particular probability of occurring given a sensory feature. In vision, for example, photons arrive at the retina at a rate governed by a random (Poisson) process, placing an external limit on visual sensitivity (Hecht, Shlaer, & Pirenne, 1942). Inside the nervous system, the transmission of information is corrupted by random processes which exist at multiple levels, from the structure of individual neurons to the complex interactions between networks of neurons (Faisal, Selen, & Wolpert, 2008).

In the BDT framework, decision-making is then defined as the process of mapping a neural response, x, onto some decision, d; for example, deciding that it was a goal.[5] Many decisions must be executed as an action, a; for example, signalling that it was a goal. This relationship may be subject to motor noise and is best described by a probability distribution, $P(a|d)$, with an action having a particular probability of being realised given a decision. In the football

[5] We use *decision* in an abstract sense to refer to any case in which the nervous system has to select among a set of alternative states. A decision need not be conscious or executed in the external world.

example, motor noise is largely irrelevant. However, in motor-based tasks, motor noise is the limiting factor on decision accuracy. Lastly, the state of the world and the action often leads to an outcome, R; for example, if the referee did not award the goal, they may be dropped for the next match. When R is not determined by C and a, then their relationship is best described by a probability distribution, $P(R|C, a)$, with an outcome having a particular probability of occurring given a state of the world and an action.

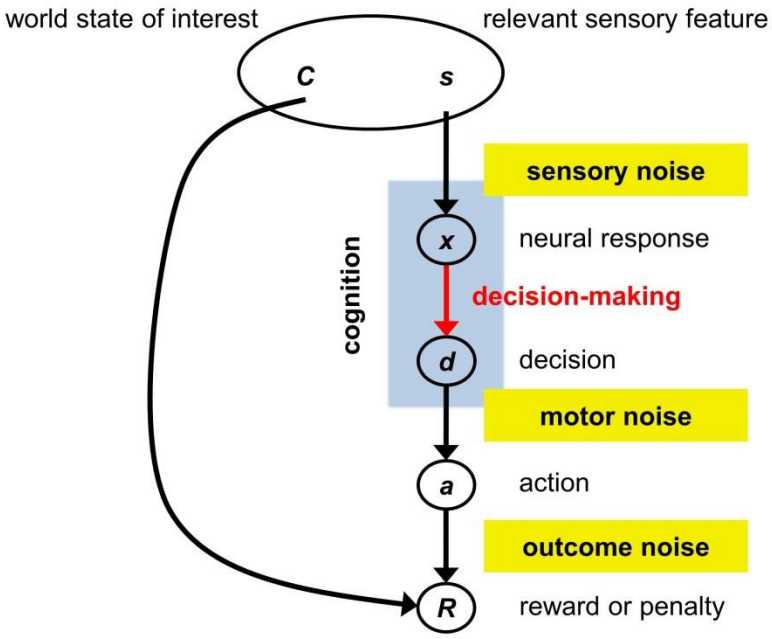


Figure 1-2. Decision-making in terms of Bayesian decision theory. Each node (circle) contains a task-relevant variable which can described by a probability distribution conditioned on the variable(s) from which the arrow(s) points. A Bayes optimal agent takes into account the full probabilistic structure of the decision problem when mapping the neural response onto a decision. See text for details. [This figure is adapted from Ma & Jazayeri, 2014.]

1.2.2.2 Bayes optimal agent

The beliefs of a Bayes optimal agent are based on the probabilistic relationship between the task-relevant variables.[6] Sensory beliefs are held over C and s and can be described by a probability distribution, $P(C, s|x)$, which represents the decision maker's knowledge about external and internal sensory noise. In the following, we will refer to the (posterior) probability distribution, $P(C|x)$, which represents the decision maker's belief about the state of the world after combining sensory information with prior knowledge as per Bayes' theorem. Further, motor beliefs are held over a and can be described by the probability distribution, $P(a|d)$, which represents the decision maker's knowledge about their motor noise. Lastly, reward beliefs are held over R and can be described by the probability distribution, $P(R|C, a)$, which represents the decision maker's knowledge about the processes which generate outcomes.

The computation of beliefs can be complex. In reward-based tasks, for example, different random processes may govern the outcome on a given trial (Payzan-LeNestour & Bossaerts, 2011). There is expected uncertainty (risk), which describes how much uncertainty there is left after complete learning. For example, despite knowing the generative model, an individual cannot predict the outcome of a coin flip. There is also unexpected uncertainty (volatility), which reflects how likely it is that the generative model will have changed on a given trial. For example, a coin might be manipulated such that the probability of heads suddenly is higher than the probability of tails. While this change would reduce the expected uncertainty, it might be hard to identify for an agent if unannounced.

BDT specifies how a Bayes optimal agent uses beliefs to make decisions. The decision maker first computes the probability of an outcome R when the neural response is x and the planned decision is d. This probability is given by

---

[6] We use *belief* (and *knowledge*) in an abstract sense to refer to an agent's representation of the probabilistic relationship between task-relevant variables. A belief need not be conscious.

$$P(R|x,d) = \iint P(R|C,a)\,P(C|x)\,P(a|d)\,\mathrm{d}C\,\mathrm{d}a \qquad \textbf{Eq. 1-1}$$

The three factors in the integrand correspond to reward beliefs, sensory beliefs and motor beliefs, respectively. The Bayes optimal agent uses these beliefs to marginalise (average) over the possible states of the world C and the possible actions a which are not known when a decision d is evaluated. Optimality can now be defined as selecting the decision which maximises reward under the distribution $P(R|x,d)$.

1.2.2.3 Evidence that humans make optimal decisions

The question is whether people use beliefs in the mapping from some neural response, x, onto some decision, d. Unfortunately, this mapping function cannot be observed but must be inferred from data. To address this issue, an experimenter would manipulate the generative model of a task in such a way that agents only can maximise reward by using beliefs in the mapping from some neural response onto some decision. We will mention a few findings which indicate that people behave as predicted by BDT (for reviews see Ernst & Bülthoff, 2004; Ma & Jazayeri, 2014; Trommershäuser et al., 2008; Wolpert & Flanagan, 2010).

In sensory tasks, the experimenter would typically seek to manipulate an individual's belief about the state of the world $P(C|x)$ by changing the stimulus s (e.g., increase the stimulus intensity), the reliability of the information that the neural response x provides about s (e.g., embed the stimulus in a set of distractors) or the prior probability of the state of the world C (e.g., increase the base rate of a certain state of the world).[7] In multisensory-integration tasks, for example, participants might be presented with visual and haptic information (e.g., about

[7] We use *reliability* to refer to some objective aspect of information; for example, the degree to which information can be trusted or is diagnostic of the true state of the world. In general, we can quantify the reliability of information in a given situation or task. However, we use *estimates / representations of reliability* and *estimates / representations of uncertainty* interchangeably.

whether some object is taller than a reference object), but one of the two sources of sensory information would be made less reliable by the experimenter. In such situations, participants have been shown to integrate the different sources of sensory information by weighting each sensory source according to its reliability, even when the less reliable sensory source generally is the more reliable one in the task at hand (Ernst & Banks, 2002). Importantly, this weighted averaging of different sources of sensory information has been found in the absence of outcomes; participants could not have learnt the appropriate stimulus-response mappings through trial-and-error (reinforcement) learning, but must have used knowledge about their sensory noise to select decisions (Ernst & Banks, 2002).

In motor tasks, the experimenter would typically seek to manipulate an individual's belief about the variability of their movements $P(a|d)$ or introduce an outcome structure which would incentivise individuals to use such a belief. Sensory beliefs would typically be irrelevant; the stimulus would not be ambiguous. In reaching tasks, for example, movements can be perturbed by linking an arm to a mechanical apparatus. In such situations, people have been shown to compensate for the variability of their movements. For example, when motor errors are large or the demand for accuracy is high, participants stiffen their effectors to stabilise their movements (Selen, Beek, & van Dieën, 2006). Moreover, in reaching tasks, participants might be asked to hit a target region while avoiding a partially overlapping penalty region under speed pressure. In such situations, participants have been shown to adjust their aiming point according to the stakes as well as the degree of overlap between the regions in a way that maximises expected gain (Trommershäuser, Landy, & Maloney, 2006). If the penalty is low, then they would aim closer towards the centre of the target region, to avoid missing the target region by mistake. In contrast, if the penalty is high, then they would aim further away from the centre of the target region, to avoid hitting the penalty region by mistake. Critically, without movement variability, no adjustments would be needed, indicating that participants indeed used knowledge about their motor noise to select decisions. Further, such adjustments

are made rapidly for novel configurations of the target region and the penalty region (Trommershäuser, Maloney, & Landy, 2003); ruling out that participants learnt the appropriate stimulus-response mappings through trial-and-error (reinforcement) learning.

In reward-based tasks, the experimenter would typically seek to manipulate an individual's beliefs about some probabilistic outcome structure $P(R|C,a)$. Sensory and motor beliefs would typically be irrelevant; the stimulus would not be ambiguous and decisions would be indicated by a simple button press. In the simplest case, for example, participants would be presented with two gambles which are perfectly matched in terms of reward magnitude but which have different probabilities of being realised (e.g., win £10 with 70% probability or win £10 with 40% probability; notice that this problem does not allow for loss aversion). In such situations, participants would choose the safer option, indicating that they can reason with probabilities and expected values (Tversky & Kahneman, 1981). In more complex tasks, the probability that each gamble is realised is not stated explicitly but has to be learnt. Further, the probability that each gamble is realised might undergo sudden but unannounced transitions, such that what used to be the better option is now the worse option. In such situations, participants learn the new outcome structure much more quickly than expected under trial-and-error (reinforcement) learning, indicating that they used knowledge about the complex probabilistic relationship between actions and outcomes to select decisions (Behrens, Woolrich, Walton, & Rushworth, 2007; Payzan-LeNestour & Bossaerts, 2011).

1.2.3 Metacognition

1.2.3.1 Conceptual definition of metacognition

Metacognition is broadly defined as a set of supervisory cognitive processes which form meta-level representations about cognition and which in turn use these representations for control

of cognition (Fernandez-Duque, Baird, & Posner, 2000; Metcalfe, 2009).[8] Metacognition thus defined initially appears indistinguishable from decision-making. In particular, we have argued that people behave as if they use representations of the uncertainty associated with not only *external* task-relevant variables (e.g., outcome noise) but also *internal* task-relevant variables (e.g., sensory and motor noise) to evaluate different decision alternatives (**Figure 1-2**). As such, an integral part of decision-making might in fact be the formation of meta-level representations about cognition and the use of these representations for control of cognition. We can, however, draw two distinctions between the canonical definitions of decision-making and metacognition – in terms of their general purpose and their respective computations.

First, the purpose of decision-making is to be able to respond to events as they unfold in the *external* world. In contrast, the purpose of metacognition is to be able to respond to events as they unfold in the *internal* world – with an internal event being the commitment to some cognitive state or action (i.e., a decision). As such, metacognition can be thought of as meta-decision-making – that is, the formation of meta-level representations about the uncertainty associated with a decision and the use of such representations to determine the best course of subsequent (cognitive and/or physical) action. More generally, metacognition is a set of *second-order* processes which make inferences about the *first-order* processes that drive behavioural responses to external stimuli (Fleming, Dolan, & Frith, 2012).

Second, decision-making uses information *in* the system – information which need not be represented – whereas metacognition constructs knowledge *for* the system – information which must necessarily be represented (Cleeremans, 2014). In some situations, decision-making and metacognition could use the same meta-level representations; a representation of the uncertainty associated with sensory information can be used not only to select the best

[8] Here, a *meta-level representation* is a representation about a property of the internal world (e.g., the reliability of a perceptual representation). In contrast, an *object-level representation* is a representation about a property of the external world (e.g., the nature, location or value of a stimulus).

decision but also to evaluate whether the decision was correct and determine the best course of subsequent (cognitive and/or physical) action. However, while metacognition always involves meta-level representations, decision-making may sometimes not do so. For example, there is evidence that meta-level information is intrinsic to the way in which the brain encodes sensory information (e.g., the notion of probabilistic population codes), and that optimal decisions can be achieved through linear operations on neural activity, without ever representing the meta-level information per se (Beck et al., 2008; Jazayeri & Movshon, 2006; Ma, Beck, Latham, & Pouget, 2006).[9] In such situations, metacognition cannot inherit a meta-level representation from the first-order decision processes but must construct a meta-level representation by itself using the available information (e.g., the time taken to make a decision or the history of past outcomes).

Metacognition defined as above is important not only to individual behaviour but also to group behaviour (Frith, 2012; Shea, Boldt, Bang, Yeung, Heyes, & Frith, 2014). In many situations, agents must make a series of interlinked decisions, where the best course of action at a given stage depends on earlier decisions. For example, it might not make sense to continue hunting a prey if it is unlikely that an animal was seen. In such cases, agents may optimise individual behaviour (e.g., preserve energy or minimise opportunity costs) by taking into account the uncertainty associated with their initial decision (i.e., their perceptual state). In addition, if the interlinked decisions are made by separate but interacting agents, then the agents may similarly optimise their group behaviour (e.g., whether to continue hunting as a pack) by sharing the uncertainty associated with their respective initial decisions (i.e., their respective perceptual states). In fact, under certain assumptions, the agents would be able to

[9] Here, *meta-level information* is information about a property of the internal world, but it need not be directly represented. In contrast, *object-level information* is information about a property of the external world, but – again – it need not be directly represented. As such, we take information to be a matter of correlation (e.g., the variance of the firing rate in a population of visual neurons typically correlates with the reliability of a percept).

make more accurate decisions together than alone – a phenomenon which has been called the Two-Heads-Better-Than-One (2HBT1) effect (Hill, 1982) and which we will explore in this thesis.

To play this dual role, metacognition must involve two types of meta-level representation: an *implicit* representation, which can be used to optimise individual behaviour, and an *explicit* representation, which can be shared and used to optimise group behaviour. The two types of representation are intimately linked with the notion of confidence. For simplicity, we will refer to the implicit representation and the explicit representation as a *confidence variable* and a *confidence report*, respectively. In everyday language, the term *confidence* has several meanings – from a feeling of whether we are right or wrong to a personality trait linked to our self-esteem and self-assurance. In this thesis, we will only use the term to refer to some meta-level representation of the uncertainty associated with a decision. It may have been more accurate to speak of *decision confidence* – a term often used in the literature on decision-making. However, for simplicity, we will only speak of *confidence*, leaving *decision* implied.

Using this terminology, we can now define metacognition in a limited sense for the current thesis: cognitive processes that are involved in the generation of a confidence variable and a confidence report and/or in the use of these two types of meta-level representation to determine the best course of subsequent (cognitive and/or physical) action. We will now seek to use BDT to specify in more formal terms how these cognitive operations may be performed – with the aim to lay the foundation for a normative theory of metacognition. We will focus on the generation of a confidence variable and a confidence report; the use of these two types of meta-level representation is assumed to follow the same principles as decision-making.

1.2.3.2 Using Bayesian decision theory to formalise metacognition

How would a Bayes optimal agent compute an estimate of the uncertainty associated with a decision? We want to define a confidence variable, z, which can be used to guide individual behaviour in different types of task and which in turn can be transformed into a confidence report, c, which can be used to guide group behaviour with different agents (**Figure 1-3**). In the BDT framework, an intuitively appealing confidence variable is the strength of the belief that a given decision maximises reward. In sensory tasks, which we will consider in this thesis, the generative model can be set up such that a Bayes optimal agent should make a decision, d, solely on the basis of their sensory belief about the state of the world, $P(C|x)$, which we will denote $\hat{C}$. This relationship is achieved by taking motor beliefs and outcome beliefs out of the equation; for example, simple button press is used to indicate a decision and a unit reward, $\hat{R}$, is administered when the true state of the world, C, is correctly inferred.

In this case, we can simplify **Eq. 1-1** and define the strength of the belief that a given decision maximises reward as the probability that the decision – which is solely determined by the sensory belief about the state of the world – is correct: $z = P(R|x,d) = P\left(\hat{R}\middle|x,d\right) = P\left(\hat{R}\middle|x,\hat{C}\right) = P\left(C = \hat{C}\middle|x\right)$. We stress that this confidence variable reflects a *subjective* probability and it need not match the *objective* probability that a given decision is correct. We note, however, that real agents might compute the confidence variable in a more heuristic (but sub-optimal) manner. In simple sensory tasks, optimal decisions can be made by comparing the neural response, x, to some decision criterion, without having to take into account the uncertainty associated with the neural response (Green & Swets, 1966). In such tasks, the neural response might be used directly as an estimate of the uncertainty associated with a decision, $z = x$.

How would a Bayes optimal agent communicate the estimate of the uncertainty associated with a decision? We next want to define a confidence report, c, which is some transformation

of the confidence variable, z (**Figure 1-3**). We will for now only assume that the confidence report should be a *monotonically increasing* function of the confidence variable – that is, c should increase strictly with z. The nature of the mapping function, $c(z)$, will depend on which format is used for communication. For example, if z were to be mapped onto a continuous probability scale (i.e., from 0 to 1), then the mapping function will be continuous. Importantly, the mapping function need not be linear or preserve an identity between z and c (e.g., $z = .90$ need not be mapped onto $c = .90$). If z were to be mapped onto a discrete probability scale (e.g., from 0 to 1 in steps of .1), then the mapping function will be made up of a series of thresholds which partition the space of z. More generally, the mapping function will be complex if z were to be mapped onto linguistic confidence expressions (e.g., "unsure", "pretty sure" and "sure") as they do not have a universally established meaning. Regardless of the target format, the optimal transformation will depend entirely on the task – an issue which we will return to shortly.

The issue of whether confidence reports reflect optimal or heuristic confidence variables has important implications for inter-personal communication. The optimal confidence variable computed using Bayesian inference is *task-independent* in the sense that it – by virtue of being a probability – can be compared between *different* tasks and even between *different* agents. In contrast, a heuristic confidence variable is *task-dependent* in the sense that its range and distribution will typically depend on not only the task but also which agent is performing the task. For example, in the BDT framework, the neural response x is not diagnostic of the true state of the world but is corrupted by both external (task-specific) and internal (agent-specific) sources of noise. As such, a heuristic confidence variable would be difficult to compare between different tasks, let alone between different agents.

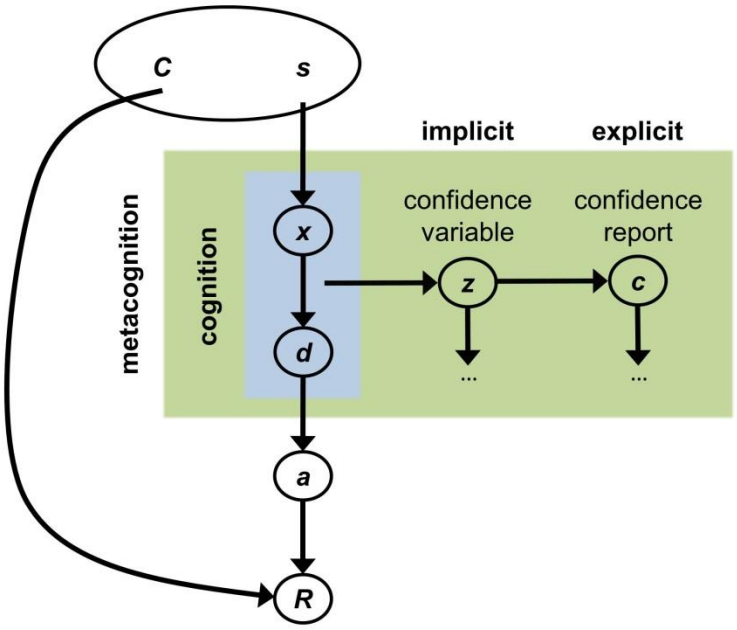


Figure 1-3. Metacognition in terms of Bayesian decision theory. There are two types of meta-level representation involved in metacognition: a confidence variable, z, which can be used to guide individual behaviour, and a confidence report, c, which can be used to guide group behaviour. A Bayes optimal agent computes the confidence variable as the probability that a given decision is correct and generates a confidence report by thresholding the confidence variable according to the task. See text for details. [This figure is adapted from Ma & Jazayeri, 2014.]

1.2.3.3 Do humans compute confidence variables in an optimal manner?

The question that we now consider is whether people can compute a confidence variable in an optimal manner – that is, compute the probability that a given decision is correct using Bayesian inference. Unfortunately, the computation of this variable cannot be observed but must be inferred from data. To our knowledge, only two studies have sought to overcome this problem and to test whether people compute a confidence variable in an optimal manner.

In one study, participants were asked to indicate which of two Gabor patches they would prefer to make an orientation judgement about (Barthelmé & Mamassian, 2009). The Gabor patches had the same level of contrast but were embedded in noise by adding random perturbations to the luminance level of each image pixel. In the experimental condition, noise was created separately for each Gabor patch, which meant that the orientation of one Gabor patch often was more ambiguous than that of the other one. In contrast, in a control condition, noise was created for only one Gabor patch, which then was mirrored (flipped along its horizontal axis) to create the other Gabor patch. While the Gabor patches were different pixel-by-pixel, their orientations were equally ambiguous. The orientation judgements were more accurate in the experimental than in the control condition, which suggests that participants estimated the probability that their decision about a given Gabor patch would be correct and then used this information to determine which Gabor patch to make an orientation judgement about. In the experimental condition – in contrast to the control condition – there was a benefit to be gained from computing and utilising such a confidence variable as the orientation of the two Gabor patches were not equally ambiguous.

In another study, participants were asked to indicate on which of two trials they felt more likely to be correct (de Gardelle & Mamassian, 2014). Critically, the two trials could involve the same task or two different tasks. In both tasks, two Gabor patches were presented in rapid succession. In one task, participants had to judge whether the second Gabor patch was tilted

to the left or to the right relative to the orientation of the first Gabor patch. In the other task, participants had to judge whether the spatial frequency (i.e., the number of "lines" in the Gabor patch) of the second Gabor patch was lower or higher relative to that of the first Gabor patch. Participants were *objectively* more likely to be correct on those trials on which they *subjectively* felt more likely to be correct. In addition, this relationship – both in terms of direction and in terms of magnitude – was observed regardless of whether the two trials involved the same or two different tasks, which suggests that participants used some task-independent confidence variable to guide their choices, such as the probability that a given decision was correct computed using Bayesian inference.

The two studies strongly suggest that people compute and utilise a confidence variable and indicate albeit indirectly that this confidence variable is computed in an optimal manner. In social interaction, however, our confidence is usually not revealed through forced-choices, but through communication, using verbal expressions (e.g., "unsure", "pretty sure" and "sure") or numerical expressions (e.g. "60%"). This additional complexity may have implications for how an underlying confidence variable is computed. For example, the generation of a confidence report involves working memory and may thus be affected by cognitive load (Lavie, 2005; Sweller, 1988) or dual-task interference (Daniel Kahneman, 1973; Pashler, 1994). It thus remains unknown whether confidence reports reflect confidence variables that are computed in an optimal manner – a question which we will address in this thesis.

1.2.3.4 Do humans report confidence in an optimal manner?

The question that we now consider is whether people report their confidence in an optimal manner. In the case of the confidence variable, optimality had a straightforward definition – that is, the probability that a given decision is correct computed using Bayesian inference. In the case of the confidence report, a general definition of optimality is hard to formulate as

optimality depends entirely on the task at hand. In fact, without incentives, there is no reason why a confidence variable should be mapped onto a confidence report in any particular manner. Here, we will mention two lines of research which have sought to overcome these problems and to test whether people report their confidence in an optimal manner.

In the judgement and decision-making literature, participants are typically asked to report their confidence in a decision as a probability (Harvey, 1997). Participants are then said to be *well-calibrated* (i.e., optimal) when the *reported* probability on average reflects the *objective* probability of being correct; for example, that participants are correct 70% of the time when they report that they feel "70%" confident (**Figure 1-4**). This behaviour can be incentivised by implementing *strictly proper scoring rules* under which participants only can maximise their reward by being well-calibrated (Brier, 1950). This line of research has generated two robust findings: the *overconfidence* and the *hard-easy* effects. In a range of tasks – from general-knowledge, through economic forecasts, to perception – participants have been found to be overconfident and overstate the probability of being correct, even when they have strong incentives not to do so (**Figure 1-4**) (Gigerenzer, Hoffrage, & Kleinbölting, 1991; Liberman & Tversky, 1993; Lichtenstein, Fischhoff, & Phillips, 1982; Moore & Healy, 2008; Zylberberg, Roelfsema, & Sigman, 2014). Paradoxically, this effect can be reversed if a given task is made very easy; in this case, participants have been found to be underconfident and understate the probability of being correct (**Figure 1-4**) (Gigerenzer et al., 1991; Lichtenstein & Fischhoff, 1977; Moore & Healy, 2008; Soll, 1996).

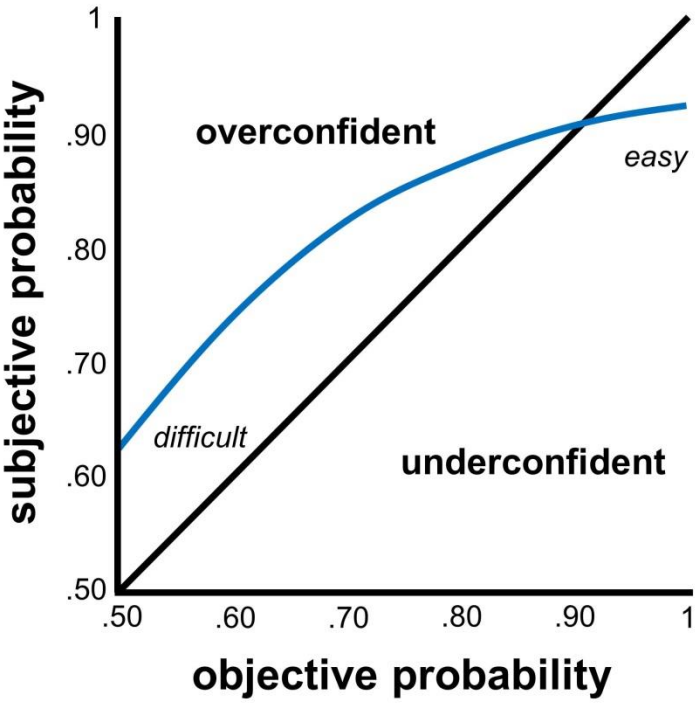


Figure 1-4. Illustration of a calibration curve and findings from the judgement and decision-making literature. The axes denote reported probability correct (y-axis) and the objective probability correct associated with a given report (x-axis). The calibration curve for a perfectly calibrated agent would fall along the identity line (black line). This agent (blue line) is, however, poorly calibrated, being generally overconfident, but underconfident when the task difficulty is very low.

Different theories have been proposed to explain inaccurate (poorly calibrated) confidence reports, with the most prominent ones proposing fundamental limitations on how the human mind represents uncertainty or a lack of insight into one's internal operations (for a review see Harvey, 1997). These accounts would seem to imply a problem all the way down to the level of the computation of the confidence variable. However, as we will show, it is in fact possible to compute the confidence variable in an optimal manner while still giving inaccurate confidence reports. More generally, the studies discussed above ignore the fact that confidence reports are meant not for *oneself* but for *others* (Frith, 2012; Shea et al., 2014); observed biases might therefore be social in nature and operate at the level of the mapping of the confidence variable onto a confidence report – an issue which will explore in this thesis. In this light, it would seem more appropriate to study the optimality of confidence reports in their natural habitat – that is, social interaction.

We now turn to a set of studies which have addressed this issue and studied the optimality of confidence reports in the context of group decision-making in the sensory domain (Bahrami, Olsen, Bang, Roepstorff, Rees, & Frith, 2012; Bahrami, Olsen, Latham, Roepstorff, Rees, & Frith, 2010; Bang et al., 2014). There, pairs of participants (a dyad) made a private decision about whether a visual target occurred in a first or a second viewing interval (a two-interval forced-choice task). If the dyad members selected the same interval, they received feedback about the accuracy of each decision and continued to the next trial. If the dyad members selected different intervals, they were asked to verbally negotiate a group decision – a process which involved sharing and combining explicit reports of confidence. In this task, to maximise the accuracy of the group decisions, dyad members should report their confidence in a *mutually consistent* manner (i.e., they should use the same subjective mapping from some confidence variable onto some confidence report) as it minimises miscommunication about whose decision is more likely to be correct in a given situation. Further, dyad members should

assign the right weights to each other's confidence reports; for example, if one of them persists with overstating their confidence, then their reports should be discounted.

Dyad members overall made better decisions together than alone. This 2HBT1 effect persisted when feedback was removed, but disappeared when communication was not allowed. These findings indicate that dyad members had not simply learnt the *objective* probability that their respective decisions were correct from the trial-by-trial feedback, but instead, when resolving disagreement, accurately reported some representation of the uncertainty associated with their respective decisions – a representation which could be compared between them – and were adequately perceptive to make efficient use of this information. Interestingly, a linguistic analysis of the verbal discussions showed that dyad members who aligned their confidence reports accrued larger 2HBT1 effects (Fusaroli et al., 2012). More specifically, dyad members who used expressions from the same linguistic set (e.g., both using gradients of "sure", such as "unsure" and "pretty sure") performed better together than those who used expressions from different linguistic sets (e.g., gradients of "sure" versus "know"). This finding indicates that the report of confidence is flexible and that people can and sometimes do learn to adapt their confidence reports to the social context – a level of flexibility which is important given that expressions of confidence do not have a universally established meaning.

The latter line of research indicates that people, at least in social interaction, can report their confidence and utilise such reports in an optimal manner. However, we still do not have a full understanding of how these processes unfold in social interaction. In the studies above, dyad members were engaged in a complex interaction in which they could influence the group decision not only by reporting their confidence in a particular manner but also by negotiating the weights assigned to their respective reports. As such, sub-optimality in the way in which confidence was reported could be offset by optimality in the way in which the reports were combined, and vice versa. In addition, the studies above did not record confidence reports in a

format that could be modelled (i.e., free verbal discussion as opposed to asking dyad members to indicate their confidence on some scale), making it harder to infer exactly what information was reported, how the information was reported and how the reports were combined. It thus remains unknown whether people can report their confidence and utilise such reports in an optimal manner. Importantly, a more mechanistic account of these processes would allow us to better understand situations in which individuals fail to make better decisions together than alone. In the studies above, dyad members did not accrue a 2HBT1 effect when they were very different in terms of their individual performance. However, it is unclear what caused this failure of social interaction; for example, a fundamental sub-optimality in the ability to share and combine confidence reports *or* social biases. More generally, an account of exactly how humans report their confidence and utilise such reports is important given the pivotal role of confidence reports in high-risk decision domains, such as financial investment (Broihanne, Merli, & Roger, 2014), medical diagnosis (Berner & Graber, 2008), jury verdicts (Tenney, MacCoun, Spellman, & Hastie, 2007) and political negotiations (Johnson, 2004).

1.2.4 Thesis questions

This thesis will seek to develop a more accurate understanding of the component processes involved in sharing and combining representations of uncertainty (**Figure 1-1**). In this section, we have introduced the concepts relevant to each component process, specified how each component processes should be performed and identified outstanding questions – questions that will guide our empirical work. In general, our approach will be to specify the optimal solution to a given problem and then to test experimentally whether people can or do solve this problem in an optimal manner – with deviations from optimality being particularly insightful. We will turn normative theories into more descriptive theories whenever they conflict with the reality of human behaviour – in all cases, the goal will be to develop a

mechanistic (computational) understanding of each of the component processes involved in sharing and combining representations of uncertainty. Here, we will outline the questions that we will seek to test empirically.

In the first two empirical chapters, we will focus on non-interacting agents, with the rationale being that non-social behaviour must be understood in order to appreciate the complexities that may arise during social interaction. In **Chapter 2**, we will focus on the generation of a decision and a confidence report. We will in particular test whether confidence reports reflect confidence variables that are computed in an optimal or a heuristic manner. We will also spell out the claim that people can compute the confidence variable in an optimal manner while still giving inaccurate confidence reports. In **Chapter 3**, we will test whether the report of confidence – that is, the mapping of a confidence variable onto a confidence report – is flexible. We will argue that such flexibility – if found – poses a problem for current theories and measures of metacognition as they assume that the report of confidence is or should be stable through time.

In the next three empirical chapters, we will turn our focus to agents who are engaged in group decision-making. In **Chapter 4**, we will test whether people can learn to report their confidence in an optimal (mutually consistent) manner. This problem is an instance of a more general problem that is inherent to most verbal communication. In particular, to minimise miscommunication, interacting agents must ensure that they intend the same meaning when they use a given expression. In **Chapter 5**, we will take a first step towards understanding how people weight each other's confidence reports and test whether they can outperform a simple algorithm which always follows the individual decision made with higher confidence. Lastly, in **Chapter 6**, we will directly test whether people can learn to weight each other's confidence reports in an optimal manner – that is, in a way that maximises the accuracy of their group decisions. This problem is an instance of a more general advice-taking problem that is inherent

to most cooperation. In particular, to ensure a good outcome, interacting agents must learn when to trust each other's opinions. The empirical chapters represent a progression through the component processes involved in sharing and combinig representations of uncertainty in the context of group decision-making (**Figure 1-1**).

1.3 Experimental and computational framework

In this section, we will outline how we will seek to address the questions specified above. We will focus on three choices: the task domain, the method for eliciting confidence reports and the approach to modelling how decisions and confidence reports are generated in the chosen task domain. The goal of this section is to motivate and provide a general background for our experimental and computational framework; the exact details will be specified in the relevant chapters.

1.3.1 Studying confidence

1.3.1.1 Task domain

The task domain and its setting will determine the generative model of the decision problem faced by participants. We would want to use a task that satisfies three requirements. First, the normative solution to the decision problem should be tractable. Second, we should have full experimental control over the task-relevant information. Lastly, the source of noise should not be correlated between participants, as this would allow us to study cooperative situations in which participants can gain a significant advantage by sharing their respective representations of uncertainty. To these ends, and in line with the group studies discussed above, we will use visual psychophysics. Psychophysicists have developed sophisticated methods for inferring the latent (unobservable) processes that govern the relationship between some stimulus and the

(neural) responses evoked by this stimulus; for example, their probabilistic relationship can be inferred by recording variation in behavioural responses while keeping the stimulus constant (Green & Swets, 1966; Macmillan & Creelman, 2005). As such, under simplifying assumptions, we can specify the normative solution to a psychophysical task despite the primary source of noise being inside the brain. Further, we can fully control the task-relevant information as participants can only rely on the presented stimuli for their responses (cf. participants may rely on complex memories in general-knowledge tasks). Lastly, sensory noise in two brains is uncorrelated, thus allowing us to study cooperative situations in which a pair of agents can gain a significant advantage by sharing their respective representations of uncertainty.

There are of course other task domains that are suitable for studying confidence. For example, there is a tradition in the judgement and decision-making literature for studying confidence reports in the context of general knowledge and forecasting (Wright & Ayton, 1987; Yates, 1990). While these task domains might be more ecologically valid, the generative model of a given decision problem is hard to specify. Without an accurately parameterised generative model, it can be impossible to tell at which level inaccurate confidence reports arise; for example, a fundamental problem in the computation of a confidence variable or a bias in the mapping of the confidence variable onto a confidence report.

1.3.1.2 Eliciting confidence

There are broadly two ways of eliciting confidence: implicit or explicit behavioural responses (Kepecs & Mainen, 2012). In a task using an implicit response mode, participants would not be required to report their confidence; instead the experimenter would create a situation in which a participant would gain an advantage by representing the uncertainty associated with a decision and using this representation to guide their individual behaviour (e.g., as in the studies by Barthelmé & Mamassian, 2009, and de Gardelle & Mamassian, 2014). In a task using

an explicit response mode, participants would be asked to report their confidence (e.g., as in the judgement and decision-making studies mentioned above) or the experimenter would create a situation in which participants would gain an advantage from reporting their confidence as part of a group (e.g., as in the group decision-making studies mentioned above). Implicit response methods have proven a powerful tool for establishing that humans and other animals compute a confidence variable and use this variable to guide their individual behaviour (Kepecs, Uchida, Zariwala, & Mainen, 2008; Kiani & Shadlen, 2009; Komura, Nikkuni, Hirashima, Uetake, & Miyamoto, 2013; Shields, Smith, & Washburn, 1997). However, as discussed earlier on, confidence in human social interaction is usually revealed through verbal communication; an additional level of complexity which may have implications for how an underlying confidence variable is computed, and therefore makes it hard to generalise insights from tasks using an implicit response mode.

For the sake of generality, we would want to use a response mode which mimics verbal communication but has a format which allows for modelling. To these ends, we will use an ordinal confidence scale (e.g., 1 to 6 in discrete steps of 1). While the different levels of the scale have an ordering, the (statistical) meaning associated with each level is unspecified. Similarly, in verbal communication, while most people would agree on the ordering of confidence expressions (e.g., “sure” > “unsure”), they would often assign different meanings to each expression (e.g., saying “unsure” when they are either 50% or 60% likely to be correct). In both cases, to minimise miscommunication, local norms must be established so that a given response has a stable (shared) meaning.

There are of course other explicit response modes for eliciting confidence. For example, as in the judgement and decision-making literature, participants could be asked to report their confidence as a probability (.5 to 1 in steps .1). However, in this case, participants would have strong a priori associations for each confidence expression, which would make it unlikely that

they would engage in a negotiation of what each confidence expression means – a key focus of this thesis. Alternatively, participants could be asked to report their confidence as a wager (e.g., bets of £0.5 to £1 in steps of £0.1), which then would be translated into a reward or a penalty depending on the accuracy of the associated decision. However, recent work has shown that people's wagers are distorted by their risk attitudes (Fleming & Dolan, 2010), thus introducing additional complexities that would make it harder to model recorded responses. We will, however, use probability judgements and wagers when the two types of response mode can be used as a control for a specific hypothesis.

1.3.2 Modelling confidence

1.3.2.1 Signal detection theory

In line with current approaches, we will assume that confidence in psychophysical tasks can be modelled using Signal Detection Theory (SDT; Green & Swets, 1966), which is a special case of BDT (Maloney, 2002). Here, we will outline how SDT describes the generative model of the decision problem in a psychophysical task and how it models the cognitive processes that are involved in the generation of a decision and a confidence report. Consider the two-interval forced-choice task which was used in the group studies discussed above and which we will also use for our empirical work (**Figure 1-5A**). An agent is presented with two viewing intervals, with each interval containing six Gabor patches. In either the first or the second interval, one of the six Gabor patches (the visual target) has a higher level of contrast. After the stimulus, s, has been presented, the agent must decide which interval they think contained the visual target and indicate their confidence in this decision (e.g., 1 to 6 in steps of 1).

In the SDT framework, the stimulus is assumed to evoke sensory evidence, x, which is usually defined as the stimulus corrupted by additive Gaussian noise: $x = s + \eta$, where $\eta \in N(0, \sigma)$. More generally, we can think of the sensory evidence as a random sample from a Gaussian

distribution, $N(s, \sigma)$, whose mean, s, is given by the stimulus and whose standard deviation, σ, relates to the level of sensory noise and must be fitted to the data (**Figure 1-5B**). In our task, the stimulus, s, can be defined as the difference in contrast between the second and the first interval at the target location, with the sign of s indicating the target interval and the magnitude of s indicating the target intensity. Together, the two parameters, s and σ, specify the fundamental generative model of the decision problem (i.e., the probabilistic relationship between the stimulus and the neural response).

To generate a decision, d, the agent first transforms the sensory evidence, x, into a decision variable, z^D, and then applies a single threshold, $d(z^D)$ (**Figure 1-5C**). The agent may compute the decision variable in a number of ways; for example, as the raw sensory evidence, $z^D = x$, or as the probability that the second interval contains the visual target, $z^D = P(s > 0|x)$. In the case of one-dimensional sensory evidence (i.e., s is scalar), the two decision variables are perfectly correlated and thus indistinguishable (Green & Swets, 1966). We will for now assume that the decision variable is computed as the raw sensory evidence. In this case, the optimal decision threshold would be 0 ($z^D < 0$: "interval 1"; $z^D > 0$: "interval 2") as it maximises *hits* (true positives) and minimises *false-alarm*s (false positives). However, if an asymmetric reward function is introduced (e.g., the agent receives a larger reward for identifying a visual target in interval 2), then it would be optimal to shift the decision threshold away from 0 (in this case, below 0). Similarly, under certain task expectations (e.g., the agent expects that the visual target will occur more often in interval 2), it would be optimal to shift the decision threshold away from 0 (again, below 0) – an issue which we will discuss in more detail in **Chapter 3**.

To generate a confidence report, c, the agent first transforms the sensory evidence, x, and the decision, d, into a confidence variable, z^C, and then applies a set of thresholds, $c(z^C)$, with the level of confidence increasing as each threshold is crossed (**Figure 1-5D**). The agent may compute the confidence variable in a number of ways; for example, as the absolute value of

the sensory evidence, $z^C = |x|$, or as the probability that the decision was correct, $z^C = P(\text{correct}|x, d)$. Again, in the case of one-dimensional sensory evidence, the two confidence variables are perfectly correlated and thus indistinguishable (de Gardelle & Mamassian, 2014) – an issue which we will explain in more detail in **Chapter 2**. As discussed previously, the optimal confidence thresholds are entirely dependent on the task – we will discuss different strategies that a single agent and interacting agents may use in **Chapter 3** and **Chapter 4**, respectively. We will for now only note that most theories and measures of metacognition assume that the confidence thresholds are stable across time and only reflect individual biases (Fleming & Lau, 2014).

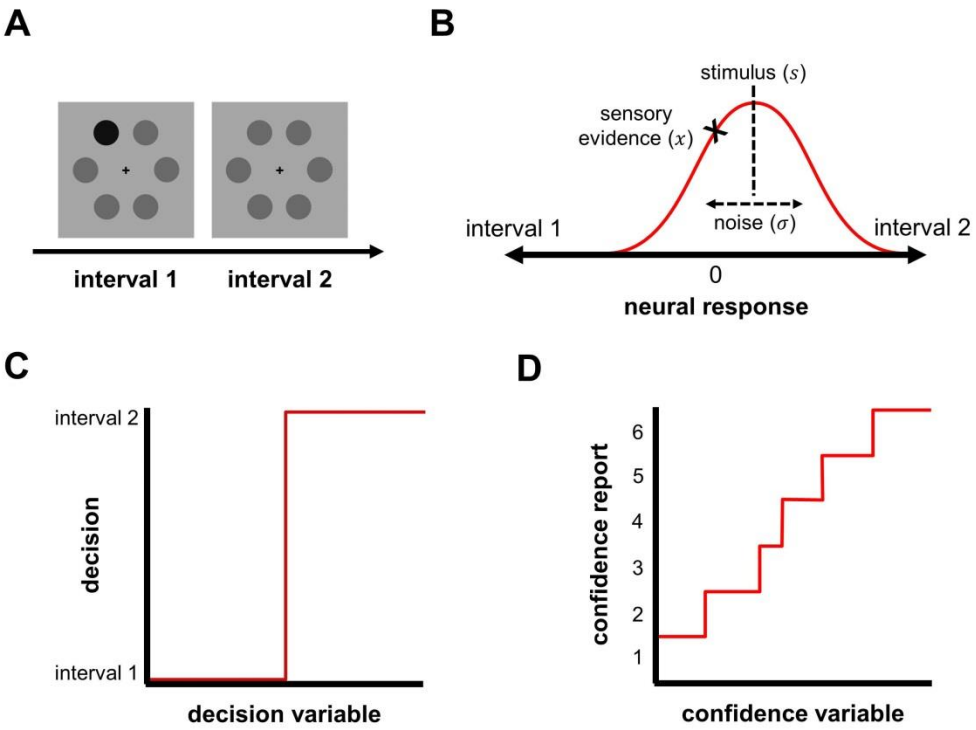


Figure 1-5. Signal detection theory model of decisions and confidence reports. **(A)** The agent performs a two-interval forced-choice contrast-discrimination task. On each trial, the agent views two consecutive intervals, each containing six contrast gratings (here dots). There is a visual target with higher contrast (darker dot) in one of the two intervals. **(B)** The agent receives on each trial sensory evidence, x, randomly drawn from a Gaussian distribution whose mean, s, is given by the stimulus and whose standard deviation, σ, specifies the level of sensory noise. Here, the stimulus sign indicates the target interval and the absolute stimulus value indicates the target intensity. **(C)** To make a decision, the agent transforms the sensory evidence into a decision variable and applies a single threshold. **(D)** To make a confidence report, the agent transforms the sensory evidence and the decision into a confidence variable and applies a set of thresholds, with the level of confidence increasing as each threshold is crossed. See text for details.

Dynamic SDT models of confidence have been developed to explain the interplay between confidence reports, accuracy and reaction time, by assuming that the sensory evidence, x, is made up of a series of samples (Kiani, Corthell, & Shadlen, 2014; Moran, Teodorescu, & Usher, 2015; Pleskac & Busemeyer, 2011; Ratcliff & Starns, 2009, 2013; Douglas Vickers, 1979). There are broadly two types of dynamic SDT models: one-accumulator and multiple-accumulator models. In both types of model, the sensory evidence and the decision variable are assumed to be equivalent; we will therefore only refer to the sensory evidence.

In one-accumulator models, the sensory evidence is encoded by a single accumulator, e. At each moment in time, a random sample is taken from a Gaussian distribution specified as in static SDT models and added to the sensory evidence accumulated thus far. A decision is made once the accumulator reaches a predetermined threshold for a given decision option. The confidence variable is usually taken to be some function of the end state of the accumulator, the decision and/or the time elapsed before making a decision. Some dynamic SDT models assume that the confidence variable also benefits from additional post-decision accumulation of sensory evidence (e.g., sampling of information maintained in working memory). As in static SDT models, a confidence report is made by thresholding the confidence variable at the relevant moment in time.

In contrast, in multiple-accumulator models (or race models), there is a separate accumulator for each decision option – here, we will only consider the case of two decision options and thus two accumulators, e_1 and e_2. The sensory evidence associated with the two decision options is accumulated at the same time; the sensory evidence associated with the correct option is sampled from a Gaussian distribution centred on the stimulus, whereas the sensory evidence associated with the incorrect decision option is sampled from a Gaussian distribution centred on zero. A decision is made once one of the accumulators reaches a predetermined threshold (i.e., wins the race). The confidence variable is usually taken to be some function of

the end states of the accumulators; for example, their absolute difference, $z^C = |e_1 - e_2|$, or the probability that the decision was correct, $z^C = P(\text{correct}|e_1, e_2, d)$. As in static SDT models, a confidence report is made by thresholding the confidence variable at the relevant moment in time.

We note that, if the accumulation process can be assumed to stop at a prespecified moment in time (e.g., because the stimulus disappears or because a decision is required), then dynamic SDT models can be reduced to static SDT models, only losing the ability to account for reaction times. In this case, the fitted level of sensory noise can be assumed to account for any noise in the accumulation process. As we are mainly interested in higher-level dynamics, we will take this simpler approach and specify our experimental tasks such that the data can be accurately modelled using static SDT models or static versions of dynamic SDT models.

Chapter 2: Confidence reports reflect Bayes optimal computations

Humans stand out from other animals in that they are able to report on the reliability of their internal operations. This ability is typically studied by asking people to report their confidence in the correctness of some decision. However, the computations underlying confidence reports remain unclear. In this chapter, we tested whether confidence reports reflect heuristic computations (e.g., the magnitude of sensory evidence) or Bayes optimal ones (i.e., how likely a decision is to be correct given the sensory evidence). In a standard design in which participants first made a decision, and only then gave their confidence, participants mostly used the Bayes optimal strategy. In contrast, in a less-commonly used design in which participants indicated their decision and their confidence simultaneously, they were roughly equally likely to use the Bayes optimal strategy or to use a heuristic strategy.

2.1 Introduction

Humans and other animals use estimates of the uncertainty associated with incoming sensory evidence to guide behaviour (Kepecs et al., 2008; Kiani & Shadlen, 2009; Komura et al., 2013; Lak, Costa, Romberg, Koulakov, Mainen, & Kepecs, 2014). For example, a monkey will wait until their sensory evidence is deemed sufficiently reliable before committing to a risky decision (Kiani & Shadlen, 2009; Komura et al., 2013). Humans can go further than other animals: they can explicitly report on estimates of the uncertainty associated with their sensory evidence, by saying, for instance, "I'm sure" – an ability that is important for effective cooperation (Bahrami et al., 2010; Fusaroli et al., 2012; Shea et al., 2014). This ability to report on the reliability of our internal operations is typically studied by asking people to report their confidence in the correctness of some decision (Fleming, Dolan, & Frith, 2012). However, the computations underlying confidence reports remain a matter of debate (Box 1 in Shea et al., 2014). For instance, in an orientation-discrimination task, confidence reports might – as a heuristic – reflect the perceived tilt of a bar. Alternatively, confidence reports might reflect more sophisticated computations, like Bayesian inference about the probability that the judgement about the bar tilt is correct.

Here, we asked: what computations do confidence reports reflect? We were in particular interested in whether confidence reports reflect heuristic or Bayes optimal computations. The latter would be consistent with a wide array of work showing that other aspects of cognition are Bayes optimal (Ma & Jazayeri, 2014). As discussed in **Chapter 1**, the question of whether confidence reports reflect Bayes optimal (or simply Bayesian) computations has implications for inter-personal communication (see **Section 1.2.3.2: Using Bayesian decision theory to formalise metacognition**). In particular, probabilities, as generated by Bayes optimal computations, can be compared across different tasks, and even different people. In contrast, heuristic computations typically lead to task-dependent representations, with ranges and distributions that depend strongly on the task, making it difficult to map them onto confidence reports consistently, or compare them between different people. However, to our knowledge, the question of whether confidence reports reflect Bayes optimal computations has not been directly tested. We used a standard psychophysical task in which participants receive sensory evidence, make a decision based on this evidence, and report how confident they are that their decision is correct. Our goal was to determine how participants transform sensory evidence into a confidence report. In essence, we were asking: if we use $\mathbf{x}$ to denote sensory evidence ($\mathbf{x}$ can be multi-dimensional) and c to denote a confidence report, what is the mapping from $\mathbf{x}$ to c? Alternatively, what is the function $c(\mathbf{x})$?[10]

To address this question, we followed an approach inspired by SDT models of confidence (see **Section 1.3.2.1: Signal detection theory**). In particular, we hypothesised that a participant first computes a decision variable, $z^{\mathrm{D}}(\mathbf{x})$, and compares this variable to a single threshold to generate a decision, *d*. We hypothesised that the participant then computes a confidence variable, $z^{\mathrm{C}}(\mathbf{x}; d)$, an internal representation of the strength of the evidence in favour of the selected decision, *d*, and then compares this variable to a set of thresholds to generate a

[10] In this chapter, we denote the sensory evidence $\mathbf{x}$ – and not x as is common SDT models – to indicate that we consider multidimensional sensory evidence – for reasons which soon will become clear.

confidence report, c.[11] Within this framework, a heuristic computation is a reasonable, but ultimately arbitrary, function of the sensory evidence. For instance, if the task is to choose the larger of two pieces of sensory evidence, x_1 or x_2, a heuristic confidence variable might be the difference between the two pieces of sensory evidence: $z_{\Delta}^{\mathrm{C}}(\mathbf{x}; d) = x_2 - x_1$ (subscript Δ denotes Difference). In contrast, the Bayes optimal confidence variable is the probability that a correct decision has been made given the two pieces of sensory evidence: $z_{\mathrm{B}}^{\mathrm{C}}(\mathbf{x}; d) = P(\mathrm{correct}|\mathbf{x}; d)$ (subscript B denotes Bayesian).

To our knowledge, it is impossible to determine directly the confidence variable, $z^{\mathrm{C}}(\mathbf{x}; d)$; instead, we can consider several models, and ask which is most consistent with experimental data. Choosing among different models for the confidence variable, $z^{\mathrm{C}}(\mathbf{x}; d)$, is possible in principle, but there are some subtleties. The most important subtlety is that if the task is "too simple", it is impossible to distinguish one model from another. Here, "too simple" means that the sensory evidence, $\mathbf{x}$, is one-dimensional (i.e., one piece of sensory evidence), which we write x to indicate that it is scalar. To see why this is a problem, suppose we wanted to distinguish between some heuristic confidence variable, $z_{\mathrm{H}}^{\mathrm{C}}(x; d) = x$ (subscript H denotes Heuristic), and the Bayes optimal confidence variable, $z_{\mathrm{B}}^{\mathrm{C}}(x; d) = P(\mathrm{correct}|x, d)$. Let's say that we found empirically that a participant always reported low confidence when the heuristic variable was less than 0.3 and high confidence when it was greater than 0.3. Can we conclude that the confidence reports reflect the heuristic variable? The answer is a clear no. For example, if the Bayesian variable is greater than 0.4 whenever the heuristic variable is greater than 0.3, then it is also true that our participant reported low confidence when the Bayesian variable was less than 0.4 and high confidence when the Bayesian variable was greater than 0.4. Thus, there is no way of knowing whether the confidence reports reflect the

[11] We note that the evidence in favour of one decision is different from the evidence in favour of the other one, so the confidence variable must not only depend on the sensory evidence, $\mathbf{x}$, but also the decision, *d*.

heuristic or the Bayesian variable. In general, there is no way to distinguish between any two functions of x that are monotonically related — one can simply map the thresholds through the relevant function (**Figure 2-1**).

The situation is very different when the sensory evidence is multi-dimensional (i.e., two or more pieces of sensory evidence), which we write $\mathbf{x}$ to indicate that it is a vector. As in the one-dimensional case, consider two models: a heuristic confidence variable, $z_{\mathrm{H}}^{\mathrm{C}}(\mathbf{x}; d)$, and a Bayes optimal confidence variable, $z_{\mathrm{B}}^{\mathrm{C}}(\mathbf{x}; d)$. In general, if $\mathbf{x}$ is a vector, it is not possible to get the same mapping from $\mathbf{x}$ to c using $z_{\mathrm{H}}^{\mathrm{C}}(\mathbf{x}; d)$ and $z_{\mathrm{B}}^{\mathrm{C}}(\mathbf{x}; d)$. In particular, when $z_{\mathrm{H}}^{\mathrm{C}}(\mathbf{x}; d)$ and $z_{\mathrm{B}}^{\mathrm{C}}(\mathbf{x}; d)$ provide a different ordering of the $\mathbf{x}$'s – that is, whenever we can have $z_{\mathrm{H}}^{\mathrm{C}}(\mathbf{x}_1; d) > z_{\mathrm{H}}^{\mathrm{C}}(\mathbf{x}_2; d)$ and simultaneously $z_{\mathrm{B}}^{\mathrm{C}}(\mathbf{x}_1; d) < z_{\mathrm{B}}^{\mathrm{C}}(\mathbf{x}_2; d)$ – then then it is not possible to find two sets of thresholds that lead to the same region in $\mathbf{x}$-space (we will illustrate this point in much more detail in the Results section). Thus, we cannot draw conclusions for one-dimensional sensory evidence, but we can draw strong conclusions for multi-dimensional sensory evidence.

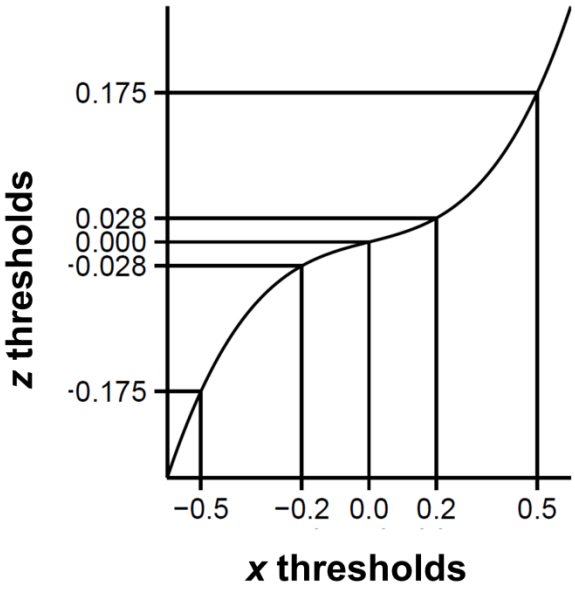


Figure 2-1. For one-dimensional sensory evidence, *x*, any monotonic transformation, *z(x)*, can give the same mapping from *x* to *c*. The best we, as experimenters, can do is to determine the mapping from x to c, which, for discrete mappings, corresponds to a set of thresholds (the vertical lines intersecting the x-axis). We can, however, get the same mapping from x to c by first transforming x to z (the curved black line), then thresholding z. The relevant thresholds are given by passing the x-thresholds through $z(x)$ (giving the horizontal lines intersecting the y-axis). Therefore, there is no way to determine the "right" $z(x)$ – any $z(x)$ will fit the data as long as $z(x)$ is a monotonic function of x.

This difference between one-dimensional and multi-dimensional sensory evidence is one of the key differences between our work and most prior work (see **Section 1.3.2.1: Signal detection theory**). Most static SDT models of confidence assume that the sensory evidence is one-dimensional, leaving them susceptible to the problem described above. In addition, in most dynamic SDT models of confidence – in which the sensory evidence is assumed to accumulate over time – the sensory evidence at a given point in time is also summarised by a single scalar value, making it impossible to tell whether participants' confidence reports reflect heuristic or Bayes optimal computations. However, such dynamic models are able to explain the interplay between confidence reports, accuracy and reaction time – something that we leave for future work.

Here, we used multi-dimensional stimuli in a way that allows us to test whether participants' confidence reports reflect heuristic or Bayes optimal computations. We asked participants to report their confidence in a decision in a psychophysical two-interval forced-choice task. This task allowed us to model the sensory evidence as having two dimensions, with one coming from the first interval and the other from the second interval. We considered three models for how participants computed their confidence variable – all three models were different "static" versions of the popular race model in which reported confidence is assumed to reflect the balance of evidence between two competing accumulators (originally proposed by Vickers, 1979, and more recently used in studies such as Kepecs et al., 2008, and De Martino, Fleming, Garrett, & Dolan, 2012). The first model, the Difference model, assumed that participants' reported confidence reflected the difference in magnitude between the sensory evidence from each interval. The second model, the Max model, assumed that participants' reported confidence reflected only the magnitude of the sensory evidence from the interval selected on a given trial – thus implementing a "winner-take-all" dynamic (Wang, 2008). The third model, the Bayes optimal model, assumed that participants' reported confidence reflected the probability that their decision was correct given the sensory evidence from each interval. We

also tested two different methods for eliciting confidence reports. In the standard two-response design, participants first indicated their decision, and only then, and on a separate scale, indicated their confidence. In the less-commonly used one-response design, participants indicated their decision and confidence simultaneously on a single scale. We were interested to see whether the more complex one-response design – in which participants, in effect, have to perform two tasks at the same time – affected the computations underlying confidence reports as expected under theories of cognitive load (Lavie, 2005; Sweller, 1988) and dual-task interference (Daniel Kahneman, 1973; Pashler, 1994).

2.2 Methods

2.2.1 Participants

Participants were recruited among undergraduate and graduate students (age range: 18-30) at the University of Oxford. In total, 26 participants took part in the study (EXP1: *N* = 11; EXP2: *N* = 15; 25 male). Participants only took part in one experiment. All participants reported normal or corrected vision. All participants provided informed consent and were reimbursed for their participation (£10/hour for a total of 2 hours). All experiments were approved by the local ethics committee.

2.2.2 Experimental details (Experiments 1 and 2)

Participants performed a two-interval forced-choice contrast-discrimination task. We will use the basic visual task for all the experiments conducted for this thesis. The experiments will differ in terms of how a response is given and in terms of the task goal. The display parameters and the stimulus presentation will only be described in detail in this chapter.

2.2.2.1 Display parameters and response device

Participants sat in a dark room. They viewed an LED monitor (ViewSonic VG2236wm-LED) at a distance of about 57 cm. The background luminance was 62.5 cd/m^2 and the resolution was 800 × 600. The monitor was connected to a personal laptop (Toshiba Satellite Pro C660-29W) and controlled by the Cogent toolbox (http://www.vislab.ucl.ac.uk/cogent.php/) for Matlab (Mathworks Inc). Participants responded using a QWERTY keyboard.

2.2.2.2 Stimulus presentation

On each trial, a central black fixation cross (width: 0.75 degrees of visual angle) appeared for a variable period, drawn uniformly from the range 500-1000 milliseconds. Two viewing intervals were then presented, separated by a blank display lasting 1000 milliseconds. Each interval lasted 83 milliseconds. In each interval, there were six vertically oriented Gabor patches (SD of the Gaussian envelope: 0.45 degrees of visual angle; spatial frequency: 1.5 cycles/degree of visual angle; baseline contrast: 0.10) organised in an imaginary circle (radius: 8 degrees of visual angle) at equal distances from each other. In either the first or the second interval, one of the six Gabor patches (the visual target) had a slightly higher level of contrast than the others. The visual target was produced by adding one of 4 possible values (0.015, 0.035, 0.07, 0.15) to the baseline contrast (0.10) of the respective Gabor patch. In total, there were two (target interval) by four (target contrast) by six (target location) stimulus combinations. The stimuli were randomised so that each of the target-interval by target-contrast combinations appeared twice for each set of 16 trials. The target location was randomised across all trials. The second interval was followed by a blank display, which lasted 500 milliseconds, and then a display which prompted participants to make a response.

2.2.2.3 Procedure

After the two viewing intervals, participants were prompted to indicate which interval they thought contained the visual target and how confident they felt about this decision. They performed one of two experiments. The difference between the experiments lay only in how a response was made. The two groups were analysed separately.

In Experiment 1, participants (*N* = 11) were first presented with a central black question mark (**Figure 1-1A**). Participants had to indicate which interval they thought contained the visual target, pressing "N" for the first interval and "M" for the second interval. After having indicated their decision, participants were presented with a central black horizontal line. A vertical white marker was displayed at the left extreme of the horizontal line. Participants indicated how confident they felt about their decision by moving the vertical marker along the line by up to six steps, with each step towards the right indicating higher confidence (1: "uncertain"; 6: "certain").[12] Participants pressed "N" or "M" to move the marker left or right, respectively, and locked the marker by pressing "B". The marker could be moved back and forth until it was locked. After having made their response, participants were presented with central white text which read "correct" if their decision about the target interval was correct or "wrong" if it was incorrect. The feedback display lasted 2000 milliseconds. Lastly, participants were presented with central white text saying "next trial" before continuing to the next trial. Participants completed 16 practice trials followed by 480 experimental trials.

In Experiment 2, participants (*N* = 15) were presented with a central black horizontal line with a fixed midpoint (**Figure 1-1B**). The region to the left of the midpoint represented the first interval; the region to the right represented the second interval. A vertical white marker was displayed on top of the midpoint. Participants were asked to indicate which interval they

[12] The scale was never explicitly labelled. Participants were only informed that their confidence should increase with each step of the scale. We discuss the motivation for this task design in Section 1.3.1.2: Eliciting confidence.

thought contained the visual target by moving the vertical marker to the left (first interval) or to the right (second interval) of the midpoint. The marker could be moved along the line by up to six steps on either side, with each step indicating higher confidence (1: “uncertain”; 6: “certain”). Participants pressed “N” or “M” to move the marker left or right, respectively, and locked the marker by pressing “B”. The marker could be moved back and forth until it was locked. After having made their response, participants were presented with central white text which read “correct” if their decision about the target interval was correct or “wrong” if it was incorrect. The feedback display lasted 2000 milliseconds. Lastly, participants were presented with central white text saying “next trial” before continuing to the next trial. Participants completed 16 practice trials followed by 480 experimental trials.

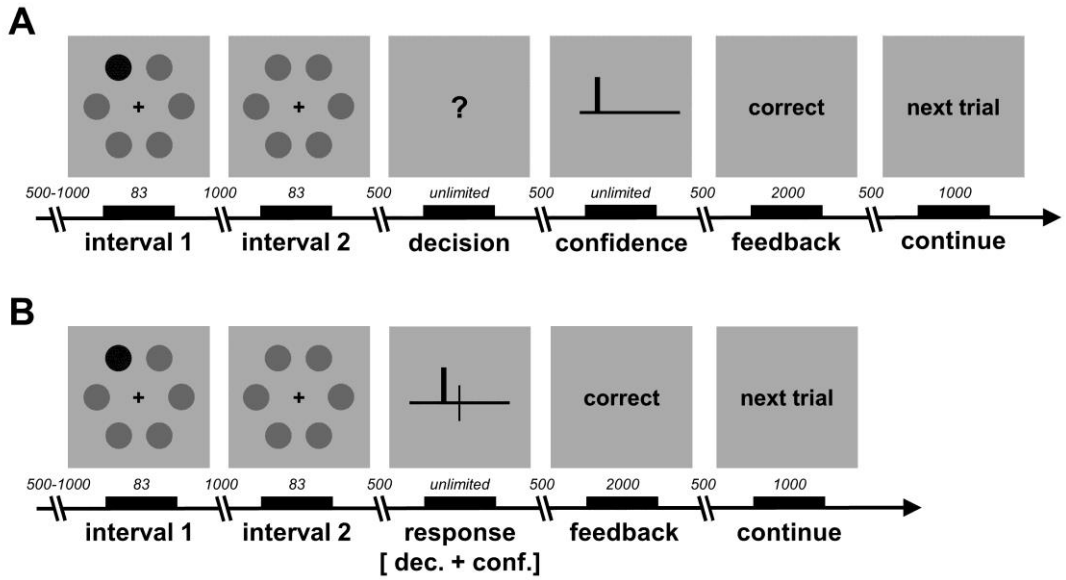


Figure 2-2. Schematic of an experimental trial. **(A)** Experiment 1 ("two response"). Participants viewed on each trial two brief intervals, each containing six Gabor patches (here dots). In either the first or the second interval, one of the six Gabor patches had a higher level of contrast (darker dot). Participants were first prompted to make a decision about which interval they thought contained the visual target. They were then then prompted to indicate how confident they felt about this decision by moving a marker along a line. The marker could be moved along the line by up to six steps, with each step indicating higher confidence. After having made their response, participants received feedback about the accuracy of their decision before continuing to the next trial. Here, the participant selected interval 1 (correct) and reported confidence level 2. **(B)** Experiment 2 ("one response"). Participants performed the same task as in Experiment 1, except that they were prompted to indicate their decision and their confidence at the same time by moving a marker along a scale with a fixed midpoint. The left-hand and the right-hand sides represented interval 1 and interval 2, respectively. The marker could be moved along the line by up to six steps on either side of the midpoint, with each step indicating higher confidence. Here, the participant selected interval 1 (correct) and reported confidence level 2. The displays have been edited for ease of illustration. All timings are shown in milliseconds.

2.2.3 Computational models of confidence reports

To model responses, we assumed the following. On each trial, participants receive two pieces of sensory evidence, $\mathbf{x}$. They transform the sensory evidence into a continuous decision variable, $z^{\mathrm{D}}(\mathbf{x})$, and compare this variable to a single threshold to make a decision, *d*. They transform the sensory evidence and the decision into a continuous confidence variable, $z^{\mathrm{C}}(\mathbf{x}; d)$, and compare this variable to a set of thresholds to obtain a confidence level, c. In this section, we will first describe our assumptions about the sensory evidence, $\mathbf{x}$. We will then detail the models for how participants might generate their decisions and their confidence reports. Lastly, we will describe the Bayesian inference technique that we used to fit the model parameters and find the most probable model.

2.2.3.1 Sensory evidence

We first assumed that a participant on each trial receives two pieces of sensory evidence, $\mathbf{x} = (x_1 , x_2)$, drawn from two different Gaussian distributions, with x_1 giving information about interval 1 and x_2 giving information about interval 2. If the visual target is in interval 1, then

$$P(x_1|s, i = 1, \sigma) = N(x_1; s, \sigma^2/2)$$
$$P(x_2|s, i = 1, \sigma) = N(x_2; 0, \sigma^2/2) \qquad \textbf{Eq. 2-1}$$

Whereas if the visual target is in interval 2, then

$$P(x_1|s, i = 2, \sigma) = N(x_1; 0, \sigma^2/2)$$
$$P(x_2|s, i = 2, \sigma) = N(x_2; s, \sigma^2/2) \qquad \textbf{Eq. 2-2}$$

Here, s specifies the contrast added to the visual target, $s \in \{0.015, 0.035, 0.07, 0.15\}$, i denotes the target interval, $i \in \{1,2\}$, and σ characterises the level of noise in a participant's

sensory system. The variance of each piece of sensory evidence is $\sigma^2/2$, which means that the variance of $x_2 - x_1$ is σ^2 as commonly assumed by psychophysical models.

2.2.3.2 Decision and confidence variables

We considered three models for how a participant computes their decision variable, $z^{\mathrm{D}}(\mathbf{x})$, and their confidence variable, $z^{\mathrm{C}}(\mathbf{x}; d)$. We refer to these models as the Difference model (Δ), the Max model (M), and the Bayesian model (B).

The Difference model proposes that the decision and the confidence variable reflect the difference between the two pieces of sensory evidence,

$$z_{\Delta}^{\mathrm{D}} = x_2 - x_1 \qquad \textbf{Eq. 2-3}$$

$$z_{\Delta}^{\mathrm{C}} = \begin{cases} x_1 - x_2 & \text{for } d = 1 \\ x_2 - x_1 & \text{for } d = 2 \end{cases} \qquad \textbf{Eq. 2-4}$$

Where $d = 1$ when interval 1 is selected and $d = 2$ when interval 2 is selected.

The Max model proposes that the decision variable reflects the difference between the two pieces of sensory evidence and that the confidence variable reflects only the piece of sensory evidence received from the selected interval,

$$z_{\mathrm{M}}^{\mathrm{D}} = x_2 - x_1 \qquad \textbf{Eq. 2-5}$$

$$z_{\mathrm{M}}^{\mathrm{C}} = \mathrm{x}_d \qquad \textbf{Eq. 2-6}$$

Finally, the Bayesian model proposes that the decision variable reflects the probability that interval 2 contained the visual target, and that the confidence variable reflects the probability that the decision about the target interval is correct,

$$z_{\mathrm{B}}^{\mathrm{D}} = P(i = 2 | x_1, x_2, \sigma) \qquad \textbf{Eq. 2-7}$$

$$z_{\mathrm{B}}^{\mathrm{C}} = P(i = d|x_1, x_2, \sigma)$$ **Eq. 2-8**

Where

$$P(i = d|x_1, x_2, \sigma) = \frac{\sum_s P(x_1|s, i = d, \sigma)\, P(x_2|s, i = d, \sigma)}{\sum_{s,i'} P(x_1|s, i = i', \sigma)\, P(x_2|s, i = i', \sigma)}$$ **Eq. 2-9**

To derive this expression, we used Bayes' theorem and assumed that the two conditions have equal prior probability ($P\ (i\ =\ 1)\ =\ P\ (i\ =\ 2)\ =\ 1/2$). The three models make different predictions about how the sensory evidence contributes to the confidence variable, $z^{\mathrm{C}}(\mathbf{x}; d)$, and therefore give rise to different confidence reports – a point which we will return to later.

2.2.3.3 Choosing decisions and confidence reports

To make a decision, the participant compares the decision variable to a single threshold, θ^{D}, and chooses interval 1 if the variable is smaller than the threshold, and interval 2 otherwise,

$$d(\mathbf{x}) = \begin{cases} 1 & \text{if } z^{\mathrm{D}}(\mathrm{x}) < \theta^{\mathrm{D}} \\ 2 & \text{if } z^{\mathrm{D}}(\mathrm{x}) > \theta^{\mathrm{D}} \end{cases}$$ **Eq. 2-10**

Likewise, to choose a confidence level, the participant compares their confidence variable to a set of thresholds, $\theta_{d,c}^{\mathrm{C}}$, and the confidence level is then determined by the pair of thresholds that the confidence variable lies between. More specifically, the mapping from confidence variables to confidence reports is determined implicitly by,

$$\theta_{d,c-1}^{\mathrm{C}} < z^{\mathrm{C}}(\mathbf{x}; d) < \theta_{d,c}^{\mathrm{C}}$$ **Eq. 2-11**

Valid confidence reports, c, run from 1 to 6; to ensure that the whole range of $z^{\mathrm{C}}(\mathbf{x}; d)$ is covered, we set $\theta_{d,0} = -\infty$ and set $\theta_{d,6} = +\infty$.

Lastly, we assumed that with some small probability b, participants lapsed and they made a random response (decision and confidence). Inclusion of this lapse rate accounts for trials in

which participants made a low-probability response (e.g., chose the first interval when they had strong evidence for the second). Such trials are probably due to some error (e.g., motor error or confusing the two intervals), and if we did not include a lapse rate to explain these trials, they could have a strong effect on the model selection.

2.2.4 Model comparison

We wish to compute the probability of each model given our data. The required probability is, via Bayes' theorem,

$$P(m|\text{data}) \propto P(m)\, P(\text{data}|m) \qquad \textbf{Eq. 2-12}$$

Where m is either Δ (Difference model), M (Max model) or B (Bayesian model). The data from participant l consists of two participant-defined variables, the participant's decisions, $\mathbf{d}_l$,and the participant's confidence reports, $\mathbf{c}_l$, and two experimenter-defined variables, the target intervals, $\mathbf{i}_l$, and the target contrasts, $\mathbf{s}_l$. Here, the bold symbols denote a vector, listing the value of a variable on every trial, for instance the target interval on the k'th trial is $i_{l,k}$. We fit different parameters to every participant, so the full likelihood, $P(\text{data}|m)$, is given by a product of single-participant likelihoods,

$$P(\text{data}|m) = \prod_l P(\mathbf{d}_l, \mathbf{c}_l, \mathbf{i}_l, \mathbf{s}_l|m) \qquad \textbf{Eq. 2-13}$$

Because $\mathbf{i}_l$ and $\mathbf{s}_l$ are independent of the model, m, we may write

$$P(\text{data}|m) \propto \prod_l P(\mathbf{d}_l, \mathbf{c}_l|\mathbf{i}_l, \mathbf{s}_l, m) \qquad \textbf{Eq. 2-14}$$

To compute the single-participant likelihood we cannot simply choose one setting for the parameters, because the data does not pin down the exact value of the parameters. Instead we must integrate over possible parameter settings,

$$P(\mathbf{d}_l, \mathbf{c}_l|\mathbf{i}_l, \mathbf{s}_l, m) = \int P(\mathbf{d}_l, \mathbf{c}_l|\mathbf{i}_l, \mathbf{s}_l, m, \boldsymbol{\theta}_l, \sigma_l, b_l)\, P(\boldsymbol{\theta}_l)\, P(\sigma_l)\, P(b_l)\, d\boldsymbol{\theta}_l\, d\sigma_l\, db_l \quad \textbf{Eq. 2-15}$$

Where $\boldsymbol{\theta}_l$ collects that participant's decision and confidence thresholds. This integral optimally combines two important factors. First, as one might expect, this integral is large if the best fitting parameters explain the data well. Second, this integral also takes into account the robustness of the model. In particular, a good model is not overly sensitive to the exact settings of the parameters – so you can perturb the parameters away from the best values, and still fit the data reasonably well.

For a single participant (dropping the participant index, l, for simplicity, but still fitting different parameters for each participant), the probability of $\mathbf{d}$ and $\mathbf{c}$ given that participant's parameters is the product of terms from each trial,

$$P(\mathbf{d}, \mathbf{c}|\mathbf{i}, \mathbf{s}, m, \boldsymbol{\theta}, \sigma, b) = \prod_k P(d_k, c_k|i_k, s_k, m, \boldsymbol{\theta}, \sigma, b) \quad \textbf{Eq. 2-16}$$

We therefore need to compute the probability of a participant making a decision, d_k, and a confidence report, c_k, given their parameters, the target interval, i_k, and the target contrast, s_k. We do this numerically by sampling. First, given a set of parameters, $\boldsymbol{\theta}$, σ and b, we generate an $\mathbf{x}$ from either **Eq. 2-1** or **Eq. 2-2** (depending on whether i_k is 1 or 2). Second, we compute $z^{\mathrm{D}}(\mathbf{x})$ from either **Eq. 2-3**, **Eq. 2-5** or **Eq. 2-7** (depending on the model), and then threshold $z^{\mathrm{D}}(\mathbf{x})$ to generate a decision, d. Lastly, we combine $\mathbf{x}$ and d to compute $z^{\mathrm{C}}(\mathbf{x}; d)$ from either **Eq. 2-4**, **Eq. 2-6** or **Eq. 2-8** (again, depending on the model), and then threshold $z^{\mathrm{C}}(\mathbf{x}; d)$ to generate a confidence report, c. We do this many times (10^5 in our simulations); $P(d_k, c_k|i_k, s_k, m, \boldsymbol{\theta}, \sigma, b)$ is then the proportion of times that the above procedure yields $d = d_k$ and $c = c_k$.

To perform the integral in **Eq. 2-15**, we must specify prior distributions over the parameters σ, b and θ. While it is possible to write down sensible priors over two of these parameters, σ and b, it is much more difficult to write down a sensible prior for the thresholds, $\boldsymbol{\theta}$. This difficulty arises because the thresholds depend on $z^{\mathrm{D}}(\mathbf{x})$ and $z^{\mathrm{C}}(\mathbf{x}; d)$, which change drastically from

model to model. To get around this difficulty, we re-parametrise the thresholds, a procedure which we will describe next.

2.2.4.1 Representation of thresholds

We re-parametrise the decision and confidence thresholds in essentially the same way, but it is helpful to start with the decision threshold as it is simpler to find. We exploit the fact that – for a given model – there is a one-to-one relationship between the threshold, θ^{D}, and the probability that the participant chooses interval 1,

$$P_{d=1} \equiv P(d = 1|m, \boldsymbol{\theta}, \sigma, b) = \int_{-\infty}^{\theta^{\mathrm{D}}} P(z^{\mathrm{D}}|m, \sigma, b)\, dz^{\mathrm{D}} \qquad \textbf{Eq. 2-17}$$

Therefore, if we specify the threshold, we specify $P_{d=1}$. Importantly, the converse is also true: if we specify $P_{d=1}$, we specify the threshold. Thus, we can use $P_{d=1}$ to parametrise the threshold. To compute the threshold from $P_{d=1}$, we represent $P\left(z^{\mathrm{D}}\middle|m, \sigma, b\right)$ using samples of z^{D}, which we can compute as described above. To find the threshold, we sweep across possible values for the threshold, until the right proportion of samples are below the threshold ($P_{d=1}$) and – thus – the right proportion of samples are above the threshold ($P_{d=2}$).

The situation is exactly the same for the confidence thresholds; the only complications are that there are multiple thresholds and that these may depend on the decision. We thus write

$$P_{c|d} \equiv P(c|d, m, \boldsymbol{\theta}, \sigma, b) = \int_{\theta^{C}_{d,c-1}}^{\theta^{C}_{d,c}} P(z^{\mathrm{C}}|d, m, \sigma, b)\, dz^{\mathrm{C}} \qquad \textbf{Eq. 2-18}$$

To compute the thresholds from $P_{c|d}$, we use the same strategy as above and represent $P\left(z^{\mathrm{C}}\middle|d, m, \sigma, b\right)$ using samples of z^{C}. To condition on a particular decision, we simply throw away those values of z^{C} associated with the other decision. As before, we compute the

thresholds by sweeping across all possible values for the thresholds, until we get $c = 1$ the right proportion of the time (i.e., $P_{c=1|d}$), and $c = 2$ the right proportion of the time (i.e., $P_{c=2|d}$), and so on for a given decision (this procedure is illustrated in **Figure 2-3**).

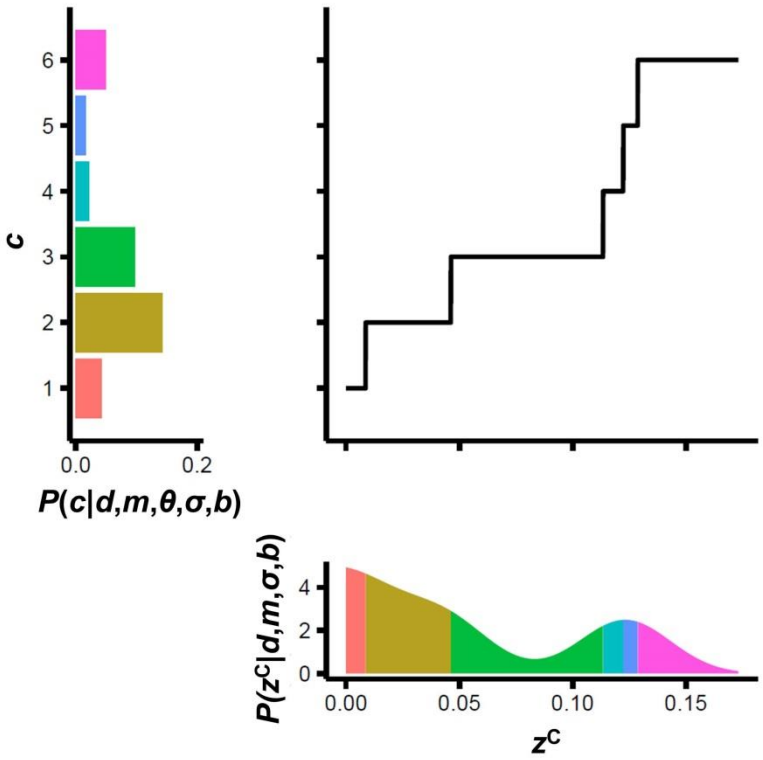


Figure 2-3. Schematic diagram of our method for mapping thresholds to confidence probabilities. The lower panel displays the distribution over the confidence variable, $P(z^C|d,m,\sigma,b)$, which does not depend on the confidence thresholds. The left panel displays the distribution over confidence reports determined by the confidence thresholds, $P(c|d,m,\boldsymbol{\theta},\sigma,b)$. The large central panel displays the fitted function mapping z^C onto c, which consists of a set of jumps, with each jump corresponding to a threshold. The thresholds are chosen so that the total probability density in $P(z^C|d,m,\sigma,b)$ between jumps is exactly equal to the probability of the corresponding confidence level (see colours).

Examining **Eq. 2-17** and **Eq. 2-18**, we see that the decision and the confidence thresholds are together fully parametrised by the joint distribution over decisions and confidence reports. We therefore specify the decision and the confidence thresholds using the joint distribution over decisions and confidence reports,

$$P_{d,c} \equiv P(d, c|m, \boldsymbol{\theta}, \sigma, b) \qquad \textbf{Eq. 2-19}$$

2.2.4.2 Computing the single-participant likelihood

Changing the representation from thresholds, $\boldsymbol{\theta}$, to probabilities, $\mathbf{p}$, allows us to rewrite the single-participant likelihood (cf. **Eq. 2-15**),

$$P(\mathbf{d}, \mathbf{c}|\mathbf{i}, \mathbf{s}, m) = \int P(\mathbf{d}, \mathbf{c}|\mathbf{i}, \mathbf{s}, m, \mathbf{p}, \sigma, b)\, P(\mathbf{p})\, P(\sigma)\, P(b)\, d\mathbf{p}\, d\sigma\, db \qquad \textbf{Eq. 2-20}$$

To perform the integral, we need to specify prior distributions over the parameters, σ, b, and $\mathbf{p}$. For σ, we use

$$P(\sigma) = \mathrm{Gamma}(2, 0.05) \propto \sigma \mathrm{e}^{-\sigma/0.05} \qquad \textbf{Eq. 2-21}$$

As this distribution broadly covers the range of plausible values for σ. We chose a very broad prior distribution for b, with values evenly distributed in log space between 10^{-3} and 10^{-1},

$$P(\log_{10} b) = \mathrm{Uniform}(-3, -1) \qquad \textbf{Eq. 2-22}$$

Finally, we chose an uninformative, uniform prior distribution over $\mathbf{p}$,

$$P(\mathbf{p}) = \mathrm{Dirichlet}(\mathbf{p}; \mathbf{1}), \qquad \textbf{Eq. 2-23}$$

Where $\mathbf{1}$ is a matrix whose elements are all 1.

The most straightforward way to compute the single-participant likelihood in **Eq. 2-20** is to find the average (expected) value of $P(\mathbf{d}, \mathbf{c}|\mathbf{i}, \mathbf{s}, m, \mathbf{p}, \sigma, b)$ when we sample values of $\mathbf{p}$, σ and b from the prior,

$$P(\text{data}|m) = \mathrm{E}_{P(\mathbf{p})P(\sigma)P(b)}[P(\mathbf{d}, \mathbf{c}|\mathbf{i}, \mathbf{s}, m, \mathbf{p}, \sigma, b)] \quad \textbf{Eq. 2-24}$$

However, the likelihood, $P(\mathbf{d}, \mathbf{c}|\mathbf{i}, \mathbf{s}, m, \mathbf{p}, \sigma, b)$ is very sharply peaked; being very high in a very small region around the participant's true parameters, and very low elsewhere. The estimated value of the integral is therefore dominated by the few samples that are close to the true parameters, and as there are only a few such samples, the sample-based estimate of $P(\mathbf{d}, \mathbf{c}|\mathbf{i}, \mathbf{s}, m, \mathbf{p}, \sigma, b)$ has high variance. Instead, we use a technique called importance sampling. The aim is to find an equivalent expectation, in which the quantity to be averaged does not vary much, allowing the distribution to be estimated using a smaller number of samples – in fact, if the term inside the expectation is constant, then the expectation can be estimated using only one sample. Importance sampling uses

$$P(\text{data}|m) = \mathrm{E}_{P(\mathbf{p})P(\sigma)P(b)}[\frac{P(\mathbf{d}, \mathbf{c}|\mathbf{i}, \mathbf{s}, m, \mathbf{p}, \sigma, b)\ P(\mathbf{p})}{Q(\mathbf{p})}] \quad \textbf{Eq. 2-25}$$

The integral form for this expectation is

$$P(\mathbf{d}, \mathbf{c}|\mathbf{i}, \mathbf{s}, m) = \int \frac{P(\mathbf{d}, \mathbf{c}|\mathbf{i}, \mathbf{s}, \mathbf{p}, \sigma, b)\ P(\mathbf{p})}{Q(\mathbf{p})}\ Q(\mathbf{p})\ P(\sigma)\ P(b)\ d\mathbf{p}\ d\sigma\ db \quad \textbf{Eq. 2-26}$$

Which is trivially equal to **Eq. 2-20**. To ensure that the term inside the expectation in **Eq. 2-25** does not vary much, we need to choose the denominator, $Q(p)$, so that it is approximately proportional to the numerator, $P(\mathbf{d}, \mathbf{c}|\mathbf{i}, \mathbf{s}, \mathbf{p}, \sigma, b)\ P(\mathbf{p})$. To do so, we exploit the fact that the numerator is proportional to a posterior distribution over $\mathbf{p}$ (considering only dependence on $\mathbf{p}$),

$$P(\mathbf{d}, \mathbf{c}|\mathbf{i}, \mathbf{s}, \mathbf{p}, \sigma, b)\, P(\mathbf{p}) \propto P(\mathbf{p}|, \mathbf{d}, \mathbf{c}, \mathbf{i}, \mathbf{s}, m, \sigma, b) \quad \textbf{Eq. 2-27}$$

Remembering that $P_{d,c}$ is just the probability of a particular decision and confidence report, aggregating across all trial types, it is straightforward to construct a good approximation to the posterior over $\mathbf{p}$. In particular, we ignore the influence of $\mathbf{i}$, $\mathbf{s}$, m, σ and b, so the only remaining information is decisions and confidence reports, $\mathbf{d}$ and $\mathbf{c}$, irrespective of trial-type. These variables can be summarised by $\mathbf{n}$, where $n_{d,c}$, is the number of times that decision d and confidence level c were made. The resulting distribution over $\mathbf{p}$ can be written,

$$Q(\mathbf{p}) = P(\mathbf{p}|, \mathbf{d}, \mathbf{c}) = \text{Dirichlet}(\mathbf{p}, \mathbf{1} + \mathbf{n}) \quad \textbf{Eq. 2-28}$$

This turns out to be a good proposal distribution for our importance sampler.

2.3 Results

2.3.1 Model selection

To compare the models, we want to compute the posterior probability of each of the models given the data, $P(\text{model}|\text{data})$. As we used a uniform prior over the models (we had no reason to prefer one model over another a priori), the posterior probability is proportional to the likelihood $P(\text{data}|\text{model})$, which we showed how to compute in **Section 2.2.4: Model comparison**. The Bayesian model is better by a factor of around 10^{25} for Experiment 1 and around 10^4 for Experiment 2 (**Figure 2-4**).

Here, we assumed that all participants used the same model to generate their responses. However, different participants might use different models. In particular, we might expect that there is some probability with which a random participant uses each model. Under this assumption, we can analyse how well the models fitted the data by inferring the probability with which participants use each model, using a variational Bayesian method presented by

Stephan, Penny, Daunizeau, Moran & Friston (2009). In agreement with the previous analysis, for Experiment 1, participants were more likely to use the Bayesian model than the Max model (**Figure 2-5A**). For the Experiment 2, on the other hand, participants were only slightly more likely to use the Bayesian model than the Max model (**Figure 2-5A**). The log-likelihood differences for individual participants tell the same story (**Figure 2-5B**). In all cases, the Difference model was the least plausible.

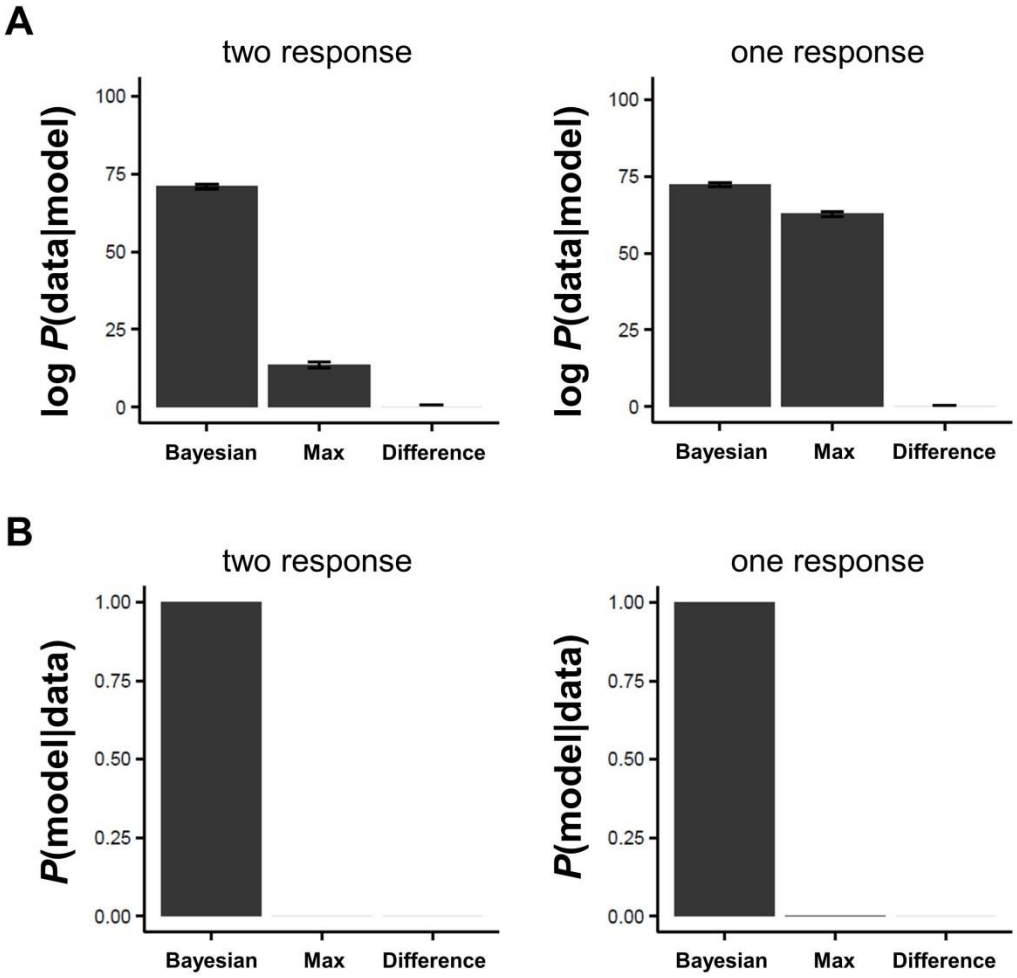


Figure 2-4. The probability of the three models given the data. **(A)** The log-likelihood differences between models – using the Difference model as baseline – for Experiment 1 (left) and Experiment 2 (right). Note the small error bars, representing two standard-errors, given by running the algorithm 10 times, and each time using 1000 samples to estimate the model evidence (as per Eq. 2-25). **(B)** The posterior probability of the models for Experiment 1 (left) and Experiment 2 (right), assuming a uniform prior.

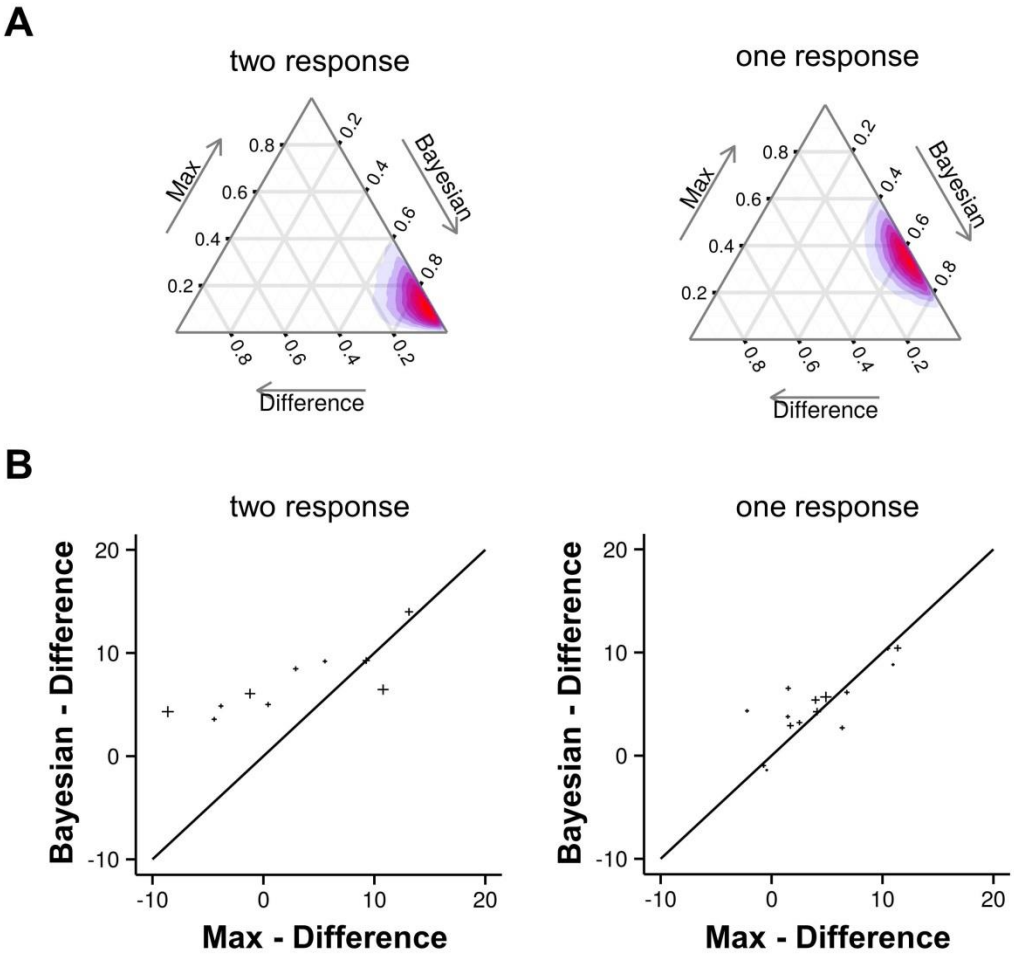


Figure 2-5. Single-participant analysis. **(A)** Probability of using each model. Participants are assumed to use each model with some probability. The coloured regions represent plausible settings for these probabilities. For Experiment 1 (left), participants were far more likely to use the Bayesian model. In contrast, for Experiment 2 (right), participants were roughly equally likely to use the Max and the Bayesian model. To read these plots, follow the grid lines in the same direction as axis ticks and labels, so for instance, lines of equal probability for the Max model run horizontally, and lines of equal probability for the Bayesian model run up and to the right. **(B)** Differences in log-likelihood of models. The difference in log-likelihood between the Max model and Difference model (x-axis) against the difference in log-likelihood between the Bayesian model and the Difference model (y-axis). For Experiment 1 (left), participants were far more likely to use the Bayesian model. In contrast, for Experiment 2 (right), participants were roughly equally likely to use the Max and Bayesian models. The size of the crosses represents the uncertainty (two standard errors) along each axis (based on the 10 runs of the model selection procedure, mentioned in Figure 2-4).

2.3.2 Model fits

While the model evidence is the right way to compare models, it is important to check that the inferred models and parameter settings are plausible. We therefore plotted the raw data — the number of times a participant reported a particular decision and confidence level for a particular target interval and target contrast — along with the predictions from the Bayesian model.

In **Figure 2-6**, we plotted for an example participant the fitted and the empirical distributions over confidence reports given a target interval and a target contrast. To make this comparison, we defined "signed confidence", whose absolute value gives the confidence level, and whose sign gives the decision,

$$\text{signed confidence} = \begin{cases} -c & \text{for } d = 1 \\ +c & \text{for } d = 2 \end{cases} \qquad \textbf{Eq. 2-29}$$

In **Figure 2-7**, we plotted for an example participant the fitted (Bayesian model) and the empirical distributions over decisions given a target interval and a target contrast. Again, to make this comparison, we defined "signed contrast", whose absolute value gives the target contrast, and whose sign gives the target interval,

$$\text{signed contrast} = \begin{cases} -s & \text{for } i = 1 \\ +s & \text{for } i = 2 \end{cases} \qquad \textbf{Eq. 2-30}$$

These plots show that the Bayesian model is, at least, plausible, and highlights the fact that our model selection procedure is able to find extremely subtle differences between models.

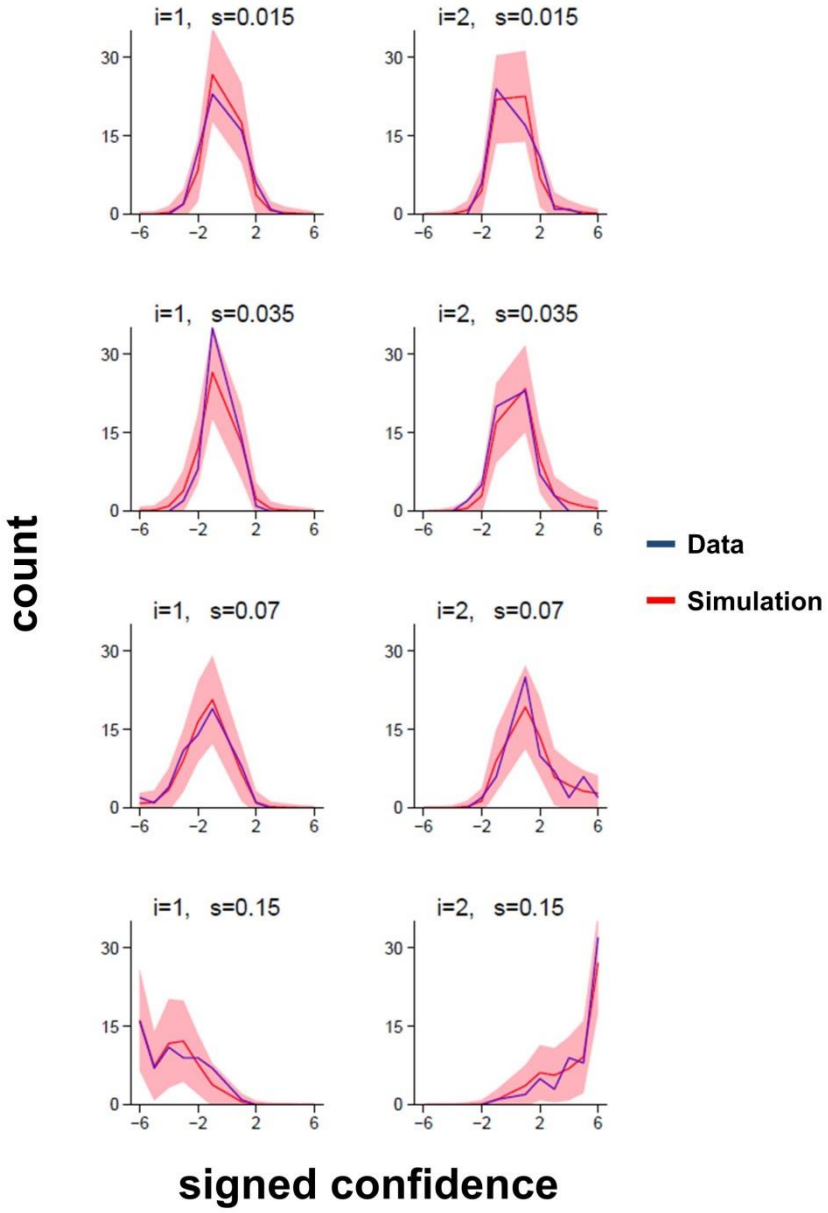


Figure 2-6. Simulated (Bayesian model) and empirical signed confidence distributions for a participant from Experiment 1. The plots on the left are for visual targets in interval 1 (i.e., $i = 1$); the plots on the right are for visual targets in interval 2 (i.e., $i = 2$). We show signed confidence on the x-axis (sign: decision; absolute value: confidence level) and the counts for each value on the y-axis. The blue line is the empirically observed distribution. The red line is the fitted mean distribution for the Bayesian model. The red shading (error bars) is the region that we would expect the data to lie within. We computed the error bars by sampling settings for the model parameters, then sampling datasets conditioned on those parameters. The error bars represent two standard deviations of those samples. This plot demonstrates that the Bayesian model is, at least, plausible.

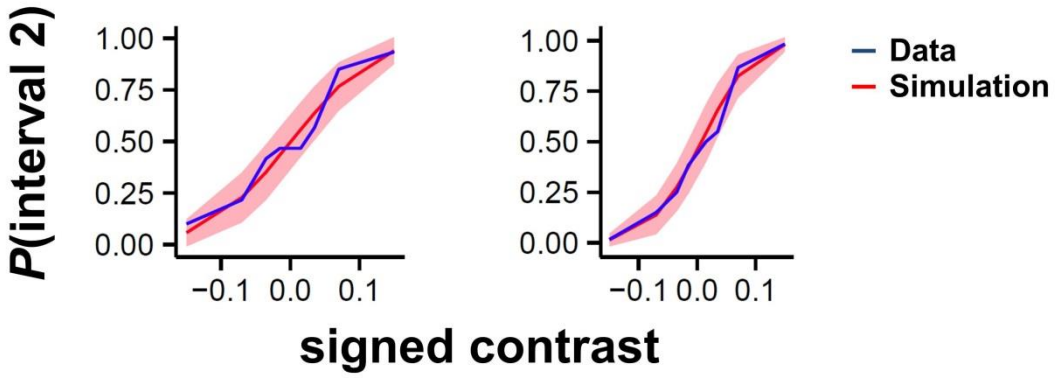


Figure 2-7. Simulated (Bayesian model) and empirical psychometric curves for two participants from Experiment 1. We show signed contrast on the x-axis (sign: target interval; absolute value: contrast level) and the probability of selecting interval 2 on the y-axis. The blue line is the empirically measured psychometric curve; the red line is the fitted mean psychometric curve for the Bayesian model; and the red area represents error bars. See Figure 2-6 for simulation details. As with Figure 2-6, this plot demonstrates the plausibility of the Bayesian model.

2.3.3 Differences between models

For our model selection to work, there need to be differences between the predictions made by the three models. Here, we show that the models do indeed make different predictions under representative parameter settings.

To understand which predictions are most relevant, we have to think about exactly what form our data takes. In our experiments, we present participants with a visual target in one of the two intervals, i, with one of four contrast levels, s, then observe their decision, d, and confidence report, c. Overall, we therefore obtain an empirical estimate of each participant's distribution over decisions and confidence reports (i.e., the signed confidence distribution) given a target interval and a target contrast. This suggests that we should examine the predictions that each model makes about this distribution. While the predicted distributions shown in **Figure 2-8** are superficially similar, closer examination reveals two interesting, albeit small, differences; we emphasise that these plots display theoretical, and hence noise-free results, so even small differences are meaningful, and are not fluctuations due to noise.

First, the Max model (blue) differs from the other models at intermediate contrast levels, especially $s = 0.07$, where the Max model displays bimodality in the signed confidence distribution. In particular, and unexpectedly, an error with confidence level 1 is less likely than an error with confidence levels 2 to 4. In contrast, the other models display smooth, unimodal behaviour across the different confidence levels. This pattern arises because the Max model uses only one of the two sensory signals. For example, when $i = 2$ and $s = 0.07$ (i.e., the visual target is fairly obvious in interval 2), then x_2 is usually large. Therefore, for x_1 to be larger than x_2, giving rise to an error, x_1 must also be large. Under the Max model, x_1 being large usually implies high confidence, and, in this case, a high confidence error.

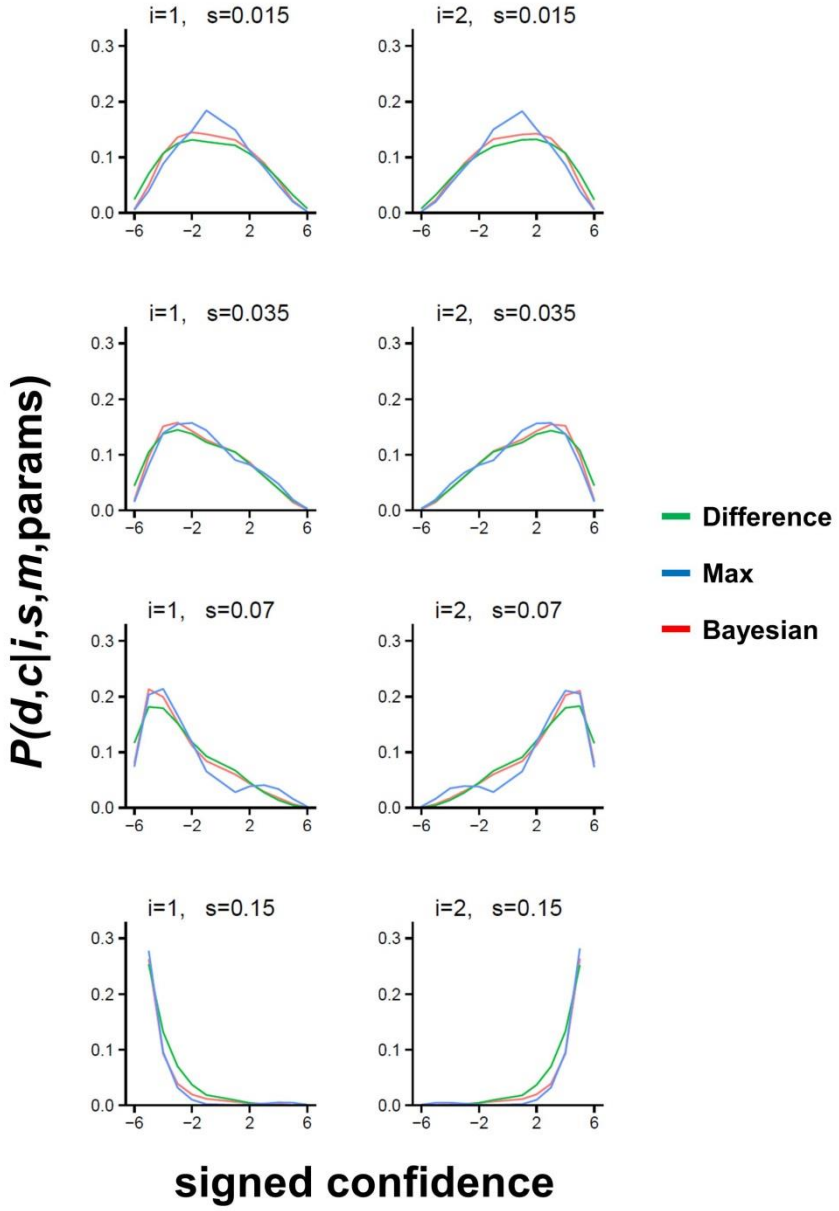


Figure 2-8. The models make different predictions about signed confidence distributions. The figure should be read as Figure 2-6. Note, however, that the model parameters were not fit to the data. Instead, they were set to fixed (but reasonable) values: $\sigma = 0.07$, $b = 0$ and $p_{d,c} = 1/12$. The models are identified by colours (green: Difference; blue: Max; red: Bayesian).

Second, the three models exist on a continuum, with the Max model using the narrowest range of confidence, the Bayesian model (red) using an intermediate range, and the Difference model (green) using the broadest range. These trends are evident at the lowest and highest contrast levels. At the lowest contrast level, $s = 0.015$, the distribution for the Max model is more peaked, whereas the distribution for the Difference model is broader, and the Bayesian model lies somewhere between them. At the highest contrast level, $s = 0.15$, the Max model decays most rapidly, followed by the Bayesian model, and then the Difference model.

To understand this apparent continuum, we need to look at how the models map sensory evidence, defined by x_1 and x_2 , onto confidence reports. We therefore traced out in **Figure 2-9** the regions of sensory space (i.e., (x_1, x_2)-space) that map to different confidence levels (black contours). This exercise highlights striking differences between the models. In particular, the Difference model has diagonal contours, whereas the Max model has contours that run horizontally, vertically or along the central diagonal at $x_1 = x_2$. In further contrast, the Bayesian model has curved contours with a shape somewhere between the Difference and the Max model. In particular, for large values of x_1 and x_2, the contours are almost diagonal, as in the Difference model. In contrast, for small values of x_1 and x_2, the contours are much more horizontally or vertically aligned, as in the Max model.

To understand how these differences translate into differences in the probability distribution over signed confidence, we consider the red dots, representing the mean values of x_1 and x_2 for a given target interval and target contrast. For instance, the mean values of x_1 and x_2 when $i = 2$ and $s = 0.15$ (i.e., a high-contrast target in interval 2) are represented by the uppermost red dot in each subplot. Importantly, the red dots lie along the horizontal and vertical axes (green lines). The angle at which the contours cross these axes therefore becomes critically important. In particular, for the Difference model, the contours pass diagonally through the axes, and therefore close to many red dots, giving a relatively broad

range of confidence levels for each target interval and target contrast. In contrast, for the Max model, the contours pass perpendicularly through the axes, minimizing the number of red dots that each contour passes close to, giving a narrower range of confidence levels for each target interval and target contrast. The contours of the Bayesian model pass through the axes at an angle between the extremes of the Difference and the Max model — as expected, giving rise to a range of confidence levels between the extremes of the Difference and the Max model.

In principle, these differences might allow us to choose between models based only on visual inspection of $P(d, c|i, s, m, \text{params})$, where "params" contains the model parameters. In practice, the distribution over decisions and confidence reports averaging over trial type, $p_{d,c}$, is not constant, as we assumed above in the theoretical simulations, but is far more complicated. This additional complexity makes it impossible to find the correct model by simple visual inspection. More powerful methods, like Bayesian model selection, are needed to pick out the differences between the models.

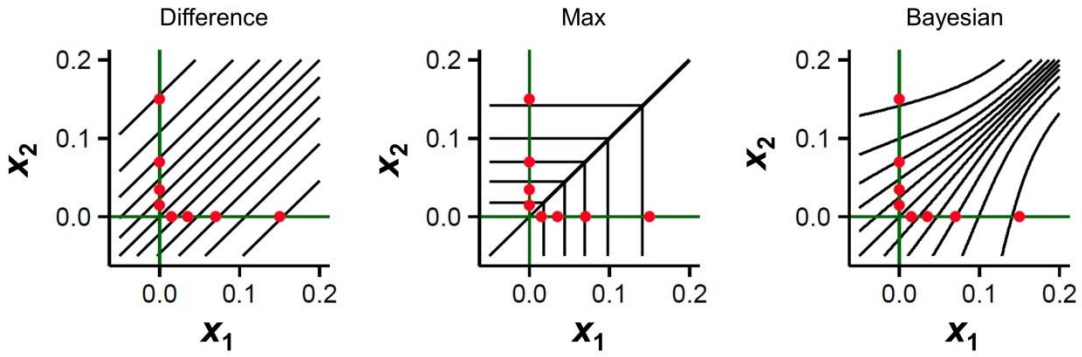


Figure 2-9. The mapping from sensory space to confidence reports induced by the different models. The axes represent the two sensory dimensions, x_1 and x_2 (cf. interval 1 and 2). The red dots represent the mean values of x_1 and x_2 for each stimulus (these values are in the limit given by the target interval and the target contrast). The black lines separate the regions in sensory space that map to a given confidence level. The model parameters are the same as in Figure 2-8.

2.4 Discussion

Using visual psychophysics we tested whether participants' confidence reports reflected heuristic or Bayes optimal computations. We assumed that participants on each trial receive sensory evidence, $\mathbf{x} = (x_1, x_2)$, and, based on that sensory evidence, generate a decision, d, and a confidence report, c. We assumed that this process is mediated by two intermediate variables: participants first transform the sensory evidence into a continuous decision variable, $z^{\mathrm{D}}(\mathbf{x})$, and then compare this variable to a single threshold to make a decision, d; participants next transform the sensory evidence and the decision into a continuous confidence variable, $z^{\mathrm{C}}(\mathbf{x}, d)$, and compare this variable to a set of thresholds to make a confidence report, c. We compared three possible ways of computing the confidence variable, $z^{\mathrm{C}}(\mathbf{x}, d)$: the Difference model, which computes the difference between the two pieces of sensory evidence; the Max model, which only uses the piece of sensory evidence supporting the selected decision; and the Bayesian model, which computes the probability that a correct decision has been made given both pieces of sensory evidence.

We used Bayesian model selection to directly compare these models. In the more standard two-response design, where subjects indicate their decision and their confidence separately (EXP1), the Bayesian model emerged as the clear winner. However, in the less standard one-response design, where subjects indicate their decision and their confidence at the same time (EXP2), the results were ambiguous — our data indicated that around half of the participants favoured the Bayesian model while the other half favoured the Max model. One possible reason for this difference is that, in the one-response design, the computations underlying confidence reports were simplified so as not to interfere with the computation of the decision, as expected under theories of cognitive load (Lavie, 2005; Sweller, 1988) and dual-task interference (Daniel Kahneman, 1973; Pashler, 1994). Alternatively, despite the instructions being the same, the two types of task design might promote different computations

(interpretations), with the one-response design promoting a judgement about the target contrast, whereas the two-response design promotes a judgement about the correctness of a decision which — perhaps critically — has already been made. Surprisingly, the commonly used Difference model was the least probable model in both types of task design.

A caveat in any model comparison is that we cannot test all possible heuristic computations. However, given our results, it seems that our models range across the continuum of sensible models – though it is certainly possible that, perhaps, the best model sits somewhere between the Bayesian and the Max models. More generally, our results indicate that subtle changes in a task can lead to changes in the computations performed – and whether participants use Bayes optimal computations. We note, however, that the differences between the models in terms of predicted behaviour were small (see **Supplementary Figure 1** and **Supplementary Figure 2** for fitted and empirical distributions over signed confidence for each participant). As such, while showing that our approach can detect subtle differences, the absence of large differences in predicted behaviour indicates that other models can be used to describe confidence reports when the exact nature of the confidence variable is not of paramount importance.

2.4.1 Relationship to other studies on the optimality of confidence

As discussed in **Chapter 1**, the studies by Barthelmé & Mamassian (2009) and de Gardelle & Mamassian (2014) went part-way towards realising the potential of multidimensional stimuli (see **Section 1.2.3.3: Do humans compute confidence variables in an optimal manner?**). In the study by Barthelmé & Mamassian (2009), participants had to indicate which of two Gabor patches they would prefer to make an orientation judgement about. In contrast to our results, their modelling slightly favoured a heuristic strategy (similar to our Max model) over the Bayes optimal strategy. In the study by de Gardelle & Mamassian (2014), participants had to indicate

on which of two trials they felt more likely to be correct. Interestingly, participants could accurately identify on which trial they were more likely to be correct regardless of whether the two trials involved the same or two different types of task (orientation discrimination versus spatial-frequency discrimination). This result provides some, albeit indirect, evidence that people compute their confidence variable in a Bayes optimal manner, in agreement with our work. However, there are two reasons why these studies are likely to be less relevant to the question of whether confidence reports reflect Bayes optimal computations. First, our model selection procedure is fully Bayesian, and therefore takes into account the uncertainty associated with model predictions, whereas their procedure was not. In particular, under some circumstances a model will make strong predictions (e.g., "the agent must make this decision"), whereas under other circumstances, the model might make weaker predictions (e.g., "the agent is most likely to make this decision, but I'm not sure – they could also do other things"). Bayesian model selection takes into account the strength or weakness of a model prediction. Second, in real life, and in our study, people tend to report their confidence using verbal expressions (e.g., "unsure", "pretty sure" and "sure") or numerical expressions (e.g. "60%"). This additional complexity may have affect how an underlying confidence variable is computed – making studies using forced-choices less relevant to the question of whether confidence reports reflect Bayes optimal computations.

2.4.2 Two types of optimality

Many studies have asked whether *confidence* is optimal. However, our work suggests that this is a not a sensible question because there are at least two types of optimality. First, the transformation of incoming evidence into an *implicit* confidence variable, $z^{\mathrm{C}}(\mathbf{x}, d)$, could be optimal. In most cases, an optimal confidence variable would be computed using Bayesian

inference. Second, the mapping of the confidence variable onto an *explicit* confidence report, $c(z^{\mathrm{C}}(\mathbf{x}, d))$, could be optimal. This time, however, optimality depends entirely on the task.

To illustrate this issue – which we raised in **Chapter 1** – we consider the *calibration* measure that is often used to evaluate confidence reports in the judgement and decision-making literature (Harvey, 1997). To recap, in a task designed to assess calibration, participants are asked to report their confidence in a decision as a probability. Participants are then said to be *well-calibrated* (i.e., optimal) when the *reported* probability on average reflects the *objective* probability of being correct; for instance, that participants are correct 70% of the time when they report that they feel "70%" confident (**Figure 1-4**). It turns out, however, that well-calibrated does not imply optimality at all levels.

To see why this is the case, consider a participant who chooses randomly, and knows that they choose randomly. If they always say that their confidence is 50%, they would be said to be well-calibrated. However, they are throwing away all the task-relevant information, so their internal computations (at least, those that give rise to this response) are clearly sub-optimal. In fact, it is not necessary to go to such an extreme case. The participant can use a moderately sub-optimal heuristic to compute their confidence variable and then adjust the mapping of this confidence variable onto a confidence report so that their reports match the objective probability of being correct. This participant would be said to be well-calibrated, but their confidence variable would not make the best use of the sensory evidence. Finally, a participant may use Bayesian inference to compute their confidence variable, but be poorly calibrated in the sense that their confidence report does not accurately scale with this confidence variable.

To illustrate this key point, we plotted one of our participant's confidence reports against the *objective* probability that they were correct (**Figure 2-10A**). If we were to reinterpret our confidence scale as a probability scale, then this participant would appear poorly calibrated. In particular, the participant rarely reported high confidence when they were almost certainly

correct. Importantly, we can find also find the range of *subjective* probabilities of being correct (computed using the Bayesian model) that were mapped onto a given confidence level and then plot these against the *objective* probability of being correct for that confidence level (**Figure 2-10B**). Of course, we showed in this chapter that the participant is indeed likely to have computed their subjective probabilities using Bayesian inference. Therefore, their subjective probabilities should, and indeed do, appear well-calibrated. The same pattern emerges when we pool the data across all participants (**Figure 2-10AB**). Admittedly, this exercise tells us nothing about the results of an experiment in which participants reported their confidence as a probability. However, it tells us that it is possible to have participants that appear under- or overconfident, despite basing their confidence reports on Bayes optimal computations. Similar concerns apply to any quantitative measure of metacognition that does not separately model, on one hand, the transformation of sensory evidence into an *implicit* confidence variable, and on the other hand, the mapping of this confidence variable onto an *explicit* confidence report.

While it often is possible to specify when a confidence variable is optimal, it is much harder to specify when the mapping of this confidence variable onto a confidence report is optimal. Without some incentive structure, there is no reason why participants should opt for any particular mapping, as long as their confidence reports increase strictly with their confidence variables. However, even when there is an incentive structure, other factors still seem to be influencing the mapping of a confidence variable onto a confidence report; participants improve their calibration but never reach perfection (Fleming & Dolan, 2010; Zylberberg, Roelfsema, & Sigman, 2014). However, as discussed in **Chapter 1**, confidence reports are meant not for *oneself* but for *others*. As such, it would seem much more appropriate to study the optimality of confidence reports in their natural habitat – that is, social interaction – a task which we will take on in **Chapter 4**.

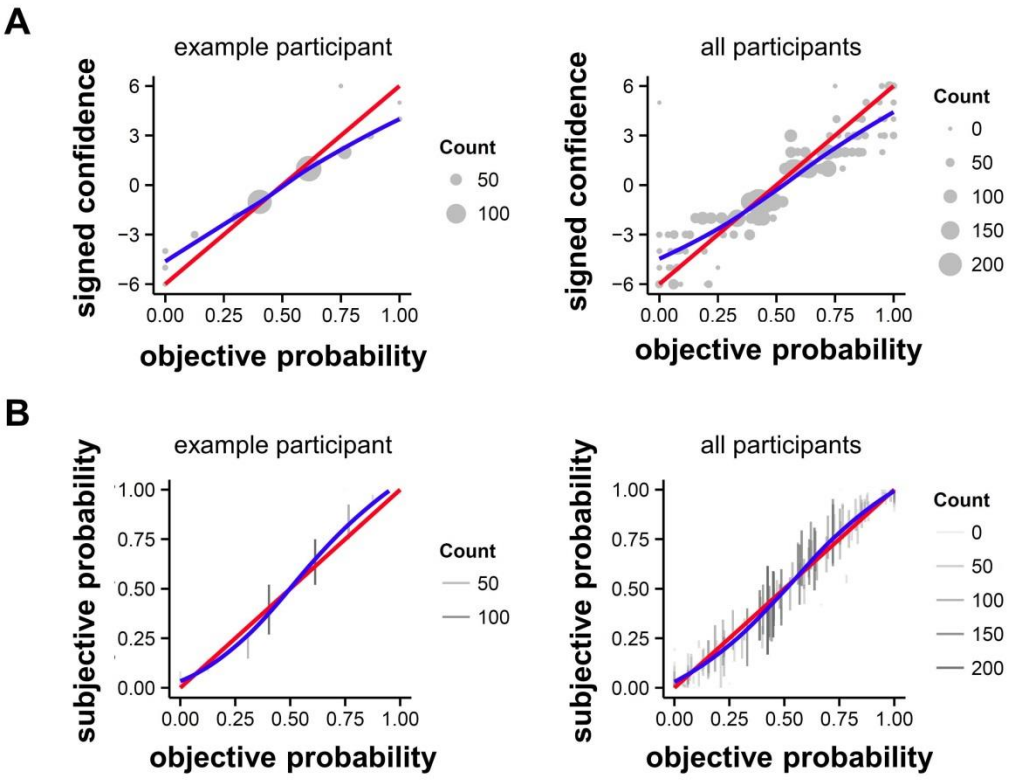


Figure 2-10. Calibration curves. **(A)** Each (grey) dot represents the probability that the visual target was in interval 2 (x-axis) given a particular confidence level (y-axis) for a single participant (left) and for all participants (right). Notice that we here used signed confidence as defined in Eq. 2-29. The size of the dots represents the number of times that the participant used a particular confidence level. The blue line is a smooth fit to the dots. The red line is a straight (identity) line running from (0,-6) to (1,6). If the example participant used the confidence scale as efficiently as possible, then we might expect the blue line to lie on top of the red line. In reality, however, this participant mostly used $c = 1$ and $c = -1$ (two large dots in the centre of the plot). **(B)** Each (grey) line represents the range of subjective probabilities (y-axis) mapped to a specific level of confidence for a single participant (left) and for all participants (right). These ranges cannot overlap, and must range from 0 to 1, so we see the ranges going up in a stepwise fashion. The subjective probabilities shown on the y-axis were calculated using the Bayesian model. Again, the x-axis shows the probability that the visual target was in interval 2 given a particular confidence level. The blue line is a smoothed representation of the line segments, and the red line is the identity line, representing a perfect match between subjective and objective probabilities. The blue and red lines match very closely, indicating that the subjective probabilities that participants mapped onto a particular confidence level were effectively the same as the objective probabilities associated with that confidence level. Notice that each legend only contains reference dots or lines.

2.4.3 Two levels of variability

Our work suggests that individual variation in confidence could be thought of as having at least two dimensions. The first, deeper, dimension relates to the transformation of incoming evidence into an *implicit* confidence variable, $z^{\mathrm{C}}(\mathbf{x}, d)$. In our data, there do indeed appear to be individual differences in how participants computed this variable – an effect which might relate to general intelligence. In addition, we found that subtle changes to the response mode could affect how (some) participants computed a confidence variable. We might therefore expect shifts in how people compute such a variable for tasks of different complexity. For example, it is not straightforward to solve general-knowledge questions, such as "what is the capital of Brazil?", using Bayesian inference. While one could in principle compute the probability that one's answer is correct, the computational load may be so high that people resort to heuristic computations (e.g., using the population size of the reported city).

The second, more superficial, dimension relates to the mapping of a confidence variable onto an *explicit* confidence report, $c(z^{\mathrm{C}}(\mathbf{x}, d))$ – with this mapping determining the average reported confidence and the distribution over reported confidence. In our data, we found considerable individual variation in how people reported their confidence (see **Supplementary Figure 1** and **Supplementary Figure 2**). We might imagine that this dimension is largely modulated by contextual factors. For example, people's reported confidence has been shown to vary with mental health (Huq, Garety, & Hemsley, 1988), personality (Campbell, Goodie, & Foster, 2004) and social role, such as profession (Broihanne et al., 2014), gender (Barber & Odean, 2001; Niederle & Vesterlund, 2011) and culture (Mann et al., 1998).

2.5 Chapter summary

In this chapter, we developed a framework for studying and modelling confidence reports – a framework which will use in our empirical work – and addressed the fundamental component

processes involved in sharing and combining representations of uncertainty (**Figure 2-11**). We tested whether confidence reports reflect confidence variables that are computed in a Bayes optimal or a heuristic (a reasonable but ultimately arbitrary) manner. When the decision and its associated confidence report were indicated separately, the confidence reports of participants overwhelmingly reflected Bayes optimal computations. However, when the decision and its associated confidence report were indicated at the same time, the confidence reports of half of the participants reflected Bayes optimal computations, whereas the confidence reports of the other half of participants reflected heuristic computations.

More generally, we argued that confidence should be thought of as having two dimensions and thus involve two types of optimality. The first, deeper, dimension relates to the computation of a confidence variable. The second, more superficial, dimension relates to the mapping of a confidence variable onto a confidence report – with this mapping determining the average reported confidence and the distribution over reported confidence. In our data, there was considerable individual variation along the second dimension. As discussed in **Chapter 1**, in the face of such individual variation, flexibility in the report of confidence is important for effective cooperation – that is, to minimise miscommunication, interacting individuals must be able to adapt to each other's use of the scale on which confidence is mutually expressed and thus establish a common metric for identifying who is more likely to be correct in a given situation. In the next chapter, we will test whether participants in the current dataset show evidence of such flexibility.

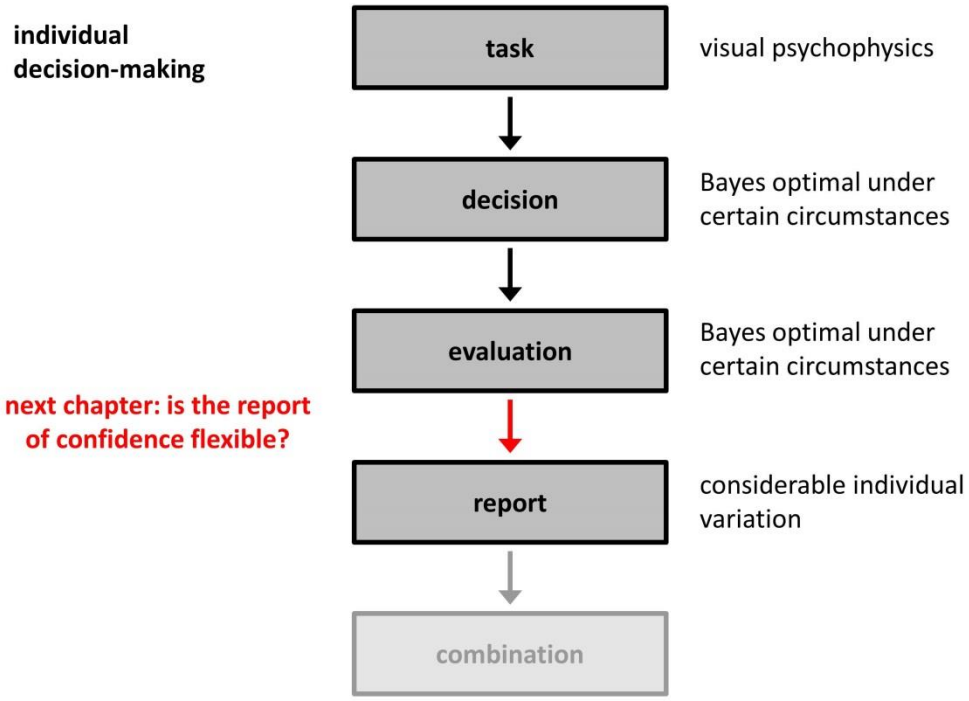


Figure 2-11. Summary of Chapter 2. The boxes denote the component processes involved in sharing and combining representations of uncertainty.

Chapter 3: Confidence reports are influenced by the trial history

Current models assume that reported confidence in a decision only reflects the strength of the evidence in favour of the decision. Trial-by-trial variation in confidence reports for a task with constant difficulty has thus been attributed to random variation (noise) in the encoding of the decision evidence. In this chapter, we tested an alternative hypothesis: that trial-by-trial variation in confidence reports is in part driven by internal factors which are sensitive to the trial history. In support of this hypothesis, combining the dataset from Chapter 2 with a dataset from a previously published study, we observed that the report of confidence varied in a state-dependent manner that was divorced from its first-order input. More precisely, the level of confidence reported on a given trial was biased toward the level of confidence reported on the previous trial, and was higher when the decision made on the previous trial was correct, independent of the stimulus, the decision, the reaction time and the motor response required on the current and on the previous trial. These history effects pose a problem for current theories and measures of metacognition which assume that the report of confidence should be stable over time. We argue that global inconsistency in the report of confidence can result from a series of locally consistent or locally informative confidence reports.

3.1 Introduction

In psychophysical tasks, reported confidence in a decision can vary substantially from one trial to another, even when the intensity of the sensory stimulus is held constant (Baranski & Petrusic, 1998; Douglas Vickers, 1979). However, the factors that contribute to this variation remain poorly understood. In the SDT framework, it is assumed that the confidence reported on a given trial only reflects the strength of the sensory evidence in favour of the decision. As such, trial-by-trial variation in confidence reports has been attributed to random variation (noise) in the encoding of the sensory evidence. Drawing on previous research on trial-by-trial effects in perceptual decision-making, we tested an alternative hypothesis: that trial-by-trial variation in confidence reports is in part driven by internal factors which are sensitive to the trial history and which bias the generation of a confidence report on a given trial. We were in particular interested in whether the report of confidence changes over and above its first-order input – a level of flexibility which is important for effective cooperation (Fusaroli et al., 2012; Shea et al., 2014).

We will use the general SDT model introduced in **Chapter 2** to illustrate the potential sources of trial-by-trial variation in confidence reports. To recap, the participant receives on each trial sensory evidence, x.[13] The participant computes a decision variable, $z^{\mathrm{D}}(x)$, and compares this variable to a single threshold, θ^{D}, to generate a decision, d – we write this mapping function as $d(z^{\mathrm{D}})$. The participant then computes a confidence variable, $z^{\mathrm{C}}(x; d)$, an internal representation of the strength of the sensory evidence in favour of the decision, d, and compares this variable to a set of thresholds, $\boldsymbol{\theta}^{\mathrm{C}}$, to generate a confidence report, c – we write this mapping function as $c(z^{\mathrm{C}})$.

In the SDT framework, trial-by-trial variation in confidence reports is typically assumed to be the consequence of variation in the sensory evidence, x. The sensory evidence is defined as some stimulus (signal) corrupted by noise. All else being equal, stronger stimuli will on average give rise to stronger confidence variables, $z^{\mathrm{C}}(x; d)$. However, even when the stimulus is held constant, sensory noise will cause the magnitude of the confidence variable to vary from one trial to another. In dynamic SDT models of confidence, sampling time (reaction time) governs the relative contribution of the stimulus and sensory noise to the sensory evidence – longer sampling time will generally average out the contribution of noise – and may thus also cause the magnitude of the confidence variable to vary from one trial to another.

More recently, it has been proposed that trial-by-trial variation in confidence reports may also result from additional noise (metacognitive noise) during the mapping of the confidence variable onto a confidence report, $c(z^{\mathrm{C}})$, with people reporting different levels of confidence for confidence variables of similar magnitudes (De Martino et al., 2012). This metacognitive noise could reflect variability in the decoding of the confidence variable ("read-out noise") or an inability to maintain a stable mapping function over time ("noisy thresholds") (Maniscalco

[13] The dimensionality of the sensory evidence is not important here; we therefore, for simplicity, assume that the sensory evidence is one-dimensional and write x as is common in SDT models.

& Lau, 2012). Regardless, the resulting trial-by-trial variation in confidence reports should be non-systematic (i.e., independent of the trial history). However, in contrast to these earlier accounts, research on perceptual decision-making suggests that trial-by-trial variation in confidence reports is in fact partly systematic. In particular, the confidence reported on a given trial might be biased by two types of history effect: first-order history effects, which are due to trial-by-trial effects in the generation of the decision, and second-order history effects, which are due to trial-by-trial effects in the generation of the confidence report and which are independent of trial-by-trial effects at the level of the decision. We will consider the two types of history effect in turn.

In psychophysical tasks, the decision made on a given trial is often found to be biased by the decision made on the previous trial (Akaishi, Umeda, Nagase, & Sakai, 2014; Fischer, Shankey, & Whitney, 2011; Fründ, Wichmann, & Macke, 2014; Green, 1964; Liberman, Fischer, & Whitney, 2014; Schüür, Tam, & Maloney, 2013; Senders & Sowards, 1953; Verplanck, Collier, & Cotton, 1952; Yu & Cohen, 2009). The nature of this so-called serial dependence depends on the task. For example, Fründ et al. (2014) observed a negative serial dependence (repulsion) mainly in forced-choice tasks, where participants have to decide which of two alternatives is the target, and a positive serial dependence (attraction) mainly in yes-no tasks, where participants have to decide whether a target (or a lure) was present. This discrepant pattern of results suggests that the serial dependence in decisions is driven by task-specific expectations held by the participant. In the SDT framework, such expectations may operate at the level of the computation of the decision variable, $z^{\mathrm{D}}(x)$, and/or the mapping of the decision variable onto a decision, $d(z^{\mathrm{D}})$. As we will explain below, trial-by-trial effects at either level would bias the generation of a confidence report on a given trial.

To illustrate how expectations might bias the computation of the decision variable, consider the two-interval forced-choice task introduced in **Chapter 2**. For a Bayes optimal agent, the

decision variable on a given trial is the posterior probability that interval 2 contains the visual target after – as per Bayes' theorem – combining the sensory evidence (likelihood) with expectations (prior). As such, a strong expectation that there is a trial-by-trial pattern to the probability that interval 2 contains the visual target would – all else being equal – bias the decision variable and thus give rise to a serial dependence in decisions: if the agent believes that the target interval is likely to alternate from one trial to another (i.e., 2-1-2-1-2-etc.), then their decisions would tend to alternate from one trial to another; in contrast, if the agent believes that the target interval is likely to stay the same from one trial to another (i.e., 2-2-2-1-1-etc.), then their decisions would tend to stay the same from one trial to another.[14] We note, however, that when the sensory evidence is particularly strong, expectations would be averaged out and have a small influence on the computation of the decision variable.

To illustrate how expectations might influence the threshold governing the mapping of the decision variable onto a decision, consider a yes-no task in which participants have to decide whether a target was present. In their trial-by-trial SDT model, Treisman & Williams (1984) distinguish between two processes which may bias the decision threshold (i.e., "no target" if $z^{\mathrm{D}}(x) < \theta^{\mathrm{D}}$ and "target" if $z^{\mathrm{D}}(x) > \theta^{\mathrm{D}}$). First, an agent may seek to maximise the consistency of their decisions. In particular, if the target base rate is expected to be stable, then it may be beneficial for an agent to adjust the decision threshold such that the same decision is made on consecutive trials. To do so, the decision threshold must be shifted *away from* the decision variable computed on recent trials, which, all else being equal, would give rise to a positive serial dependence in decisions (**Figure 3-1**). Theoretically, this strategy increases the robustness of a given decision to variation in the decision variable due to sensory noise, and it may serve to maintain some desired reward rate. Second, an agent may seek to maximise the information carried by their decisions. Again, if the target base rate is expected to be stable,

[14] The Bayes optimal model described in Chapter 2 assumed that the target occurred in interval 1 or 2 with equal probability; we therefore did not discuss the role of priors.

then it may be beneficial for an agent to adjust the decision threshold relative to the current flux of sensory evidence. To do so, the decision threshold must be shifted *toward* the decision variables computed on recent trials, which, all else being equal, would give rise to a negative serial dependence in decisions (**Figure 3-1**). Theoretically, this strategy maximises the (Fisher) information carried by decisions about fluctuations in the decision variable. The two strategies are believed to operate over and above the computation of the decision variable and should therefore be thought of as behavioural strategies.

In the SDT framework, trial-by-trial effects in the computation of the decision variable and/or the mapping of the decision variable onto a decision should "indirectly" bias the generation of a confidence report on a given trial. First, the computation of the decision variable, $z^{\mathrm{D}}(x)$, and the computation of the confidence variable, $z^{\mathrm{C}}(x, d)$ are related; expectations affect not only the current belief about the state of the world (decision variable) but also the strength of this belief (confidence variable). As such, all else being equal, strong expectations would give rise to a serial dependence in confidence reports. Second, the confidence variable reflects the strength of the evidence in favour of the *selected* decision and is thus affected by behavioural strategies which operate on the decision threshold. To see why, consider a Bayes optimal agent who performs a yes-no task in which the target appears on 50% of the trials. For some complex reason they gradually shift the decision threshold towards negative infinity; while the probability of a "target" decision would increase, the probability that this response is correct would decrease. Again, all else being equal, this adjustment to the decision threshold would give rise to a serial dependence in confidence reports.

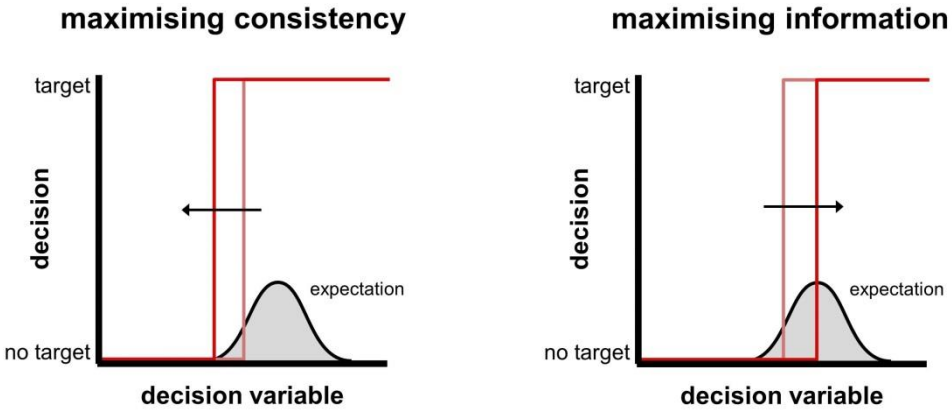


Figure 3-1. Different strategies for placing the decision threshold. An agent may move their decision threshold (*left*) away from expected decision variables (cf. recently computed) so as to maximise the consistency of their decisions in the face of sensory noise or (*right*) towards expected decision variables (cf. recently computed) so as to be maximally sensitive to fluctuations in the decision variable.

Here, we considered another, and for this thesis more relevant, hypothesis about systematic trial-by-trial variation in confidence reports: that participants directly adjust the thresholds governing the mapping of the confidence variable onto a confidence report. What internal factors might drive such second-order adjustments over and above the first-order input? As in the case of decisions, participants might seek to maximise the consistency of their confidence reports in the face of variation in the confidence variable due to sensory noise, or to maximise the information carried by their confidence reports about fluctuations in the confidence variable – making confidence reports an "efficient code" (Barlow, 1961). Alternatively, and more in line with the role of confidence reports in social interaction, participants might seek to calibrate their confidence reports so that their current level of confidence accurately reflects their current level of performance (Ferrell & McGoey, 1980; Yeung & Summerfield, 2012). Without feedback, recent high (low) confidence variables can be taken as an index of recent good (poor) performance and thus be used to adjust the confidence thresholds accordingly for the subsequent trial. Further, with feedback, recent outcomes offer additional information about whether a set of confidence thresholds are too biased toward low or high confidence. This calibration may be subserved by simple reinforcement learning (Sutton & Barto, 1998). On error trials, a difference between the lowest level of confidence and the reported level of confidence (negative prediction error) would indicate that the next report should be less biased toward high confidence. In contrast, on correct trials, a difference between the highest level of confidence and the reported level of confidence (positive prediction error) would indicate that the next report should be more biased toward high confidence.

To our knowledge, only one study has sought to dissociate the different potential sources of systematic trial-by-trial variation in confidence reports. Treisman & Faulkner (1984) observed a serial dependence in confidence reports in a yes-no task, but concluded that this effect was entirely driven by adjustments to the decision threshold. However, their conclusion might be premature. In particular, subtle adjustments to the confidence thresholds might not have

been detected as participants indicated their responses on a limited scale – from “sure no”, through “possibly no” and “possibly yes”, to “sure yes” – with only two levels of confidence. In addition, participants might have focused exclusively on optimal placement of the decision threshold because of the arbitrariness of the response scale (i.e., what does “possibly” and “sure” mean in terms of probability correct?) as well as the absence of incentives for accurate confidence reports. It therefore remains unknown whether participants can or do adjust the confidence thresholds on a trial-by-trial basis. We addressed this question using the dataset described in **Chapter 2** and a dataset provided by Moran et al. (2015; their Experiment 3).

We sought in particular to test whether the confidence reported on a given trial was biased by past reports and/or feedback obtained – such history effects would be consistent with the hypothesis that participants continuously calibrate the thresholds governing the mapping of the confidence variable onto a confidence report. To this end, we constructed a trial-by-trial regression model which allowed us to control for classic sources of trial-by-trial variation in confidence reports, such as the stimulus, sensory noise and reaction time, and trial-by-trial effects in the generation of a decision, at the level of the decision variable and at the level of the decision threshold. The datasets included also allowed us to control for a set of potential confounds. In the first experiment, participants indicated their decision and their confidence level separately – with the latter indicated by moving a marker along a scale with six steps (1: “uncertain”; 6: “certain”). This design is often used in confidence research, but it introduces a motor confound; participants might report the same level of confidence on two consecutive trials simply because they have a tendency to repeat previous motor movements. However, in the second experiment, participants indicated their decision and their confidence at the same time on a scale from -6 to -1 and 1 to 6 – with the sign indicating the decision and the absolute value indicating the confidence level. In this task design, different motor movements were often required to indicate the same confidence level (e.g., -2 versus +2). While arbitrary scales without incentives for accurate confidence reports are often used in confidence research, both

factors might dampen adjustments to the thresholds governing the mapping of the confidence variable onto a confidence report. To control for this confound, we included a third experiment in which participants indicated their confidence on a probability scale (i.e., from 50% to 100% in steps of 10%) and received a monetary reward for an accurate report.

3.2 Methods

3.2.1 Datasets

Experiments 1 and 2 are based on the dataset from **Chapter 2** (also Experiments 1 and 2; see **Section 2.2.2: Experimental details (Experiments 1 and 2)** for details). As the number of participants in the two experiments were not the same (EXP1: *N* = 11; EXP2: *N* = 15), we recruited six participants for the current study so as to make it easier to compare effect sizes across the two experiments (EXP1: *N* = 16; EXP2: *N* = 16; 26 male). Experiment 3 is based on a dataset provided by Moran et al. (2015, their Experiment 3; EXP3: *N* = 9; information about gender not available). Participants only took part in one experiment. All participants reported normal or corrected vision. All participants provided informed consent. All experiments were approved by the local ethics committee. In Experiments 1 and 2, participants were reimbursed for their participation (£10/hour for a total of 2 hours). In Experiment 3, participants received course credits and a performance-based monetary bonus (mean bonus: 35NIS ≈ £7).

3.2.2 Experimental details (Experiments 1 to 3)

3.2.2.1 Experiment 1

Participants performed the same task as in Experiment 1 of **Chapter 2** where the decision and the confidence report were made separately (**Figure 3-2A**). After 16 practice trials, participants completed 480 experimental trials.

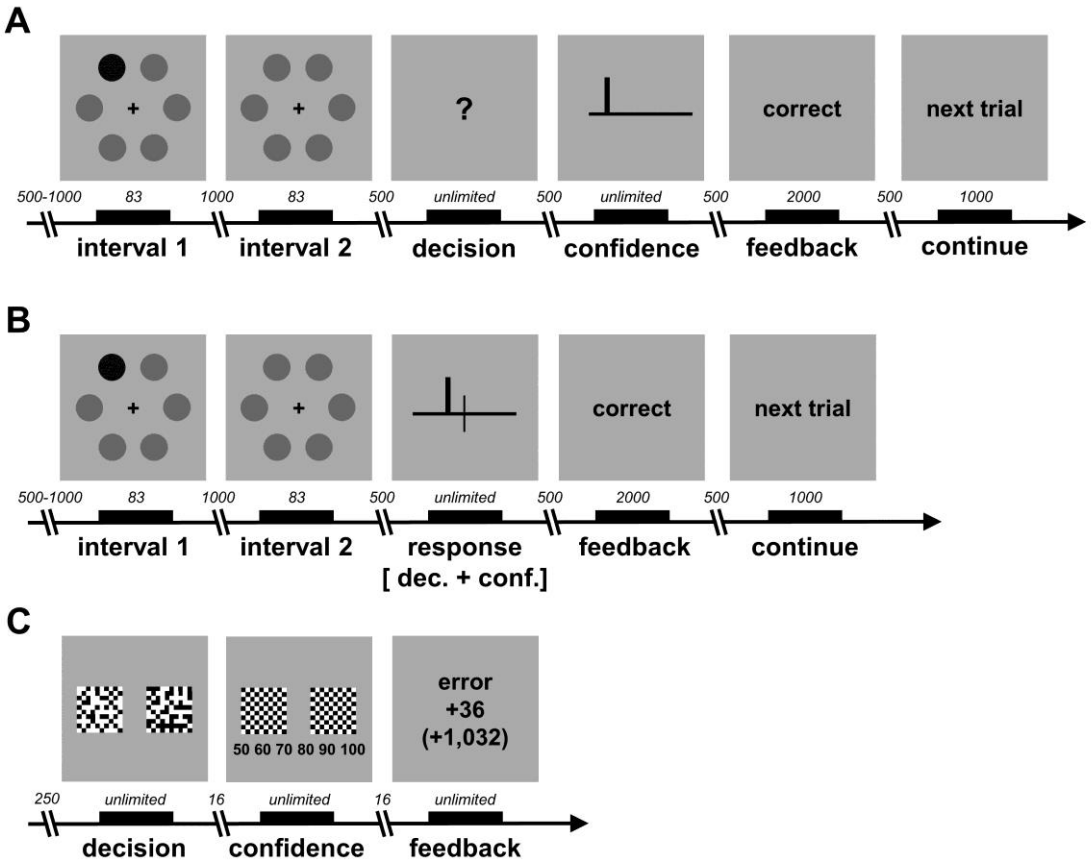


Figure 3-2. Schematic of an experimental trial. **(A)** Experiment 1 ("two response"). Participants viewed on each trial two brief intervals, each containing six Gabor patches (here dots). In either the first or the second interval, one of the six Gabor patches had a higher level of contrast (darker dot). Participants were first prompted to make a decision about which interval they thought contained the visual target. They were then then prompted to indicate how confident they felt about this decision by moving a marker along a line. The marker could be moved along the line by up to six steps, with each step indicating higher confidence. After having made their response, participants received feedback about the accuracy of their decision before continuing to the next trial. Here, the participant selected interval 1 (correct) and reported confidence level 2. **(B)** Experiment 2 ("one response"). Participants performed the same task as in Experiment 1, except that they were prompted to indicate their decision and their confidence at the same time by moving a marker along a scale with a fixed midpoint. The left-hand and the right-hand sides represented interval 1 and 2, respectively. The marker could be moved along the line by up to six steps on either side of the midpoint, with each step indicating higher confidence. Here, the participant selected interval 1 (correct) and reported confidence level 2. **(C)** Experiment 3. On each trial, participants viewed two square black-and-white grids. In the reference grid, the ratio of black-to-white squares was 50:50. In the target grid, the majority of the squares were black. Participants could at any point indicate which grid they thought was the visual target. After having made their decision, the two grids were masked and a probability scale appeared, prompting the participants to indicate how likely their decision was to be correct (from "50%" to "100%" in steps of "10%). After having made their response, participants received feedback. In a speed condition, they were informed whether their decision was "too slow" (> 750 milliseconds). In an accuracy condition, they were informed whether their decision was an "error". In both types of condition, they were informed about the points earned in the current trial and the points accumulated in the current block (see text for details about scoring). Here, the participant – who was in the accuracy condition – selected the left grid (incorrect) and reported confidence "80%". The participant accrued 36 points in the current trial and had accumulated 1,032 points thus far in the current block. The displays have been edited for ease of illustration. All timings are shown in milliseconds.

3.2.2.2 Experiment 2

Participants performed the same task as in Experiment 2 of **Chapter 2** where the decision and the confidence report were made at the same time (**Figure 3-2B**). After 16 practice trials, participants completed 480 experimental trials.

3.2.2.3 Experiment 3

Participants performed a two-interval forced-choice luminance-discrimination task (**Figure 3-2C**). On each trial, two black-and-white grids were presented with a horizontal offset from the centre of the screen. Each grid was made up of 65 by 65 small squares. In the reference grid, half of the squares were black and the other half were white (50:50). In the target grid, the majority of the squares were black, with the proportion drawn from one of three ratios (≈51:49, ≈52:48, ≈53:47). The location of the target grid (left or right to the centre of the screen) and the ratio of black-to-white squares were randomised for each block of 72 trials. The arrangement of the squares within each grid was randomised across all trials.

After stimulus onset, participants were free to indicate which grid they thought contained more black squares (see details about response mode below). After they had made their decision, the two grids were masked (black-and-white checkerboards), and a probability scale (from “50%” to “100%” in steps of “10%”) appeared below the masked grids, prompting the participants to indicate how likely their decision was to be correct (see details about response mode below). Note that, in contrast to Experiments 1 and 2, participants did not indicate their confidence on an arbitrary scale. After having indicated their confidence, participants received feedback (see details about feedback below) before continuing to the next trial.

Participants indicated their decision and their confidence on a QWERTY keyboard. Throughout, they placed their index, middle, medicinal and little fingers of the left hand on the keys “G”,

"F", "D" and "S". They placed their corresponding right-hand fingers on the keys "H", "J", "K" and "L". Both thumbs were placed on the spacebar key. Participants indicated their decision by pressing "S" (left-side grid) or "L" (right-side grid). Participants indicated their confidence by pressing one of the keys located between "D" to "K". If they had selected the left-side grid, then "D" corresponded to "50%" and "K" corresponded to "100%". If they had selected the right-side grid, then "K" corresponded to "50%" and "D" corresponded to "100%". The displayed probability scale increased from left to right or from right to left according to the decision. Participants received training with the input mode in a separate input-entry task.

For each block of 72 trials, either decision speed or decision accuracy was emphasised. In the speed blocks, participants were instructed to indicate their decisions as quickly as possible. If they took longer than 750 milliseconds to indicate their decision, then the feedback display would read "too slow". In the accuracy blocks, participants were instructed to make as many accurate decisions as possible. If they made an incorrect decision, then the feedback display would read "error". In both conditions, participants were instructed to report accurately how likely their decision was to be correct. The two types of condition alternated.

On each trial, participants accrued points, g, according to a quadratic scoring rule, $g = 100 * (1 - (a - c)^2)$, where a indicates the accuracy of the decision ($a = 0$ if incorrect; $a = 1$ if correct) and c indicates the confidence level ($c \in \{.5, .6, .7, .8, .9, 1\}$). The scoring rule is strictly proper: participants would maximise their earnings if they maximised the accuracy of both their decisions and their confidence reports (i.e., participants must report their confidence so that the proportion of trials assigned a given confidence level matches the proportion of times that the decisions made on those trials were correct). Participants were informed about the scoring rule and shown a table that demonstrated why they would fare best if they maximised the accuracy of both their decisions and their confidence reports. In both conditions, the feedback display always informed participants of the points accrued in the current trial as well

as the points accumulated in the current block. In the speed blocks, the points accrued on the current trial were halved if the decision was slower than the 750-milliseconds deadline. Note that, in the speed blocks, participants could most of the time infer whether their decision was accurate (if $g > .75$, then correct; if $g < .75$, then incorrect; however, if $g = .75$, then the feedback is ambiguous). For every 1,600 points, participants earned 1NIS (≈ £0.2). There were 1,152 experimental trials.

3.2.3 Analysis

3.2.3.1 Regression

3.2.3.1.1 Trial-by-trial regression model of confidence reports

To test whether confidence reports are subject to history effects, we created a trial-by-trial regression model in which we sought to predict the confidence report, c, reported on a given trial, t, using a set of predictors meant to reflect the encoding of the sensory evidence and internal factors which are sensitive to the trial history (**Table 3-1**).

Predictor	Description
k_t	Stimulus on current trial.[1] This term relates to the sensory signal.
a_t	Accuracy on current trial (-.5: incorrect; .5: correct). This term relates to sensory noise.
i_t	Target interval on current trial (-.5: interval 1; .5: interval 2).
$RT1_t$	Time taken to indicate decision on current trial.[2] Only EXP1 and EXP3.
$RT2_t$	Time taken to indicate confidence report on current trial.
k_{t-1}	Stimulus on previous trial. See k_t.
a_{t-1}	Accuracy on previous trial. See a_t.
i_{t-1}	Target interval on previous trial. See i_t.
$RT1_{t-1}$	Time taken to indicate decision on previous trial. See $RT1_t$.
$RT2_{t-1}$	Time taken to indicate confidence report on previous trial. See $RT2_t$.
$i_t * d_{t-1}$	Congruency between current interval and previous decision (-.5: incongruent, .5: congruent).
$d_t * d_{t-1}$	Congruency between current and previous decision (-.5: incongruent, .5: congruent).
c_{t-1}	Confidence report on previous trial.
$c_{t-1} * a_{t-1}$	Interaction between confidence report and accuracy on previous trial.
s_t	Condition on current trial (-.5: speed stress, .5: accuracy stress).[3] Only EXP3.

Table 3-1. Trial-by-trial regression model of confidence reports. We sought to predict confidence report, c, on a given trial, t, using the predictors shown in the table. See text for details. [1]EXP1 and EXP2: the contrast level of the visual target (values: .015, .035, .07, .15). EXP3: the number of black dots added to the target array (values: 48, 81, 113). [2]Time is entered in seconds. [3]Every block of 72 trials involved the same condition.

To capture classic sources of trial-by-trial variation in confidence reports, we first predicted that the current confidence report (i.e., c_t) would depend on the strength of the current stimulus (i.e., k_t) as well as the current accuracy (i.e., a_t) – with the two terms capturing the relative contribution of the stimulus and sensory noise to the sensory evidence. Further, we predicted that the current confidence report would depend on the time taken to indicate the current decision (i.e., $RT1_t$; only Experiments 1 and 3) and the current confidence report (i.e., $RT2_t$) – with the two terms capturing modulation of the relative contribution of the stimulus and sensory noise to the sensory evidence due to sampling time (stimulus or information held in working memory). To capture asymmetry in visual sensitivity to visual targets in interval 1 versus 2, we included a term indicating the target interval (i.e., i_t). To capture any low-level trial-by-trial effects, we also included the data from the previous trial for each of the above variables (i.e., k_{t-1}, a_{t-1}, i_{t-1}, $RT1_{t-1}$ and $RT2_{t-1}$).

To capture trial-by-trial effects in the generation of a decision, we first predicted that the current confidence report would depend on the congruency between the current target interval and the previous decision (i.e., $i_t * d_{t-1}$) – with this term capturing prior expectations about which interval contains the visual target. In particular, if participants expect the target interval to alternate from one trial to another, then their confidence reports should tend to be higher when the current target interval and the previous decision are incongruent (i.e., the fitted effect for the term is negative). In contrast, if participants expect the target interval to stay the same from one trial to another, then their confidence reports should tend to be higher when the current target interval and the previous decision are congruent (i.e., the fitted effect for the term is positive). Further, we predicted that the current confidence report would depend on the congruency between the current and the previous decision (i.e., $d_t * d_{t-1}$) – with this term capturing any adjustments to the decision threshold.

To test for trial-by-trial effects in the generation of a confidence report, we first predicted that the current confidence report would depend on the previous confidence report (i.e., c_{t-1}). Further, that the current confidence report would depend on the previous accuracy (i.e., a_{t-1}) – or, at least, that the influence of the previous confidence report would be modulated by the previous accuracy (i.e., $c_{t-1} * a_{t-1}$).

Lastly, to control for the design of Experiment 3, we included the condition on the current trial (i.e., s_t; either emphasis on speed or accuracy). We note that we re-coded the confidence levels in Experiment 3 to the range 1 to 6 in steps of 1 for consistency across the experiments (cf. the scale originally ranged from "50%" to "100%" in steps of "10%").

We note that, for variables which have data included from both the current and the previous trial, the effect of the data on the current trial effect will reflect any serial dependence in a given variable. To illustrate this point, consider accuracy, a. The inclusion of both a_t on a_{t-1} means that the fitted coefficient for a_t will reflect not only the effect of a_t but also the effect of a_{t-1} on a_t; participants might, for example, pay more attention after an incorrect decision and thus have stronger sensory evidence on a given trial. Importantly, this relationship – which falls out of the regression approach – means that the fitted coefficient for a_{t-1} will reflect an effect over and above any serial dependence between a_t and a_{t-1}.

3.2.3.1.2 Ordered probit model

Given the discrete nature of our confidence scales, a Linear Regression Model (LRM) is not appropriate as a trial-by-trial regression model of confidence reports. For example, an LRM can predict values that are outside the range of a discrete scale (e.g., smaller than 1) or that cannot be realised (e.g., 1.5). Further, an LRM cannot capture non-linearity in the use of a discrete scale (e.g., a participant might have a bias toward reporting 1 or treat the difference

between 1 and 2 as subjectively larger than the difference between 5 and 6). We therefore used an Ordered Regression Model (ORM), which only predicts observable values and which allows for non-linearity in the use of a discrete scale. In this approach, the confidence scale is treated as an ordinal, rather than an interval, scale. Here, we will briefly outline the intuition behind an ORM (see Long, 1997, for all mathematical details).

In broad terms, an ORM can be derived from a model in which the observed variable (dependent variable), y, reflects a latent variable, y^*, which ranges from $-\infty$ to $+\infty$. The latent variable on a given trial, t, is generated by $y_{\mathrm{t}}^* = \mathbf{x}_{\mathrm{t}}\boldsymbol{\beta} + e_{\mathrm{t}}$. Here, $\mathbf{x}_t$ is a row vector with the observation of a given input, x_{i}, on trial t in column i (independent variables), $\boldsymbol{\beta}$ is a column vector of weights governing the influence of each input (coefficients), and e_{t} is random noise (error term). The observed variable on a given trial is then generated by comparing the latent variable, y_{t}^*, to a set of subjective thresholds, $\boldsymbol{\tau}$,

$$y_{\mathrm{i}} = m \quad \text{if } \tau_{\mathrm{m}-1} \leq y_{\mathrm{t}}^* < \tau_{\mathrm{m}} \quad \text{for } m = 1 \text{ to } J \qquad \textbf{Eq. 3-1}$$

Here, $\tau_0 = -\infty$ and $\tau_J = +\infty$ and J denotes the number of steps on the scale in question. The distribution of the error term depends on the specific ORM. The Ordered Logit Model (OLM) assumes that the error term has a logistic distribution, with mean 0 and variance $\pi^2/3$. In contrast, the Ordered Probit Model (OPM) assumes that the error term has a standard normal distribution, with mean 0 and variance 1. Once we have defined the variance of the error term and made a series of observations, we can fit $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ with maximum likelihood estimation. Here, we used the OPM as it provided a better fit to our data.

3.2.3.1.3 Meta-analysis

We report the results of the trial-by-trial regression model of confidence reports at the group level – we adopted this approach because of the relatively small number of participants in

each experiment. We first conducted a separate regression for each participant. This analysis generated the subject-level statistics – that is, the fitted coefficients (i.e., $\boldsymbol{\beta}$) and subjective thresholds (i.e., $\boldsymbol{\tau}$) – and an estimate of the variance associated with each subject-level statistic. We then computed the group-level statistics using a meta-analysis approach (Borenstein, Hedges, Higgins, & Rothstein, 2009; Shadish & Haddock, 1994). In broad terms, we computed the average coefficients and subjective thresholds after weighting the subject-level statistics of each participant by the inverse of the variance of the respective statistics. In addition, for each group-level statistic, we computed the associated group-level standard error and a z-statistic to test the hypothesis that the group-level statistic is non-zero. While it is common to conduct inferential tests on raw coefficients pooled across participants (Lorch & Myers, 1990), such tests increase the risk of errors because they do not take into account the error in the estimation of a given coefficient and/or subject-level differences in the size of this error. In contrast, the meta-analysis approach takes into account both types of confound.

3.2.3.1.4 Predicted probabilities at selected input values

The coefficients (i.e., $\boldsymbol{\beta}$) and subjective thresholds (i.e., $\boldsymbol{\tau}$) are arbitrary in the sense that they cannot be identified without assuming a specific distribution for the error term. However, even when the error term is assumed to have a specific distribution, the coefficients and subjective thresholds might change when the independent variables have been transformed (e.g., re-scaled). However, transformations of the independent variables do not affect the predicted probability of a given response (i.e., whether the participant responds “1”, “2”, “3”, “4”, “5” or “6”) for a given set of independent variables, $P(y = m|\mathbf{x})$ – that is, the model prediction stays the same. As such, predicted probabilities at selected input values offer a more intuitive interpretation of the effects identified by an ORM.

In addition to the group-level statistics described above, we therefore show the change in the predicted probability of a given response for a discrete change in an independent variable of interest. Let $P(y = m|\mathbf{x}_{-\mathrm{i}}, x_i)$ be the predicted probability of a given response for a given set of independent variables, with x_i being the independent variable of interest. Then, $P(y = m|\mathbf{x}_{-\mathrm{i}}, x_i + \delta)$ is the probability of a given response when x_i is increased by δ, while holding all other independent variables constant. The change in the predicted probability of a given response for a discrete change in the independent variable of interest can then be computed as

$$\Delta P(y = m|\mathbf{x}) = P(y = m|\mathbf{x}_{-\mathrm{i}}, x_i + \delta) - P(y = m|\mathbf{x}_{-\mathrm{i}}, x_i) \quad \textbf{Eq. 3-2}$$

As an example, we could compute the change in the predicted probability of a given response when the current stimulus is at its maximum compared to its minimum value as

$$\Delta P(y = m|\mathbf{x}) = P(y = m|\mathbf{x}_{-\mathrm{i}}, \max k) - P(y = m|\mathbf{x}_{-\mathrm{i}}, \min k) \quad \textbf{Eq. 3-3}$$

The resulting value will tell us whether the predicted probability of the response is higher or lower when the current stimulus is at its maximum compared to its minimum value. Critically, since the change in the predicted probability as defined above also depends on the values of the independent variables that are not of interest (i.e. $\mathbf{x}_{-\mathrm{i}}$), then the latter must be held constant at sensible values. We fixed non-binary independent variables at their average values (across participants in a given experiment) and binary independent variables at 0. If a variable can only be realised as -.5 and .5, then fixing it at 0 for this calculation is equivalent to averaging across both realisations of the variable.

3.2.3.2 Conditional probability of each response pairing

To anticipate our results, we found, in all three experiments, that the confidence reported on a given trial was biased by the confidence reported on the previous trial. To further unpack

this effect, we performed a permutation test of the observed conditional probability of each response pairing (e.g., the conditional probability of responding "6" on the current trial given that the response on the previous trial was "6").

We first computed for each participant the conditional probability of each response pairing, $P(c_t = m|c_{t-1} = n)$, where m is the confidence reported on the current trial, n is the confidence reported on the previous trial and the conditional probabilities of all m for a given value of n sum to 1. Thus, we obtained a subject-level m-by-n (square) matrix with the conditional probabilities of all possible response pairings, $\mathbf{G}_{\text{subject}} = f(c_t, c_{t-1})$. To compute the corresponding group-level matrix, $\mathbf{G}_{\text{group}} = f(c_t, c_{t-1})$, we averaged across the subject-level matrices. Next, to test whether the conditional probability of a given response pairing was higher than expected by chance, we compared it to a distribution of conditional probabilities expected under (shuffled) data without serial dependencies. More precisely, we shuffled the trial structure for each participant, computed the subject-level matrices of the conditional probability of each response pairing under the shuffled data, $\mathbf{G}_{\text{shuffle:subject}}$, and then computed the corresponding group-level matrix, $\mathbf{G}_{\text{shuffle:group}}$, by averaging across the subject-level matrices. We repeated this procedure 1,000 times, giving rise to a distribution of group-level matrices expected under shuffled data. Lastly, we calculated whether the value of each element in the empirically observed group-level matrix was higher than the 975th value of its respective distribution of values expected under shuffled data (values sorted in ascending order). This procedure amounts to a two-sided test of the null hypothesis that the conditional probability of a given response paring is not higher than expected by chance at the 5%-significance level.

3.3 Results

3.3.1 Regression

3.3.1.1 Group-level results

We used our trial-by-trial regression model of confidence reports (**Table 3-1**) to dissociate the different sources of trial-by-trial variation in reported confidence. We show in **Table 3-2** the group-level statistics from the meta-analysis of the subject-level statistics generated by separate regression models for each participant. Further, to facilitate interpretation of the group-level effects and comparison between experiments, we show in **Table 3-3** how the predicted probability of a given response (i.e., the prediction of the trial-by-trial regression model of confidence) varies with a discrete change in a given predictor of interest. We only show this analysis for predictors which were significant in at least two experiments or which sensibly could be compared between experiments.

Predictor or threshold	Estimated effect ± SE		
	EXP1	*EXP2*	*EXP3*
k_t	10.60 ± 0.91 ***	13.04 ± 0.94 ***	0.00 ± 0.00 ***
a_t	0.41 ± 0.04 ***	0.45 ± 0.04 ***	0.85 ± 0.13 ***
i_t	0.04 ± 0.04	–0.17 ± 0.04 ***	0.02 ± 0.06
$RT1_t$	–0.55 ± 0.07 ***	N/A	–0.73 ± 0.20 ***
$RT2_t$	0.87 ± 0.15 ***	0.16 ± 0.06 **	–3.12 ± 0.56 ***
k_{t-1}	–0.79 ± 0.32 *	–0.98 ± 0.33 **	0.00 ± 0.00
a_{t-1}	0.11 ± 0.06	0.14 ± 0.06 *	0.08 ± 0.06
i_{t-1}	0.01 ± 0.03	0.01 ± 0.03	–0.02 ± 0.03
$RT1_{t-1}$	0.06 ± 0.03	N/A	0.45 ± 0.08 ***
$RT2_{t-1}$	–0.02 ± 0.04	0.01 ± 0.01	0.38 ± 0.19 *
$i_t * d_{t-1}$	0.01 ± 0.03	0.03 ± 0.03	–0.08 ± 0.05
$d_t * d_{t-1}$	0.03 ± 0.03	0.00 ± 0.03	0.04 ± 0.05
c_{t-1}	0.09 ± 0.03 ***	0.14 ± 0.03 ***	0.14 ± 0.06 *
$c_{t-1} * a_{t-1}$	0.08 ± 0.04 *	0.07 ± 0.05	0.08 ± 0.04 *
s_t	N/A	N/A	0.22 ± 0.09 *
τ_1	–2.20 ± 0.11	–2.30 ± 0.21	–0.50 ± 0.23
τ_2	–1.40 ± 0.10	–1.92 ± 0.22	0.24 ± 0.19
τ_3	–0.64 ± 0.12	–1.40 ± 0.21	0.93 ± 0.22
τ_4	0.19 ± 0.15	–0.61 ± 0.20	1.39 ± 0.23
τ_5	1.23 ± 0.14	0.49 ± 0.20	1.88 ± 0.21

Table 3-2. Results of the trial-by-trial regression model of confidence reports. The table shows the group-level coefficients and subjective thresholds and the standard error associated with each group-level statistic as estimated by the meta-analysis. We used the z-statistics obtained from the meta-analysis to test the hypothesis that a given group-level coefficient was non-zero. All tests were two-sided. *, **, *** indicate $p < .05, .01, .001$, respectively. We only show five subjective thresholds as $\tau_0 = -\infty$ and $\tau_6 = +\infty$. All non-binary variables were mean-centred. Mean confidence was 3.093, 2.255 and 4.641 in EXP1, EXP2 and EXP3, respectively. Mean RT1 was .739 and .697 in EXP1 and EXP3, respectively. Mean RT2 was .981, 1.434 and .541 in EXP1, EXP2 and EXP3, respectively. See Table 3-1 for details about each predictor.

Exp.	Predictor	Change	$\bar{\Delta}$	Effect on predicted probability					
				1	*2*	*3*	*4*	*5*	*6*
EXP1	k_t	min : max	1.01	–0.25	–0.26	0.04	0.21	0.18	0.08
	a_t	–.5 : +.5	0.32	–0.08	–0.08	0.03	0.07	0.04	0.01
	k_{t-1}	min : max	0.08	0.02	0.02	–0.01	–0.02	-0.01	0.00
	c_{t-1}	min : max	0.35	–0.08	–0.09	0.03	0.08	0.05	0.02
	a_{t-1}	–.5 : +.5	0.09	–0.02	–0.02	0.01	0.02	0.01	0.00
	$c_{t-1}\|\text{incorrect}_{t-1}$	min : max	0.27	–0.07	–0.07	0.03	0.06	0.04	0.01
	$c_{t-1}\|\text{correct}_{t-1}$	min : max	0.56	–0.12	–0.16	0.02	0.12	0.09	0.04
EXP2	k_t	min : max	1.12	–0.54	–0.02	0.24	0.15	0.08	0.09
	a_t	–.5 : +.5	0.33	–0.17	0.03	0.08	0.03	0.01	0.01
	k_{t-1}	min : max	0.10	0.05	–0.01	–0.02	–0.01	0.00	0.00
	c_{t-1}	min : max	0.48	–0.24	0.01	0.12	0.06	0.03	0.02
	a_{t-1}	–.5 : +.5	0.11	–0.05	0.01	0.03	0.01	0.00	0.00
	$c_{t-1}\|\text{incorrect}_{t-1}$	min : max	0.40	–0.20	0.03	0.10	0.04	0.02	0.01
	$c_{t-1}\|\text{correct}_{t-1}$	min : max	0.62	–0.29	–0.02	0.15	0.08	0.04	0.04
EXP3	k_t	min : max	0.21	–0.02	–0.02	–0.03	–0.04	0.01	0.10
	a_t	–.5 : +.5	0.64	–0.06	–0.07	–0.09	–0.10	0.03	0.29
	k_{t-1}	min : max	0.01	0.00	0.00	0.00	0.00	0.00	0.00
	c_{t-1}	min : max	0.53	–0.06	–0.06	–0.07	–0.07	0.04	0.22
	a_{t-1}	–.5 : +.5	0.06	–0.01	–0.01	–0.01	–0.01	0.00	0.03
	$c_{t-1}\|\text{incorrect}_{t-1}$	min : max	0.16	–0.02	–0.02	–0.02	–0.02	0.01	0.07
	$c_{t-1}\|\text{correct}_{t-1}$	min : max	0.46	–0.05	–0.05	–0.06	–0.07	0.03	0.20

Table 3-3. Predicted probabilities at selected input values. The table shows how the predicted probability of a given response varies as a function of a discrete change in a given predictor of interest. We performed the analysis using the group-level statistics shown in Table 3-2. We held predictors that were not of interest at 0 (i.e., their mean value or average influence). For a non-binary predictor of interest, we computed the difference in the predicted probability of a given response when the predictor was at its maximum compared to its minimum (mean-centred) value. For a binary predictor of interest, we computed the difference in the predicted probability of a given response when the predictor was at its positive (+.5) compared to its negative (-.5) value. Thus, a positive/negative value means that the predicted probability of a given response increases/decreases with the change in the predictor of interest. To ease visual inspection, we show negative values and positive values in blue and red, respectively. As a measure of effect size, we show, for each predictor of interest, the sum of the absolute differences across all responses ($\bar{\Delta}$) – note that this quantity was computed using exact values before rounding to two decimals. To unpack the interaction between the previous confidence and the previous accuracy, we show how the previous confidence affects the predicted probability of a given response when the previous decision was incorrect ($c_{t-1}|\text{incorrect}$) and when it was correct ($c_{t-1}|\text{correct}$).

3.3.1.1.1 Classic sources of trial-by-trial variation – current trial

In Experiments 1 to 3, the confidence reported on a given trial was positively dependent on the sensory stimulation, with participants being more confident for strong stimuli (EXP1: β = 10.60, z = 11.68, p < .001; EXP2: β = 13.04, z = 13.80, p < .001; EXP3: β = 0.00, z = 5.13, p < .001) and for correct decisions (EXP1: β = 0.41, z = 9.66, p < .001; EXP2: β = 0.45, z = 10.22, p < .001; EXP3: β = 0.85, z = 6.53, p < .001). This local dependence was strong, as indicated by large shifts in predicted response probabilities for changes in the stimulus strength (EXP1: $\bar{\Delta}$ = 1.01; EXP2: $\bar{\Delta}$ = 1.12; EXP3: $\bar{\Delta}$ = 0.21) and for changes in accuracy (EXP1: $\bar{\Delta}$ = 0.32; EXP2: $\bar{\Delta}$ = 0.33; EXP3: $\bar{\Delta}$= 0.64). Further, but in Experiment 2 only, the confidence reported on a given trial was dependent on the target interval, with participants being more confident when the visual target appeared in interval 1 (EXP1: β = 0.04, z = 1.06, p = .291; EXP2: β = -0.17, z = -4.32, p < .001; EXP3: β = 0.02, z = 0.33, p =.740). As for the design of Experiment 3, the confidence reported on a given trial was dependent on the current condition, with participants being more confident in accuracy blocks than in speed blocks (EXP3: β = 0.22, z = 2.64, p = .008).

We could only analyse RT1 and RT2 separately in Experiments 1 and 3. In both experiments, the confidence reported on a given trial was negatively dependent on RT1, with participants being more confident for faster decisions (EXP1: β = -.55, z = -8.24, p < .001; EXP3: β = -0.73, z = -3.71, p < .001). In Experiment 1, the confidence reported on a given trial was positively dependent on RT2, with participants being more confident for slower confidence reports (EXP1: β = 0.87, z = 5.90, p < .001). In contrast, in Experiment 3, the confidence reported on a given trial was negatively dependent on RT2, with participants being more confident for faster confidence reports (EXP1: β = -3.12, z = -5.61, p < .001). In Experiment 2, the confidence reported on a given trial was positively dependent on RT2 (now decision and confidence report), with participants being more confident for slower responses (EXP1: β = 0.16, z = 2.84, p = .005). This discrepant pattern of results – where longer RT2 is associated with higher

confidence in Experiment 1 and 2 but lower confidence in Experiment 3 – could be due to the simple fact that participants had to make more key presses to indicate high confidence in Experiment 1 and 2 and thus often would take longer to do so.

3.3.1.1.2 Classic sources of trial-by-trial variation – previous trial

In Experiments 1 and 2, the confidence reported on a given trial was negatively dependent on the sensory stimulation on the previous trial, with participants being less confident following strong stimuli (EXP1: β = -0.79, z = -2.49, p = .013; EXP2: β = -0.98, z = -2.94, p = .003). This serial dependence was weak, as indicated by small shifts in predicted response probabilities for changes in the strength of the previous stimulus (EXP1: $\bar{\Delta}$ = 0.08; EXP2: $\bar{\Delta}$ = 0.10). In Experiment 3, there was no effect of the sensory stimulation on the previous trial (EXP3: β = 0.00, z = -0.17, p = .868; $\bar{\Delta}$ = 0.01). We note that the task design in all three experiments (i.e., stimuli were only randomised within a block of trials and not across the entire experiment) gave rise to a negative serial dependence between the current and the previous stimulus strength (meta-analysis of statistics estimated by linear regression model; EXP1: β = -0.03, z = -2.57, p = .010; EXP2: β = -0.02, z = -1.68, p = .094; EXP3: β = -0.04, z = -3.34, p < .001). However, since we included both predictors, any serial dependence between the current confidence report and the previous stimulus strength would reflect an effect over and above the stimulus contingencies – such as sensory adaptation or sensory expectations. There was no effect of the previous target interval in any of the three experiments (EXP1: β = 0.01, z = 0.30, p = .761; EXP2: β = 0.01, z = 0.37, p = .713; EXP3: β = -0.02, z = -0.51, p = .612).

Again, we could only analyse RT1 and RT2 separately in Experiments 1 and 3. In Experiment 3 only, the confidence reported on a given trial was positively dependent on both RT1 and RT2 on the previous trial, with participants reporting higher confidence following longer RT1s (EXP1: β = -0.06, z = 1.91, p = .056; EXP3: β = 0.45, z = 5.40, p < .001) as well as longer RT2s

(EXP1: β = -0.02, z = -0.58, p = .565; EXP3: β = 0.38, z = 1.96, p = .050). There was no effect of RT2 on the previous trial in Experiment 2 (EXP2: β = 0.01, z = 0.99, p = .323). We will not consider these results further as our goal was only to test whether any systematic variation in trial-by-trial confidence report would remain once we took into account classic first-order effects.

3.3.1.1.3 Trial-by-trial effects in the generation of a decision

In Experiments 1 to 3, the confidence reported on a given trial was not dependent on the congruency between the current target interval and the previous decision (EXP1: β = 0.01, z = 0.21, p = .834; EXP2: β = 0.03, z = 1.04, p = .297; EXP3: β = -0.08, z = -1.64, p = .101). Further, there was no effect of the congruency between the current and the previous decision (EXP1: β = 0.03, z = 0.77, p = .441; EXP2: β = 0.00, z = -0.12, p = .908; EXP3: β = 0.04, z = 0.87, p = .386).

3.3.1.1.4 Trial-by-trial effects in the generation of a confidence report

Turning to our main question of interest, in all experiments we observed that the confidence reported on a given trial was dependent on the response and the feedback history. Here, we consider separately the main effect of the previous confidence report, the main effect of the previous accuracy and their interaction effect. In Experiments 1 to 3, the confidence reported on a given trial was positively dependent on the confidence reported on the previous trial, with participants being more confident following reports of high confidence (EXP1: β = 0.09, z = 3.51, p < .001; EXP2: β = 0.14, z = 4.22, p < .001; EXP3: β = 0.14, z = 2.43, p = .015). This serial dependence was strong, as indicated by large shifts in predicted response probabilities for changes in the previous confidence (EXP1: $\bar{\Delta}$ = 0.35; EXP2: $\bar{\Delta}$ = 0.48; EXP3: $\bar{\Delta}$ = 0.53). In Experiment 2, the confidence reported on a given trial was also positively dependent on the

accuracy of the decision made on the previous trial, with participants being more confident following correct decisions (EXP2: β = 0.14, *z* = 2.50, *p* = .012). This serial dependence was weak, as indicated by small shifts in predicted response probabilities for changes in the previous accuracy (EXP1: $\bar{\Delta}$ = 0.11). In Experiments 1 and 3, there was no main effect of the accuracy of the decision made on the previous trial (EXP1: β = 0.11, *z* = 1.85, *p* = .065, $\bar{\Delta}$ = 0.09; EXP3: β = 0.08, *z* = 1.28, *p* = .201; $\bar{\Delta}$ = 0.06). However, in Experiments 1 and 3, the positive serial dependence in confidence reports was positively modulated by the accuracy of the decision made on the previous trial (i.e., an interaction effect), with the influence of past reports of high confidence being stronger when their associated decisions were correct (EXP1: β = 0.08, z = 2.09, p = .037; EXP3: β = 0.08, z = 1.99, p = .047). This relationship was reflected in larger shifts in predicted response probabilities for changes in the previous confidence when the associated decision was correct compared to incorrect (ratio of $\bar{\Delta}$ for $c_{t-1}|\text{correct}$ relative to $\bar{\Delta}$ for $c_{t-1}|\text{incorrect}$; EXP1: 2.07; EXP3: 2.88). A simple-effects analysis (for $c_{t-1}|\text{incorrect}$, $a_{t-1} = +1$; for $c_{t-1}|\text{correct}$, $a_{t-1} = -1$) indicated that the positive serial dependence in confidence reports was limited to trials that were preceded by a correct as opposed to an incorrect decision (simple effect of $c_{t-1}|\text{incorrect}$; EXP1: β = 0.05, z = 1.24, p = .222; EXP3: β = 0.09, z = 1.65, p = .099; simple effect of $c_{t-1}|\text{correct}$; EXP1: β = 0.13, z = 6.20, p < .001; EXP3: β = 0.18, z = 2.83, p = .005). In Experiment 2, the positive serial dependence in confidence reports was not affected by the accuracy of the decision made on the previous trial (EXP2: β = 0.07, z = 1.52, p = .128, ratio of $\bar{\Delta}$ for $c_{t-1}|\text{incorrect}$ relative to $\bar{\Delta}$ for $c_{t-1}|\text{incorrect}$: 1.55).

3.3.1.2 Conditional probability of each response pairing

To further unpack the positive serial dependence in confidence reports, we performed a permutation test of the conditional probability of each response pairing (e.g., the conditional probability of responding "6" on the current trial given that the response on the previous trial was "6"). More specifically, we tested whether the conditional probability of a given response

pairing was higher than expected under shuffled data without any trial-by-trial effects. **Figure 3-3** shows, for each experiment, the results of the permutation test, with diagonal elements corresponding to consecutive instances of the same response. We first sought to determine whether participants only had a tendency to repeat their previous response regardless of its meaning (i.e., "response priming"). As can be seen in **Figure 3-3** (left column), participants repeated their previous response more often than expected by chance (diagonal elements). However, participants also made a response in the vicinity of their previous response more often than expected by chance (off-diagonal elements). Therefore, our effect was unlikely to be due simply to response priming.

We then sought to determine whether participants had a tendency to press the same buttons as on the previous trial regardless of their response association (i.e., "motor priming"). To address this issue, we performed the permutation test separately for "stay" trials (i.e., the current decision was the same as the previous one) and "switch" trials (i.e., the current decision was not the same as the previous one). Critically, in Experiments 2 and 3, the motor contingencies involved in two consecutive instances of the same response were different for the two types of trial. On stay trials, participants would have to press the same buttons to repeat their previous response. In contrast, on switch trials, participants would have to press different buttons. As can be seen in **Figure 3-3** (middle and right columns), the diagonal and the off-diagonal patterns were similar for the two types of trial, with participants repeating a previous response as well as responding in the vicinity of a previous response more often than expected by chance. Therefore, our effect was unlikely to be due simply to motor priming.

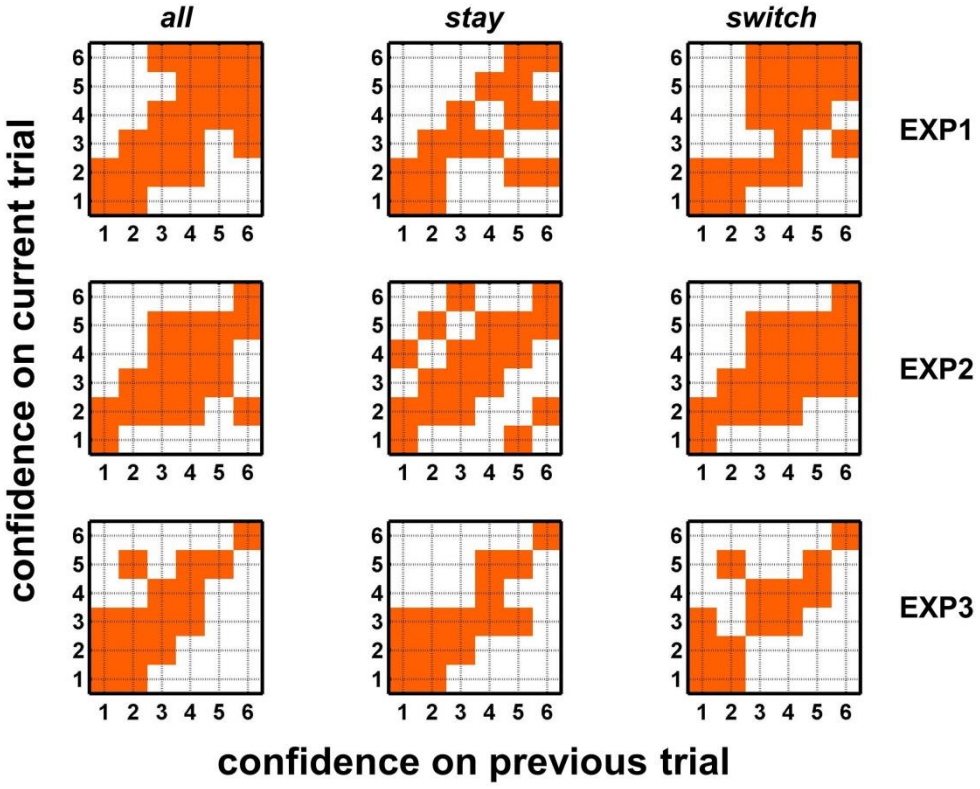


Figure 3-3. Permutation test of the conditional probability of each response pairing. Subplots in the same row are based on data from the same experiment. Subplots in the same column are based on the same analysis ("all": all trials; "stay": when the current decision was the same as the previous one; "switch": when the current decision was not the same as the previous one). The axes of each subplot indicate the confidence reported on the previous trial (x-axis) and the confidence reported on the current trial (y-axis). Thus, each element (coordinate) corresponds to a given response pairing. A coloured element indicates that the conditional probability of the respective response pairing was higher than expected by chance (shuffled data) at the 5%-significance level in a two-sided test.

3.4 Discussion

We considered the novel hypothesis that trial-by-trial variation in confidence reports is in part systematic, with the confidence reported on a given trial influenced by the history of past reports and feedback obtained. Using a trial-by-trial regression model, we found that the confidence reported on a given trial was indeed biased toward the confidence reported on the previous trial. In addition, that the confidence reported on a given trial was higher when the decision made on the previous trial was correct – through either an additive effect (EXP2) or a modulation of the positive serial dependence in confidence reports (EXP1 and EXP3). Importantly, the effect of the history of past reports and feedback obtained were found (a) after controlling for the sensory evidence, (b) after controlling for trial-by-trial effects in sensory encoding, (c) after controlling for trial-by-trial effects in the generation of a decision, (d) when two consecutive instances of the same confidence report required different button presses, (e) when the decision and its associated confidence report were entered sequentially (EXP1 and EXP3) and simultaneously (EXP2), (f) for two types of task (EXP1 and EXP2: two-interval forced-choice contrast discrimination; EXP3: two-interval forced-choice luminance discrimination), (g) for two types of confidence scale (EXP1 and EXP2: arbitrary scale from 1 to 6 in steps of 1; EXP3: probability scale from 50% to 100% in steps of 10%), and (h) without (EXP1 and EXP2) and with (EXP3) incentives for accurate decisions and confidence reports.

3.4.1 What drives the effects of the report and the feedback history?

In the SDT framework, there are three potential sources of systematic trial-by-trial variation in confidence reports: trial-by-trial effects in the encoding of the sensory evidence; trial-by-trial effects in the generation of a decision, either at the level of the computation of the decision variable or at the level of the mapping of the decision variable onto a decision; and trial-by-trial effects in the generation of a confidence report, with participants directly adjusting the

thresholds governing the mapping of the confidence variable onto a confidence report (see **Section 1.3.2.1: Signal detection theory**). Given that we controlled for the stimulus, the reaction time and the decision history, the effects of the report and the feedback history are most likely to be the result of direct trial-by-trial adjustments to the confidence thresholds. Here, we consider different processes that might drive such adjustments. We emphasise that the processes need not be mutually exclusive; they might operate simultaneously or at different moments in time.

3.4.1.1 Maximising consistency or information transmission given expectations

As discussed previously, task-specific expectations might drive trial-by-trial adjustments to the confidence thresholds (Treisman & Williams, 1984; **Figure 3-1**). First, participants might seek to maximise the consistency of their confidence reports. In particular, if the stimulus intensity is expected to be stable, then it may be beneficial for a participant to adjust the confidence criteria such that the same response is made on consecutive trials. Theoretically, this strategy would increase the robustness of a given response to variation in the confidence variable due to sensory noise, and it could serve to maintain some desired reward rate. Under this account, a positive serial dependence in confidence reports is expected. Second, participants might seek to maximise the information carried by their confidence reports. Again, if the stimulus intensity is expected to be stable, then it may be beneficial to adjust the confidence criteria so that the reported confidence is sensitive to fluctuations in the confidence variable. For example, if the stimulus intensity is increased to a new but stable value, then a participant may adjust the confidence criteria so that the weakest confidence variables are reported with low confidence and the strongest confidence variables with high confidence – regardless of their global rank across the experiment. Theoretically, this strategy maximises the (Fisher) information carried by confidence reports – thus making confidence an "efficient code"

(Barlow, 1961). Under this account, a negative serial dependence in confidence reports is expected. On both accounts, feedback about decision accuracy may be used to control a planned adjustment; an error would indicate that the sensory evidence on the trial was not diagnostic of the stimulus intensity and therefore that it might be beneficial to dampen the planned adjustment.

The positive serial dependence in confidence reports would initially appear to be consistent with participants maximising consistency as opposed to maximising information transmission. However, because the stimulus intensities were randomised in all three experiments, we cannot tell apart the two accounts. On one hand, participants might have mistaken local "streaks" (i.e., consecutive instances of the same stimulus intensity) for stability and built up an expectation that the stimulus intensity will be stable at least for a period of time. If so, the positive serial dependence in confidence reports would be consistent with participants seeking to maximise consistency. On the other hand, because of the randomisation of the stimulus intensities, participants might have expected the stimulus intensity to alternate from one trial to another (i.e., low-high-low-high-low-etc.). In fact, the randomisation procedure gave rise to a negative serial dependence in stimulus intensity. This effect was relatively negligible, but participants might have over-generalised this trial-by-trial pattern. If so, the positive serial dependence in confidence reports would be consistent with participants seeking to maximise information transmission.

3.4.1.2 Calibration or hot-hand signal

As discussed previously, trial-by-trial adjustments to the confidence thresholds might be the result of a continuous evaluation of whether the current level of confidence meets the task goal or reflects current performance (Ferrell & McGoey, 1980; Yeung & Summerfield, 2012). For example, participants may have sought to adjust the confidence threshold such that their

reports of low and high confidence accurately discriminate between their incorrect and correct decisions, respectively. This type of calibration process may be subserved by reinforcement learning (Sutton & Barto, 1998). On error trials, a difference between the lowest level of confidence and the reported level of confidence (negative prediction error) would indicate that the next report should be less biased toward high confidence. In contrast, on correct trials, a difference between the highest level of confidence and the reported level of confidence (positive prediction error) would indicate that the next report should be more biased toward high confidence.

Under simple calibration, reported confidence would be expected to decrease after incorrect trials but increase after correct trials. In addition, under more sophisticated calibration, the magnitude of this change would be expected to scale with the confidence reported on the previous trial. While we did not observe the latter pattern of responses after incorrect trials (i.e., negative serial dependence), this might have been due to the facts that participants were incorrect less often than they were correct and/or that their level of confidence varied much less when they were incorrect. Alternatively, instead of calibration, the effect of the history of past reports and feedback obtained might simply have been driven by some internal "hot-hand" signal (Ayton & Fischer, 2004; Sundali & Croson, 2006), which reinforces past reports of confidence and which may be reset whenever an error occurs. We note that neither account requires that participants have direct access to the space of sensory evidence (e.g., track history of sensory evidence or use this information to build up expectations). A trial-by-trial adjustment to the level of confidence could be implemented through a simple bias term that operates during the later stages of the generation of a confidence report.

3.4.2 Implications for research on metacognition

The ability to report on the reliability of one's internal operations – that is, metacognition – is often studied by asking participants to report how confident they feel about some sensory decision (Fleming & Lau, 2014). The basic assumption in such studies is that, for participants with high metacognitive sensitivity, their reports of low and high confidence should accurately discriminate between their incorrect and correct decisions, respectively. To quantify this pattern, data is typically averaged across an entire experiment, and an attempt is made to isolate metacognitive sensitivity (trial-by-trial correlation between fluctuations in the reported level of confidence and accuracy) from metacognitive bias (a bias toward reporting low or high confidence). As such, measures of metacognition assume that participants should maintain a stable set of confidence thresholds – with global inconsistency in the mapping of a confidence variable onto a confidence report attributed to random variation (e.g., noisy read-out or noisy thresholds) and thus taken to reflect low metacognitive sensitivity. The theories and results presented here suggest that this assumption is flawed.

Even if participants sought to maximise consistency, this might require local adjustments to the confidence thresholds in accordance with expectations. Importantly, a drive for local consistency might give rise to global inconsistency – not because of random variation but because of systematic variation. Similarly, if participants sought to maximise information transmission, this might require local adjustments to the confidence thresholds in accordance with expectations. For example, if "1" was associated with an objective accuracy of 50% and "6" with an objective accuracy of 75% under low stimulus intensities, then "1" might become associated with an objective accuracy of "75" and "6" with an objective accuracy of 100% if the stimulus intensities were increased. Again, a drive for information transmission might give rise to global inconsistency – not because of random variation but because of systematic variation. The expectations that underpin the local adjustments to the confidence thresholds might be wrong or superstitious – with participants picking up on local patterns in a random

but highly artificial environment or mistaking variation in the sensory evidence due to sensory noise for variation in the stimulus intensity (Schüür et al., 2013; Yu & Cohen, 2009). However, given such expectations, the local adjustments themselves would be rational (Beck, Ma, Pitkow, Latham, & Pouget, 2012). Thus, as has been done for measures of visual sensitivity (Fründ et al., 2014), future research should seek to develop measures of metacognitive sensitivity that partial out history effects.

3.4.3 Neural correlates of trial-by-trial confidence

Recent studies have identified the right rostrolateral prefrontal cortex (rlPFC) as playing a key role in the generation of confidence reports (De Martino et al., 2012; Fleming, Huijgen, & Dolan, 2012; Fleming, Weil, Nagy, Dolan, & Rees, 2010). For example, rlPFC shows greater activity during active report of confidence compared to a matched control condition (forced-response entry). Further, rlPFC activity during active report correlates with the reported confidence level. Lastly, the size of the rlPFC (grey-matter volume) correlates with the extent to which participants' reports of low and high confidence accurately discriminate between their incorrect and correct decisions, respectively. However, it still remains unclear exactly what computations rlPDF activity reflect. Does activity in this region reflect the computation of the confidence variable? If so, rlPFC activity should be directly proportional to the strength of the sensory evidence in favour of the chosen decision? Alternatively, does activity in this region reflect the mapping of the confidence variable onto a confidence report? If so, rlPFC activity should be directly proportional to the reported confidence level. In most situations, the strength of the sensory evidence and the reported confidence level will be tightly coupled. However, our results suggest that it should, in principle, be possible to de-couple them, by slowly increasing the stimulus intensity (stronger sensory evidence) while providing false

negative feedback (no decrease in confidence), for example. As such, our approach provides a framework for comparing competing explanations of rlPFC activity.

3.5 Chapter summary

In this chapter, we used our framework to test whether the report of confidence is flexible – a key ingredient for effective cooperation (**Figure 3-4**). Analysing the dataset from **Chapter 2** and a dataset provided by Moran et al. (2015), we found that the confidence reported on a given trial was biased toward the confidence reported on the previous trial, and was higher when the decision made on the previous trial was correct. We discussed two different explanations of the observed history effects. First, that participants adjusted their confidence reports so as to maximise consistency or information transmission. Second, that participants adjusted their confidence reports according to some calibration process or internal hot-hand signal.

Our findings pose a problem for current theories and measures of metacognition which assume that the report of confidence should be stable over time. We argued that global inconsistency in the report of confidence can in fact result from a series of locally consistent or locally informative confidence reports. More generally, the observed flexibility in the report of confidence is important for effective cooperation. To minimise miscommunication, interacting individuals must adapt to each other's use of the scale on which confidence is mutually expressed and thus establish a common metric for identifying who is more likely to be correct in a given situation. In the next chapter, we will test whether interacting participants can solve this communication problem.

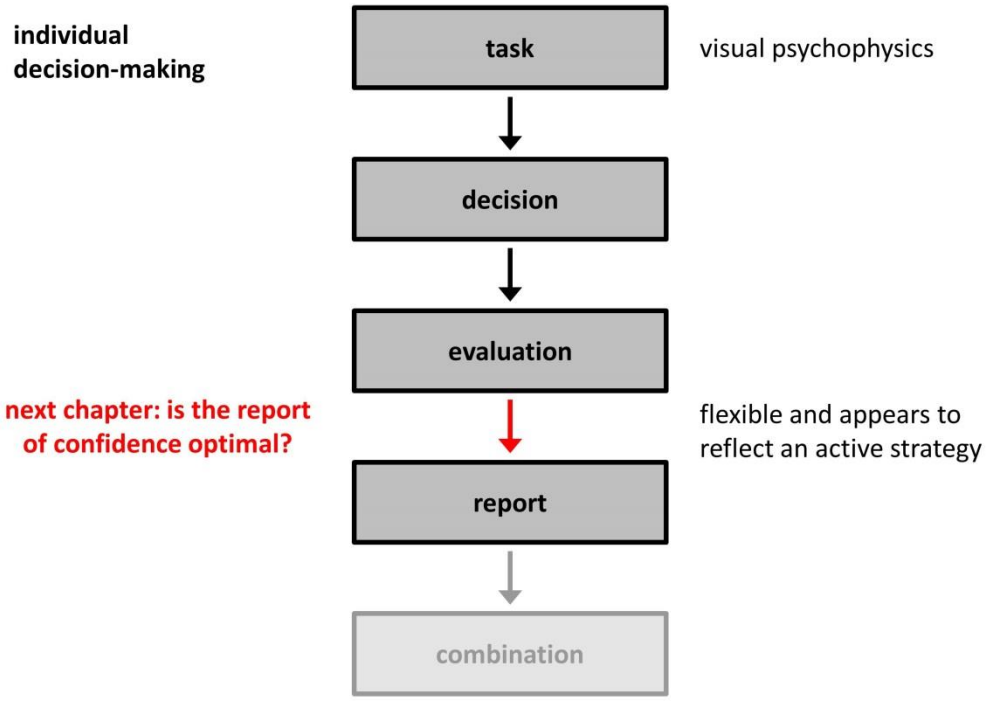


Figure 3-4. Summary of Chapter 3. The boxes denote the component processes involved in sharing and combining representations of uncertainty.

Chapter 4: Confidence matching as a strategy for communication of uncertainty

Most important decisions in government, business and science are made by groups of people. When group members disagree about which course of action to take, opinions expressed with higher confidence often carry more weight. However, pooling information in this way is prone to error, as people's reported confidence correlates with a range of contextual (irrelevant) factors, such as gender, profession and mental health. To avoid miscommunication, group members must learn to take each other's individual biases into account and establish a common metric for identifying who is more likely to be correct in a given situation. Solving this communication problem is, however, cognitively difficult. In this chapter, studying the simplest case in which a group consists of two people, we show that group members instead use a heuristic strategy – confidence matching – whereby they match their mean confidence and confidence distributions. Using computational modelling, we quantify the efficacy of this strategy and show that it, despite not being optimal, is a good solution to the communication problem when group members are equally competent (accurate), or when they are poor judges of their relative competence. More generally, we show that confidence reports are highly malleable – rapidly adapting to the social context – and we raise the possibility that well-known biases in the report of confidence might in fact reflect particular norms for social interaction.

4.1 Introduction

We often share and discuss how confident we feel about our sensations, beliefs or decisions – saying "I am doubtful" or "I am certain". Communicating our confidence in this way allows us to elicit feedback from others to aid our learning and engage in novel forms of cooperation (Frith, 2012; Shea, Boldt, Bang, Yeung, Heyes, & Frith, 2014). However, conventions for how confidence should be communicated are inherently ill-defined as it is hard to verify what we truly sense or believe at any given moment in time (Frith & Wentzer, 2013; Wittgenstein, 1958). Consequently, people's reported confidence rarely reflects a common metric (**Figure 4-1**). In particular, while reported confidence typically is proportional to the probability that a given sensation, belief or decision is correct (Harvey, 1997), the shape of this relationship for an individual (e.g., a bias towards high confidence) correlates with a range of contextual (irrelevant) factors, including mental health (Huq et al., 1988), personality (Campbell et al., 2004) and social role, such as profession (Broihanne et al., 2014), gender (Barber & Odean,

2001; Niederle & Vesterlund, 2011) and culture (Mann et al., 1998). During group decision-making, these individual differences can be costly: we might, for example, agree to pursue a course of action that one of us actually thought was unlikely to be successful. Despite the apparent importance, there has been little research investigating whether people can learn to minimise such miscommunication. Drawing on psychophysics and game theory, we present a framework that addresses these issues.

First, psychophysicists have developed methods for inferring the latent processes that govern the relationship between some stimulus and the responses evoked by this stimulus. This approach has revealed that single individuals use statistically optimal estimates of sensory uncertainty to guide their decisions (**Chapter 1**), and that their reported confidence can reflect this information (**Chapter 3**), albeit in an idiosyncratic manner. Second, game theorists have developed sophisticated tasks in which it is in people's own interest to minimise misunderstandings about each other's true intentions or beliefs. These tasks include, for example, signalling games with common interests, where the "sender" and the "receiver" fare best if they establish a convention for which signal denotes the correct action (Lewis, 1969). Building on these two lines of research, we asked pairs of individuals to perform a joint psychophysical task in which the individual decision made with higher confidence was always selected as the joint decision. By means of our task, we could model how participants communicated some internal representation of uncertainty. In addition, the joint-decision rule, which mimics the process of group decision-making (Laughlin & Ellis, 1986; Zarnoth & Sniezek, 1997), required participants to communicate their internal representations of uncertainty in a mutually consistent (optimal) manner.

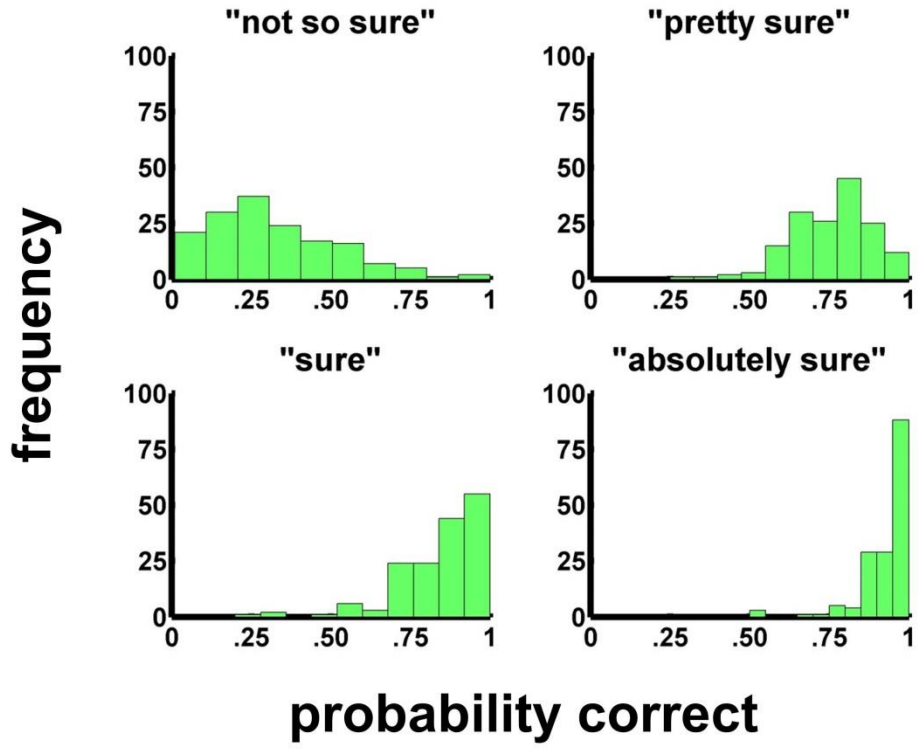


Figure 4-1. Individual differences in the report of confidence. The data is taken from an online survey (conducted for this thesis) in which people were asked to indicate how likely they think they are to be correct (using probabilities from 0% to 100%) when they use a given confidence expression in everyday conversation. The histograms show the distribution of responses for each confidence expression binned into 10 bins. While the confidence expressions can be rank ordered (average of each distribution; “not so sure” < “pretty sure” < “sure” < “absolutely sure”), people intend quite different meanings when they use a given confidence expression (variance of each distribution). 154 people took part in the survey.

To illustrate the communication problem faced by group members, consider two football referees who disagree about whether the ball briefly crossed the goal line (**Figure 4-2A**). Each referee states their opinion with a certain level of confidence (y-axis). This level of confidence reflects some internal representation of uncertainty that scales with the probability that their opinion is correct (x-axis). The two referees have, however, different subjective mappings (blue and red lines), with the blue referee more biased towards reporting high confidence. Consequently, the group decision is dominated by the blue referee who is in fact less likely to be correct (arrows from x-axis). To avoid such miscommunication, the two referees must align their subjective mappings and thus establish a common metric for identifying who is more likely to be correct in a given situation (**Figure 4-2B**).

The communication problem is, however, hard for group members to solve. First, the optimal solution is impossible to achieve instantly. Without prior interaction, group members can only make general guesses about each other's subjective mappings. Further, group members must adjust their subjective mappings at the same time, introducing an infinite regress: to adjust my subjective mapping appropriately, I have to estimate yours, your estimate of mine, your estimate of my estimate of yours, and so on. Second, the optimal solution is difficult to achieve over time. To estimate precisely how likely their partner is to be correct when they express a given level of confidence, group members would need extensive experience. However, even with such experience, this learning process may place too high demands on memory or related processes.

When faced with cognitively complex problems, people tend to resort to heuristics – that is, use a simple strategy which, despite not being optimal, works most of the time (Gigerenzer & Brighton, 2009). Although group members cannot directly access each other's covert subjective mappings, they have access to each other's overt confidence reports. Intuitively, one heuristic strategy that group members might adopt is to match their use of the scale on

which confidence is mutually expressed (**Figure 4-2C**). In line with this intuition, people have been shown to mimic each other's communicative behaviours, including prosody, syntax and vocabulary (Fusaroli et al., 2012; Pickering & Garrod, 2004), and it has been proposed that this type of mimicry reflects a general solution to the problem of miscommunication (Friston & Frith, 2015; Friston & Frith, 2015; Fusaroli et al., 2012). Indeed, if group members are, on average, equally likely to be correct, then their subjective mappings should gradually converge under confidence matching (**Figure 4-2D**). However, if group members are, on average, not equally likely to be correct, then their subjective mappings might in fact diverge under confidence matching (**Figure 4-2E**).

This chapter will involve three studies. In the first study, we tested whether pairs of individuals engage in confidence matching in the context of group-decision making. In the second study, having established confidence matching as a genuine behavioural phenomenon, we used computational modelling to evaluate the efficacy of this solution to the communication problem. In the last study, we tested how confidence matching affects group accuracy in stereotypical cooperative situations (e.g., when participants interact with partners who are more accurate but less confident than themselves).

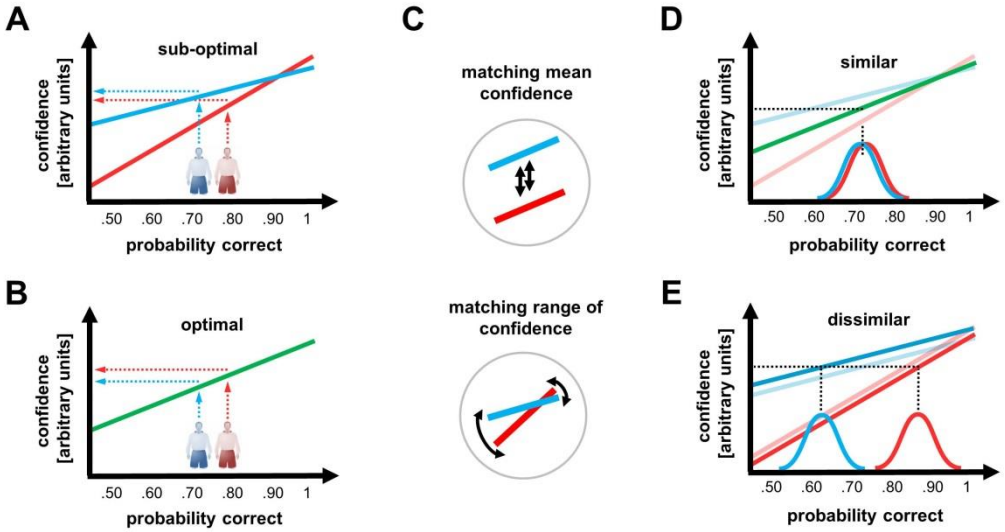


Figure 4-2. Theoretical framework. **(A)** Communication problem. Two referees have different subjective mappings (blue and red lines) from some internal representation related to the probability that their opinion is correct (x-axis) onto confidence (y-axis). **(B)** Optimal strategy. To solve the communication problem, the referees must align their subjective mappings (green line). **(C)** Heuristic strategy. The intercept (*top*) and the slope (*bottom*) of the referees' subjective mappings would change if they matched their mean confidence and their range of confidence. **(D)** Similar group members. If the referees' respective internal representations reflect similar ranges of probability correct, then their subjective mappings should converge under confidence matching. **(E)** Dissimilar group members. If the referees' respective internal representations reflect different ranges of probability correct, then their subjective mappings might diverge under confidence matching. We use a continuous confidence scale and assume linear mapping functions for ease of illustration.

4.2 Confidence matching as a behavioural phenomenon

To test whether confidence matching is a robust behavioural phenomenon, we conducted three experiments. In Experiment 1, we tested whether pairs of individuals (a dyad) showed evidence of confidence matching when they performed a psychophysical task together as compared to alone. Having established that dyad members engaged in confidence matching in the social version of the psychophysical task, we conducted two control experiments. In Experiment 2, we tested whether confidence matching would dissipate with extensive task experience. In Experiment 3, we tested whether confidence matching would dissipate with a non-categorical (less language-like) confidence scale.

4.2.1 Methods

4.2.1.1 Participants

Participants were recruited among students (age range: 18-30) at the University of Tehran (EXP1) and the University of Oxford (EXP2 and EXP3). In total, 120 participants took part in the study (EXP1: *N* = 60; EXP2: *N* = 30; EXP3: *N* = 30; all male). For each experiment, participants were recruited in pairs. Participants only took part in one experiment. All participants reported normal or corrected vision. All participants provided informed consent and were reimbursed for their participation (Iran: 125000R/hour ≈ £5/hour for a total of two hours; UK: £10/hour for a total of two hours). All experiments were approved by the local ethics committee.

4.2.1.2 Experimental details (Experiments 1 to 3)

4.2.1.2.1 Experiment 1

4.2.1.2.1.1 Display parameters and response devices

Participants sat at right angles to each other in a dark room, each with their own monitor and response device. They were separated by a screen to prevent both non-verbal and verbal communication. The monitors (ViewSonic VG2236wm-LED) were linked to a personal laptop (Toshiba Satellite Pro C660-29W) via a video splitter and controlled by the Cogent toolbox (http://www.vislab.ucl.ac.uk/cogent.php/) for MATLAB (Mathworks Inc). The background luminance was 62.5 cd/m^2 and the resolution was 800 × 600. Participants viewed the monitors at a distance of about 57 cm. One participant responded using a QWERTY keyboard and the other responded using a standard mouse. Participants switched response mode half-way through the experiment. A piece of black cardboard was used to occlude half of the monitor for each participant. This manipulation ensured that participants could make their individual responses in private despite receiving input from the same computer.

4.2.1.2.1.2 Basic task (psychophysics)

In Experiment 1 (EXP1), participants performed the psychophysical task from **Chapter 2** (see **Section 2.2.2: Experimental details (Experiments 1 and 2)** for details) in two conditions: a social condition (social task; EXP1-S) and an isolated condition (isolated task; EXP1-I). The order of the conditions was counterbalanced across dyads. In both conditions, participants privately indicated their decision and their confidence level at the same time (**Figure 4-3A**). On each trial, after the two viewing intervals, participants were presented with a horizontal line bisected at its midpoint. A vertical marker was placed on top of the midpoint. The marker could be moved along the line by up to six steps on either side of the midpoint. Here, we take the left hand steps to be negative values (-6 to -1), and the right hand steps to be positive values (1 to 6). The sign of the response indicates the decision (negative: interval 1; positive:

interval 2) and the absolute value of the response indicates the confidence level (1: “unsure”; 6: “certain”). We will use *response* and *confidence* to refer to signed response values and unsigned response values, respectively. We adopted the “two-response” design to make the decision rule in the social task graphically intuitive.

4.2.1.2.1.3 Social task (social condition)

Participants performed the psychophysical task together (EXP1-S; **Figure 4-3B**). Their private responses were first made public (colour-coded; 2000 milliseconds) and the response made with higher confidence was then automatically selected as the joint decision (white box; 1000 milliseconds). In the case of a “tie” (i.e., they selected different intervals but indicated the same confidence level), one of the two responses was randomly selected as the joint decision. Next, participants received feedback about the accuracy of each decision (color-coded; joint feedback shown in white; 2000 milliseconds). The feedback text read “correct” or “wrong”; the vertical order of the individual feedback was randomised, whereas the joint feedback always appeared at the centre. Participants were then presented with white text reading “next trial” (1000 milliseconds) before continuing to the next trial. Participants were instructed to maximise the accuracy of their joint decisions; but note that they could only influence the joint decision by means of their confidence reports. After 16 practice trials, participants completed 160 experimental trials.

4.2.1.2.1.4 Isolated task (isolated condition)

Participants performed the psychophysical task without any social interaction (EXP1-I; **Figure 4-3C**). After having made their response, they received feedback about the accuracy of their decision (2000 milliseconds). The feedback text read “correct” or “wrong”. They were then

presented with white text reading “next trial” (1000 milliseconds) before continuing to the next trial. Participants were instructed to maximise the accuracy of their individual decisions. After 16 practice trials, participants completed 160 experimental trials.

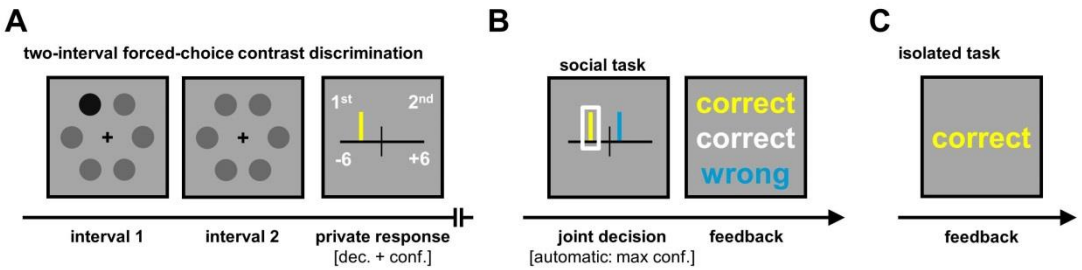


Figure 4-3. Experimental framework. **(A)** Psychophysical task. All experiments involved the same task. On each trial, participants viewed two consecutive intervals, each containing six contrast gratings (here dots). There was a visual target with higher contrast (darker dot) in one of the two intervals. Participants made a response by moving a marker along a scale with a fixed midpoint. There were six steps on either side of the midpoint; the left hand steps were negative values (-6 to -1), whereas the right hand steps were positive values (1 to 6). The sign of a response indicated the decision (negative: 1st; positive: 2nd), and its absolute value indicated the confidence level (1: "unsure"; 6: "certain"). We use response and confidence to refer to signed response values and unsigned response values, respectively. **(B)** The social task. Participants' private responses (colour-coded) were made public, and the response made with higher confidence was automatically selected as the joint decision (white box). They received feedback (colour-coded) before the next trial. Here, the participant of focus (yellow) selected interval 1 (correct) and reported confidence level 3, whereas their partner (blue) selected interval 2 (wrong) and reported confidence level 1. The decision made by the participant of focus was selected as the joint decision (white). **(C)** The isolated task. Participants performed the visual task on their own. Here, the participant of focus (yellow) selected interval 1 (correct) and reported confidence level 3. The displays have been edited for ease of illustration.

4.2.1.2.2 Experiment 2

In Experiment 2 (EXP2), participants only performed the social task from Experiment 1 (**Figure 4-3B**). Participants completed 16 practice trials followed by 384 experimental trials.

4.2.1.2.3 Experiment 3

In Experiment 3 (EXP3), participants only performed the social task from Experiment 1 (**Figure 4-3B**). In contrast to Experiments 1 and 2, they indicated their response using a continuous confidence scale. Participants completed 16 practice trials followed by 384 experimental trials.

4.2.2 Results

4.2.2.1.1 Experiment 1

To address our hypothesis – and following from the results presented in **Chapter 3** – we tested whether there was a (stronger) trial-by-trial relationship between dyad members' confidence reports when they performed the task together as compared to alone. Indeed, there was a serial dependence between dyad members' confidence reports in the social condition only, with their level of confidence on a given trial depending on their partner's level of confidence on the previous trial, after taking into account a set of confounds (**Figure 4-4A**; EXP1-I: $t(59) = .19$, $p = .848$, one-sample *t*-test; EXP1-S: $t(59) = 4.32$, $p < .001$, one-sample *t*-test; comparison of regression coefficients: $t(59) = -3.72$, $p < .001$, paired *t*-test). We then tested whether this trial-by-trial effect was reflected in more global aspects of dyad members' confidence reports – in particular, the relationship between their mean confidence and the proportion of times that they expressed each level of confidence. Indeed, the difference between dyad members' mean confidence was smaller in the social condition (**Figure 4-4B**; $t(29) = 4.20$, $p < .001$, paired *t*-test). Further, the magnitudes of their mean confidence were more correlated in the social

condition, after partialling out their individual accuracy (**Figure 4-4C**; EXP1-I: $r(28) = .28$, $p = .136$; EXP1-S: $r(28) = .69$, $p < .001$; comparison of correlation coefficients: $p = .043$, Fisher transformation). Lastly, the difference between their confidence distributions was smaller in the social condition (**Figure 4-4D**; $t(29) = 3.57$, $p = .001$, paired t-test). Taken together, these results indicated that dyad members – locally and more globally – engaged in confidence matching as a solution to the communication problem.

Next, we considered two alternative explanations of the results of Experiment 1. First, given the well-known correlation between reported confidence and accuracy (Fleming & Lau, 2014), one concern was that the observed confidence matching could be driven by dyad members becoming more similar in terms of their individual performance during social interaction (i.e., "accuracy matching"). However, we could rule out this explanation. While the difference between dyad members' mean confidence was smaller in the social condition (**Figure 4-4B**), the difference between dyad members' accuracy was in fact larger in the social condition (EXP1-I: *mean* ± *SD* = .05 ± .05; EXP1-S: *mean* ± *SD* = .07 ± .06; $t(29) = -2.08$, $p = .046$, paired *t*-test). Further, our analyses took into account or partialled out dyad members' individual accuracy when relevant (**Figure 4-4A** and **Figure 4-4C**). Second, another concern was that the observed confidence matching could be driven by a systematic drift in reported confidence as dyad members vied to influence the joint decision during social interaction (i.e., "confidence escalation"; e.g., Mahmoodi, Bang, Ahmadabadi, & Bahrami, 2013). However, we could also rule out this explanation. Dyad members were on average equally confident in the social and the isolated conditions (EXP1-I: *mean* ± *SD* = 3.00 ± .66; EXP1-S: *mean* ± *SD* = 2.94 ± .56; $t(59) = .88$, $p = .384$, paired *t*-test). Taken together, these results indicated that the social effects reflected a pure second-order phenomenon – that is, direct changes to the mapping of some internal representation of uncertainty onto a confidence report.

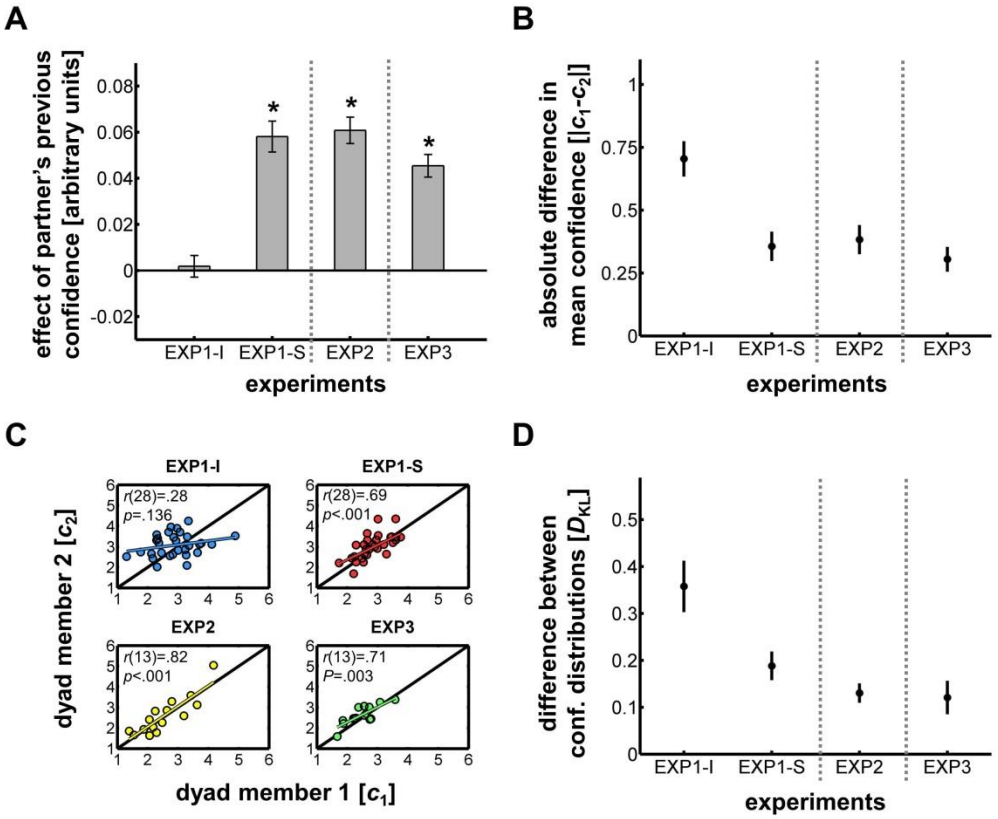


Figure 4-4. Evidence for confidence matching in Experiments 1 to 3. **(A)** Serial dependence in confidence reports. Linear regression coefficients (β-weights; y-axis) encoding how participants' confidence report on trial t was influenced by their partner's confidence report on trial $t-1$ (x-axis); we included the stimulus strength on trial t and $t-1$, the accuracy of each decision on trial $t-1$, and the participant's confidence report on trial $t-1$, as nuisance predictors (see Supplementary Table 1). We calculated significance by testing (one-sample *t*-test) whether the fitted coefficients pooled across all participants were different from 0: * $p < .001$. We used a linear regression model because the responses in Experiment 3 were continuous. However, as discussed in Chapter 3, linear regression models are not suitable for ordinal responses as collected in Experiment 1 and 2. Critically, we found the same effects as above when we used an ordered probit model (see Supplementary Table 2). **(B)** Convergence in mean confidence. The y-axis shows the absolute difference between dyad members' mean confidence, $|c_1 - c_2|$. **(C)** Correlation in mean confidence. The axes of each scatterplot show dyad members' mean confidence (sorted a priori into 1 and 2). We computed correlations after partialling out each dyad member's accuracy from their mean confidence. **(D)** Convergence in confidence distributions. The y-axis shows the Kullback-Leibler divergence, D_{KL}, between dyad members' confidence distributions. In Experiment 3, the continuous confidence values were, for the analyses shown in panel B and C, normalised to the range 1 to 6, and, for the analysis shown in panel D, discretised to the range 1 to 6 in steps of 1 by applying a set of linearly spaced thresholds. In panels A, B and D, each bar/dot is averaged data, and error bars are 1 SEM. In panel C, each dot is a dyad, and each line is the best-fitting line of a robust regression.

4.2.2.1.2 Experiment 2

In Experiment 2 (EXP2), we sought to test whether the observed confidence matching would dissipate with extensive task experience. Participants only performed the social task from Experiment 1, but over a longer period (384 compared 160 trials). None of the social effects were statistically different from those obtained in Experiment 1 (**Figure 4-4ABD**; all $t < 1.30$, all $p > .200$, independent *t*-tests; **Figure 4-4C**: comparison of correlation coefficients: $p = .371$, Fisher transformation). To test the effect of task experience, we divided the data into three time bins (3 x 128 trials). None of the "time-bin" social effects were statistically different from the "overall" social effects obtained in Experiment 1 (**Figure 4-4ABD**: all $t < 1.00$, all $p > .330$, independent *t*-tests; **Figure 4-4C**: comparison of correlation coefficients: all $p > .290$, Fisher transformations). Limiting our analysis to Experiment 2, the level of confidence matching did decrease over time, but only very marginally and only for a subset of the measures (cf. **Figure 4-4A**, $\beta^{\mathrm{bin}=3} - \beta^{\mathrm{bin}=1}$: *mean* ± *SD* = -0.03 ± 0.17, $F(2,14) = 0.36$, $p = .555$, ANOVA; cf. **Figure 4-4B**, $|c_1 - c_2|^{\mathrm{bin}=3} - |c_1 - c_2|^{\mathrm{bin}=1}$: *mean* ± *SD* = 0.10 ± 0.24, $F(2,14) = 2.68$, $p = .124$, ANOVA; cf. **Figure 4-4C**, $r^{\mathrm{bin}=3} - r^{\mathrm{bin}=1}$: .079, all $p > .569$, Fisher transformations; cf. **Figure 4-4D**, $D_{\mathrm{KL}}^{\mathrm{bin}=3} - D_{\mathrm{KL}}^{\mathrm{bin}=1}$: *mean* ± *SD* = 0.08 ± 0.14, $F(2,14) = 5.06$, $p = .041$, ANOVA). Taken together, the results indicated that dyad members, even with extensive task experience, persisted with confidence matching as a solution to the communication problem.

4.2.2.1.3 Experiment 3

In Experiment 3 (EXP3), we sought to test whether the observed confidence matching would disappear with a non-categorical (less language-like) scale. Participants only performed the social task from Experiment 1 (384 trials as in EXP2), but over a longer period (384 compared 160 trials), and they indicated their responses using a continuous scale. None of the social effects were statistically different from those obtained in Experiment 1 or in Experiment 2

(**Figure 4-4ABD**; all $t < 1.40$, all $p > .180$, independent t-tests; **Figure 4-4C**: comparison of correlation coefficients: all $p > .370$, Fisher transformations). Taken together, these results indicated that the engagement in confidence matching was not dependent on the scale on which confidence is mutually expressed.

4.3 Efficacy of confidence matching

We next sought to quantify how confidence matching affects joint accuracy. To address this question, we used a SDT model of confidence to compute how far each dyad – pooling data from Experiments 1 to 3 – was from reaching optimal performance. In broad terms, our model assumed that a participant, on each trial, received noisy sensory evidence, computed some internal estimate of the strength of the sensory evidence in favour of a given decision, and then mapped this internal estimate onto the confidence scale by applying a set of thresholds (subjective mappings). Our model, whose only free parameter is the level of sensory noise, provided a good fit to the data of individual dyad members. Critically, by changing the dyad members' inferred subjective mappings, we could derive for each dyad a landscape of the joint accuracy expected under hypothetical pairs of subjective mappings and thus identify the pair of subjective mappings that would maximise joint accuracy (optimal performance). By means of this analysis, we could compare the observed joint accuracy with that expected under the optimal strategy (see intuition illustrated in **Figure 4-2B**). Here, we will first explain all the steps involved in this analysis and then present the results.

4.3.1 Methods

4.3.1.1 Computational model

We used the general SDT model introduced in **Chapter 2** to model the behaviour of a single agent and to derive the joint accuracy of a pair of agents. However, for analytical simplicity,

we simplified the SDT model in two respects. First, we assumed that the sensory evidence, x, was scalar. Second, as the decision and the confidence level were determined by a single response in our task (response sign and absolute response value), we did not distinguish between a decision variable and a confidence variable. Instead, we assumed that a response was made by thresholding a *single* variable, z, which encoded the evidence in favour of interval 2 (decision variable) as well as the strength of this evidence (confidence variable). For simplicity, we will refer to this variable as the confidence variable.

We assumed that an agent on each trial receives sensory evidence, x, which we modelled as a random sample from a Gaussian distribution, $x \in N(s, \sigma)$. The mean of the distribution, s, is given by the stimulus. Here, we defined the stimulus, s, as the difference in contrast between the second and the first interval at the target location. As such, s will be drawn from the set, $s \in \{-.15,\ -.07,\ -.035,\ -.015,\ .015,\ .035,\ .07,\ .15\}$ – the sign of s indicates the target display ($s < 0$: interval 1; $s > 0$: interval 2) and its absolute value indicates the value that was added to the baseline contrast of the visual target. The standard deviation of the Gaussian distribution, σ, describes the level of sensory noise and is the same for all stimuli. If σ is high (low), then the sensory evidence, x, would be sampled from a broad (narrow) distribution and thus have high (low) variability for any given value of s.

We modelled the confidence variable, z, as the raw value of the sensory evidence, $z = x$. The confidence variable thus ran from low (negative) values, indicating strong evidence that the visual target was in interval 1, through intermediate values, indicating uncertainty, to high (positive) values, indicating strong evidence that the visual target was in interval 2. We note that we assumed that the *actual* confidence variable was computed as the probability that the visual target was in the second interval, $z^* = P(s > 0|x)$, which is a monotonic transformation of the sensory evidence, x. However, for analytical simplicity (calculations and derivations shown below), we modelled the confidence variable as $z = x$.

We assumed that the agent maps the confidence variable onto a response, r, by applying a set of thresholds, $\boldsymbol{\theta}$. As in our experiments, the responses range from -6 to -1 and 1 to 6 – with the sign of the response indicating the decision ($r < 0$: interval 1; $r > 0$: interval 2) and its absolute value indicating the confidence level ($c = |r|$). The thresholds, $\boldsymbol{\theta}$, were set in such a way that each response, r, is made a desired proportion of times, $p(r)$.

We assumed that the probability that the agent makes a particular response, r, equals the probability of all confidence variables that map to that particular response,

$$p_i = P(r = i) = \begin{cases} P(z \leq \theta_{-6}) & i = -6 \\ P(\theta_{i-1} < z \leq \theta_i) & -6 < i \leq -1, \quad 2 \leq i < 6 \\ P(\theta_{-1} < z \leq \theta_1) & i = 1 \\ P(z > \theta_5) & i = 6 \end{cases} \qquad \textbf{Eq. 4-1}$$

Note that there is no criterion θ_6, because $r = 6$ is determined by z exceeding θ_5. It follows that, $P(r \leq i) = P(z \leq \theta_i)$, where $i = -6, \ldots -1, 1, \ldots 5$, which can be rewritten, $cdf_{(r)}(i) = cdf_{(z)}(\theta_i)$ – where cdf is the cumulative distribution function (CDF). Computing the thresholds for a response distribution is then straightforward, $\theta_i = invcdf_{(z)}(cdf_{(r)}(i))$, where $invcdf$ is the inverse of the CDF. Specifically, $cdf_{(z)}(x) = \sum_{k=1}^{8} \Phi(\frac{x-\mu_k}{\sigma})$ and $cdf_{(r)}(i) = \sum_{l \leq i} p_l$, where μ_k denotes the mean of the k'th Gaussian distribution (cf. the eight different stimuli, s, corrupted by sensory noise, σ), where $k = 1,2, \ldots ,8$, and Φ denotes the CDF of a standard Gaussian distribution. We thus calculated the thresholds using MATLAB's (Mathworks Inc) "fzero" function as the unique zero x of the function $\sum_{j=1}^{8} \Phi(\frac{x-\mu_j}{\sigma}) - \sum_{l \leq i} p_l$.

4.3.1.2 Deriving accuracy

We calculated the accuracy rate of an agent, given a level of sensory noise, σ, and a response distribution, p_i (where $i = \pm 1,2, \ldots , 6$), across all trials, according to the following steps: We first calculated the thresholds, $\boldsymbol{\theta}$, as described above. We then calculated for, each Gaussian

distribution (indexed k – cf. the eight different stimuli, *s*, corrupted by sensory noise, σ) the predicted response distribution as

$$p_{i,k} = \begin{cases} \Phi\left(\frac{\theta_{-6} - \mu_k}{\sigma}\right) & i = -6 \\ \Phi\left(\frac{\theta_i - \mu_k}{\sigma}\right) - \Phi\left(\frac{\theta_{i-1} - \mu_k}{\sigma}\right) & -6 < i < 6 \\ 1 - \Phi\left(\frac{\theta_5 - \mu_k}{\sigma}\right) & i = 6 \end{cases} \quad \textbf{Eq. 4-2}$$

The accuracy rates for the Gaussian distributions with positive and negative mean were calculated as $\sum_{i=1}^{6} p_{i,k}$ and $\sum_{i=-6}^{-1} p_{i,k}$, respectively. The overall accuracy rate for an agent is then the arithmetic mean of the accuracy rates across the eight Gaussian distributions.

We calculated the joint accuracy rate for a pair of agents (a dyad) according to the following steps: We first computed the predicted response distributions for each Gaussian distribution as described above; here denoted as $p_{i,k,m}$, where $m = 1{,}2$ denotes the dyad member. For each Gaussian distribution we then calculated the probability that dyad member 1 makes response i_1 and dyad member 2 makes response i_2 as $\tilde{p}_{i_1,i_2,k} = p_{i_1,k,1} * p_{i_2,k,2}$. The response combinations that necessarily (with probability 1) yield a correct response for a positive-mean Gaussian distribution is given by: $A_+ = \{(i_1, i_2) |\ (i_1 > 0 \text{ and } i_2 > 0) \text{ or } (i_1 > 0 \text{ and } -i_1 < i_2 < 0\ ) \text{ or } (i_2 > 0 \text{ and } -i_2 < i_1 < 0\ )\}$. The response combinations that necessarily yield a correct response for a negative-mean Gaussian distribution is given by $A_- = \{(i_1, i_2) | - (i_1, i_2) \epsilon A_+\}$. Additionally, response combinations $A_{+/-} = \{i, i_2 |\ (i_1 = -i_1)\}$ – that is, trials in which the dyad members select different intervals but with the same level of confidence – the dyad is correct with chance probability 50%. Thus, the joint accuracy rate for a positive or a negative mean Gaussian distribution is $\sum_{A+} \tilde{p}_{i_1,i_2,k} + .5 * \sum_{A_{+/-}} \tilde{p}_{i_1,i_2,k}$, $\sum_{A_-} \tilde{p}_{i_1,i_2,k} + .5 * \sum_{A_{+/-}} \tilde{p}_{i_1,i_2,k}$, respectively. The joint accuracy rate for a pair of agents is then the arithmetic mean across the eight Gaussian distributions.

4.3.1.3 Model fitting

We fitted our model to each participant by searching exhaustively through parameter space (steps of .001) for the level of sensory noise that minimised the squared error between the observed accuracy and that predicted by the model. We note that our model only has one free parameter; for each participant, we parameterised the thresholds governing the mapping from the confidence variable onto a response directly using their observed response distribution. To evaluate the goodness of the model fit for participants, we computed psychometric functions (**Figure 4-5**) and response distributions (**Figure 4-6**) for each target contrast as predicted by the model. Taking the simplicity of our model into account, there was good agreement between the empirical data and that predicted by the model. Critically, the observed individual accuracy, a_{sbj}, was not different from that predicted by the model, a_{agent} ($a_{\mathrm{sbj}} - a_{\mathrm{agent}}$: *mean* ± *SD* = .00 ± .00). To evaluate the goodness of the model fit for dyads, we computed the joint accuracy expected under dyad members' observed response distributions, a_{fit}. Our model slightly overestimated the observed joint accuracy, a_{emp} ($a_{\mathrm{emp}} - a_{\mathrm{fit}}$: mean ± SD = -.01 ± .02). We take this model misestimation into account when quantifying how far each dyad was from reaching optimal performance.

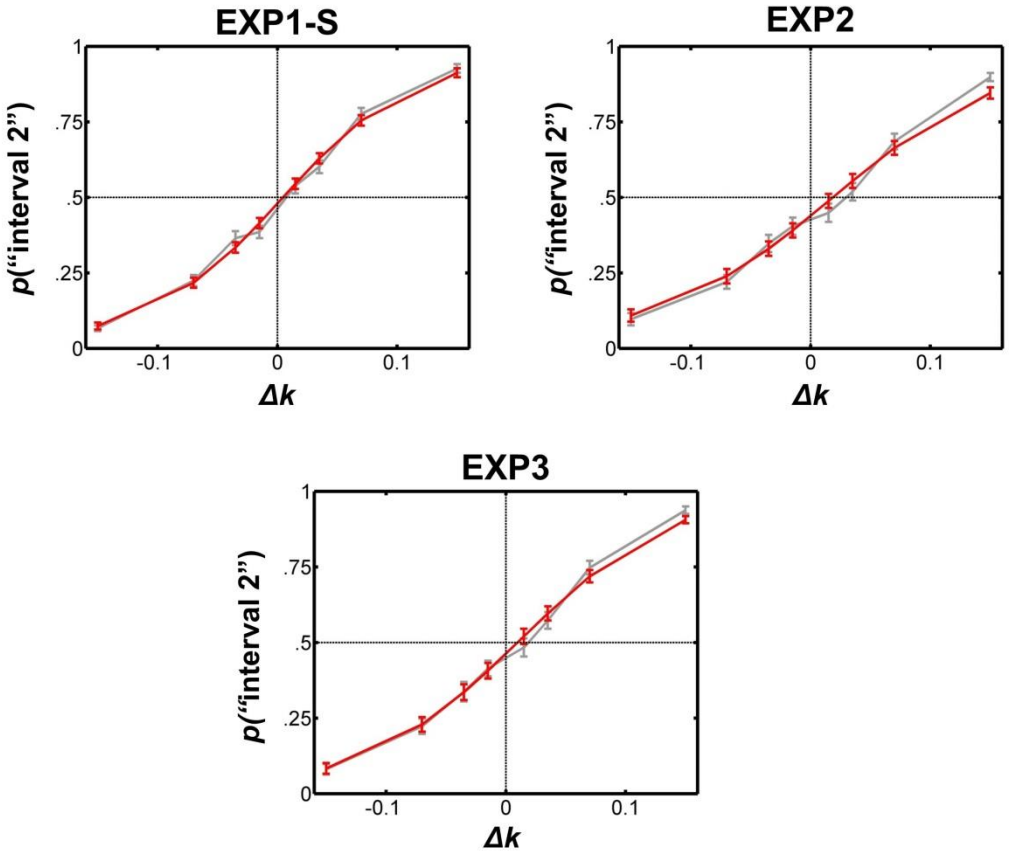


Figure 4-5. Comparison of empirical psychometric functions with those of the model for Experiments 1 to 3. The x-axis shows the contrast difference between the second and the first interval at the target location, Δk. The y-axis shows the proportion of times that the interval 2, $p(\text{"interval 2"})$, was selected for a given value of Δk. We computed R^2-values across participants (*mean* ± *SD*) to evaluate the model fits. EXP1-S: R^2 = .87 ± .17; EXP2: R^2 = .89 ± .09; EXP3: R^2 = .95 ± .030. Grey: average empirical data. Red: average model data (best fits). Error bars are 1 SEM.

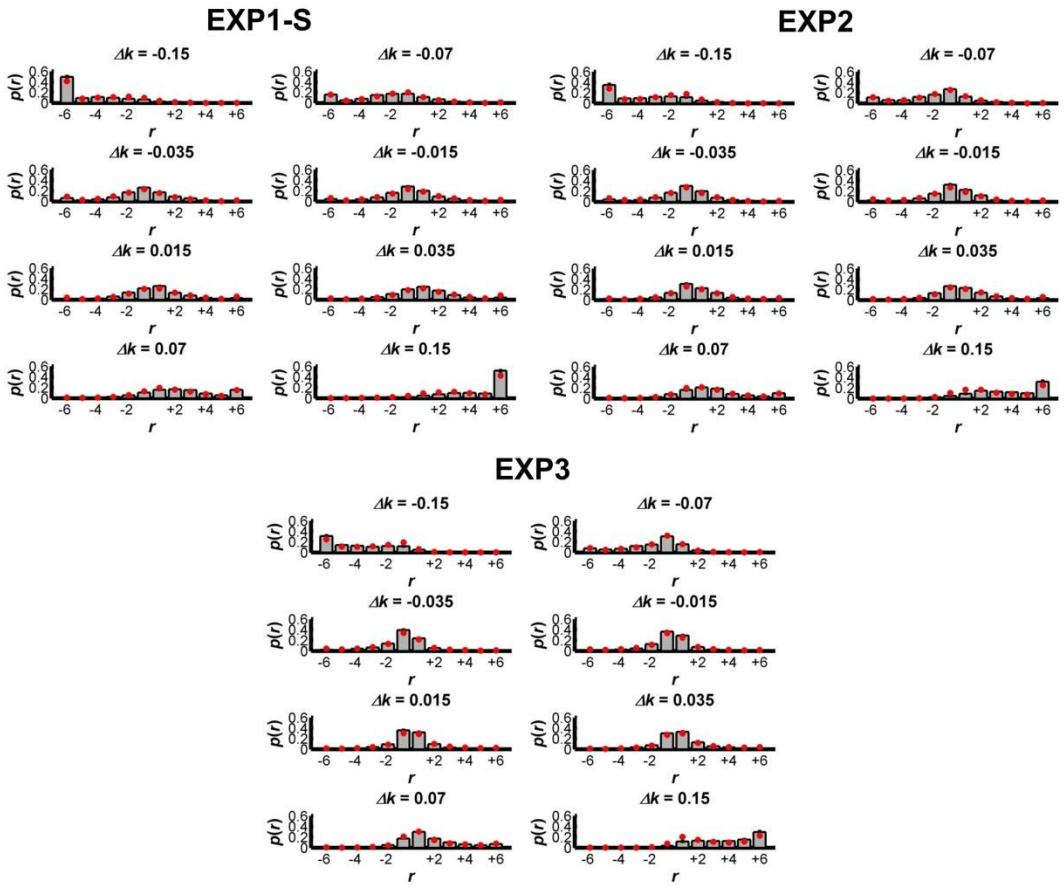


Figure 4-6. Comparison of empirical response distributions for each stimulus with those of the model for Experiments 1 to 3. Each plot shows the response distribution for a particular value of Δk (i.e., the contrast difference between the second and the first display at the target location). The x-axis shows the response values (from -6 to 1 and 1 to 6). The y-axis shows the proportion of times that each response value was selected, $p(r)$. We computed R^2-values across participants (*mean* ± *SD*) to evaluate the model fits. EXP1-S: R^2 = .55 ± .27; EXP2: R^2 = .71 ± .20; EXP3: R^2 = .81 ± .14. Grey: average empirical data. Red: average model data (best fits). Error bars are 1 SEM.

4.3.1.4 Confidence landscapes

We used our model to calculate how the expected joint accuracy of a pair of dyad members varies as a function of their subjective mappings. We first set their levels of sensory noise as fitted to the data. We then derived their expected joint accuracy under different pairs of response distributions, with each distribution associated with a specific mean confidence. In our model, for a given level of sensory noise, specifying a response distribution is equivalent to specifying a subjective mapping. While there is an infinite number of distributions that can give arise to the same mean confidence, this is not the case when only considering one family of distributions. We therefore limited our analysis to maximum entropy distributions, from *mean confidence* = 1 to *mean confidence* = 6 in steps of .1 (see below for mathematical details; the set of maximum entropy distributions are shown in **Supplementary Figure 3**).

We defined the joint accuracy expected under the optimal strategy as the maximum value of a given confidence landscape (see **Figure 4-7A** in the Results section for two examples; separate confidence landscapes for each dyad are shown in **Supplementary Figure 4**). We note that this value does not index the pair of response distributions under which dyad members' subjective mappings are perfectly aligned. First, given our task design, dyad members must maximise the alignment of their subjective mappings while minimising the number of confidence ties (i.e., the number of times that they select different displays but report the same confidence level). Dyad members' subjective mappings are – for trivial reasons – perfectly aligned under response distributions that give rise to a mean confidence of 6 (i.e., their respective thresholds are placed at $\pm\infty$ so that they always respond ± 6). Under this pair of subjective mappings, despite being fully aligned, all disagreements would be resolved randomly and thus yield a low expected joint accuracy. Thus, the maximum value of a given landscape indicates the pair of response distributions under which dyad members would strike an optimal balance between maximising the alignment of their subjective mappings and minimising the number of confidence ties. Second, we did not derive the joint accuracy expected under all possible

subjective mappings (we only considered 51 maximum entropy distributions). Thus, there will almost certainly always be another pair of response distributions, even if in close vicinity to those indexed by the maximum value of a given confidence landscape, under which dyad members' subjective mappings would be more aligned.

4.3.1.5 Maximum entropy distributions

The entropy of a distribution, $p_i, i = 1,2,..,6$ is defined by $-\sum_{i=1}^{6} p_i ln p_i$. Therefore, a maximum entropy distribution constrained to have a mean c is defined as the distribution that solves $\max(-\sum_{i=1}^{6} p_i ln p_i)$, subject to the constraints $\sum_{i=1}^{6} p_i = 1$, $\sum_{i=1}^{6} i p_i = c$ and $p_i \geq 0$. For the case $c = 1$ the unique distribution that satisfies the constrains is $\underline{p} = (1,0,0,0,0,0)$. Similarly, for $c = 6$, $\underline{p} = (0,0,0,0,0,1)$ is the solution. For any intermediate value for the mean, $1 \leq c \leq 6$, the solution can be found using Lagrange multipliers: $\forall i: \ln(p_i) + 1 = \lambda_1 + i\lambda_2$. It follows that $\ln\left(\frac{p_i}{p_{i-1}}\right) = \lambda_2$ or that $p_i = e^{\lambda_2} p_{i-1}$. Thus, $\{p_i\}_{i=1}^{6}$ is a geometric sequence – that is, $p_i = e^{(i-1)\lambda_2} p_1$. The constraint $\sum_{i=1}^{6} p_i = 1$, implies that $p_1 = (\sum_{i=0}^{5} e^{i\lambda_2})^{-1}$ – that is, that $p_i = \frac{e^{(i-1)\lambda_2}}{(\sum_{j=0}^{5} e^{j\lambda_2})}$ for some constant λ_2. Thus, λ_2 is the only parameter needed to specify a set of maximum entropy distributions. We found λ_2 by solving the constraint $\frac{\sum_{j=0}^{5}(j+1)e^{j\lambda_2}}{\sum_{j=0}^{5} e^{j\lambda_2}} = c$ using MATLAB's (Mathworks Inc) "fzero" function. We note that we first created the maximum entropy distributions in unsigned space (confidence values) and then transformed the distributions into signed space (response values) for the reported analyses.

4.3.1.6 A note on comparing observed to optimal joint accuracy

For the analysis shown in **Figure 4-8B** – in which we compare the observed joint accuracy to that expected under the optimal strategy, $a_{\mathrm{emp}}/a_{\mathrm{opt}}$, as a function of how similar dyad

members were in terms of their accuracy, $a_{\min}/a_{\max}$ – we computed a_{opt} as: $a_{\text{opt}} - (a_{\text{fit}} - a_{\text{emp}})$, where a_{fit} is the joint accuracy expected under dyad members' observed response distributions. First, this correction anchors a_{opt} to the empirical data – thus making a_{emp} comparable to a_{opt} – by taking into account any model misestimation. Second, this correction allows us to rule out that a_{opt} diverged from a_{emp} simply because a_{opt} was derived under a model which never makes "errors". We note that the correction term, $a_{\text{fit}} - a_{\text{emp}}$, did not correlate with how similar dyad members were in terms of their accuracy, $a_{\min}/a_{\max}$ ($r(58) = .23$, $p = .072$).

The analysis shown in **Figure 4-8B** may, however, be subject to two additional confounds. First, a_{opt} might diverge from a_{emp} simply because a_{opt} was computed under maximum entropy distributions, which – for any given mean confidence – minimise the number of ties (i.e., the number of times that dyad members select different displays but report the same confidence level). Second, the level of optimality itself, $a_{\text{emp}}/a_{\text{opt}}$, might be smaller for similar dyad members, $a_{\min}/a_{\max} \approx 1$, than dissimilar dyad members, $a_{\min}/a_{\max} \ll 1$, due to range effects. For the most dissimilar dyad members, the difference between the minimum and the maximum value of a given confidence landscape is about .1. In contrast, for the most similar dyad members, the difference between the minimum and the maximum value of a given confidence landscape is only about .05. Thus, there was less room for the latter to deviate from the joint accuracy expected under the optimal strategy.

We therefore repeated the analysis shown in **Figure 4-8B** but re-computed the level of optimality as: $(a_{\text{maxent}} - a_{\text{anti}})/(a_{\text{opt}} - a_{\text{anti}})$, where a_{anti} is the lowest value of a confidence landscape and a_{maxent} is the value of a landscape coordinate indexed by dyad member's observed mean confidence. First, by using a_{maxent} instead of a_{emp} in the numerator, we control not only for model misestimation but also for the use of maximum entropy distributions. Second, the subtraction of a_{anti} in both the numerator and

denominator controls for range effects. In line with the analysis shown in **Figure 4-8B**, we found a positive correlation between this corrected measure of optimality and how similar dyad members were in terms of their accuracy, $a_{\min}/a_{\max}$ ($r(58) = .52$, $p < .001$).

4.3.2 Results

Our simulations confirmed that confidence matching is sub-optimal: the degree of confidence matching should depend on how similar dyad members are in terms of their individual accuracy (compare left and right inset in **Figure 4-7A**). Theoretically, this sub-optimality arises because, to match their use of the confidence scale, dissimilar dyad members must in fact misalign their subjective mappings, thus increasing miscommunication about who is more likely to be correct on a given trial (see the intuition illustrated in **Figure 4-2E**). Indeed, in our simulations, dyad members' subjective mappings become less aligned under confidence matching as the difference in their levels of sensory noise increases (**Figure 4-7B**).

While we observed that the more accurate dyad members were (slightly) more confident than their less accurate partners (more dots above vertical line in **Figure 4-8A**; $t = 8.73$, $p < .001$, paired *t*-test), their differences in mean confidence did not scale with their differences in individual accuracy as expected under the optimal strategy (**Figure 4-8A**, $r(58) = .16$, $p = .213$). Given our simulation results, we would therefore predict that similar dyad members were closer to reaching optimal performance. Indeed, the divergence between the observed joint accuracy and that expected under the optimal strategy was smaller for similar dyad members (**Figure 4-8B**; $r(58) = .68$, $p < .001$). Taken together, these results indicated that confidence matching was mainly costly for dissimilar dyad members.

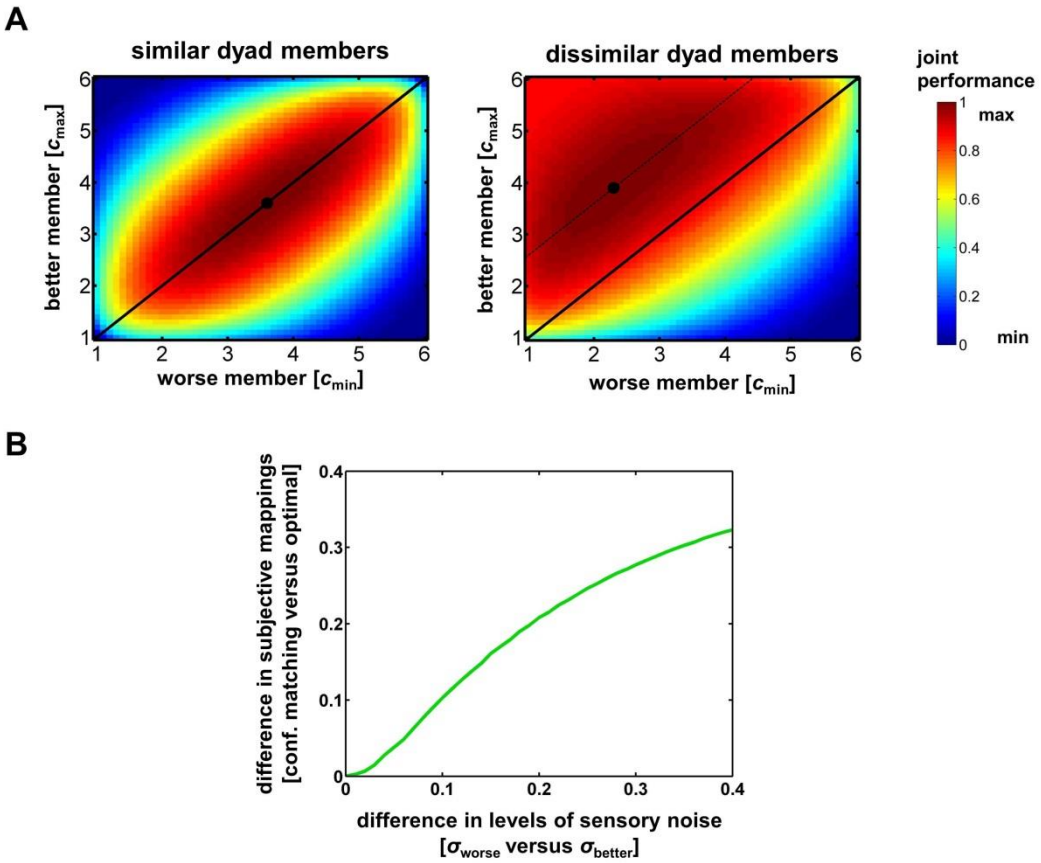


Figure 4-7. Efficacy of confidence matching: simulations. **(A)** Expected joint accuracy (colour scale – normalised to the range 0-1) varies as a function of the mean confidence of the less accurate ($c_{\min}$; x-axis) and the more accurate ($c_{\max}$; y-axis) dyad member. For dyad members with similar levels of accuracy (*left*), the expected joint accuracy reaches its maximum value (black dot: optimal strategy) when their mean confidence is matched; not so for dyad members with different levels of accuracy (*right*). We could specify a subjective mapping using a single value (mean confidence) by limiting our analysis to one family of response distributions (maximum entropy). **(B)** Difference in the subjective mappings under confidence matching and the optimal strategy varies as a function of dyad members' levels of sensory noise. For a given pair of dyad members, we computed the sum of squared differences between their subjective mappings (the location of their respective thresholds in $P(s > 0|x)$-space) under the optimal strategy (the pair of maximum entropy distributions indexed by the maximum value of a given confidence landscape) and under confidence matching (here, as an approximation, a pair of maximum entropy distributions associated with a mean confidence of 3.5). The x-axis shows the difference (optimal minus confidence matching) between the sum of squared differences computed separately for each strategy. We fixed the level of sensory noise for the better dyad member, σ_{better}, at .1, and then, starting at .1, increased the level of sensory noise for the worse dyad member, σ_{worse}, so as to cover differences in sensory noise, $\sigma_{\text{worse}} - \sigma_{\text{better}}$, up to .4. The (0,0)-coordinate corresponds to a pair of dyad members whose levels of sensory noise are: σ_{worse} = .1 and σ_{better} = .1. The plots show simulated data. See Section 4.3.1: Methods for full details.

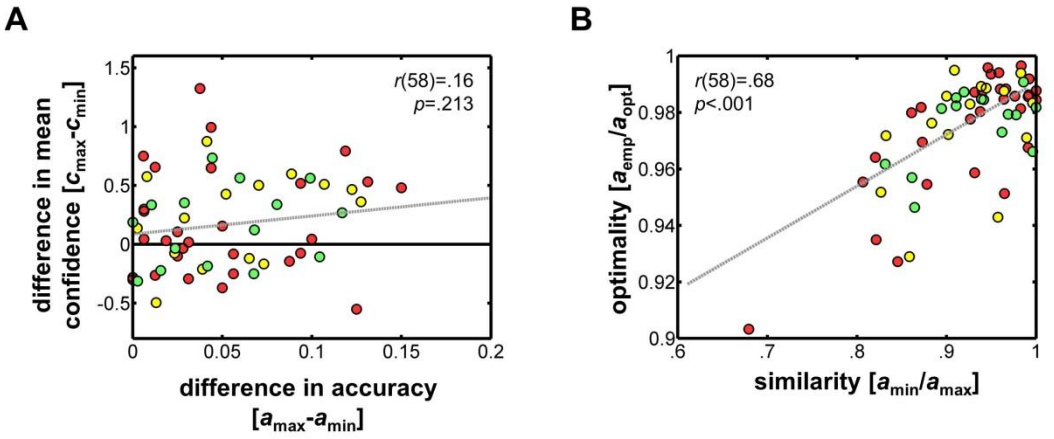


Figure 4-8. Efficacy of confidence matching: results from Experiments 1 to 3. **(A)** Scaling relative confidence with relative accuracy. The axes show the difference in accuracy ($a_{\max} - a_{\min}$; x-axis) and in mean confidence ($c_{\max} - c_{\min}$; y-axis) between the better (max) and the worse (min) dyad member. The pattern was the same when the correlation was computed separately for each experiment (all $r <$ .340, all $p >$ .200). **(B)** Optimality. The x-axis shows how similar dyad members were in terms of their accuracy, $a_{\min}/a_{\max}$. The y-axis shows the observed joint accuracy (a_{emp}) normalised by the joint accuracy expected under the optimal strategy (a_{opt}), $a_{\text{emp}}/a_{\text{opt}}$. For full details, see Section 4.3.1: Methods; in particular, for a control analysis, see Section 4.3.1.6: A note on comparing observed to optimal joint accuracy. In Experiment 3, the continuous confidence values were, for the analysis shown in panel A, normalised to the range 1 to 6, and, for the analysis shown in panel B, discretised to the range 1 to 6 in steps of 1 by applying a set of linearly spaced thresholds. In both panels, each dot is a dyad, and each line is the best-fitting line of a robust regression. As in Figure 4-4C, red: EXP1-S, yellow: EXP-2, green: EXP-3.

4.4 Confidence matching in stereotypical cooperative situations

We next sought to unpack how confidence matching affects joint accuracy for dissimilar dyad members. Our simulations predicted that confidence matching, while costly for dyad members who are "well-calibrated" to each other (i.e., when the less accurate is the less confident; squares in **Figure 4-9**), should in fact be beneficial for dyad members who are "poorly calibrated" to each other (i.e., when the less accurate is the more confident; circles in **Figure 4-9**).[15] To test this prediction, we conducted a fourth experiment in which such cooperative situations were created by pairing *real* participants with stereotypical *virtual* partners.

[15] In this chapter, we will use *well-calibrated* and *poorly calibrated* in a social sense: that is, to describe the relationship between the individual accuracy and the mean confidence of a pair of individuals.

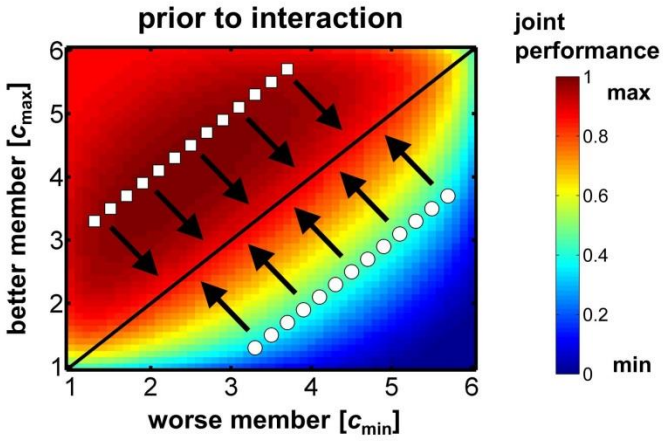


Figure 4-9. Stereotypical social scenarios. The symbols indicate dyad members who prior to interaction are well-calibrated to each other (squares above diagonal) or poorly calibrated to each other (circles below diagonal). The arrows indicate the movement through the confidence landscape expected under confidence matching: well-calibrated dyad members would incur a relative cost, whereas poorly calibrated dyad members would gain a relative benefit. The plot shows simulated data.

4.4.1 Methods

4.4.1.1 Participants

Participants were recruited among students (age range: 18-30) at the University of Oxford. In total, 38 participants (19 male) took part in the study. All participants reported normal or corrected vision. All participants provided informed consent and were reimbursed for their participation (£10/hour for a total of 5 hours). All experiments were approved by the local ethics committee.

4.4.1.2 Experimental details (Experiment 4)

Participants took part in the experiment as part of a large group (2 x 19 people). Participants were randomly allocated to private work stations in a large computer laboratory. The experiment consisted of two sessions. In the first session, participants performed the isolated task (EXP4-I; **Figure 4-3C**). In the second session, participants performed the social task (EXP4-S; **Figure 4-3C**) over four blocks. For each block, they were told that they were randomly paired with one of the other participants. In reality, they were, for each block, paired with a virtual (computer-generated) partner. We created the four virtual partners using participants' data from the isolated task (the *baseline* data), and we varied their accuracy – Accuracy Low (AL) or High (AH) – and confidence – Confidence Low (CL) or High (CH) – in a 2-by-2 within-subject design (**Figure 4-10**). This procedure gave rise to four stereotypical virtual partners: ALCL, ALCH, AHCL and AHCH. Overall, there were thus two conditions in which dyad members prior to interaction were meant to be poorly calibrated to each other (ALCH, AHCL), and two conditions in which they were meant to be well-calibrated to each other (ALCL, AHCH). Participants performed 1160 experimental trials (isolated: 200 trials; social: 4 x 240 trials). The order of the social conditions (i.e., virtual agents) was randomised across participants. At the beginning of each session, participants completed 16 practice trials. Participants were

debriefed about the deception after the experiment; no one had noticed the deception or decided to withdraw from the study.

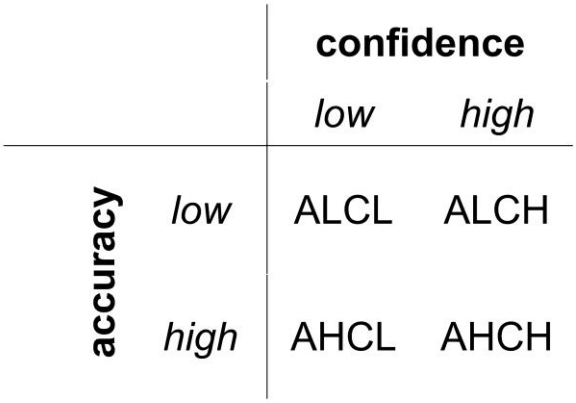


Figure 4-10. Virtual partners in Experiment 4. The virtual partners were tuned to reflect a 2 (Accuracy Low vs. High) by 2 (Confidence Low vs. High) within-subject design.

4.4.1.3 Virtual partners

4.4.1.3.1 Computational model

We used our model described in **Section 4.3.1.1: Computational model** to create the four virtual partners. To specify the model for the virtual partners of a given participant, we fitted our model to each participant's data from the isolated task (i.e., the baseline data) while participants were waiting to start the social task. As in Experiments 1 to 3, there was good agreement between the empirical data and that predicted by the model (**Figure 4-11** and **Figure 4-12**). We then used the fitted level of sensory noise, σ_{sbj}, to specify the accuracy of the respective virtual partners. For the "low accuracy" virtual partners, we set the level of sensory noise to be 50% higher than that of the participant, $\sigma_{\mathrm{low}} = \sigma_{\mathrm{sbj}} + .5 * \sigma_{\mathrm{sbj}}$. Conversely, for the "high accuracy" virtual partners, we set the level of sensory noise to be 50% lower than that of the participant, $\sigma_{\mathrm{high}} = \sigma_{\mathrm{sbj}} - .5 * \sigma_{\mathrm{sbj}}$. We then used two generic confidence distributions to specify the confidence of the respective virtual partners. For the first group of participants (19 in total), we used the following confidence distributions: $p_{\mathrm{low}} = [.35,\ .27,\ .15,\ .11,\ .08,\ .04]$ with *mean* = 2.42 and $p_{\mathrm{high}} = [.04,\ .08,\ .11,\ .15,\ .27,\ .35]$ with *mean* = 4.56 – where the first element corresponds to confidence level 1, the second element corresponds to confidence level 2, and so forth. For the second group of participants (19 in total), we used the following confidence distributions: $p_{\mathrm{low}} = [.44,\ .30,\ .14,\ .06,\ .02,\ .04]$ with *mean* = 2.04 and $p_{\mathrm{high}} = [.10,\ .12,\ .14,\ .16,\ .18,\ .26]$. The confidence distributions for the second group of participants were set so as to better fit those of the first group of participants. We transformed the distributions from unsigned space (confidence values) to signed space (response values) so as to generate the virtual partners' responses. We used generic confidence distributions because we have observed that some participants only report either low (1-2) or high confidence (5-6) when tested in isolation. In such cases, tuning the virtual partners' confidence to the participant would have resulted in the former almost exclusively reporting either 1 ("low confidence" partners) or 6 ("high confidence" partners).

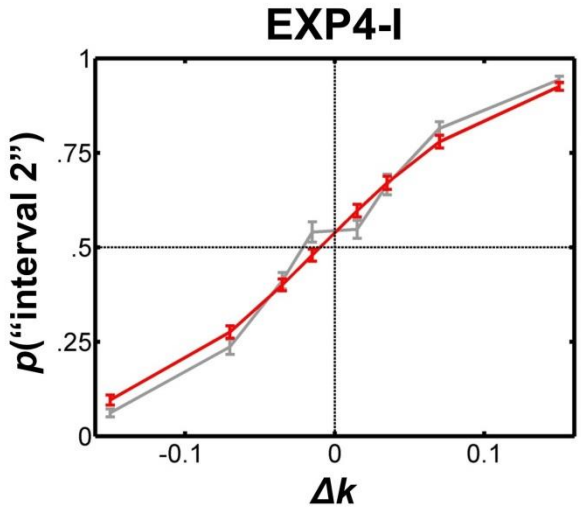


Figure 4-11. Comparison of empirical psychometric function with that of the model for Experiment 4. The x-axis shows the contrast difference between the second and the first interval at the target location, Δk. The y-axis shows the proportion of times that the interval 2, $p(\text{"interval 2"})$, was selected for a given value of Δk. We computed R^2-values across participants (*mean* ± *SD*) to evaluate the model fits. EXP4-I: R^2 = .844± .14. Grey: average empirical data. Red: average model data (best fits). Error bars are 1 SEM.

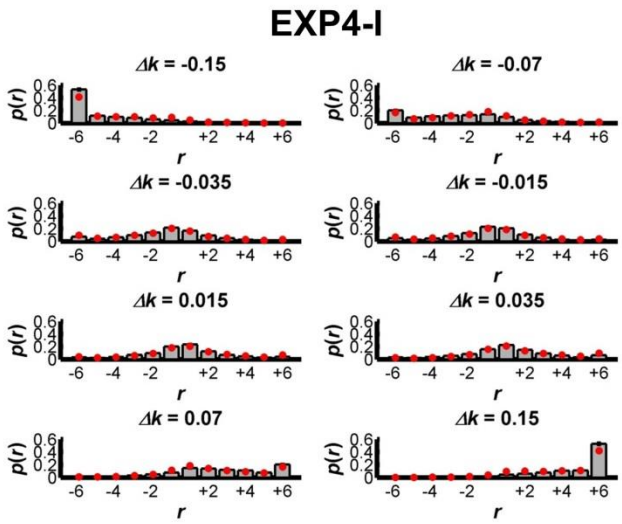


Figure 4-12. Comparison of empirical response distributions for each stimulus with those of the model for Experiments 4. Each plot shows the response distribution for a particular value of Δk (i.e., the contrast difference between the second and the first display at the target location). The x-axis shows the response values (from -6 to 1 and 1 to 6). The y-axis shows the proportion of times that each response value was selected, $p(r)$. We computed R^2-values across participants (*mean* ± *SD*) to evaluate the model fits. EXP4-I: R^2 = .55 ± .27. Grey: average empirical data. Red: average model data (best fits). Error bars are 1 SEM.

4.4.1.3.2 Trial-by-trial responses

To generate the trial-by-trial responses of a given virtual partner, we first created the trial-by-trial sequence of stimuli to be shown to the participant in a given social block. We then created a trial-by-trial sequence of random values, with each value randomly drawn from a Gaussian distribution whose mean was given by the stimulus on the corresponding trial and whose standard deviation was set as above (cf. sensory evidence). Next, we transformed the sequence of random values into trial-by-trial responses by applying – post-hoc – a set of thresholds that gave rise to a desired response distribution set as above. Lastly, to mimic lapses of attention and response errors, we randomly selected a response (-6 to -1 and 1 to 6) on 5% of the trials (12 out 240 trials). On half of the trials, an agent's response arrived earlier than that of the participant (i.e., the participant saw their partner's response as soon as they had made their own response). On the other half of the trials, an agent's response arrived later than that of the participant (i.e., the participant had to wait for their partner's response). This delay was drawn uniformly from the range 500-1500 milliseconds.

4.4.1.3.3 Effect of manipulation

We computed summary statistics to test whether our manipulation worked (cf. the accuracy and the mean confidence of a participant at baseline relative to the accuracy and the mean confidence of a given virtual partner). We obtained the desired effects (**Table 3-3**).

	data	condition	*mean* accuracy (*SD*)	*mean* confidence (*SD*)
GROUP 1	participant	*isolated*	.711 (.052)	2.697 (.613)
		ALCL	.727 (.062)	3.079 (.649)
		ALCH	.717 (.068)	3.566 (.675)
		AHCL	.714 (.065)	2.659 (.538)
		AHCH	.720 (.053)	3.313 (.610)
	virtual partner	*ALCL*	.653 (.054) *$	2.427 (.092) *$
		ALCH	.659 (.053) *$	4.582 (.096) *$
		AHCL	.792 (.059) *$	2.417 (.069) *$
		AHCH	.784 (.054) *$	4.572 (.097) *$
GROUP 2	participant	*isolated*	.715 (.042)	2.832 (.702)
		ALCL	.713 (.056)	3.121 (.776)
		ALCH	.730 (.055)	3.685 (.669)
		AHCL	.726 (.045)	2.796 (.512)
		AHCH	.719 (.063)	3.412 (.646)
	virtual partner	*ALCL*	.639 (.046) *$	2.043 (.081) *$
		ALCH	.666 (.051) *$	4.083 (.100) *$
		AHCL	.800 (.033) *$	2.046 (.056) *$
		AHCH	.789 (.041) *$	4.023 (.067) *$

Table 4-1. Summary statistics for Experiment 4. The summary statistics are shown separately for the first and the second group of participants (19 in each group). We used paired *t*-tests (one-tailed) to test whether the data of participants were significantly different from that of the virtual partners. *: data of virtual partners significantly different from that of participants at baseline (isolated task). $: data of virtual partners significantly different from that of participants in a given social condition (social task).

4.4.1.4 Joint accuracy expected prior to interaction

We computed the joint accuracy expected prior to interaction, a_{iso}, by – iteratively – playing out responses of a given virtual partner – for whom we knew the generative model – against those of a given participant in the isolated task. Briefly, for each iteration, we provided the virtual partner with the observed sequence of stimuli, simulated a sequence of trial-by-trial responses, and then determined the joint decision on each trial by selecting the decision made with higher confidence. In the case of a tie (i.e., they reported the same level of confidence but selected different displays), one of the two decisions was randomly selected as the joint decision. We averaged a_{iso} across $10{,}000$ iterations ("experiments") to arrive at a stable value (the virtual agents' response on a given trial varied due to sensory noise and the small number of random responses). To control for fluctuations in individual performance between conditions, we normalised the joint accuracy expected prior to interaction, a_{iso}, and the joint accuracy observed in a given social condition, a_{soc}, using the mean of the accuracy of the individual responses (participant and virtual partner) in the respective datasets: a_{iso}: $a_{\mathrm{iso}}/(.5 * (a_{\mathrm{sbj}}^{\mathrm{isolated}} + a_{\mathrm{agent}}^{\mathrm{iteration}}))$ and a_{soc}: $a_{\mathrm{soc}}/(.5 * (a_{\mathrm{sbj}}^{\mathrm{social}} + a_{\mathrm{agent}}^{\mathrm{social}}))$ – note that we computed the former term for each iteration ("experiment") and then averaged across iterations.

4.4.2 Results

To test our hypothesis, we used robust linear regression to estimate the relative influence of confidence matching (Δm), calibration (cal), and their interaction ($\Delta m * cal$) on a measure of joint accuracy (Δa_{dyad}) that had been normalised using participants' baseline data (**Figure 4-13**). As predicted, the effect of confidence matching on joint accuracy depended on dyad members' initial calibration (**Figure 4-13**; $\Delta m * cal$: $t(148) = 4.68$, $p < .001$). Confidence matching had a negative effect for well-calibrated dyad members (blue line; Δm: $t(148) = -2.06$, $p = .041$, simple effect), but a positive effect for poorly calibrated dyad members (red

line; Δm: $t(148) = 4.42$, $p < .001$, simple effect). Intriguingly, while participants showed evidence of confidence matching when they were well-calibrated to their partner (averaged across ALCL and AHCH; Δm: *mean* ± SD = 0.18 ± 0.41; $t(37) = 2.66$, $p = .012$, one-sample *t*-test) as well as when they were poorly calibrated to their partner (averaged across ALCH and AHCL; Δm: *mean* ± *SD* = 0.43 ± 0.34; $t(37) = 7.82$, $p < .001$, one-sample *t*-test), they showed a larger degree of confidence matching in the latter scenarios ($t(75) = -4.04$, $p < .001$, paired *t*-test). Taken together, these results indicated that dyad members did not blindly match each other's confidence, but had some insight into the extent to which confidence matching solves the communication problem and adjusted their use of the strategy accordingly.

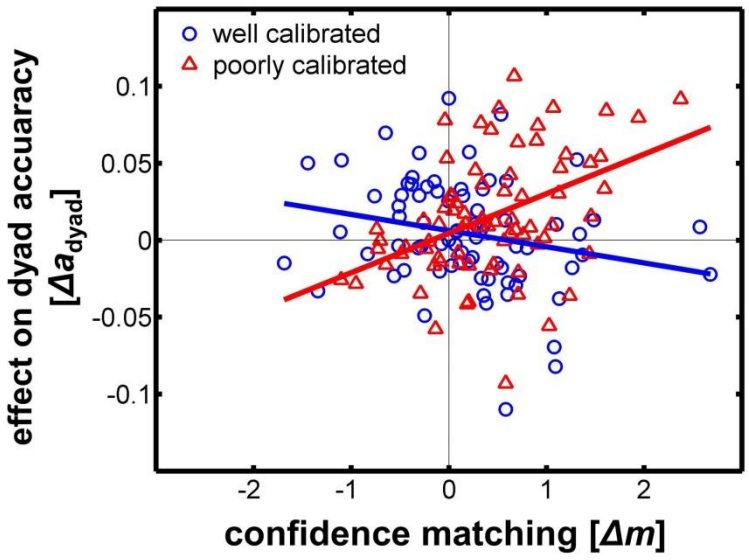


Figure 4-13. Effect of confidence matching on joint accuracy in stereotypical cooperative situations: results from Experiment 4. The x-axis shows the degree of confidence matching, $\Delta m = |c_{\text{sbj}}^{\text{isolated}} - c_{\text{agent}}^{\text{social}}| - |c_{\text{sbj}}^{\text{social}} - c_{\text{agent}}^{\text{social}}|$. This measure tells us whether the absolute difference between a participant's mean confidence and that of their respective partner was larger prior to ($\Delta m > 0$: $|c_{\text{sbj}}^{\text{isolated}} - c_{\text{agent}}^{\text{social}}| > |c_{\text{sbj}}^{\text{social}} - c_{\text{agent}}^{\text{social}}|$) or during interaction ($\Delta m < 0$: $|c_{\text{sbj}}^{\text{isolated}} - c_{\text{agent}}^{\text{social}}| < |c_{\text{sbj}}^{\text{social}} - c_{\text{agent}}^{\text{social}}|$). The y-axis shows the difference between the observed joint accuracy and that expected prior to interaction, $\Delta a_{\text{dyad}} = a_{\text{soc}} - a_{\text{iso}}$. The latter value ($a_{\text{iso}}$) was calculated by playing out (simulated) responses of the partner against those of the participant in the isolated task (for details see Section 4.4.1.4: Joint accuracy expected prior to interaction). Calibration, *cal*, was coded as follows: well-calibrated = .5 (blue circles); poorly calibrated = -.5 (red triangles). The data is pooled across conditions (4 data points per participant). Each line is the best-fitting line of a robust linear regression (simple effects; see Supplementary Table 3 for full regression model). We found the same results (both pattern and significance) using a linear mixed-effects model with a random intercept for each participant.

4.5 Discussion

Groups should in principle benefit from combining the opinions of their members (Bovens & Hartmann, 2003; Condorcet, 1785). Indeed, cultural and technological advances have largely depended on our ability to pool information across individuals in pursuit of a common goal (Richardson & Boyd, 2005; Shea et al., 2014). However, this process can be susceptible to individuals whose confidence outstrips their competence. In such cases, group members must learn to take each other's individual biases into account and establish a common metric for identifying who is more likely to be correct in a given situation. Here, studying the simplest case in which a group consists of two people, we found that group members solve this communication problem by matching their use of the scale on which confidence is mutually expressed (mean confidence and confidence distributions) – regardless of task experience and the scale at hand. We usually assume that "speaking the same language" facilitates successful communication. Our study highlights, however, that we might at times be intending very different meanings despite using a shared set of expressions. As in our task as well as most verbal interaction, misalignment may arise whenever communication involves an arbitrary mapping between an intended meaning and its expression. Indeed, such misalignment may underpin catastrophic failures of communication about risk and uncertainty, such as in the build-up to the 2007 financial crisis.

4.5.1 Why confidence matching?

Confidence matching may, however, be a sensible solution to the communication problem, for several reasons. First, confidence matching is computationally cheap. People do not have to infer each other's subjective mappings, but simply have to track each other's overt confidence reports. Second, confidence matching fares best when people are equally competent. Indeed, we tend to associate with similar others – friends, partners or colleagues – with whom we are

more likely to share common levels of competence (McPherson, Smith-Lovin, & Cook, 2001). Further, even if people were randomly paired from a large population, they would, for any normally-distributed skill, be more likely to be similar (close to the distribution mean) than not (from the opposite tails of the distribution). Lastly, even when people differ in competence, confidence matching is beneficial when their confidence does not reflect this difference. In such situations, people – inadvertently – come to report their confidence in a way that better reflects their relative competence. Using the virtual partners in Experiment 4 as an illustration, miscommunication is minimised because the "more confident but less accurate" group member becomes less confident, whereas the "less confident but more accurate" group member becomes more confident. Critically, this "equilibrium" does not require that people have any insight into their own or others' competence; an insight that cannot always be taken for granted (Fleming et al., 2010; Kruger & Dunning, 1999), or can be hard to establish in a novel task domain or when there is no immediate feedback. We note, however, that our participants appeared to differ with respect to their "strategy" when faced with the virtual partners in Experiment 4 (see spread of dots in **Figure 4-13**); some responded as expected under the optimal strategy, some responded as expected under confidence matching, and others always sought to dominate the joint decision or fully opted out. Future research should seek to identify the nature and the sources of this individual variation.

4.5.2 Implications for the study of confidence reports

Our findings have implications for research concerned with the psychological determinants of confidence reports. As discussed in **Chapter 3**, trial-by-trial variation in confidence reports is usually assumed to reflect random variation in some internal representation of uncertainty (e.g., due to sensory noise) or noise in the read out of this internal representation (e.g., due to unstable confidence thresholds or a lack of direct access to the internal representation). In this chapter, we once again showed that trial-by-trial variation in confidence reports can be

systematic, with people changing the way in which they report their confidence according to the *social* context. In addition, we showed that this socially induced trial-by-trial variation could result in global changes in mean confidence and confidence distributions. Our findings thus raise the possibility that well-known biases in the report of confidence, such as *overconfidence* (**Figure 1-4**), might reflect particular social norms, moulded through specific patterns of interaction in specific social contexts, rather than limitations on how the human mind represents uncertainty or a lack of insight into one's internal operations. This interpretation is consistent with findings that the degree of overconfidence has been found to correlate with contextual factors, including mental health (Huq et al., 1988), personality (Campbell et al., 2004) and social role, such as profession (Broihanne et al., 2014), gender (Barber & Odean, 2001; Niederle & Vesterlund, 2011) and culture (Mann et al., 1998). An implication of this interpretation is that it should be possible to alter biases in confidence reports by introducing another set of norms for how confidence should be reported.

4.6 Chapter summary

In this chapter, we extended our basic framework to *study* and *model* how the *sharing* of representations of uncertainty unfolds in the context of group decision-making (**Figure 3-4**). We tested whether pairs of participants could learn to report their confidence in an optimal (mutually consistent) manner. In separate experiments, we found that participants did not report their confidence in an optimal manner. Instead, they engaged in confidence matching – at the level of their trial-by-trial confidence and at the level of their mean confidence and confidence distributions – regardless of task experience and the scale on which confidence was mutually expressed. Theoretically, confidence matching is sub-optimal because, to match their use of a confidence scale, group members who are not equally competent must report their confidence in a mutually *inconsistent* manner, thus lowering the group accuracy. While

confidence matching is a sub-optimal solution to the problem of establish a common metric for identifying who is more likely to be correct in a given situation, it may nevertheless be a sensible solution to this communication problem; the strategy is cognitively simple, and it works well when group members are equally competent, or when they are poor judges of their relative competence.

The question remains, however, whether our participants were aware of mismatches between their relative confidence and their relative competence. One possibility is that confidence matching reflects conformity. In particular, despite knowing that they differ in competence, participants might still match each other's confidence out of a desire for cohesion (Asch, 1951). While we have shown that such a social norm is in fact adaptive in a range of scenarios, the motivation for confidence matching (cohesion) is different from the one emphasised here (computational simplicity). In the next chapter, to evaluate whether people take into account the credibility of each other's confidence reports, we will modify our task and give participants the option to ignore each other's confidence reports and make a joint decision as they see fit.

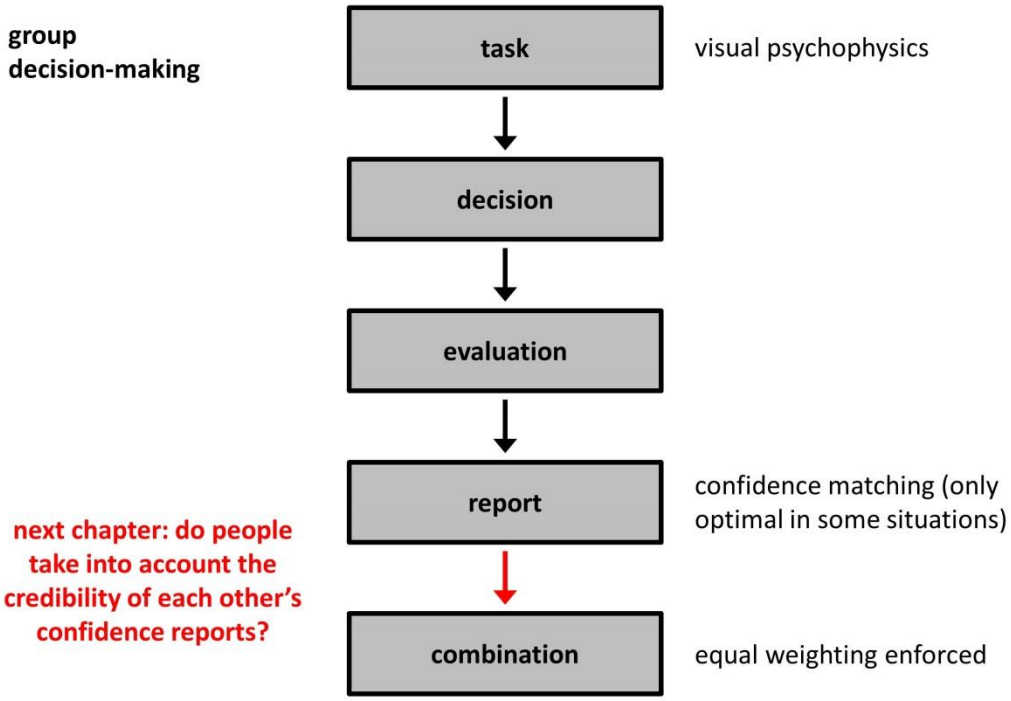


Figure 4-14. Summary of Chapter 4. The boxes denote the component processes involved in sharing and combining representations of uncertainty.

Chapter 5: Interaction outperforms simple algorithms for group choice

In a range of contexts, individuals arrive at group decisions by sharing confidence in their judgements. This tendency to evaluate the reliability of information by the confidence with which it is expressed has been termed the "confidence heuristic". In this chapter, we tested two ways of implementing the confidence heuristic: directly, by opting for the judgement made with higher confidence, or indirectly, by opting for the faster judgement, exploiting an inverse correlation between reported confidence and reaction time. We found that the success of these heuristics depends on how similar individuals are in terms of individual performance and, more importantly, that for dissimilar individuals such heuristics are dramatically inferior to interaction. Interaction allows individuals to alleviate, but not fully resolve, differences in their individual performance. We discuss the implications of these findings for models of confidence and group decision-making.

5.1 Introduction

There is a growing interest in the mechanisms underlying the 2HBT1 effect, which refers to the ability of a pair of individuals (a dyad) to make more accurate decisions than either of their members (Hill, 1982). One study (Bahrami et al., 2010), using a psychophysical task in which a pair of individuals had to detect a visual target, showed that two heads become better than one by communicating and combining explicit reports of confidence, thus allowing them to identify who is more likely to be correct in a given situation (discussed in **Section 1.2.3.4: Do humans report confidence in an optimal manner?**). Sharing of confidence reports as a strategy for combining individual opinions into a group decision has also been established in non-perceptual domains (Laughlin & Ellis, 1986; Zarnoth & Sniezek, 1997). This tendency to evaluate the reliability of information by the confidence with which it is expressed has been termed the "confidence heuristic" (Thomas & McFadyen, 1995).

A recent study has shown that a simple algorithm inspired by the confidence heuristic – here, always opting for the opinion made with higher confidence – can yield a 2HBT1 effect in the absence of any interaction between individuals (Koriat, 2012). Intrigued by this finding, we asked whether this algorithm could in practice replace interaction in group decision-making.

Importantly, such a formula for group choice – if effective – would not be susceptible to the egocentric biases that may impair interaction (Gilovich, Savitsky, & Medvec, 1998), and could readily be used by decision makers, such as jurors, medical doctors or financial investors, who have to combine different opinions in limited time. Indeed, the implementation of heuristics inspired by individual decision-making has proved useful in professional contexts (Gigerenzer & Brighton, 2009).

5.1.1 Circumventing social interaction

Building on Bahrami et al.'s (2010) study, Koriat (2012) asked isolated individuals to estimate the degree of confidence in their perceptual decisions. Participants, all of whom had received the same sequence of stimuli, were afterwards paired into virtual dyads so that they matched each other in terms of their individual performance (i.e., the accuracy of their decisions about the target interval). To remove individual biases, their confidence reports were normalised, so that they shared the same mean and standard deviation, before being submitted to the Maximum Confidence Slating (MCS) algorithm, which selected the decision made with higher confidence on every trial. While circumventing interaction, the MCS algorithm yielded a robust 2HBT1 effect. Interestingly, the confidence reported in a decision has been found to be negatively correlated with reaction times when the decision is made in the absence of time pressure (Patel, Fleming, & Kilner, 2012; Vickers & Packer, 1982), suggesting that a Minimum Reaction Time Slating (MRTS) algorithm may be sufficient to yield a 2HBT1 effect.

Here, we implemented the MCS and MRTS algorithms among *interacting* dyad members who were *not matched* in terms of their individual performance; we sought to evaluate the efficacy of the algorithms by comparing their advised responses with those reached by dyad members through interaction (henceforth *dummy* versus *empirical* dyads/group decisions). We sought to address three questions. First, does the success of the MCS and MRTS algorithms depend

on how similar dyad members are in terms of their individual performance? Bahrami et al. (2010) found that the accuracy of group decisions was constrained by the similarity of dyad members' individual performance. For similar dyad members, two heads were better than one. In contrast, for dissimilar dyad members, two heads were worse than the better one. We predicted that the efficacy of the MCS and the MRTS algorithms would also depend on the similarity of dyad members' individual performance.

Second, do the algorithms fare just as well as interacting dyad members? People vary in their ability to report on the reliability of their decisions (Fleming & Lau, 2014); an ability which is typically referred to as *metacognitive ability* and, in social contexts, determines the *credibility* of people's confidence reports (i.e., the extent to which their low and high confidence reports discriminate between their incorrect and correct decisions). While the algorithms are prone to error when people misestimate the reliability of their own decisions, interacting individuals may take such misestimates into account (Tenney et al., 2007). We therefore predicted that interacting dyad members would take into account the credibility of each other's confidence reports. Further, that interaction would be more beneficial than the algorithms for dissimilar dyad members, as they would have more to lose from following the more confident but less competent of the two.

Third, what is the effect of normalising confidence reports before selecting the decision made with higher confidence? Koriat (2012) reported that the MCS algorithm performed equally well when using *raw* and *normalised* confidence reports as its input. However, this analysis was limited to (virtual) dyad members with similar levels of individual performance. Even though people vary in the ability to evaluate the reliability of their own decisions, confidence reports are rarely uninformative about underlying performance (Fleming & Lau, 2014). As a consequence, normalising confidence reports may remove statistical moments that reflect actual differences in underlying performance (e.g., differences in mean confidence due to

differences in individual performance). We therefore predicted that submitting normalised confidence reports to the MCS algorithm would be relatively more costly for dissimilar dyad members. In a sense, normalising confidence reports is equivalent to full-fledged confidence matching as described in **Chapter 5** – another reason to expect that the MCS algorithm would be relatively more costly for dissimilar dyad members.

5.2 Methods

5.2.1 Participants

Participants were recruited among students (age range: 18-30) at the University of Aarhus. In total, 58 participants took part in the study (EXP1: *N* = 28; EXP2: *N* = 30; all male). Participants only took part in one experiment. All participants reported normal or corrected vision. All participants provided informed consent and were reimbursed for their participation (100DKR/hour ≈ £10/hour for a total of 2 hours). All experiments were approved by the local ethics committee.

5.2.2 Experimental details (Experiments 1 and 2)

Participants took part in one of two experiments. The experiments only differed as to whether dyad members were allowed to communicate (EXP1) or not allowed to communicate (EXP2). All other details of the experiments were the same, and we therefore describe them together.

5.2.2.1 Display parameters and response devices

Participants sat at right angles to each other in a dark room, each with their own monitor and response device. They were separated by a screen to prevent both non-verbal and verbal

communication. The monitors (ViewSonic VG2236wm-LED) were linked to a personal laptop (Toshiba Satellite Pro C660-29W) via a video splitter and controlled by the Cogent toolbox (http://www.vislab.ucl.ac.uk/cogent.php/) for MATLAB (Mathworks Inc). The background luminance was 62.5 cd/m^2 and the resolution was 800 × 600. Participants viewed the monitors at a distance of about 57 cm. One participant responded using a QWERTY keyboard and the other responded using a standard mouse. Participants switched response mode halfway through the experiment. A piece of black cardboard was used to occlude half of the monitor for each participant. This manipulation ensured that participants could make their individual responses in private despite receiving input from the same computer.

5.2.2.2 Basic task (psychophysics)

The basic psychophysical task was the same as in **Chapter 2** (see **Section 2.2.2: Experimental details (Experiments 1 and 2)** for details). On each trial, after the two viewing intervals, participants were presented with a horizontal line bisected at its midpoint. A vertical marker was placed on top of the midpoint. The marker could be moved along the line by up to five steps on either side of the midpoint. Here, we take the left hand steps to be negative values (-5 to -1), and the right hand steps to be positive values (1 to 5). The sign of the response indicates the decision (negative: interval 1; positive: interval 2) and the absolute value of the response indicates the confidence level (1: "unsure"; 5: "certain"). We will use response and confidence to refer to signed response values and unsigned response values, respectively. We adopted the "two-response" design to make the joint-decision stage graphically intuitive.

5.2.2.3 Social task (arbitration)

Their private responses were first made public (colour-coded; 2000 milliseconds; **Figure 4-3**). On agreement trials (i.e., if participants privately selected the same interval), the joint decision

was automatically taken to be the same as their private decisions. On disagreement trials (i.e., if participants privately selected different intervals), one participant (the *arbitrator*) was randomly prompted to make a joint decision on behalf of the dyad. In Experiment 1 (non-verbal), the arbitrator made the joint decision only having access to the declared responses. In Experiment 2 (verbal), the arbitrator could also negotiate the joint decision with their partner. Crucially, in contrast to the social task in **Chapter 4**, participants could override each other's confidence reports at this stage of the experiment. After a group decision had been made (automatically or through arbitration), the participants received feedback about the accuracy of each decision (color-coded; joint feedback shown in white; 2000 milliseconds). The feedback text read "correct" or "wrong"; the vertical order of the individual feedback was randomised, whereas the joint feedback always appeared at the centre. Participants were then presented with white text reading "next trial" (1000 milliseconds) before continuing to the next trial. Participants were instructed to maximise the accuracy of their joint decisions. After 16 practice trials, participants completed 256 experimental trials.

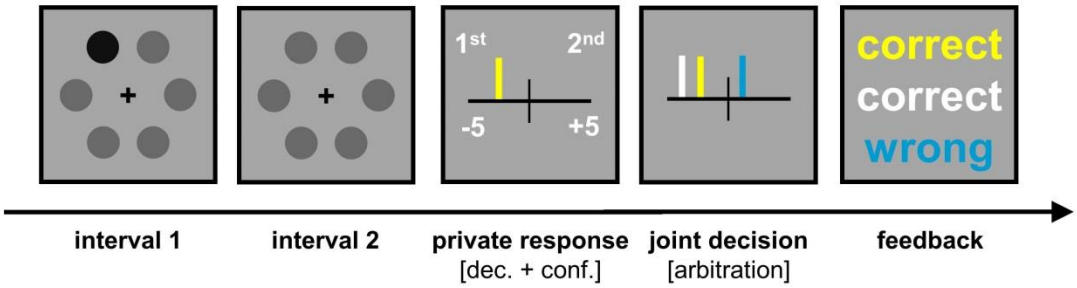


Figure 5-1. Schematic of an experimental trial. On each trial, participants viewed two consecutive intervals, each containing six contrast gratings (here dots). There was a visual target with higher contrast (darker dot) in one of the two intervals. Participants made a response by moving a marker along a scale with a fixed midpoint. There were five steps on either side of the midpoint; the left hand steps were negative values (-5 to -1), whereas the right hand steps were positive values (1 to 5). The sign of a response indicated the decision (negative: 1st; positive: 2nd), and its absolute value indicated the confidence level (1: "unsure"; 6: "certain"). We use *response* and *confidence* to refer to signed response values and unsigned response values, respectively. Participants' private responses (colour-coded) were then made public. If they independently selected the same interval, they received feedback (colour-coded) and continued to the next trial. If they independently selected different intervals, one of the two participants (the arbitrator) was randomly prompted to make a joint decision, using an additional white marker. Here, the participant of focus (yellow) selected interval 1 (correct) and reported confidence level 3, whereas their partner (blue) selected interval 2 (wrong) and reported confidence level 1. The participant of focus was prompted to make the joint decision and selected interval 1 (correct) and reported confidence level 4. The displays have been edited for ease of illustration.

5.2.3 Analysis

5.2.3.1 Computing MCS responses

We used Goodman & Kruskal's gamma to test whether high (or low) confidence reports were associated with correct (or wrong) decisions. As expected, the gamma coefficients (*mean* = .16) were significantly positive across dyad members ($t(57) = 14.04$, $p < .001$, one-sample *t*-test). For each dyad, we then derived the joint decisions advised by the MCS algorithm. In line with Koriat (2012), we first normalised the dyad members' confidence reports and then selected the decision of the more confident dyad member on each trial; the normalised confidence reports, c^*, were related to the raw confidence reports, c, via $c^* = (c - \mu_{1,2})/\sigma_{1,2}$, where $\mu_{1,2}$ and $\sigma_{1,2}$ were the mean and the standard deviation of the raw confidence reports pooled from both dyad members, 1 and 2.

5.2.3.2 Computing MRTS responses

Due to a programming error, the experimental code continued to sample the mouse response time until both mouse and keyboard responses had been registered. This error meant that mouse response times were either identical to keyboard response times (i.e., those trials on which the dyad member using the mouse actually made a faster response than the dyad member using the keyboard) or slower than keyboard response times (i.e., those trials on which the dyad member using the mouse made a slower response than the dyad member using the keyboard). Only including those trials in which dyad members responded with the keyboard (i.e., the first 128 trials for dyad member A and the last 128 trials for dyad member B), we regressed reaction times (measured in milliseconds) against reported confidence to test whether faster decisions were associated with higher confidence. As expected, the regression coefficients (*mean* = -156.29) were significantly negative across dyad members ($t(57) = -5.13$, $p < .001$, one-sample *t*-test). Including all trials, we then, for each dyad, derived the joint

decisions advised by the MRTS algorithm; as described above, we could infer which dyad member made the faster decision on each trial (identical response times: mouse faster; different response times: keyboard faster). Due to the programming error, we could not test the effect of normalising reaction times.

5.2.3.3 Computing sensitivity

Psychometric functions were created for each dyad member and for the dyad (empirical or dummy) by plotting the proportion of trials in which the visual target was reported to be in the second interval against the contrast difference between the second and the first interval at the target location. The steepness of the slope provides an estimate of sensitivity. See **Figure 5-2** for an example psychometric function.

The psychometric curves were fit with a cumulative Gaussian function whose parameters were bias, b, and variance, σ^2. To estimate these parameters, we used a probit regression model. A dyad member with bias b and variance σ^2 would have a psychometric curve, denoted $P(\Delta k)$ where Δk is the contrast difference between the second and first interval at the target location, given by

$$P(\Delta k) = H(\frac{\Delta k + b}{\sigma}) \qquad \textbf{Eq. 5-1}$$

Where $H(z)$ is the cumulative normal function

$$H(z) = \int_{-\infty}^{z} \frac{dt}{(2\pi)^{1/2}} \exp[-t^2/2] \qquad \textbf{Eq. 5-2}$$

The psychometric curve, $P(\Delta k)$, corresponds to the probability of reporting that the second interval contained the visual target. Given the above definitions for $P(\Delta k)$, the variance is related to the maximum slope of the psychometric curve, denoted S, via

$$S = \frac{1}{(2\pi\sigma^2)^{1/2}}$$ **Eq. 5-3**

A steep slope indicates small variance and thus highly sensitive performance. We used this measure to quantify individual and dyad (empirical or dummy) performance. We computed the similarity of dyad members' individual performances as the ratio of the sensitivity of the worse dyad member to that of the better dyad member (i.e., $S_{\min}/S_{\max}$), with values near zero corresponding to dyad members with very different sensitivities and values near one corresponding to dyad members of nearly equal sensitivity. We computed the group outcome (empirical or dummy) as the ratio of the sensitivity of the dyad (empirical or dummy) to that of the more sensitive dyad member (i.e., $S_{\mathrm{emp}}/S_{\max}$, $S_{\mathrm{MCS}}/S_{\max}$, and $S_{\mathrm{MRTS}}/S_{\max}$), with values below 1 indicating a collective loss and values above 1 indicating a collective benefit. We used *sensitivity* (as inferred from a psychometric function) instead of *accuracy* (proportion correct) for consistency with the study by Bahrami et al. (2010). We note that the decisions of more sensitive individuals are by definition more reliable.

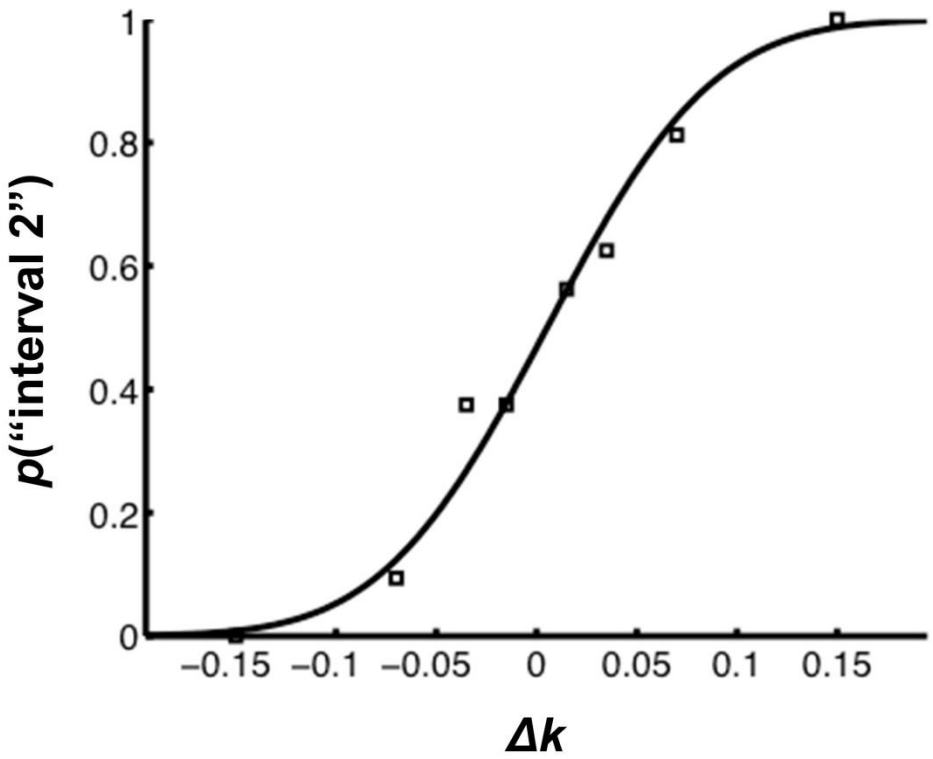


Figure 5-2. Example of a psychometric function. The y-axis shows the proportion of trials on which the visual target was reported to be in the second interval, $p(\text{"interval 2"})$. The x-axis shows the contrast difference between the second and the first interval at the target location, Δk, with negative values corresponding to visual targets in the first interval and positive values corresponding to visual targets in the second interval. A highly sensitive observer would produce a steeply rising psychometric function with a large slope.

5.2.3.4 Computing credibility of confidence reports

In line with previous studies (Fleming et al., 2010, and Song et al., 2011, who build on Galvin et al., 2003, and Kornbrot, 2006), we used the measure, A_{ROC}, to quantify each dyad member's *metacognitive accuracy*, and thereby the *credibility* of their confidence reports. We first calculated the probabilities $p(confidence = i \mid correct)$ and $p(confidence = i \mid incorrect)$ for each confidence level i. We then transformed these probabilities into cumulative probabilities, and plotted them against each other, thus yielding a Receiver Operating Characteristic (ROC) curve, with anchors at [0,0] and [1,1]. See **Figure 5-3** for an example ROC curve. Briefly, as the ROC curve shifts toward the upper left corner, the probability of high confidence given correct rises more rapidly than the probability of high confidence given incorrect. The area under the ROC curve, A_{ROC}, thus provides an estimate of metacognitive accuracy. This area can be calculated geometrically as the sum of the area between the upper bound on the ROC curve and the diagonal (dark grey) and the area of the half-square triangle below the diagonal (light grey):

$$A_{\mathrm{ROC}} = \sum_{i=1}^{K-1} \frac{(X_{i+1} - X_i) + (Y_{i+1} - Y_i)}{2} + \frac{1}{2} \qquad \textbf{Eq. 5-4}$$

where X and Y are the co-ordinates of the data points (squares in **Figure 5-3**) and K is the number of confidence levels k (number of squares in **Figure 5-3**).

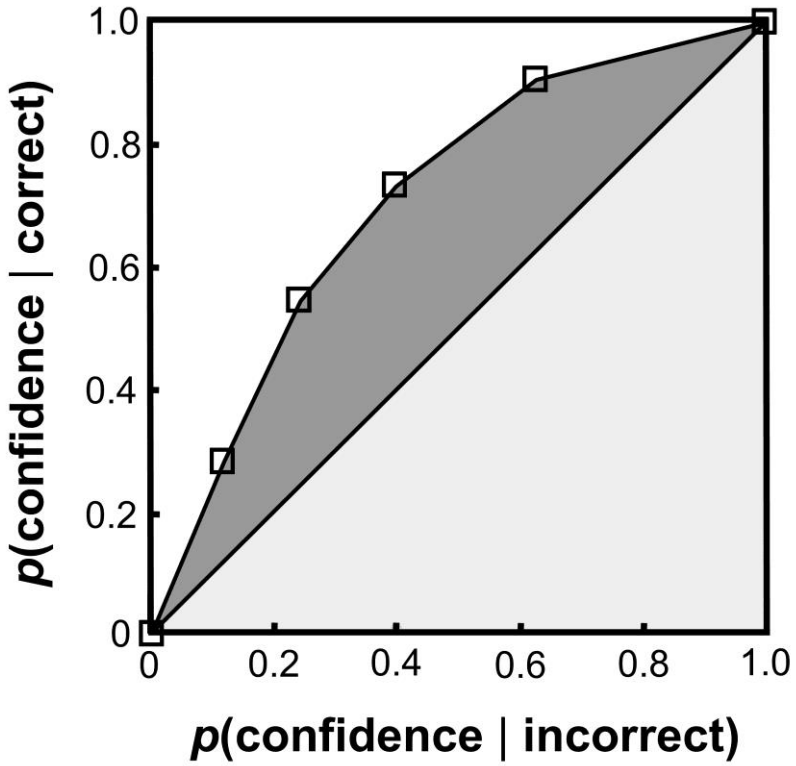


Figure 5-3. Example of an ROC curve. The y-axis and the x-axis show the cumulative probabilities for each level of confidence conditional on being correct and incorrect, respectively. The total area under the curve (A_{ROC}) provides an estimate of metacognitive accuracy. The more bowed the curve, the higher the metacognitive accuracy (i.e., the probability of high confidence conditional on being correct rises more rapidly than the probability of high confidence given incorrect). We used A_{ROC} as an index of the credibility of a dyad member's confidence reports.

5.2.3.5 Normalised versus raw confidence reports

To test the effect of normalising confidence reports before selecting the decision made with higher confidence, we also submitted raw confidence reports to the MCS algorithm. The trials in which dyad members reported the same level of confidence but selected different intervals were resolved by randomly selecting the decision of one of the two dyad members; we note that there were no such confidence ties when using normalised confidence reports. To ensure that the random selection did not favour one of the dyad members by chance, for each dyad, we generated one hundred dummy dyads and used their mean sensitivity (i.e., S_{rawMCS}) to test the effect of normalising confidence reports (i.e., $S_{\mathrm{MCS}}/S_{\mathrm{rawMCS}}$). See **Table 5-1** for a summary of the different strategies for group choice.

Strategy	Abbreviation	Explanation
Normalised Maximum Confidence Slating	MCS	The joint decision was based on the dyad member with the higher normalised confidence report.
Raw Maximum Confidence Slating	rawMCS	The joint decision was based on the dyad member with the higher raw confidence report.
Minimum Reaction Time Slating	MRTS	The joint decision was based on the dyad member with the faster reaction time.
Empirical dyad	emp	A randomly chosen dyad member made the joint decision after having access to the other dyad member's raw confidence report (EXP1) or also having the opportunity to discuss with the other dyad member (EXP2).

Table 5-1. Strategies for group choice.

5.3 Results

As none of our measures of interest showed significant differences between Experiment 1 (non-verbal) and Experiment 2 (verbal), we only report results based on data collapsed across experiments. In addition, as none of our measures of interest changed over time (i.e., from the first 128 trials to the last 128 trials), we only report data collapsed across time.

5.3.1 The efficacy of the algorithms

The similarity of dyad members' sensitivities significantly predicted the group outcome of the MCS algorithm (i.e., $S_{\mathrm{MCS}}/S_{\mathrm{max}}$; $t(27) = 5.17$, $p < .001$, one-sample *t*-test; **Figure 5-4A**). For similar dyad members, the MCS algorithm yielded a collective benefit (i.e., $S_{\mathrm{MCS}}/S_{\mathrm{max}} > 1$ when $S_{\mathrm{min}}/S_{\mathrm{max}}$ is near 1). But for dissimilar dyad members, the MCS algorithm yielded a collective loss (i.e., $S_{\mathrm{MCS}}/S_{\mathrm{max}} < 1$ when $S_{\mathrm{min}}/S_{\mathrm{max}} < 0.6$). The similarity of dyad members' sensitivities also significantly predicted the group outcome of the MRTS algorithm (i.e., $S_{\mathrm{MRTS}}/S_{\mathrm{max}}$; $t(27) = 4.69$, $p < .001$, one-sample *t*-test; **Figure 5-4B**). The MRTS only yielded a collective benefit for five dyads. The performance of the MCS algorithm was superior to that of the MRTS algorithm, both when using normalised confidence reports ($t(28) = 4.99$, $p < .001$, paired *t*-test) and raw confidence reports ($t(28) = 6.26$, $p < .001$, paired *t*-test) as its input.

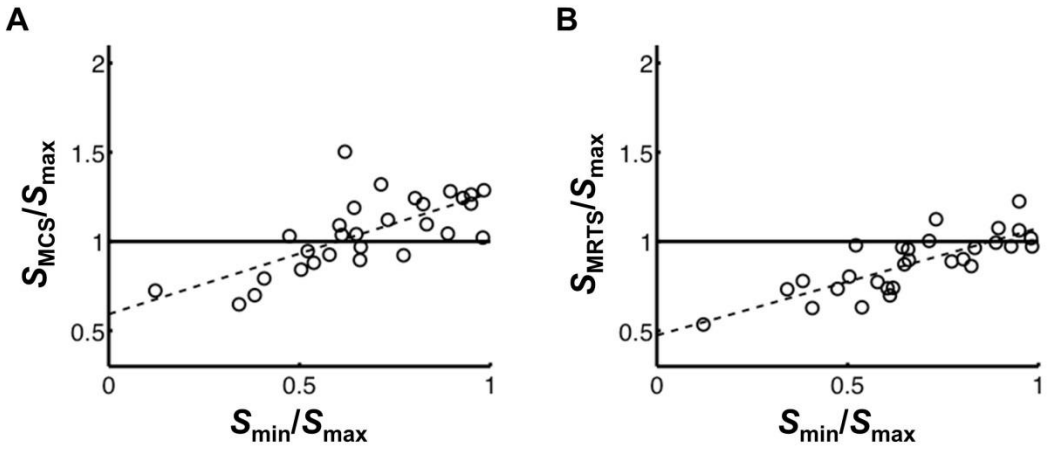


Figure 5-4. The group outcome obtained from the MCS and MRTS algorithms depended on the similarity of dyad members' sensitivities. The *y*-axis shows **(A)** the ratio of the sensitivity of the MCS algorithm relative to that of the more sensitive dyad member ($S_{\mathrm{MCS}}/S_{\mathrm{max}}$) and **(B)** the ratio of the sensitivity of the MRTS algorithm relative to that of the more sensitive dyad member ($S_{\mathrm{MRTS}}/S_{\mathrm{max}}$), with values above 1 indicating a collective benefit over the more sensitive dyad member. The *x*-axis shows the ratio of the sensitivity of the worse dyad member relative to that of the better dyad member ($S_{\mathrm{min}}/S_{\mathrm{max}}$), with values near one corresponding to dyad members of nearly equal sensitivity. Each dot is a dyad, and each line is the best-fitting line of a linear regression. We calculated significance by testing (one-sample *t*-test) whether the fitted coefficients pooled across all dyads were different from 0.

5.3.2 The relative benefit of interaction over the algorithms

To test whether interacting dyad members took into account the credibility of each other's confidence reports, we regressed the percentage of disagreement trials in which the dyad followed the decision of dyad member A instead of that of dyad member B (the choice ratio) against the ratio of the A_{ROC} for dyad member A relative to that of dyad member B (the A_{ROC} ratio). The A_{ROC} ratio significantly predicted the choice ratio ($t(27) = 3.58$, $p < .001$, one-sample *t*-test; **Figure 5-5**).

The similarity of dyad members' sensitivities significantly predicted the group outcome of the empirical dyads relative to that of the MCS algorithm (i.e., $S_{\mathrm{emp}}/S_{\mathrm{MCS}}$; $t(27) = -4.21$, $p < .001$, one-sample *t*-test; **Figure 5-6A**). For similar dyad members, there was no relative benefit for interaction over the MCS algorithm (i.e., $S_{\mathrm{emp}}/S_{\mathrm{MCS}} < 1$ when $S_{\mathrm{min}}/S_{\mathrm{max}}$ is near 1). But for dissimilar dyad members, there was a relative benefit for interaction over the MCS algorithm (i.e., $S_{\mathrm{emp}}/S_{\mathrm{MCS}} > 1$ when $S_{\mathrm{min}}/S_{\mathrm{max}} < 0.8$). The similarity of dyad members' sensitivities also significantly predicted the group outcome of the empirical dyads relative to that of the MRTS algorithm (i.e., $S_{\mathrm{emp}}/S_{\mathrm{MRTS}}$; $t(27) = -2.20$ $p = .037$, one-sample *t*-test; **Figure 5-6B**). However, the MRTS algorithm only (marginally) outperformed two of the empirical dyads.

5.3.3 The effect of normalising confidence reports

The similarity of dyad members' sensitivities significantly predicted the effect of normalising confidence reports (i.e., $S_{\mathrm{MCS}}/S_{\mathrm{rawMCS}}$; $t(27) = 2.70$, $p = .012$, one-sample *t*-test; **Figure 5-7**). For similar dyad members, there was a relative benefit for normalising confidence reports (i.e., $S_{\mathrm{MCS}}/S_{\mathrm{rawMCS}} > 1$ when $S_{\mathrm{min}}/S_{\mathrm{max}}$ is near 1). But for dissimilar dyad members, there was a relative cost to normalising confidence reports (i.e., $S_{\mathrm{MCS}}/S_{\mathrm{rawMCS}} < 1$ when $S_{\mathrm{min}}/S_{\mathrm{max}} < 0.6$).

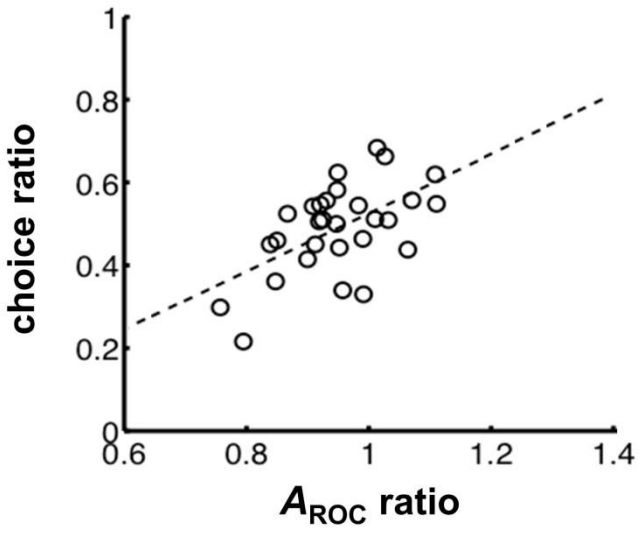


Figure 5-5. Interacting dyad members took into account the credibility of each other's confidence reports. The y-axis shows the percentage of disagreement trials in which the dyad followed the decision of dyad member A instead of that of dyad member B. The x-axis shows the ratio of the A_{ROC} of dyad member A relative to that of dyad member B. Each dot is a dyad, and the line is the best-fitting line of a linear regression. We calculated significance by testing (one-sample *t*-test) whether the fitted coefficients pooled across all dyads were different from 0.

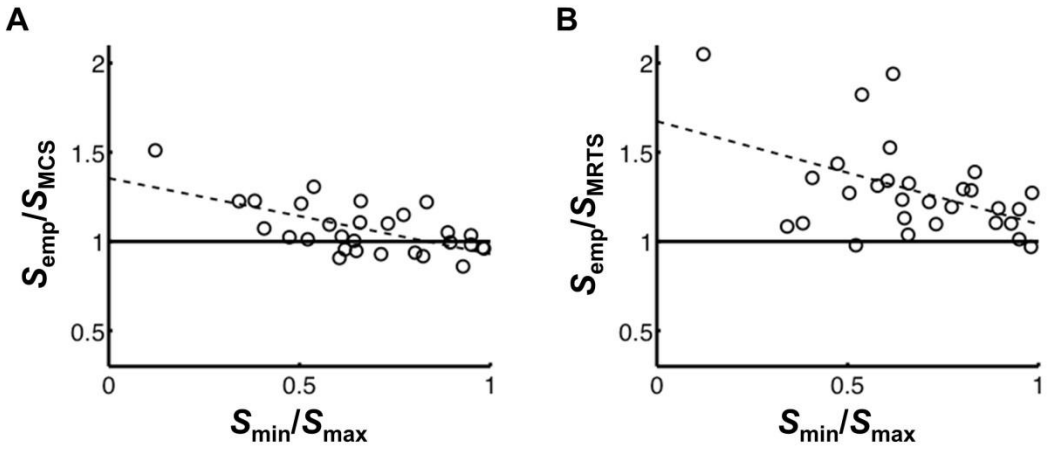


Figure 5-6. The relative benefit for interaction over the MCS and MRTS algorithms depended on the similarity of dyad members' sensitivities. The y-axis shows **(A)** the ratio of the sensitivity of the empirical dyad relative to that of the MCS algorithm ($S_{\mathrm{emp}}/S_{\mathrm{MCS}}$) and **(B)** the ratio of the sensitivity of the empirical dyad relative to that of the MRTS algorithm ($S_{\mathrm{emp}}/S_{\mathrm{MRTS}}$), with values above 1 indicating a relative benefit for interaction. The *x*-axis shows the ratio of the sensitivity of the worse dyad member relative to that of the better dyad member ($S_{\mathrm{min}}/S_{\mathrm{max}}$), with values near one corresponding to dyad members of nearly equal sensitivity. Each dot is a dyad, and each line is the best-fitting line of a linear regression. We calculated significance by testing (one-sample *t*-test) whether the fitted coefficients pooled across all dyads were different from 0.

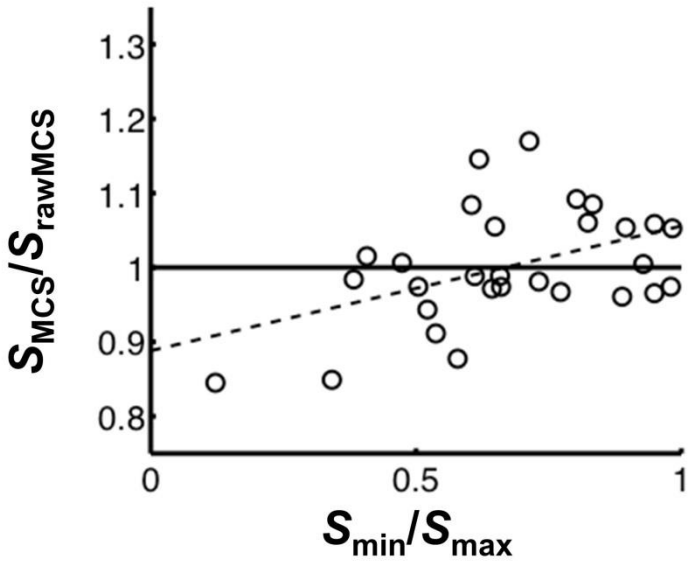


Figure 5-7. The effect of normalising confidence reports depended on the similarity of dyad members' sensitivities. The *y*-axis shows the ratio of the sensitivity of the MCS algorithm using normalised confidence reports relative to that of the MCS algorithm using raw confidence reports ($S_{\mathrm{MCS}}/S_{\mathrm{rawMCS}}$) with values above the horizontal line indicating a relative benefit for normalising confidence reports. The x-axis shows the ratio of dyad members' sensitivities ($S_{\mathrm{min}}/S_{\mathrm{max}}$), with values near one corresponding to dyad members of nearly equal sensitivity. Each dot is a dyad, and the line is the best-fitting line of a linear regression. We calculated significance by testing (one-sample *t*-test) whether the fitted coefficients pooled across all dyads were different from 0.

5.4 Discussion

Sharing of confidence as a strategy for combining individual opinions into group decisions has been established in a wide range of contexts. This tendency to evaluate the reliability of information by the confidence with which it is expressed has been termed the "confidence heuristic". Here, we tested two simple ways of implementing the confidence heuristic: the MCS algorithm, which opts for the decision made with higher confidence, and the MRTS algorithm, which opts for the faster decision, exploiting a negative correlation between reported confidence and reaction time. Our findings have important implications for the use of heuristics for group choice and for models of confidence and group decision-making.

5.4.1 Computational background

To unpack our findings, we will briefly explain the SDT model that Bahrami et al. (2010) used in their original study and that partially inspired the study by Koriat (2012). In SDT models (see **Section 1.3.2.1: Signal detection theory**), the perception of a visual event (sensory evidence), x, can be defined as a random sample from a Gaussian distribution, $x \in N(s, \sigma)$. In the current task, the mean of the Gaussian distribution, s, would be given by the contrast difference between the two intervals at the target location (Δk; **Figure 5-2**) and the standard deviation of the Gaussian distribution, σ, specifies the level of sensory noise.

In their study, Bahrami et al. (2010) proposed that the decision variable, z^D, is computed as the raw sensory evidence, $z^D = x$, and that there is no bias, $\theta^D = 0$. In this case, the sign of the sensory evidence determines whether a participant selects the first interval ($x < 0$) or the second interval ($x > 0$). As such, a *reliable* decision is characterised by a large x and a small σ. Bahrami et al. (2010) proposed that dyad members computed their confidence variable, z^C, as their respective x/σ-ratios (a z-score), $z^C = x/\sigma$, and that their confidence report, c, was a monotonic function of the confidence variable, $c \propto z^C$, with dyad members having the *same*

mapping function, $c(z^C)$. Lastly, Bahrami et al. (2010) proposed that dyad members made a group decision by comparing the confidence reported in their respective decisions and then selecting the decision made with higher confidence – a strategy which is optimal under the assumption that dyad members have the same mapping function. Under this model – called the Weighted Confidence Sharing (WCS) model – group performance is determined by the similarity of dyad members' sensitivities (see **Section 5.2.3.3: Computing sensitivity**), with low similarity leading to a collective loss.[16] In support of the WCS model, Bahrami et al. (2010) observed – as discussed previously – that group performance was indeed constrained by the similarity of dyad member's sensitivities. For similar dyad members, two heads were better than one. For dissimilar dyad members, two heads were worse than the better one.

5.4.2 The efficacy of the MCS and the MRTS algorithms

The MCS algorithm effectively implements the WCS model by selecting the decision made with higher confidence on each trial. We therefore predicted that the MCS algorithm would also depend on the similarity of dyad members' sensitivities. In addition, noting that confidence reports typically are negatively correlated with reaction time in the absence of time pressure (Patel et al., 2012; Vickers & Packer, 1982), we predicted that the MRTS algorithm would also depend on the similarity of dyad members' sensitivities.

As expected, the MCS and the MRTS algorithms only yielded 2HBT1 effects for dyad members of nearly equal sensitivity. However, despite a strong negative correlation between confidence reports and reaction time, the MCS algorithm markedly outperformed the MRTS algorithm. The superiority of the MCS algorithm to the MRTS algorithm could be due to differences in the pre-processing of their input. While we could submit both normalised and raw confidence

[16] Note that Bahrami et al. (2010) did not record explicit confidence reports; their model only seeks to infer what might have happened during the discussions leading to a group decision.

reports to the MCS algorithm, we could only submit raw reaction times to the MRTS algorithm because of a programming error (see **Section 5.2.3.2: Computing MRTS responses**). Since individuals vary with respect to the speed of their responses, pooling raw reaction times from two individuals might corrupt the link between confidence reports and reaction time at the level of the dyad. However, the MCS algorithm outperformed the MRTS algorithm even when raw confidence reports were used as its input, suggesting that reaction times are a very poor substitute for confidence reports.

We note that, outside the context of the current study, the normalisation of confidence reports may have a more straightforward interpretation than the normalisation of reaction times. The normalisation of confidence reports was intended to remove biases in the use of the scale – that is, how an implicit variable was mapped onto an explicit variable – and could relate to important aspects of communication (see **Section 5.4.4: The effect of normalising confidence reports**). It is less clear how the normalisation of reaction times would translate into other contexts. Taken together, our findings show that heuristics for group choice are susceptible to individual differences in performance, and suggests that reaction time cannot be substituted for confidence reports without incurring a considerable cost. In this light, we will limit the remainder of our discussion to the MCS algorithm.

5.4.3 The relative benefit of interaction over the MCS algorithm

If the WCS model is correct, the responses advised by the MCS algorithm should be just as accurate as those reached through interaction. However, the ability to estimate the reliability of one's own decisions (cf. estimating σ in the SDT model) is subject to substantial individual variation (Fleming et al., 2010; Maniscalco & Lau, 2012; Song et al., 2011); this ability is typically referred to as metacognitive ability and can be quantified to determine the *credibility* of people's confidence reports (A_{ROC}; see **Section 5.2.3.4: Computing credibility of confidence**

reports). While the MCS algorithm is prone to error when people misestimate the reliability of their own decisions, interacting individuals may take such misestimates into account. For example, Tenney et al. (2007) showed that mock jurors found witnesses who were confident about erroneous testimony less credible than witnesses who were not confident about it. We therefore predicted that interacting dyad members would take into account the credibility of each other's confidence reports, and, in turn, that interaction would be more beneficial than the algorithms for dissimilar dyad members, as they would have more to lose from following the more confident but less competent of the two.

As for the first prediction, the fraction of disagreement trials in which the dyad followed dyad member A instead of dyad member B depended on the credibility of their confidence reports. As for the second prediction, interaction was more robust in the face of individual differences in sensitivity compared to the MCS algorithm. For dyad members of nearly equal sensitivity, the decisions reached through interaction were no more accurate than those advised by the MCS algorithm. In contrast, for dyad members with different levels of sensitivity, the decisions reached through interaction were more accurate than those advised by the MCS algorithm. This was true regardless of whether normalised or raw confidence reports were submitted to the MCS algorithm. While previous work has assumed that confidence reports are weighted equally (Bahrami et al., 2010; Koriat, 2012), our findings suggest that this might not always be the case. Without taking into account the credibility of each other's confidence reports, dyad members would not have outperformed the MCS algorithm.

5.4.4 The effect of normalising confidence reports

Even if dyad members had direct access to their respective x/σ-ratios as assumed by the WCS model, they still had to solve the problem of how to map specific x/σ-ratios onto specific levels of confidence. As discussed and shown in **Chapter 4**, the assumption of the WCS model

that dyad members have the *same* mapping function is almost certainly false (see survey data shown in **Figure 4-1**). This additional communication problem (see illustration in **Figure 4-2**) may explain why the relative benefit of normalising confidence reports depended on the similarity of dyad members' sensitivities. If dyad members mapped *similar* x/σ-ratios (x/σ-ratios of similar magnitudes) onto *different* levels of confidence, then normalising confidence reports would re-map those ratios onto *similar* levels of confidence, thus improving the performance of the MCS algorithm. In contrast, if dyad members mapped *different* x/σ-ratios (x/σ-ratios of different magnitudes) onto *similar* levels of confidence, then normalising confidence reports would strengthen this erroneous mapping (mutually inconsistent), thus decreasing the performance of the MCS algorithm. On this account – and in line with the findings presented in **Chapter 4** – similar dyad members (who overall would have had x/σ ratios of similar magnitudes) did not map their respective x/σ ratios onto sufficiently *similar* confidence distributions, whereas dissimilar dyad members (who overall would have had x/σ ratios of different magnitudes) did not map their respective x/σ ratios onto sufficiently *different* confidence distributions. More generally, the normalisation involved in the MCS algorithm can be thought of as an extreme form of the confidence-matching phenomenon described in **Chapter 4**.

A linguistic analysis by Fusaroli et al. (2012) of the conversations in Bahrami et al.'s (2010) study showed that dyad members who aligned their confidence reports accrued larger 2HBT1 effects. More specifically, dyad members who used expressions from the same linguistic set (e.g., both using gradients of "sure", such as "unsure" and "pretty sure") performed better together than those who used expressions from different linguistic sets (e.g., gradients of "sure" versus "know"). If linguistic alignment involves confidence matching – or normalisation in the statistical sense described here – then the positive effect of linguistic alignment might depend on the similarity of dyad members' sensitivities. However, linguistic interaction may offer other ways of solving the communication problem. For example, interacting individuals

may carve a fine-grained and custom-made scale of confidence that fits the continuous nature and the range of their internal representations of uncertainty (e.g., the linguistic sets in Fusaroli et al., 2012, contained on average 18 items). Further, interacting individuals may discuss each other's use of a given confidence scale (e.g., what it means to be "very sure"), thus preventing any confusion about what each expression means. Lastly, individuals have more experience with linguistics expressions than numerical estimates. Indeed, when probed with linguistic expressions instead of numerical estimates, people's metacognitive accuracy is higher (Overgaard & Sandberg, 2012).

5.5 Chapter summary

In this chapter, we extended our basic framework to *study* and *quantify* how the *combining* of representations of uncertainty unfolds in the context of group decision-making (**Figure 3-4**). We tested whether a confidence heuristic could in fact replace interaction among pairs of interacting participants – either directly, by opting for the judgement made with higher confidence, or indirectly, by opting for the faster judgement. We found that reaction time could not be substituted for confidence reports without incurring a considerable cost to group accuracy. Further, we found that, for participants with similar levels of individual performance, the group decisions advised by the confidence heuristic were as accurate as those reached through interaction, but for individuals with different levels of individual performance, the group decisions advised by the confidence heuristic were less accurate than those reached through interaction. Our results suggested that, in the latter case, interacting individuals outperformed the confidence heuristic because they took into account the credibility of each other's confidence reports, making them less susceptible to those situations in which the more confident was the less competent group member. Lastly, we found that normalising confidence reports increased the efficacy of the confidence heuristic for participants with

similar levels of individual performance but had the opposite effect for individuals with different levels of individual performance – a result which is consistent with the observed confidence matching in **Chapter 4**.

The question remains, however, whether people combine their opinions once expressed in an optimal manner. Some studies suggest that we have a poor insight into how good we are relative to others (Kruger & Dunning, 1999), whereas others suggest that we are can combine different (non-social and social) sources of information in an optimal manner (Behrens, Hunt, & Rushworth, 2009; Behrens, Hunt, Woolrich, & Rushworth, 2008). In the next chapter, we will use the same task as in the current chapter and use computational modelling to quantify exactly how good people are weighting their partner's opinions relative to their own.

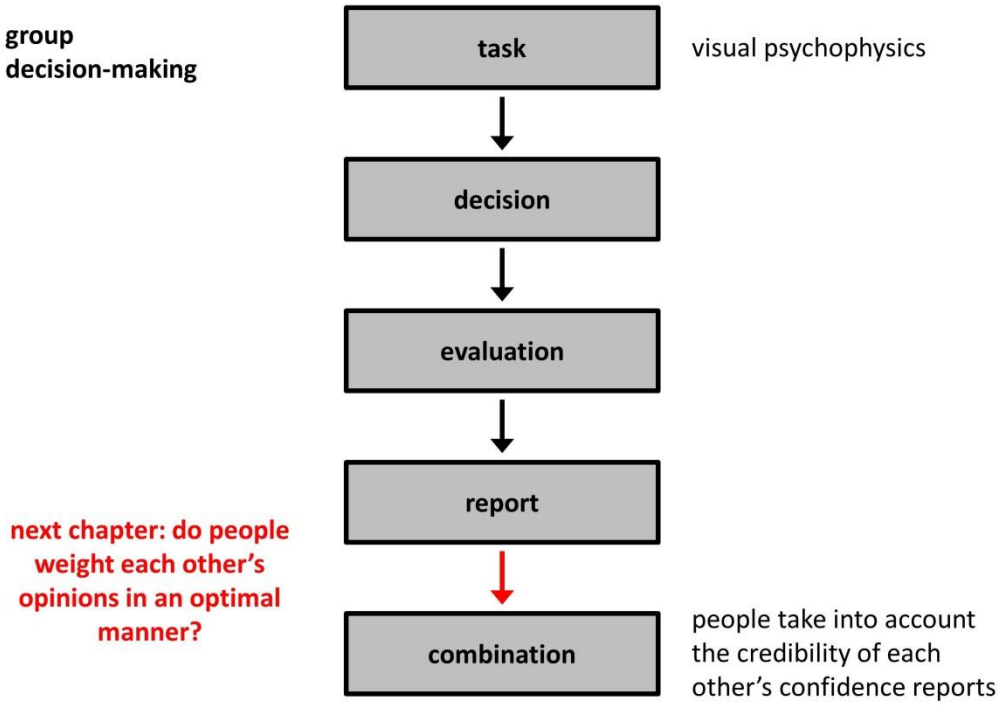


Figure 5-8. Summary of Chapter 5. The boxes denote the component processes involved in sharing and combining representations of uncertainty.

Chapter 6: Equality bias in the weighting of other people's confidence reports

We tend to think that everyone deserves an equal say in a discussion. However, this seemingly innocuous assumption can be damaging when we make decisions together as part of a group. To make optimal decisions, group members should weight their differing opinions according to how competent they are relative to one another; whenever they differ in competence, an equal weighting is sub-optimal. In this chapter, we asked how people deal with individual differences in competence (sensitivity) in the context of group decision-making. We developed a novel metric for estimating how participants weight their partner's opinion relative to their own and compared this weighting to an optimal benchmark. Replicated across three countries (Denmark, Iran and China), we show that participants assigned nearly equal weights to each other's opinions regardless of true differences in their competence. This "equality" bias – whereby people behave as if they are as good or as bad as their partner – was also found for separate groups of participants who received explicit feedback about their relative competence, who differed significantly in terms of their individual competence or who received monetary rewards for making accurate group decisions.

6.1 Introduction

When it comes to making decisions together, we tend to give everyone an equal chance to voice their opinion. This is not just good manners but reflects a long-standing consensus on the "wisdom of crowds" (Condorcet, 1785; Galton, 1907b; Surowiecki, 2004). But a question much contested (Galton, 1907a; Perry-Coste, 1907) is, once everyone has voiced their opinion, how should they be combined so as to make the most of them? One suggestion is to weight each opinion by the confidence with which it is expressed (Bahrami et al., 2010; Koriat, 2012). However, this strategy may fail (Bang et al., 2014; **Chapter 4** and **Chapter 5**) when "the fool who thinks he is wise" is paired with "the wise who [thinks] himself to be a fool" (Shakespeare, 1993). In the face of such a *competence gap*, knowing whose opinion to trust is critical (Nitzan & Paroush, 1985).

A wealth of research suggests that people are poor judges of their own competence – not only when judged in isolation but also when judged relative to others. For example, people tend to overestimate their own performance on hard tasks; paradoxically, when given an easy task,

they tend to underestimate their own performance (the *hard-easy* effect; e.g., Gigerenzer et al., 1991; Soll & Larrick, 2009; see **Figure 1-4**). Relatedly, when comparing themselves to others, people with low competence tend to think they are as good as everyone else, whereas people with high competence tend to think they are as bad as everyone else (the *Dunning-Kruger* effect; e.g., Kruger & Dunning, 1999). In addition, when presented with expert advice, people tend to insist on their own opinion, even though they would have benefitted from following the advisor's recommendation (*egocentric advice discounting*; e.g., Yaniv & Kleinberger, 2000; Yaniv, 2004). These and similar findings suggest that individual differences in competence may not feature saliently in social interaction. However, it remains unknown whether – or to what extent – people take into account such individual differences when resolving disagreements in group decision-making.

To address this issue, we developed a computational framework inspired by recent work on how people learn the reliability of (non-social and social) information (Behrens et al., 2009, 2008, 2007). We used this framework to (a) quantify how members of a group weight each other's opinions and (b) compare this weighting to that of an optimal model in the *arbitration* task described in **Chapter 5** (non-verbal; **Figure 6-1**). Briefly, on each trial, a pair of participants (a dyad) viewed two intervals, with a visual target in either the first or the second one. They privately indicated which interval they thought contained the visual target, and how confident they felt about this decision. In the case of disagreement (i.e., they privately selected different intervals), one of the two participants (the *arbitrator*) was asked to make a joint decision, only having access to their own as well as their partner's responses. In the case of agreement (i.e., they privately selected the same interval), the joint decision was taken to be the same as their private decisions. Lastly, participants received feedback about the accuracy of each decision before continuing to the next trial. In this task, *competence* can be quantified as *sensitivity* – a measure which reflects the reliability of decisions about the visual target.

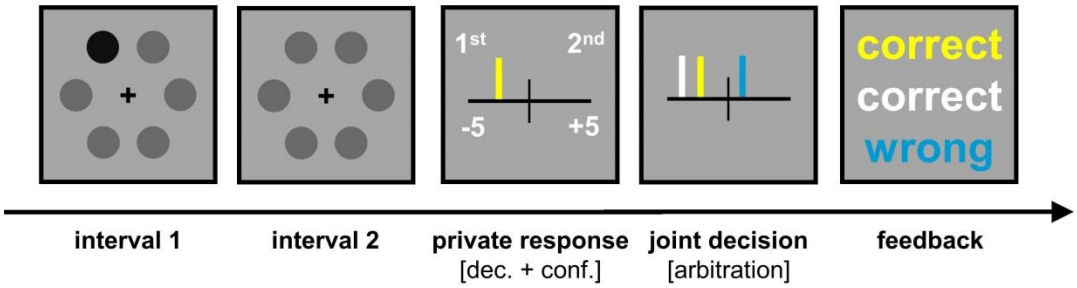


Figure 6-1. Schematic of an experimental trial. On each trial, participants viewed two consecutive intervals, each containing six contrast gratings (here dots). There was a visual target with higher contrast (darker dot) in one of the two intervals. Participants made a response by moving a marker along a scale with a fixed midpoint. There were five steps on either side of the midpoint; the left hand steps were negative values (-5 to -1), whereas the right hand steps were positive values (1 to 5). The sign of a response indicated the decision (negative: 1st; positive: 2nd), and its absolute value indicated the confidence level (1: "unsure"; 6: "certain"). We use *response* and *confidence* to refer to signed response values and unsigned response values, respectively. Participants' private responses (colour-coded) were then made public. If they independently selected the same interval, they received feedback (colour-coded) and continued to the next trial. If they independently selected different intervals, one of the two participants (the arbitrator) was randomly prompted to make a joint decision, using an additional white marker. Here, the participant of focus (yellow) selected interval 1 (correct) and reported confidence level 3, whereas their partner (blue) selected interval 2 (wrong) and reported confidence level 1. The participant of focus was prompted to make the joint decision and selected interval 1 (correct) and reported confidence level 4. The displays have been edited for ease of illustration.

Previous work predicts that participants would be able to use the trial-by-trial feedback to track the probability that their own decision is correct as well as that of their partner (Behrens et al., 2009, 2008). We hypothesised that the arbitrator would make a joint decision by scaling these probabilities, p_{s} and p_{o} (s: self; o: other), by their responses, r_{s} and r_{o} ("signed confidence"; **Figure 6-1**) – thus making full use of the available information – and combining these scaled probabilities into a (joint) decision variable, $DV = p_{\mathrm{s}} * r_{\mathrm{s}} + p_{\mathrm{o}} * r_{\mathrm{o}}$, selecting interval 1 if $DV < 0$ and interval 2 if $DV > 0$. In addition, to capture any bias in how the arbitrator weighted their partner's opinion relative to their own, we included a free parameter, γ, that modulated the influence of the partner in the computation of the decision variable, $DV = p_{\mathrm{s}} * r_{\mathrm{s}} + \gamma * p_{\mathrm{o}} * r_{\mathrm{o}}$.

In four experiments, we tested whether – and to what extent – advice taking (quantified as γ) varied with differences in competence between dyad members. To anticipate our findings, we found that the worse members of each dyad *under-weighted* their partner's opinion (i.e., assigned less weight to their partner's opinion than recommended by the optimal model), whereas the better members of each dyad *over-weighted* their partner's opinion (i.e., assigned less weight to their partner's opinion than recommended by the optimal model). Remarkably, dyad members exhibited this *equality* bias – behaving *as if* they were as good as or as bad as their partner – even when they (a) received a running score of their own and their partner's performance, (b) differed dramatically in terms of their individual performance, and (c) had a monetary incentive to assign the appropriate weight to their partner's opinion.

Recently, psychological phenomena previously believed to be universal have been shown to vary across cultures – where culture is understood as a set of behavioural practices specific to a particular population (Roepstorff, Niewöhner, & Beck, 2010) – calling the generalizability of studies based on Western, Educated, Industrialised, Rich and Democratic (WEIRD) participants into question (Henrich, Heine, & Norenzayan, 2010a, 2010b). To test whether the pattern of

advice taking observed here was culture-specific, we conducted our experiments in Denmark, China and Iran. Coarsely speaking, we take these three countries to represent "Northern European" (Denmark), "East Asian" (China) and "Middle-Eastern" (Iran) cultures. According to the latest World Values Survey[17], 71.1% of Northern European respondents (data from Norway and Sweden), 52.3% of Chinese respondents but only 10.6% of Iranian respondents favoured the sentence "most people can be trusted" over "you can never be too careful when dealing with people". Reflecting this pattern of responses, research using public goods games has shown that Danish participants contribute more than Chinese and that Chinese participants in turn contribute more than Middle-Eastern (data available from Turkey, Saudi Arabia and Oman) (Gachter, Herrmann, & Thoni, 2010; Herrmann, Thoni, & Gachter, 2008). Our sample should thus be sensitive to any cultural commonalities or differences in advice taking.

6.2 Methods

6.2.1 Participants

Participants were recruited among students (age range: 18-30) at the University of Aarhus (EXP1: *N* = 30; EXP3: *N* = 26), Peking University (EXP1: *N* = 34), the University of Tehran (EXP1: *N* = 30; EXP2: *N* = 22; EXP3: *N* = 20) and University College London (EXP4: *N* = 34). In total, 196 participants (all male) took part in the study. For each experiment, participants were recruited in pairs. Participants only took part in one experiment. All participants reported normal or corrected vision. All participants provided informed consent. All experiments were approved by the local ethics committee. In all experiments, participants were reimbursed for their participation (Iran: 125000R/hour ≈ £5/hour for a total of two hours; DK: 100DKR/hour ≈ £10/hour; China: 20RMB/hour ≈ £2/hour for a total of two hours; UK: £7.5/hour for a total of

[17] A cross-country project coordinated by the Institute for Social Research of the University of Michigan: http://www.worldvaluessurvey.org/.

2 hours). In Experiment 4, participants also received a performance-based monetary bonus (mean bonus: £2.5).

6.2.2 Experimental details (Experiments 1 to 4)

All experiments involved the arbitration task (non-verbal) and the same experimental setup as described in **Chapter 5**. Here, we will only highlight the key differences (see **Section 5.2.2: Experimental details (Experiments 1 and 2)** for details and references to earlier chapters).

6.2.2.1 Experiment 1

Participants (Denmark, Iran, China) performed the arbitration task (non-verbal) without any modifications. After 16 practice trials, participants completed 256 experimental trials.

6.2.2.2 Experiment 2

Participants (Iran) performed the arbitration task (non-verbal), except that they, in addition to the trial feedback, were shown a running score of their own and their partner's accuracy at the end of each trial (i.e., proportion of correct responses from the first trial up until the current trial). By means of this design, we could test the alternative hypothesis that the equality bias observed in Experiment 1 was due to limitations on memory capacity. After 16 practice trials, participants completed 256 experimental trials.

6.2.2.3 Experiment 3

Participants (Iran and Denmark) took part in two experimental sessions. In the first session, the participants performed the psychophysical task on their own (no sharing of responses, no joint decision, and no shared feedback). At the end of this session, we computed the sensitivity of each participant, and for the second session, white noise was added to all Gabor patches for the less sensitive participant. In the second session, the participants performed the arbitration task (non-verbal). Participants were told that the first session served as practice and were not informed about the noise manipulation until at end of the experiment. By means of this design, we could test the alternative hypothesis that the equality bias observed in Experiment 1 was due to small or happenstance differences in performance. In each session, after 16 practice trials, participants completed 128 experimental trials.

6.2.2.4 Experiment 4

Participants (UK) performed the arbitration task (non-verbal), except that, instead of rating their confidence, participants placed post-decision wagers (Persaud, McLeod, & Cowey, 2007). The possible wagers were: £0.2, £0.4, £0.6, £0.8 and £1. Feedback was shown as the sum of each participant's wins/losses from the individual and the group decision. To illustrate: (1) participant A bet £0.40 on interval 1 and participant B bet £0.8 on interval 2; (2) the arbitrator then bet £0.6 on interval 2 on behalf of the dyad; (3) the correct answer was interval 2; (4) participant A won: -£0.4 + £0.6 = £0.4; and participant B won £0.8 + £0.6 = £1.4. Apart from the base reimbursement rate for taking part in the study, each participant received a bonus payment calculated by randomly selecting 5 trials and adding up the earnings in those trials. After 16 practice trials, participants completed 256 experimental trials. By means of this design, we could test the alternative hypothesis that the equality bias observed in Experiment 1 was not due to a lack of incentives for accurate joint decisions.

6.2.3 Cross-cultural considerations

We used the same experimental procedure across all three locations, except that all instructions and task text were given in the local language. This procedure was chosen to eliminate any *foreign-language* effects on behavior (Keysar, Hayakawa, & An, 2012). We observed no difference in performance (sensitivity) between the different locations (all $t < 1.60$; all $p > .100$, independent *t*-tests).

6.2.4 Analysis

6.2.4.1 Computing sensitivity

For consistency with **Chapter 5**, we used *sensitivity* (as inferred from a psychometric function) instead of *accuracy* (proportion correct) as our measure of individual and group performance (see **Section 5.2.3.3: Computing sensitivity** for details). As in **Chapter 5**, we computed the group outcome as the ratio of the sensitivity of the dyad to that of the more sensitive dyad member (i.e., $S_{\mathrm{dyad}}/S_{\mathrm{max}}$), with values below 1 indicating a collective loss and values above 1 indicating a collective benefit. By definition, the decisions of more sensitive individuals are more reliable. As such, we take sensitivity to be a measure of *competence* in our task.

6.2.4.2 Computational model

6.2.4.2.1 Model details

We modelled the beliefs of a participant about the reliability of their own decisions and their partner's decisions using a previously introduced computational model (Behrens et al., 2008, 2007). We refer the reader to the original studies for mathematical details; here, we will only explain the model in broad terms. In its original formulation, the model employs Bayesian reinforcement learning to track the probability of obtaining a reward (*risk*) from taking a

choice alternative. Broadly, it assumes that the choice outcome is generated by an underlying probability, p, and after having observed the outcome on a given trial, the model updates its estimate of p. The model scales this update (*learning rate*) according to an estimate of the stability (*volatility*) of the environment. In a changing environment, updates should be large so that p only reflects the outcome history of *recent* trials. In a stable environment, updates should be small so that p incorporates the outcome history of more *distant* trials.

We assumed that participants in our task tracked the probability that their own decision is correct, p_{s}, and the probability that their partner's decision is correct, p_{o}, in a similar manner (cf. risk and choice alternatives). In particular, on the first trial, the estimate of probability correct for each participant is set to 50%. The model then trial-by-trial updates each estimate according to the discrepancy between the predicted probability correct and the observed accuracy (prediction error). If the prediction error is negative (i.e., the participant was wrong), the estimate of probability correct is decreased, but if the prediction error is positive (i.e., the participant was correct), then the estimate of probability correct is increased. As in the original model, the magnitude of this update (learning rate) varies according to the estimated volatility of each participant's performance. Possible sources of volatility in our task are attention, arousal and/or perceptual learning. We note that it has been shown that people can indeed simultaneously track risk and volatility for multiple choice alternatives (both non-social and social information) (Behrens et al., 2008). We obtained the trial-by-trial probabilities for each participant by fitting (maximum likelihood estimation) the model to the decisions about the target interval and their associated feedback.

6.2.4.2.2 Estimating social influence

To estimate how a participant weighted their partner's opinion relative to their own, we focused on the disagreement trials on which the participant made the joint decision on behalf

of the dyad. We assumed that, on a disagreement trial i, the participant made a joint decision by (a) scaling the estimates of probability correct associated with their own and their partner's decision, $p_{\mathrm{s,i}}$ and $p_{\mathrm{o,i}}$, with the respective responses, $r_{\mathrm{s,i}}$ and $r_{\mathrm{o,i}}$ ("signed confidence"; **Figure 6-1**), (b) combining these scaled estimates into a decision variable, $DV_{\mathrm{i}} = p_{\mathrm{s,i}} * r_{\mathrm{s,i}} + \gamma * p_{\mathrm{o,i}} * r_{\mathrm{o,i}}$ and (c) thresholding the decision variable, selecting interval 1 if $DV_{\mathrm{i}} < 0$ and interval 2 if $DV_{\mathrm{i}} > 0$. The computation of the decision variable includes a free parameter, γ, which controls the influence of the partner. As such, $\gamma > 1\gamma$ indicates that the participant is *strongly* influenced by their partner's opinion. In contrast, $\gamma < 1$ indicates that the participant is *weakly* influenced by their partner's opinion. We defined γ_{fit} as the γ-value that maximised the fit (maximum likelihood estimation) between the model-derived and the empirically observed joint decisions. We defined γ_{opt} as the γ-value that maximised the sensitivity of the model-derived joint decision (exhaustive search). We used the discrepancy between these two values, $\gamma_{\mathrm{fit}} - \gamma_{\mathrm{opt}}$, to determine whether the participant *under-weighted* ($\gamma_{\mathrm{fit}} < \gamma_{\mathrm{opt}}$) or *over-weighted* ($\gamma_{\mathrm{fit}} > \gamma_{\mathrm{opt}}$) their partner's opinion relative to their own.

6.2.4.2.3 Model fits

To evaluate the model fit, we computed, for each participant, the proportion of joint decisions that were correctly predicted by the model. Overall the model provided a good fit to the data. In Experiment 1, the model predicted 83% ± 7% (*mean* ± *SD*) of the joint decisions (Denmark: 85% ± 7%; Iran: 83% ± 7%; China: 82% ± 7%). In Experiment 2, the model predicted 83% ± 8% of the joint decisions. In Experiment 3, the model predicted 82% ± 6% of the joint decisions (Denmark: 83% ± 7%; Iran: 83% ± 7%; China: 81% ± 5%). In Experiment 4, the model predicted 84% ± 6% of the joint decisions. There were no between-country differences (all $t < 1.00$, all $p > .350$, independent *t*-tests).

6.3 Results

Surprisingly, none of our measures of interest showed significant differences between the different cultural contexts and the same qualitative pattern was observed in all three cultural contexts. We will therefore ignore cultural differences but stress that the observed effects were found irrespective of the cultural context.

6.3.1 Experiment 1 (Denmark, Iran and China)

To address the question of how individual differences in competence affect joint decisions, we first asked if the group outcome depended on who arbitrated a disagreement trial. More precisely, we calculated the sensitivity of group decisions separately for the trials arbitrated by the less and the more sensitive member of each dyad ($S_{\min}$ and $S_{\max}$) – with group outcome defined as the ratio of the sensitivity of the dyad to that of the more sensitive dyad member (i.e., $S_{\mathrm{dyad}}/S_{\max}$). The group outcome was significantly higher on the trials arbitrated by the more sensitive dyad members ($t(46) = -4.18$, $p < .001$, paired *t*-test). We next asked if the differences in group outcome were due to different advice-taking strategies. Overall, there was evidence of a small egocentric bias, with participants only following their partner's decision on 45% ± 16% (*mean* ± *SD*) of disagreement trials – a finding which is consistent with previous work on egocentric advice discounting (Yaniv & Kleinberger, 2000; Yaniv, 2004). However, splitting the data as above showed that, whereas the more sensitive dyad members followed their partners' decision less often (41% ± 11%), the less sensitive dyad members followed their partner's decision just as often as their own decision (50% ± 20%) (**Figure 6-2A**; $t(46) = -2.35$, $p = .023$, paired *t*-test). Taken together, the results indicated that the more sensitive dyad members had a larger – more positive – influence on the joint decisions.

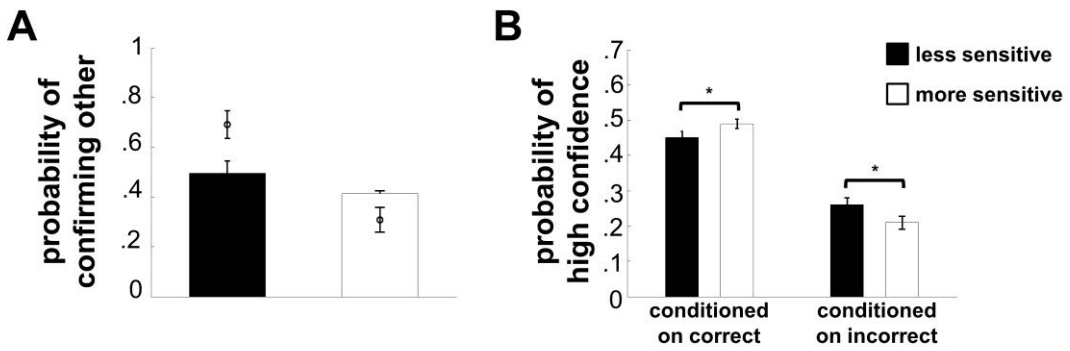


Figure 6-2. Results of Experiment 1: behaviour. **(A)** The proportion of arbitration trials on which the participant followed their partner's decision instead of their own. The dots indicate what the optimal model would have done. **(B)** The proportion of trials on which participants reported high confidence (i.e., above their individual mean confidence) calculated separately for correct and incorrect trials. Black: less sensitive dyad members. White: more sensitive dyad members. Each bar is averaged data, and error bars are 1 SEM.

A critical insight came from evaluating participants' reported confidence and the accuracy of their individual decisions. For each participant, we calculated the probability that a participant reported high confidence in a correct decision (proportion of correct trials on which the participant reported high confidence) and in an incorrect decision (proportion of incorrect trials on which the participant reported high confidence). Here, confidence was defined as a binary variable (below mean confidence: low; above mean confidence: high). A 2 (dyad member: less versus more sensitive) by 2 (accuracy: wrong versus correct decision) mixed-ANOVA (dependent variable: probability high confidence) showed an expected significant main effect of accuracy (**Figure 6-2B**; $F(1,92) = 491.20$, $p < .001$) but, importantly, no main effect of dyad member (**Figure 6-2B**; $F(1,92) = .001$, $p = .967$). However, there was a significant interaction between dyad member and accuracy (**Figure 6-2B**; $F(1,92) = 18.31$, $p < .001$). In particular, the less sensitive dyad members were more likely (compared to their partner) to report high confidence in their incorrect decisions (**Figure 6-2B**; $t(46) = -2.15$, $p = .037$, paired *t*-test,). Conversely, the more sensitive members were more likely (compared to their partner) to report high confidence in their correct decision (**Figure 6-2B**; $t(46) = 2.39$, $p = .021$, paired *t*-test). The interaction suggests that the less sensitive dyad members were more likely (than their partner) to lead the group astray by expressing high confidence in their errors; in contrast, the more sensitive dyad members were more likely (than their partner) to lead the group to the right answer by expressing high confidence when they were correct.

To estimate how participants weighted their partner's opinion relative to their own, we used our computational model (see **Section 6.2.4.2: Computational model** for details). Briefly, we assumed that a participant – when making a joint decision on behalf of the dyad – computed and thresholded a decision variable, $DV = p_s * r_s + \gamma * p_o * r_o$, where s denotes the participant, o denotes the partner, p denotes the estimate of probability correct, r denotes the response ("signed confidence") and γ determines the influence of the partner. We calculated γ for each participant under two constraints: (1) the value that maximised the fit

between the model-derived and the empirically observed decisions, γ_{fit}, and (b) the value that maximised the accuracy of the model-derived decisions, γ_{opt}. Critically, by comparing γ_{fit} with γ_{opt}, we could judge whether a participant *over-weighted* (i.e., $\gamma_{\mathrm{fit}} > \gamma_{\mathrm{opt}}$) or *under-weighted* (i.e., $\gamma_{\mathrm{fit}} < \gamma_{\mathrm{opt}}$) their partner's opinion relative to their own.

We first tested whether the less sensitive and the more sensitive dyad members differed in terms of how they weighted their partner's opinion relative to their own. Interestingly, the less sensitive dyad members under-weighted (i.e., $\gamma_{\mathrm{fit}} < \gamma_{\mathrm{opt}}$) the opinion of their better-performing partners (**Figure 6-3A**, black bars; $t(46) = -5.78$, $p < .001$, paired *t*-test). In contrast, the more sensitive dyad members over-weighted (i.e., $\gamma_{\mathrm{fit}} > \gamma_{\mathrm{opt}}$) the opinion of their poorer-performing partner (**Figure 6-3A**, white bars $t(46) = 2.27$, $p < .028$, paired *t*-test). Critically, the optimal model (γ_{opt}) allowed us to calculate the proportion of trials on which the participant *should* have taken their partner's advice. Echoing the results above, the less sensitive dyad members took their partner's advice significantly less often than they should have (compare the black bar, empirical data, and the circle above it, optimal model, in **Figure 6-2A**; $t(46) = -4.96$, $p < .001$, paired *t*-test). In contrast, the more sensitive dyad members took their partner's advice significantly more often than they should have (compare white bar, empirical data, and the circle below it, optimal model, in **Figure 6-2A**; $t(46) = -3.66$, $p < .001$, paired *t*-test). We return to the implication of these findings for the literature on egocentric advice discounting in **Section 6.4.4: Relationship to studies on advice taking**.

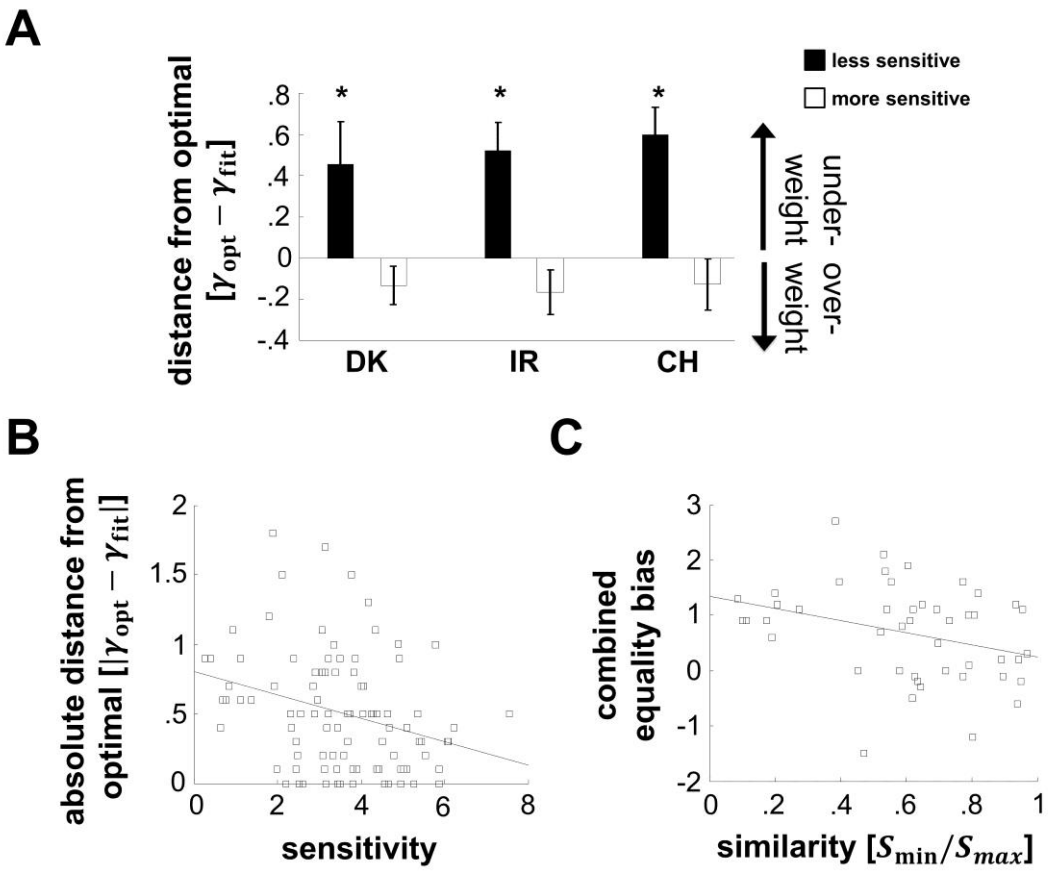


Figure 6-3. Results of Experiment 1: computational model. **(A)** Deviation from optimal performance ($\gamma_{\mathrm{opt}} - \gamma_{\mathrm{fit}}$). Data plotted separately for each country. Positive and negative values indicate under- and over-weighting, respectively. Black: less sensitive dyad members. White: more sensitive dyad members. Each bar is averaged data, and error bars are 1 SEM. **(B)** Absolute deviation from optimal weighting ($|\gamma_{\mathrm{opt}} - \gamma_{\mathrm{fit}}|$; y-axis) plotted against sensitivity (S; x-axis). Each square is a participant. **(C)** Combined equality bias ($(\gamma_{\mathrm{opt,min}} - \gamma_{\mathrm{fit,min}}) - (\gamma_{\mathrm{opt,max}} - \gamma_{\mathrm{fit,max}})$; y-axis) plotted against the similarity of dyad members sensitivities ($S_{\mathrm{min}}/S_{\mathrm{max}}$; x-axis). Each square is a dyad. For panels B and C, data are collapsed across countries, and each line is the best-fitting line of a linear regression.

We then tested if less and more sensitive dyad members differed in terms of how effectively they weighted their partner's opinion relative to their own (i.e., the absolute difference between the fitted and the optimal weight, $|\gamma_{\mathrm{fit}} - \gamma_{\mathrm{opt}}|$). Interestingly, the less sensitive dyad members were significantly worse at weighting their partner's opinion (i.e., $|\gamma_{\mathrm{fit}} - \gamma_{\mathrm{opt}}|$ larger for S_{min} than S_{max}; t(46) = 4.58, p < .001, paired t-test). In addition, we found that, across participants, the absolute difference between the optimal and the fitted weight was negatively correlated with individual sensitivity (**Figure 6-3B**; *r*(92) = -.29, p = .005, Pearson's correlation). In other words, the more sensitive dyad members were better at judging their own competence relative to their partner. This set of findings is in line with the earlier finding that misjudgements of one's own competence is more severe in individuals with low competence than those with high competence (the Dunning-Kruger effect; Kruger & Dunning, 1999). However, our results go beyond the earlier finding by probing people's implicit beliefs as revealed through their decisions rather than eliciting explicit judgements.

The above pattern of behaviour can be described as an *equality* bias – that is, participants appeared to arbitrate the disagreement trials *as if* they were as good as or as bad as their partner. In the context of our model, such an equality bias would entail that the fitted weights, γ_{fit}, should cluster around 1, whereas the optimal weights, γ_{opt}, should be scattered away from 1. Indeed, we found that the fitted weights were significantly closer to 1 (i.e. $|1 - \gamma_{\mathrm{fit}}| < |1 - \gamma_{\mathrm{opt}}|$; *t*(93) = -4.19, *p* < .001, paired *t*-test). Critically, an equality bias should be useful when true. More precisely, if dyad members are indeed equally sensitive, then the fitted and the optimal weights should be close to 1. However, for dyad members with different levels of sensitivity, the fitted and the optimal weights should diverge dramatically as an equality bias would be counterproductive. This entails that the similarity of dyad members' sensitivities, $S_{\mathrm{min}}/S_{\mathrm{max}}$, should be negatively correlated with their combined equality bias, which we defined as $(\gamma_{\mathrm{opt,min}} - \gamma_{\mathrm{fit,min}}) - (\gamma_{\mathrm{opt,max}} - \gamma_{\mathrm{fit,max}})$. In particular, the equality bias is

expected to drive the less and more sensitive dyad members in opposite directions, meaning that $(\gamma_{\text{opt,min}} - \gamma_{\text{fit,min}}) > 0$ and $(\gamma_{\text{opt,max}} - \gamma_{\text{fit,max}}) < 0$. We therefore reasoned that the combined equality bias would be best estimated by subtracting these two quantities from one another. Under this calculation, more positive values indicate a higher equality bias. In support of our prediction, there was a negative correlation between the similarity of dyad members' sensitivities and their combined equality bias (**Figure 6-3C**; $r(45) = -.32$, $p = .032$, Pearson's correlation).

It is important to note that only the positive values of $\gamma_{\text{opt}} - \gamma_{\text{fit}}$ indicate egocentric advice discounting. Indeed, when collapsed across all participants, $\gamma_{\text{opt}} - \gamma_{\text{fit}}$ (*mean* ± *SD* = .19 ± .63) was significantly positive ($t(93) = 2.95$, $p = .004$, one-sample *t*-test). Had we followed the convention established by previous studies on advice taking – that is, only looking at the average of $\gamma_{\text{opt}} - \gamma_{\text{fit}}$ across all participants rather than splitting the data according to S_{min} and S_{max} – we might have taken our results as evidence of egocentric advice discounting. As shown above, however, our results indicate that two types of advice discounting, egocentric and self-discarding, may in fact coexist in social interaction.

6.3.2 Experiment 2 (Iran)

One critique of Experiment 1 is that limitations on memory capacity might have made it too difficult for participants to track the history of trial outcomes (i.e., p_{s} and p_{o}) – leading to an inability to decide whose opinion is more accurate on a given trial. We tested this hypothesis in Experiment 2 in which participants performed the same task as in Experiment 1, except that they, in addition to the *current trial* accuracy, were shown their own and their partner's *cumulative* accuracy up until the current trial. Despite this memory aid, we found the same effect as in Experiment 1. In particular, the less sensitive dyad members under-weighted the

opinion of their better-performing partners (i.e., $\gamma_{\mathrm{fit}} < \gamma_{\mathrm{opt}}$; $t(10) = -4.69$, $p < .001$, paired t-test). In contrast, the more sensitive dyad members over-weighted the opinion of their poorer-performing partners (i.e., $\gamma_{\mathrm{fit}} > \gamma_{\mathrm{opt}}$; $t(10) = 3.73$, $p = .004$, paired t-test).

6.3.3 Experiment 3 (Denmark and Iran)

Another critique of Experiment 1 is that participants might have exhibited an equality bias because their individual differences in sensitivity were (at least in some cases) small or accidental. We tested this hypothesis in Experiment 3 in which participants performed the same task as in Experiment 1, except that we now covertly made the task much harder for one of the two participants – thus creating a significant difference in performance between the less sensitive and the more sensitive dyad members ($t(22) = 8.85$, $p < .001$, paired t-test). Despite this conspicuous difference in individual performance, we found the same effect as in Experiment 1. In particular, the less sensitive dyad members under-weighted the opinion of their better-performing partners (i.e., $\gamma_{\mathrm{fit}} < \gamma_{\mathrm{opt}}$; $t(22) = -3.38$, $p = .003$, paired t-test). In contrast, the more sensitive dyad members over-weighted the opinion of their poorer-performing partners (i.e., $\gamma_{\mathrm{fit}} > \gamma_{\mathrm{opt}}$; $t(22) = 2.11$, $p = .039$, paired t-test).

6.3.4 Experiment 4 (UK)

Finally, another critique of Experiment 1 is that participants might have exhibited an equality bias because they had no incentives to override each other's opinions. We tested this hypothesis in Experiment 4 in which participants performed the same task as in Experiment 1, except that they, instead of rating their confidence, placed post-decision wagers (max £1) on their own decision and on the joint decision when arbitrating. The payoff on a given trial determined by the sum of the placed bets (correct: positive values; incorrect: negative values).

Despite monetary incentives being introduced, we found the same effect as in Experiment 1. In particular, the less sensitive dyad members under-weighted the opinion of their better-performing partners (i.e., $\gamma_{\mathrm{fit}} < \gamma_{\mathrm{opt}}$; $t(16) = -3.91$, $p = .001$, paired t-test). In contrast, the more sensitive dyad members over-weighted the opinion of their poorer-performing partners (i.e., $\gamma_{\mathrm{fit}} > \gamma_{\mathrm{opt}}$; $t(16) = 5.04$, $p < .001$, paired t-test).

6.4 Discussion

Individual differences in competence are a key challenge for group decision-making (e.g., jury voting, policy making and medical diagnosis). While theory tells us that the opinion of each group member should be weighted by its reliability (Nitzan & Paroush, 1985), empirical research cautions that this is easier said than done. For example, we tend to grossly misestimate our own competence – not only when judged in isolation but also relative to others (Kruger & Dunning, 1999). This raises the question whether people can or do take into account individual differences in competence in group decision-making.

To address this question, we developed a computational framework inspired by previous work on how people learn about the reliability of (non-social and social) information (Behrens et al., 2009, 2008, 2007). By means of this approach, we could quantify how members of a group weighted each other's opinions in our psychophysical task. Our participants exhibited an *equality* bias – behaving *as if* they were *as good* or *as bad* as their partner. We could exclude several alternative explanations. In Experiment 2, we found the equality bias even when participants were presented with a running score of each other's performance. We could thus rule out that the observed effects were due to poor memory for past outcomes. Further, we could in part rule out that the observed effects were due to a previously reported *self-enhancement* or *self-superiority* bias (Hoorens, 1993) as these were (at least for the less

sensitive group members) explicitly contradicted by the running score. In Experiment 3, we observed the equality bias even when participants' individual performance was covertly manipulated so as to differ significantly. We could thus rule out that the observed effects were due to small or happenstance differences in performance (i.e., due to random noise and not competence per se). In Experiment 4, we found the equality bias even when it was in participants' monetary interest to make as many correct group decisions as possible. We could thus rule out that the observed effects were limited to situations without financial incentives. Finally, we found the equality bias in three cultures (Denmark, Iran, and China) that differ widely in terms of their norms for interpersonal trust (see **Section 6.1: Introduction**), suggesting that the bias may reflect a general strategy for group decision-making rather than a culturally specific norm.

6.4.1 Can our results be explained by regression to the mean?

Regression to the mean may happen in situations in which (a) one variable is measured at two time points or (b) two variables are measured at the same time point. In the first situation, if a variable is extreme (relative to the population distribution) on its first measurement, then it tends to be less extreme on its second measurement. Similarly, if a variable is extreme on its second measurement, it tends to have been less extreme on its first measurement. In the second situation, if one variable is measured to be extreme, then another variable measured at the same time will tend to be less extreme, and vice versa. Therefore, regression to the mean should be considered as a potential confound whenever data is split according to where data points fall along some distribution – as is (roughly) the case with our split of participants into less sensitive ($S_{\min}$) and more sensitive ($S_{\max}$) group members. However, we believe that there are two key reasons why regression to the mean cannot explain our finding that less

sensitive group members under-weight their partner's opinion (i.e., $\gamma_{\mathrm{fit}} < \gamma_{\mathrm{opt}}$), whereas more sensitive group members over-weight their partner's opinion (i.e., $\gamma_{\mathrm{fit}} > \gamma_{\mathrm{opt}}$).

First, if sensitivity (i.e., S) and optimality (i.e., $\gamma_{\mathrm{fit}} - \gamma_{\mathrm{opt}}$) were two (uncorrelated) random variables, then – under regression to the mean – we would expect participants who were extreme in terms of sensitivity (e.g., very low or very high) to be less extreme in terms of optimality (i.e., $\gamma_{\mathrm{fit}} - \gamma_{\mathrm{opt}}$ approaching 0), and vice versa. However, we found – as indicated by the correlation shown in **Figure 6-3A** – that participants who were extreme in terms of sensitivity (in particular – very low) also were extreme in terms of optimality (i.e., $\gamma_{\mathrm{fit}} - \gamma_{\mathrm{opt}}$ far from 0). Second, regression to the mean is more expected when data is split according to the distribution of values across the population. However, we did not categorise participants according to the distribution of sensitivity across participants – but according to the within-dyad relationship in sensitivity. Therefore, a participant who was the less sensitive member of one dyad could have been the more sensitive member of another dyad, again making it unlikely that our results were due to regression to the mean.

6.4.2 Cause of equality bias

Participants may have displayed an equality bias, for social or strategic reasons. At the social level, participants may have tried to diffuse the responsibility for difficult group decisions (Harvey & Fischer, 1997). On this account, group members would be more likely to alternate between their own opinion and that of their partner when the decision is particularly difficult (high uncertainty), thus sharing the responsibility for potential errors. To test this hypothesis directly, one could vary the potential outcome of group decisions and quantify the degree of equality bias when (a) the reward is small and the cost of an error is high versus (b) the reward is large and the cost is low. The shared responsibility hypothesis would predict a stronger equality bias under the former condition. Alternatively, participants may have tried to avoid

social exclusion (Williams, 2007). Social exclusion invokes strong aversive emotions that – some have even argued – may resemble physical pain (Eisenberger, Lieberman, & Williams, 2003). On this account, worse group members might under-weight their partner's opinion in order to stay relevant and socially included. Conversely, better group members might over-weight their partner's opinion in order to maintain some pro-social reputation.

At the strategic level, an equality bias may arise because group members "normalise" their partner's confidence to their own. In our task, while the better-performing members of each group were more confident, they also over-weighted the opinion of their respective partners, and vice versa. This suggests that participants may have rescaled their partners' confidence to their own and made the joint decisions on the basis of these re-scaled values. Following on from **Chapter 5**, another possibility is that participants in the current task may have engaged at last partially in confidence matching. In particular, the optimal weights might diverge from the fitted weights because the confidence reports were too similar but taken at face value – for social reasons or because participants genuinely believed that confidence matching is a good solution to the communication problem (**Figure 4-2**). Indeed, pooling data across countries in Experiment 1, we found a correlation between the magnitudes of dyad members' mean confidence ($r(45) = .51$, $p < .001$, Pearson's correlation).

6.4.3 Benefits of equality bias

It is also constructive to consider what benefits an equality bias may confer; indeed, cognitive biases often serve as heuristic solutions to complex problems (Gigerenzer & Brighton, 2009). The equality bias may facilitate social coordination, with participants trying to get the task done with minimal effort but at the potential expense of group accuracy. Indeed, it has been argued that interacting individuals rapidly converge on social norms (i.e., coordination devices) to reduce "disharmony" and "chaos" in joint tasks (Schelling, 1980) – the equality bias may be

one such coordination device. As in the case of confidence matching described in **Chapter 4**, the equality bias should work when group members have similar levels of competence. In such cases, the equality bias reduces cognitive load, with little cost to group accuracy. Further, as discussed in **Chapter 4**, we tend to associate with similar others with whom we are more likely to share common levels of competence (McPherson et al., 2001). In addition, even if people were randomly paired from a large population, they would, for any normally-distributed skill, be more likely to be similar (close to the distribution mean) than not (from the opposite tails of the distribution). Future research could refine our understanding of the utility of cognitive biases such as confidence matching and the equality bias by assessing the probability that two (randomly-paired) individuals are *too different* in terms of competence.

6.4.4 Relationship to studies on advice taking

Research on advice taking (Yaniv & Kleinberger, 2000; Yaniv, 2004) predicts that group members would tend to discount the advice of their partner. Only looking at our behavioural results, one may have concluded that only the better-performing members of each group displayed such egocentric advice discounting (white bars in **Figure 6-2A**). However, the model-based analysis showed that the better-performing group members *should* have followed their partner's opinion *even less often* than they actually did (compare white bars and circles in **Figure 6-2A**). Conversely, their poorer-performing counterparts should have followed their partner's opinion *even more often* than they actually did (compare black bars and circles in **Figure 6-2A**). This divergence between the conclusions from the behavioural data and the model-based analysis underscores the power of the latter approach for understanding social behaviour. More generally, thinking of advice taking as purely "egocentric" may miss out subtle nuances in actual social interaction (e.g., people with high competence should generally be egocentric in group decision-making).

6.5 Chapter summary

In this chapter, we extended our basic framework to *study* and *model* how the *combining* of representations of uncertainty unfolds in the context of group decision-making (**Figure 3-4**). We tested whether pairs of participants engaged in group decision-making could learn to weight each other's opinions in an optimal manner. In separate experiments, we found that participants did not weight each other's opinions in an optimal manner. Instead, they displayed an equality bias – behaving as if they were as good as or as bad as their partner – regardless of cultural context, explicit feedback about their relative competence, conspicuous differences in competence and monetary incentives. The equality bias could be a strictly social phenomenon in the sense that it mitigates the responsibility for group errors and social exclusion. Alternatively, the equality bias may be an active strategy and a natural consequence of deliberate confidence matching.

In the early years of the twentieth century, Marcel Proust, a sick man in bed but armed with a keen observer's vision, wrote "Inexactitude, incompetence do not modify their assurance, quite the contrary" (Proust, 2006). Indeed, our results show that people fail to take into account their own "inexactitudes". But are people able to *learn* or are they, as implied by Marcel Proust's melancholic tone, forever trapped in their incompetence?[18]

[18] We did address this question by splitting our data (Experiment 1) into two sessions (2 x 128 trials) to test whether participants' moved closer towards the optimal weight over time. However, we found no statistically reliable difference between the two sessions. This could be due genuine inflexible behaviour or that we did not have sufficient data (power) to address the issue of learning.

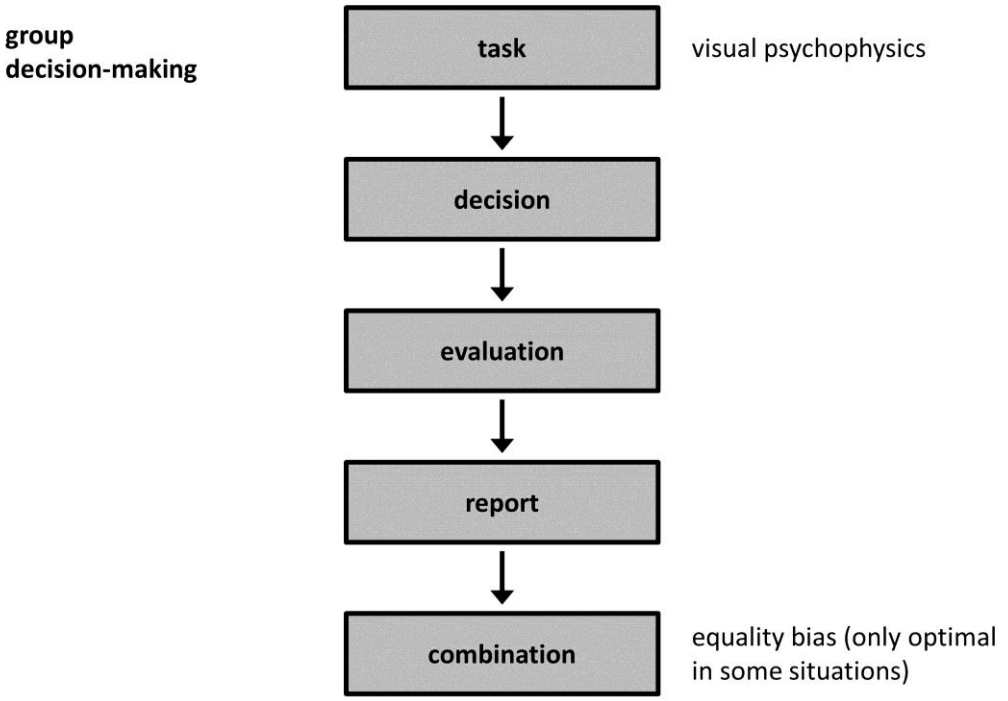


Figure 6-4. Summary of Chapter 6. The boxes denote the component processes involved in sharing and combining representations of uncertainty.

Chapter 7: General discussion

The aim of this chapter is to situate the thesis within research on group decision-making and confidence more generally. We summarise our key findings. We address recurrent criticisms of our work: familiarity among group members; the presence of feedback; the task domain; and the confidence scale. Lastly, we highlight directions for future research: the multiple factors which may bias the generation of a confidence report; the dissociation between a confidence variable and a confidence report; and the radical hypothesis that inaccurate confidence reports are largely of a social nature.

7.1 Thesis summary

Many risky decisions are made by groups of people. Exam decisions, for example, are typically made by a panel of experts. Groups should in principle benefit from combining the opinions of their members (Bovens & Hartmann, 2003; Condorcet, 1785), but this process can easily be jeopardised by social dynamics; for example, people's desire for cohesion might lead them to suppress valid but unpopular viewpoints (Packer, 2009) or isolate themselves from outside influences (Minson & Mueller, 2012). It is therefore not surprising that a recent survey of academic opinions listed "*how can we aggregate information possessed by individuals to make the best decisions?*" as one of the most pressing questions in the 21st century (Giles, 2011).

Numerous disciplines have indeed studied group decision-making – from the humanities and the social sciences to the life sciences. The so-called *soft* disciplines have studied the topic as a window into social life, using texts, observation and psychological testing. The so-called *hard* disciplines have studied group decision-making as a problem of information aggregation, using mathematical models and computer simulations. In this thesis, we sought to bridge this divide, using *hard* tools to ask *soft* questions. In particular, we combined visual psychophysics with computational modelling to study how the sharing and the combining of representations of uncertainty unfold in *real* social interaction. We used this approach to address a series of outstanding questions in the literature:

How do people generate a confidence report? In **Chapter 2**, we showed that three processes govern the generation of a confidence report. First, people accumulate evidence from memory or the senses towards a decision (e.g., that a football crossed the goal line). Second, people transform the accumulated evidence into a confidence variable that relates to the probability that their decision is correct. This process is largely internal, and proceeds, at least in some tasks, in an optimal manner, with people *directly* computing the probability that their decision is correct. Lastly, people map their confidence variable onto a confidence report (e.g., "I'm sure it was a goal"). This process turns out to be subject to external, and especially social, considerations. We demonstrate that *confidence* can be optimal in at least two senses. First, the confidence variable can be computed in an optimal manner. Second, the confidence variable can be mapped onto a confidence report in an optimal manner. Critically, optimality at one level does not imply optimality at the other level. More generally, our findings pose a problem for any measure of metacognition that does not dissociate the computation of the confidence variable and the mapping of this variable onto a confidence report.

How do people's confidence reports change with experience? In **Chapter 3**, we showed that people adjust the mapping of the confidence variable onto a confidence report in a flexible manner that is divorced from the first-order input – with people adjusting their confidence reports according to the history of reports given and feedback obtained. More generally, our findings pose a problem for current theories (and measures) of metacognition which assume that the report of confidence should be stable over time. Further, in **Chapter 4**, we showed that people engaged in group decision-making match each other's confidence reports – both at the level of their trial-by-trial confidence and at the level of their mean confidence and confidence distributions. This strategy – *confidence matching* – turns out to be useful when differences in competence are uncommon and/or people have no insight into their relative competence. First, confidence matching is beneficial for group members of nearly equal competence, but costly for group members with different levels of competence. Second,

confidence matching is beneficial when there is an initial mismatch between the relative confidence and the relative competence of group members. In contrast, confidence matching is costly when there is an initial match between the relative confidence and the relative competence of group members. More generally, our findings demonstrate that confidence reports are socially malleable, and suggest that well-known biases, such as overconfidence, might in fact reflect particular norms for social interaction.

How do people evaluate each other's confidence reports once expressed? In **Chapter 5**, we showed that people take into account the credibility of each other's confidence reports when making decisions as part of a group. This "weighting-by-reliability" allows group members to outperform simple heuristics for group choice that simply take confidence reports at their face value. However, in **Chapter 6**, we showed that the weighting strategy used by people engaged in group decision-making falls short of optimal; people display an equality bias and behave as if they were as good or as bad as their partners. While less competent group members tend to under-weight the confidence reports of more competent partners, more competent group members tend to over-weight the confidence reports of less competent partners. More generally, our findings demonstrate that our evaluation of other people's confidence reports is strongly egocentric.

The work presented in this thesis (a) has been subject to a set of recurrent criticisms and (b) raises important questions for future research. We will deal with these two issues in turn.

7.2 Recurrent criticisms

7.2.1 Familiarity

One recurrent criticism is whether the shared history between participants may have biased their choice of strategy. In particular, in our social studies, for each dyad, one participant was

recruited, and then asked to bring along a friend. In **Chapter 4** and **Chapter 6**, participants may therefore have engaged in confidence matching or exhibited an equality bias so as to maintain cohesion (or a smooth relationship) outside the experimental situation. Further, in **Chapter 5**, participants may have relied on the history of past interactions to establish the credibility of each other's confidence reports. We note, however, that we observed confidence matching even when participants were blind to the identity of their partner (virtual agents). Further, in **Chapter 6**, we observed the equality bias even when participants had strong financial incentives to sacrifice cohesion for group accuracy. Lastly, in relation to **Chapter 5**, Bahrami et al. (2013), using a similar task, found that familiarity had no impact on group performance, supporting that our participants evaluated the credibility of each other's confidence reports in the context of the current task. However, while Bahrami et al. (2013) found no *main effect* of familiarity on group performance, it may be that a shared history is more helpful (or harmful) for participants with different levels of competence.

7.2.2 Feedback

Another recurrent criticism is whether the current findings would generalise to situations in which there is no (immediate) feedback about decision accuracy. For example, in **Chapter 5**, participants may only have used the trial-by-trial feedback to evaluate the credibility of each other's confidence reports. Indeed, previous research has shown that feedback helps groups of individuals to identify their more accurate members (Henry, Strickland, Yorges, & Ladd, 1996). The relative benefit of interaction over the MCS algorithm reported in **Chapter 5** might thus not persist in the absence of feedback. However, using a similar task, Bahrami et al. (2012) found that feedback was not necessary for the accumulation of a 2HBT1 effect – feedback only accelerated the process – indicating that interacting individuals may rely on other signals when they learn about the credibility of each other's confidence reports. The

identification and incorporation of these signals will pose a major challenge for models of social learning. We note that, in SDT models of confidence, participants do indeed have access to some subjective representation of uncertainty. As such, they may have compared their own responses to those of their partner, or paid attention to the number of disagreement trials, to establish whether their partner could be trusted. More generally, we would in fact predict that the confidence matching observed in **Chapter 4** and the equality bias observed in **Chapter 6** would be more pronounced in situations without feedback. In particular, confidence matching as well as the equality bias can be thought of as "equilibrium" strategies (not best but also not worst) when an insight into one's own or other's competence is hard to establish.

7.2.3 Task domain

Another recurrent criticism is whether the current findings would generalise to other task domains. We used sensory tasks for three reasons. First, sensory tasks have an objectively true answer, thus allowing us to specify how an optimal agent or a pair of such agents would solve a given decision problem. Second, we can fully control the task-relevant information (cf. participants can rely on complex memories in general-knowledge tasks). Lastly, sensory noise in two brains is uncorrelated, allowing us to study situations in which participants can gain a significant advantage by sharing their representations of uncertainty (cf. participants might share too many task-irrelevant memories in general-knowledge tasks). This experimental setup allowed us not only to isolate but also to develop a mechanistic understanding of the component processes involved in sharing and combining representations of uncertainty in the context of group decision-making. However, it may have removed factors which play a critical role in everyday life. For example, in **Chapter 4** and **Chapter 6**, participants may have engaged in confidence matching or shown an equality bias because they had no expectations about their own or their partner's visual sensitivity – or had a strong expectation that there is no

individual variation in visual sensitivity (e.g., “everyone has normal or corrected vision”). In contrast, in general-knowledge or forecasting tasks, participants might more readily accept differences in competence and respond accordingly. An important question for future research is therefore whether the current findings would generalise to situations with strong a priori expectations about competence.

We note, however, that the illusory feeling of superiority displayed by relatively incompetent individuals (the Dunning-Kruger effect; Kruger & Dunning, 1999) – which we also observed among our less-sensitive participants in **Chapter 6** – has been documented in tasks in which participants should have had sufficient experience to build up accurate expectations about their own competence (e.g., logical reasoning and grammatical skills). Further, as with the absence of feedback, we would in fact predict that the confidence matching observed in **Chapter 4** and the equality bias observed in **Chapter 6** would be more pronounced in domains with multiple sources of uncertainty. For example, pollsters adjust their estimates to those of other pollsters when predicting the outcome of an upcoming election (Silver, 2013). While this *herding* (Raafat, Chater, & Frith, 2009) reduces the accuracy of the *average* prediction across pollsters (correlated noise), it improves the accuracy of the *individual* predictions. In other words, an estimate of the likelihood of an event can often be improved by adjusting it to the estimates made by others. The confidence matching observed in **Chapter 4** may have reflected this type of *informational* conformity (Sherif, 1935) rather than *normative* conformity (Asch, 1951); we might therefore expect this type of behaviour to be more pronounced when the sources of uncertainty are difficult to identity and larger advantages could be gained from adjusting confidence reports to those made by others (Toelch & Dolan, 2015).

7.2.4 Confidence scale

Another recurrent criticism is whether the current findings are limited to our confidence scale. We used an ordinal confidence scale (e.g., 1 to 6 in discrete steps of 1) for two reasons. First, our confidence scale mimics relevant aspects of language. In verbal communication, while most people would agree on the ordering of confidence expressions (e.g., “sure” > “unsure”), they would often assign different meanings to each expression (e.g., saying “unsure” when they are either 50% or 60% likely to be correct; **Figure 4-1**). Similarly, while the different levels of our confidence scale have an ordering, the (statistical) meaning associated with each level is unspecified. Second, our confidence scale allows us to record responses in a format that allows for modelling. However, despite these advantages, the scale may have introduced at least one confound that should not pertain to non-arbitrary scales. In **Chapter 3**, participants may have reported their confidence in a globally inconsistent manner because of the vague nature of our confidence scale, using their past reports as some anchor for current reports. In **Chapter 4**, participants may similarly have used their partner’s reports as some anchor for current reports. An important question for future research is therefore whether the current findings would generalise to other formats for communication.

We note, however, that we observed global inconsistency for a probability scale in **Chapter 3**; in this case, the (statistical) meaning associated with each level was fully specified. Further, as discussed in **Chapter 1**, probability scales by no means guarantee accurate confidence reports (**Figure 1-4**). More generally, people’s explicit understanding of probabilities is rather poor – mirroring the hard-easy effect (**Figure 1-4**) – with people over-estimating the likelihood of low-probability events and under-estimating the likelihood of high-probability events (Hertwig & Erev, 2009; Tversky & Kahneman, 1992). In this light, we would therefore also expect to find confidence matching for a probability scale; in fact, the introduction of a probability scale may induce more herding behaviour for the reasons described above. Lastly, we note that, in the linguistic analysis by Fusaroli et al. (2012) of the conversations in Bahrami et al.'s (2010) study,

participants were found not only to converge on some global set of confidence expressions but also to recycle each other's confidence expressions from recent trials; as such, confidence matching as described in **Chapter 4** does not seem to be limited to any particular format for communication. It remains to be seen whether the linguistic recycling observed by Fusaroli et al. (2012) – like the confidence matching observed in **Chapter 4** – has a negative effect on the group performance of individuals with different levels of competence.

7.3 Future directions

7.3.1 What factors influence confidence reports?

We observed that our participants' confidence reports reflected not only the history of their own past reports and feedback obtained but also the history of their partner's past reports and feedback obtained. However, confidence reports in everyday life are likely to reflect more factors than we considered here. An important direction for future research is to tease apart and model the different factors which may be at play in social interaction.

One important factor might be *status*. For example, Bayarri & Degroot (1989) considered a (hypothetical) scenario in which a number of advisors report the strength of their belief in some event (e.g., the probability that a stock goes up) to a client. The client has assigned a prior weight (status) to each advisor but will update these weights according to the accuracy of their prediction (e.g., the stock went up) and the strength of this prediction (e.g., predicted with high certainty) once the event has been observed. In this scenario, the advisors should report the strength of their belief in a manner that maximises their posterior weight (e.g., a future bonus) – which need not be the true strength of their belief. Bayarri & Degroot (1989) showed mathematically that low-weight advisors should give extreme predictions, whereas high-weight advisors should give conservative predictions. While low-weight advisors have little to lose, high-weight advisors would want to preserve their current status for the future.

Another important factor might be *reputation*. In particular, we seem to have a preference for decisive individuals (Penrod & Cutler, 1995). Individuals might therefore seek to avoid middle-range confidence reports. Indeed, our participants in **Chapter 2** tended to show bimodal confidence distributions (see **Supplementary Figure 1** and **Supplementary Figure 2**). As a potentially insightful example, it has been shown that weather forecasters tend to avoid predicting that the probability of rain is 50% (Silver, 2013) – even if this is their true belief. Instead, if the accuracy of their forecasting is evaluated over say a 10-day period with a predicted 50% probability of rain on all days, they would diversify their individual predictions so as not to appear indecisive, while ensuring that the average of their individual predictions is accurate (e.g., predict that the probability of rain is 70% on 5 days and 30% on 5 days).

Yet another important factor might be *stereotypes*. For example, in **Chapter 4**, we asked our participants in the virtual-agent experiment whether they thought their partners were male or female. Interestingly, despite being gender-balanced, 70% (30%) of our participants thought that the "high-confidence" agents were male (female), whereas 25% (75%) of our participants thought that the "low-confidence" agents were male (female). Because of the anonymisation of data for large-group experiments, we could not test directly whether this expectation was mirrored in the behaviour of our participants. However, it would be surprising if these strong stereotypes had no basis in reality.

We note that we observed significant individual variation in mean confidence and confidence distributions among our participants (see **Supplementary Figure 1** and **Supplementary Figure 2**). Another important direction for future research is to test whether this variation reflects *personality traits* or *general attitudes* towards risk and uncertainty – over and above the factors discussed above. One line of research would be to test whether the *surface* statistics of people's confidence reports extend across formats (e.g., arbitrary and probability scales) and domains (e.g., sensory and general-knowledge tasks). A related line of research would be to

test the *depth* of these statistics; that is, whether implicit response behaviours (e.g., willingness to wait for a reward or the degree of risk-taking in reward-based tasks; e.g., Kepecs & Mainen, 2012) show the same statistical trends as explicit response behaviours. If so, individual biases may operate not only at the level of the confidence report but also at the level of the confidence variable upon which the confidence report is based.

We note, however, that it is hard to dissociate the computation of a confidence variable and the mapping of this variable onto a confidence report. For example, in **Chapter 3**, we cannot definitively rule out that the observed serial dependence in confidence reports arose during the computation of the confidence variable through some prior expectations rather than the mapping of this variable onto a confidence report through some calibration process. On the former account, participants truly *felt* more likely to be correct – rather than simply *expressed* that they were more likely to be correct – following a high confidence report or positive feedback. Similarly, in **Chapter 4**, we cannot definitively rule out that participants truly *felt* more (or less) likely to be correct – rather than simply *expressed* that they were more (or less) likely to be correct – when they interacted with more (or less) confident partners. Another important direction for future research is to test whether the changes at the level of explicit response behaviours (induced by feedback or computer-generated partners) trickle down to the level of implicit response behaviours (e.g., changes in the willingness to wait for a reward or the degree of risk-taking in reward-based tasks).

7.3.2 What causes inaccurate confidence reports?

As discussed in **Chapter 1**, confidence reports can be remarkably inaccurate, with people being underconfident or overconfident about the objective accuracy of their responses (**Figure 1-4**). The dominant view in the sciences of the mind and brain is that such inaccurate confidence reports are due to biases in the way in which the human mind represents uncertainty, or a lack

of insight into one's own internal operations (Nigel Harvey, 1997). We propose, instead, that such errors are largely of a social nature: that is, people do not suffer from a biased mind, or a lack of insight, but follow, implicitly or explicitly, different norms for how confidence should be communicated. Our hypothesis resonates with the proposal that human faculties commonly believed to be "hard-wired" (e.g., colour perception, spatial cognition and social cognition) are under a much stronger social or cultural influence than previously thought (Davidoff, 2001; Heyes & Frith, 2014; Majid, Bowerman, Kita, Haun, & Levinson, 2004).

Our computational framework allows us to formalise the two dominant hypotheses about the source of inaccurate confidence reports and specify their predictions. The first (biased-mind) hypothesis would state that the confidence variable itself is computed in a biased or imprecise manner. This hypothesis would predict that people's implicit and explicit response behaviours are equally biased. The second (no-insight) hypothesis would state that, while the confidence variable is computed precisely, the read out of this information is noisy. This hypothesis would predict that people's responses will be unbiased when the confidence variable is probed implicitly but inconsistent when the confidence variable is probed explicitly.

In contrast to these two hypotheses, our (social) hypothesis is that the confidence variable is computed precisely and that there is no noise in the read out of this information; inaccurate confidence reports arise because people map the confidence variable onto a confidence report in a highly idiosyncratic manner. Our hypothesis predicts that people's responses will be unbiased when the confidence variable is probed implicitly and that their responses will be consistent but idiosyncratic when it is probed explicitly. As such, we expect the mapping of a confidence variable onto a confidence report to be *monotonically increasing* and *deterministic* – once history-effects, which we have argued could result from rational strategies, have been partialled out. Further, we expect the shape of this mapping to correlate with measures of social dispositions. An implication of our hypothesis is that it should be possible to *correct*

inaccurate confidence reports by introducing a new set of norms through incentives, training or social interaction. An important direction for future research is to compare directly these competing explanations of inaccurate confidence reports.

References

Akaishi, R., Umeda, K., Nagase, A., & Sakai, K. (2014). Autonomous mechanism of internal choice estimate underlies decision inertia. *Neuron*, *81*(1), 195–206.

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgment. In H. Guetzkow (Ed.), *Groups, Leadership and Men*. Pittsburgh, PA: Carnegie Press.

Ayton, P., & Fischer, I. (2004). The hot hand fallacy and the gambler's fallacy: two faces of subjective randomness? *Memory & Cognition*, *32*(8), 1369–1378.

Bahrami, B., Didino, D., Frith, C. D., Butterworth, B., & Rees, G. (2013). Collective enumeration. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(2), 338–347.

Bahrami, B., Olsen, K., Bang, D., Roepstorff, A., Rees, G., & Frith, C. D. (2012). Together, slowly but surely: the role of social interaction and feedback on the build-up of benefit in collective decision-making. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(1), 3–8.

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science*, *329*(5995), 1081–1085.

Bang, D., Mahmoodi, A., Olsen, K., Roepstorff, A., Rees, G., Frith, C. D., & Bahrami, B. (2014). What failure in collective decision-making tells us about metacognition. In S. M. Fleming & C. D. Frith (Eds.), *The Cognitive Neuroscience of Metacognition*. Berlin: Springer.

Baranski, J. V, & Petrusic, W. M. (1998). Probing the locus of confidence judgments: experiments on the time to determine confidence. *Journal of Experimental Psychology. Human Perception and Performance*, *24*(3), 929–945.

Barber, B. M., & Odean, T. (2001). Boys will be boys: gender, overconfidence, and common stock investment. *The Quarterly Journal of Economics*, *116*(1), 261–292.

Barlow, H. B. (1961). Possible principles underlying the transformations of sensory messages. In W. A. Rosenblith (Ed.), *Sensory Communication* (pp. 216–234). Cambridge, MA: MIT Press.

Barthelmé, S., & Mamassian, P. (2009). Evaluation of objective uncertainty in the visual system. *PLoS Computational Biology*, *5*(9), e1000504.

Bayarri, M. J., & Degroot, M. H. (1989). Optimal Reporting of Predictions. *Journal of the American Statistical Association*, *84*(405), 214–222.

Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J., Shadlen, M. N., Latham, P. E., & Pouget, A. (2008). Probabilistic Population Codes for Bayesian Decision Making. *Neuron*, *60*(6), 1142–1152.

Beck, J. M., Ma, W. J., Pitkow, X., Latham, P. E., & Pouget, A. (2012). Not noisy, just wrong: the role of suboptimal inference in behavioral variability. *Neuron*, *74*(1), 30–39.

Behrens, T. E. J., Hunt, L. T., & Rushworth, M. F. S. (2009). The computation of social behavior. *Science*, *324*(5931), 1160–1164.

Behrens, T. E. J., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. S. (2008). Associative learning of social value. *Nature*, *456*(7219), 245–249.

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221.

Berlin, B., & Kay, P. (1969). *Basic Color Terms: Their Universality and Evolution*. Berkeley, CA: University of California Press.

Berner, E. S., & Graber, M. L. (2008). Overconfidence as a cause of diagnostic error in medicine. *The American Journal of Medicine*, *121*(5), S2–S23.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. Chichester, UK: John Wiley & Sons, Ltd.

Bovens, L., & Hartmann, S. (2003). *Bayesian Epistemology*. Oxford, UK: Oxford University Press.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3.

Broihanne, M. H., Merli, M., & Roger, P. (2014). Overconfidence, risk perception and the risk-taking behavior of finance professionals. *Finance Research Letters*, *11*(2), 64–73.

Campbell, W. K., Goodie, A. S., & Foster, J. D. (2004). Narcissism, confidence, and risk attitude. *Journal of Behavioral Decision Making*, *17*(4), 297–311.

Cleeremans, A. (2014). Connecting conscious and unconscious processing. *Cognitive Science*, *38*(6), 1286–1315.

Condorcet, M. (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix (Essay on the Application of Analysis to the Probability of Majority Decisions)*. Paris, France: L'Imprimerie Royale.

Davidoff, J. (2001). Language and perceptual categorisation. *Trends in Cognitive Sciences*, *5*(9), 382–387.

Dayan, P., & Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, *8*(4), 429–453.

de Gardelle, V., & Mamassian, P. (2014). Does confidence use a common currency across two visual tasks? *Psychological Science*, *25*(6), 1286–1288.

De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2012). Confidence in value-based choice. *Nature Neuroscience*, *16*(1), 105–110.

Eisenberger, N. I., Lieberman, M. D., & Williams, K. D. (2003). Does rejection hurt? An fMRI study of social exclusion. *Science*, *302*(5643), 290–292.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433.

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences*, *8*(4), 162–169.

Faisal, A. A., Selen, L. P. J., & Wolpert, D. M. (2008). Noise in the nervous system. *Nature Reviews Neuroscience*, *9*(4), 292–303.

Fernandez-Duque, D., Baird, J. A., & Posner, M. I. (2000). Executive attention and metacognitive regulation. *Consciousness and Cognition*, *9*(2), 288–307.

Ferrell, W. R., & McGoey, P. J. (1980). A model of calibration for subjective probabilities. *Organizational Behavior and Human Performance*, *26*(1), 32–53.

Fischer, J., Shankey, J., & Whitney, D. (2011). Serial dependence in visual perception. *Journal of Vision*, *11*(11), 1201–1201.

Fleming, S. M., & Dolan, R. J. (2010). Effects of loss aversion on post-decision wagering: implications for measures of awareness. *Consciousness and Cognition*, *19*(1), 352–363.

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.

Fleming, S. M., Huijgen, J., & Dolan, R. J. (2012). Prefrontal contributions to metacognition in perceptual decision making. *Journal of Neuroscience*, *32*(18), 6117–6125.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*.

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science*, *329*(5998), 1541–1543.

Friston, K. J., & Frith, C. D. (2015). A duet for one. *Consciousness and Cognition*, *36*, 390–405.

Friston, K. J., & Frith, C. D. (2015). Active inference, communication and hermeneutics. *Cortex*, *68*(12), 129–143.

Frith, C. D. (2012). The role of metacognition in human social interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1599), 2213–2223.

Frith, C. D., & Wentzer, T. (2013). Neural hermeneutics. In *Encyclopedia of Philosophy and the Social Sciences*. Thousand Oaks, CA: SAGE Publications.

Fründ, I., Wichmann, F. A., & Macke, J. H. (2014). Quantifying the effect of intertrial dependence on perceptual decisions. *Journal of Vision*, *14*(7), 1–16.

Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C. D., & Tylen, K. (2012). Coming to terms: quantifying the benefits of linguistic coordination. *Psychological Science*, *23*(8), 931–939.

Gachter, S., Herrmann, B., & Thoni, C. (2010). Culture and cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1553), 2651–2661.

Galton, F. (1907a). One vote, One value. *Nature*, *75*(1948), 414–414.

Galton, F. (1907b). Vox populi. *Nature*, *75*(1949), 450–451.

Galvin, S. J., Podd, J. V, Drga, V., & Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: discrimination between correct and incorrect decisions. *Psychonomic Bulletin & Review*, *10*(4), 843–876.

Gigerenzer, G., & Brighton, H. (2009). Homo Heuristicus: why biased minds make better inferences. *Topics in Cognitive Science*, *1*(1), 107–143.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: a Brunswikian theory of confidence. *Psychological Review*, *98*(4), 506–528.

Giles, J. (2011). Social science lines up its biggest challenges. *Nature*, *470*(7332), 18–19.

Gilovich, T., Savitsky, K., & Medvec, V. H. (1998). The illusion of transparency: biased assessments of others' ability to read one's emotional states. *Journal of Personality and Social Psychology*, *75*(2), 332–346.

Green, D. M. (1964). Consistency of auditory detection judgments. *Psychological Review*,

71(5), 392–407.

Green, D. M., & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. New York, NY: Wiley.

Harvey, N. (1997). Confidence in judgment. *Trends in Cognitive Sciences*, *1*(2), 78–82.

Harvey, N., & Fischer, I. (1997). Taking advice: accepting help, improving judgment, and sharing responsibility. *Organizational Behavior and Human Decision Processes*, *70*(2), 117–133.

Hecht, S., Shlaer, S., & Pirenne, M. H. (1942). Energy, quanta, and vision. *The Journal of General Physiology*, *25*(6), 819–840.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010a). Most people are not WEIRD. *Nature*, *466*(7302), 29–29.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010b). The weirdest people in the world? *The Behavioral and Brain Sciences*, *33*(2-3), 61–83; discussion 83–135.

Henry, R. A., Strickland, O. J., Yorges, S. L., & Ladd, D. (1996). Helping groups determine their most accurate member: the role of outcome feedback. *Journal of Applied Social Psychology*, *26*(13), 1153–1170.

Herrmann, B., Thoni, C., & Gachter, S. (2008). Antisocial punishment across societies. *Science*, *319*(5868), 1362–1367.

Hertwig, R., & Erev, I. (2009). The description-experience gap in risky choice. *Trends in Cognitive Sciences*, *13*(12), 517–523.

Heyes, C., & Frith, C. D. (2014). The cultural evolution of mind reading. *Science*, *344*(6190), 1243091.

Hill, G. W. (1982). Group versus individual performance: are N + 1 heads better than one? *Psychological Bulletin*, *91*(3), 517–539.

Hoorens, V. (1993). Self-enhancement and superiority biases in social comparison. *European Review of Social Psychology*, *4*(1), 113–139.

Huq, S. F., Garety, P. A., & Hemsley, D. R. (1988). Probabilistic judgements in deluded and non-deluded subjects. *The Quarterly Journal of Experimental Psychology Section A*, *40*(4), 801–812.

Janis, I. (1982). *Groupthink*. Boston, MA: Houghton Mifflin.

Jazayeri, M., & Movshon, J. A. (2006). Optimal representation of sensory information by neural populations. *Nature Neuroscience*, *9*(5), 690–696.

Johnson, D. (2004). *Overconfidence and War: The Havoc and Glory of Positive Illusions*. Cambridge, MA: Harvard University Press.

Kahneman, D. (1973). *Attention and Effort*. Englewood Cliffs, NJ: Prentice Hall.

Kepecs, A., & Mainen, Z. F. (2012). A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1322–1337.

Kepecs, A., Uchida, N., Zariwala, H. a, & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature*, *455*(7210), 227–231.

Keysar, B., Hayakawa, S. L., & An, S. G. (2012). The foreign-language effect: thinking in a

foreign tongue reduces decision biases. *Psychological Science*, *23*(6), 661–668.

Kiani, R., Corthell, L., & Shadlen, M. N. (2014). Choice certainty is informed by both evidence and decision time. *Neuron*, *84*(6), 1329–1342.

Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science*, *324*(5928), 759–764.

Komura, Y., Nikkuni, A., Hirashima, N., Uetake, T., & Miyamoto, A. (2013). Responses of pulvinar neurons reflect a subject's confidence in visual categorization. *Nature Neuroscience*, *16*(6), 749–55.

Kording, K. (2007). Decision theory: what "should" the nervous system do? *Science*, *318*(5850), 606–610.

Koriat, A. (2012). When are two heads better than one and why? *Science*, *336*(6079), 360–362.

Kornbrot, D. E. (2006). Signal detection theory, the approach of choice: model-based and distribution-free measures and evaluation. *Perception & Psychophysics*, *68*(3), 393–414.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121–1134.

Lak, A., Costa, G. M., Romberg, E., Koulakov, A. A., Mainen, Z. F., & Kepecs, A. (2014). Orbitofrontal cortex is required for optimal waiting based on decision confidence. *Neuron*, *84*(1), 190–201.

Laughlin, P. R., & Ellis, A. L. (1986). Demonstrability and social combination processes on mathematical intellective tasks. *Journal of Experimental Social Psychology*, *22*(3), 177–189.

Lavie, N. (2005). Distracted and confused? Selective attention under load. *Trends in Cognitive Sciences*, *9*(2), 75–82.

Lewis, D. (1969). *Convention. A Philosophical Study*. Cambridge, MA: Harvard University Press.

Liberman, A., Fischer, J., & Whitney, D. (2014). Serial dependence in the perception of faces. *Current Biology*, *24*(21), 1–6.

Liberman, V., & Tversky, A. (1993). On the evaluation of probability judgments: calibration, resolution, and monotonicity. *Psychological Bulletin*, *114*(1), 162–173.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, *20*(2), 159–183.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: the state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgement under Uncertainty: Heuristics and Biases*. Cambridge, UK: Cambridge University Press.

Long, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: SAGE Publications.

Lorch, R. F., & Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(1), 149–157.

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*(11), 1432–1438.

Ma, W. J., & Jazayeri, M. (2014). Neural coding of uncertainty and probability. *Annual Review*

of Neuroscience, *37*(1), 205–220.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection Theory: A User's Guide*. New York, NY: Erlbaum.

Mahmoodi, A., Bang, D., Ahmadabadi, M. N., & Bahrami, B. (2013). Learning to make collective decisions: the impact of confidence escalation. *PLoS ONE*, *8*(12), e81195.

Majid, A., Bowerman, M., Kita, S., Haun, D. B. M., & Levinson, S. C. (2004). Can language restructure cognition? The case for space. *Trends in Cognitive Sciences*, *8*(3), 108–114.

Maloney, L. T. (2002). Statistical decision theory and biological vision. In D. Heyer & R. Mausfeld (Eds.), *Perception and the Physical World: Psychological and Philosophical Issues in Perception*. New York, NY: Wiley.

Maloney, L. T., & Zhang, H. (2010). Decision-theoretic models of visual perception and action. *Vision Research*, *50*(23), 2362–2374.

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422–430.

Mann, L., Radford, M., Burnett, P., Ford, S., Bond, M., Leung, K., Nakamura, H., Vaughan, G., & Yang, K.-S. (1998). Cross-cultural differences in self-reported decision-making style and confidence. *International Journal of Psychology*, *33*(5), 325–335.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: homophily in social networks. *Annual Review of Sociology*, *27*(1), 415–444.

Metcalfe, J. (2009). Metacognitive judgments and control of study. *Current Directions in Psychological Science*, *18*(3), 159–163.

Minson, J. A., & Mueller, J. S. (2012). The cost of collaboration: why joint decision making exacerbates rejection of outside information. *Psychological Science*, *23*(3), 219–224.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502–517.

Moran, R., Teodorescu, A. R., & Usher, M. (2015). Post choice information integration as a causal determinant of confidence: novel data and a computational account. *Cognitive Psychology*, *78*, 99–147.

Niederle, M., & Vesterlund, L. (2011). Gender and competition. *Annual Review of Economics*, *3*(1), 601–630.

Nitzan, S., & Paroush, J. (1985). *Collective Decision Making: An Economic Outlook*. Cambridge, UK: Cambridge University Press.

Overgaard, M., & Sandberg, K. (2012). Kinds of access: different methods for report reveal different kinds of metacognitive access. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1287–1296.

Packer, D. J. (2009). Avoiding groupthink. *Psychological Science*, *20*(5), 546–548.

Pashler, H. (1994). Dual-task interference in simple tasks: data and theory. *Psychological Bulletin*, *116*(2), 220–244.

Patel, D., Fleming, S. M., & Kilner, J. M. (2012). Inferring subjective states through the observation of actions. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1748), 4853–4860.

Payzan-LeNestour, E., & Bossaerts, P. (2011). Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. *PLoS Computational Biology*, *7*(1), e1001048.

Penrod, S., & Cutler, B. (1995). Witness confidence and witness accuracy: assessing their forensic relation. *Psychology, Public Policy, and Law*, *1*(4), 817–845.

Perry-Coste, F. H. (1907). The ballot-box. *Nature*, *75*(1952), 509–509.

Persaud, N., McLeod, P., & Cowey, A. (2007). Post-decision wagering objectively measures awareness. *Nature Neuroscience*, *10*(2), 257–261.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *The Behavioral and Brain Sciences*, *27*(2), 169–190; discussion 190–226.

Pleskac, T. J., & Busemeyer, J. R. (2011). Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. *Psychological Review*, *118*(1), 56.

Prelec, D. (2004). A Bayesian truth serum for subjective data. *Science*, *306*(5695), 462–466.

Proust, M. (2006). *Remembrance of Things Past*. London, UK: Wordsworth.

Raafat, R. M., Chater, N., & Frith, C. D. (2009). Herding in humans. *Trends in Cognitive Sciences*, *13*(10), 420–428.

Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological Review*, *116*(1), 59–83.

Ratcliff, R., & Starns, J. J. (2013). Modeling confidence judgments, response times, and multiple choices in decision making: Recognition memory and motion discrimination. *Psychological Review*, *120*(3), 697–719.

Reed, N., McLeod, P., & Dienes, Z. (2010). Implicit knowledge and motor skill: what people who know how to catch don't know. *Consciousness and Cognition*, *19*(1), 63–76.

Richardson, P. J., & Boyd, R. (2005). *Not by Genes Alone: How Culture Transformed Human Evolution*. Chicago, IL: University of Chicago Press.

Roepstorff, A., Niewöhner, J., & Beck, S. (2010). Enculturing brains through patterned practices. *Neural Networks*, *23*(8-9), 1051–1059.

Schelling, T. C. (1980). *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Schüür, F., Tam, B. P., & Maloney, L. T. (2013). Learning patterns in noise: environmental statistics explain the sequential effect. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Selen, L. P. J., Beek, P. J., & van Dieën, J. H. (2006). Impedance is modulated to meet accuracy demands during goal-directed arm movements. *Experimental Brain Research*, *172*(1), 129–138.

Senders, V. L., & Sowards, A. (1953). Further analysis of response sequences in the setting of a psychophysical experiment. *The American Journal of Psychology*, *66*(2), 215–228.

Shadish, W. R., & Haddock, K. C. (1994). Combining estimates of effect size. In H. Cooper & L. V. Hedges (Eds.), *The Handbook of Research Synthesis*. New York, NY: Russell Sage Foundation.

Shakespeare, W. (1993). *As You Like It*. London, UK: Wordsworth Editions.

Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., & Frith, C. D. (2014). Supra-personal cognitive control and metacognition. *Trends in Cognitive Sciences*, *18*(4), 186–193.

Sherif, M. (1935). A study of some social factors in perception. *Archives of Psychology*, *187*, 5–61.

Shields, W. E., Smith, J. D., & Washburn, D. a. (1997). Uncertain responses by humans and rhesus monkeys (Macaca mulatta) in a psychophysical same-different task. *Journal of Experimental Psychology. General*, *126*(2), 147–164.

Silver, N. (2013). *The Signal and the Noise: The Art and Science of Prediction*. London, UK: Penguin.

Soll, J. B. (1996). Determinants of overconfidence and miscalibration: the roles of random error and ecological structure. *Organizational Behavior and Human Decision Processes*, *65*(2), 117–137.

Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: how (and how well) people use others' opinions. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *35*(3), 780–805.

Song, C., Kanai, R., Fleming, S. M., Weil, R. S., Schwarzkopf, D. S., & Rees, G. (2011). Relating inter-individual differences in metacognitive performance on different perceptual tasks. *Consciousness and Cognition*, *20*(4), 1787–92.

Stasser, G., & Titus, W. (1985). Pooling of unshared information in group decision making: biased information sampling during discussion. *Journal of Personality and Social Psychology*, *48*(6), 1467–1478.

Stasser, G., & Titus, W. (2003). Hidden profiles: a brief history. *Psychological Inquiry*, *14*(3), 304–313.

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *NeuroImage*, *46*(4), 1004–1017.

Sundali, J., & Croson, R. (2006). Biases in casino betting : the hot hand and the gambler's fallacy. *Judgment and Decision Making*, *1*(1), 1–12.

Surowiecki, J. (2004). *The Wisdom of Crowds*. New York, NY: Anchor Books.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Sweller, J. (1988). Cognitive load during problem solving: effects on learning. *Cognitive Science*, *12*(2), 257–285.

Tenney, E. R., MacCoun, R. J., Spellman, B. A., & Hastie, R. (2007). Calibration trumps confidence as a basis for witness credibility: Research report. *Psychological Science*, *18*(1), 46–50.

Thomas, J. P., & McFadyen, R. G. (1995). The confidence heuristic: a game-theoretic analysis. *Journal of Economic Psychology*, *16*(1), 97–113.

Toelch, U., & Dolan, R. J. (2015). Informational and normative influences in conformity from a neurocomputational perspective. *Trends in Cognitive Sciences*, *19*(10), 579–89.

Treisman, M., & Faulkner, A. (1984). The setting and maintenance of criteria representing levels of confidence. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(1), 119–139.

Treisman, M., & Williams, T. C. (1984). A theory of criterion setting with an application to sequential dependencies. *Psychological Review*, *91*(1), 68–111.

Trommershäuser, J., Landy, M. S., & Maloney, L. T. (2006). Humans rapidly estimate expected gain in movement planning. *Psychological Science*, *17*(11), 981–988.

Trommershäuser, J., Maloney, L. T., & Landy, M. S. (2003). Statistical decision theory and the selection of rapid, goal-directed movements. *Journal of the Optical Society of America A*, *20*(7), 1419.

Trommershäuser, J., Maloney, L. T., & Landy, M. S. (2008). Decision making, movement planning and statistical decision theory. *Trends in Cognitive Sciences*, *12*(8), 291–297.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323.

Verplanck, W. S., Collier, G. H., & Cotton, J. W. (1952). Nonindependence of successive responses in measurements of the visual threshold. *Journal of Experimental Psychology*, *44*(4), 273–282.

Vickers, D. (1979). *Decision Processes in Visual Perception*. London, UK: Academic Press.

Vickers, D., & Packer, J. (1982). Effects of alternating set for speed or accuracy on response time, accuracy and confidence in a unidimensional discrimination task. *Acta Psychologica*, *50*(2), 179–197.

Wang, X.-J. (2008). Decision making in recurrent neuronal circuits. *Neuron*, *60*(2), 215–234.

Williams, K. D. (2007). Ostracism. *Annual Review of Psychology*, *58*(1), 425–452.

Wittgenstein, L. (1958). *Philosophical Investigations*. Oxford, UK: Blackwell.

Wolpert, D. M., & Flanagan, J. R. (2010). Motor learning. *Current Biology*, *20*(11), 467–472.

Wright, G., & Ayton, P. (1987). *Judgemental Forecasting*. Chichester, UK: John Wiley & Sons, Ltd.

Yaniv, I. (2004). Receiving other people's advice: influence and benefit. *Organizational Behavior and Human Decision Processes*, *93*(1), 1–13.

Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes*, *83*(2), 260–281.

Yates, J. F. (1990). *Judgement and Decision Making*. Englewood Cliffs, NJ: Prentice Hall.

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.

Yu, A. J., & Cohen, J. D. (2009). Sequential effects: superstition or rational behavior? In D. Koller, D. Schuurmans, & Y. Bengio (Eds.), *Advances in Neural Information Processing Systems 21 (NIPS 2008)* (Vol. 24, pp. 1–8). Cambridge, MA: MIT Press.

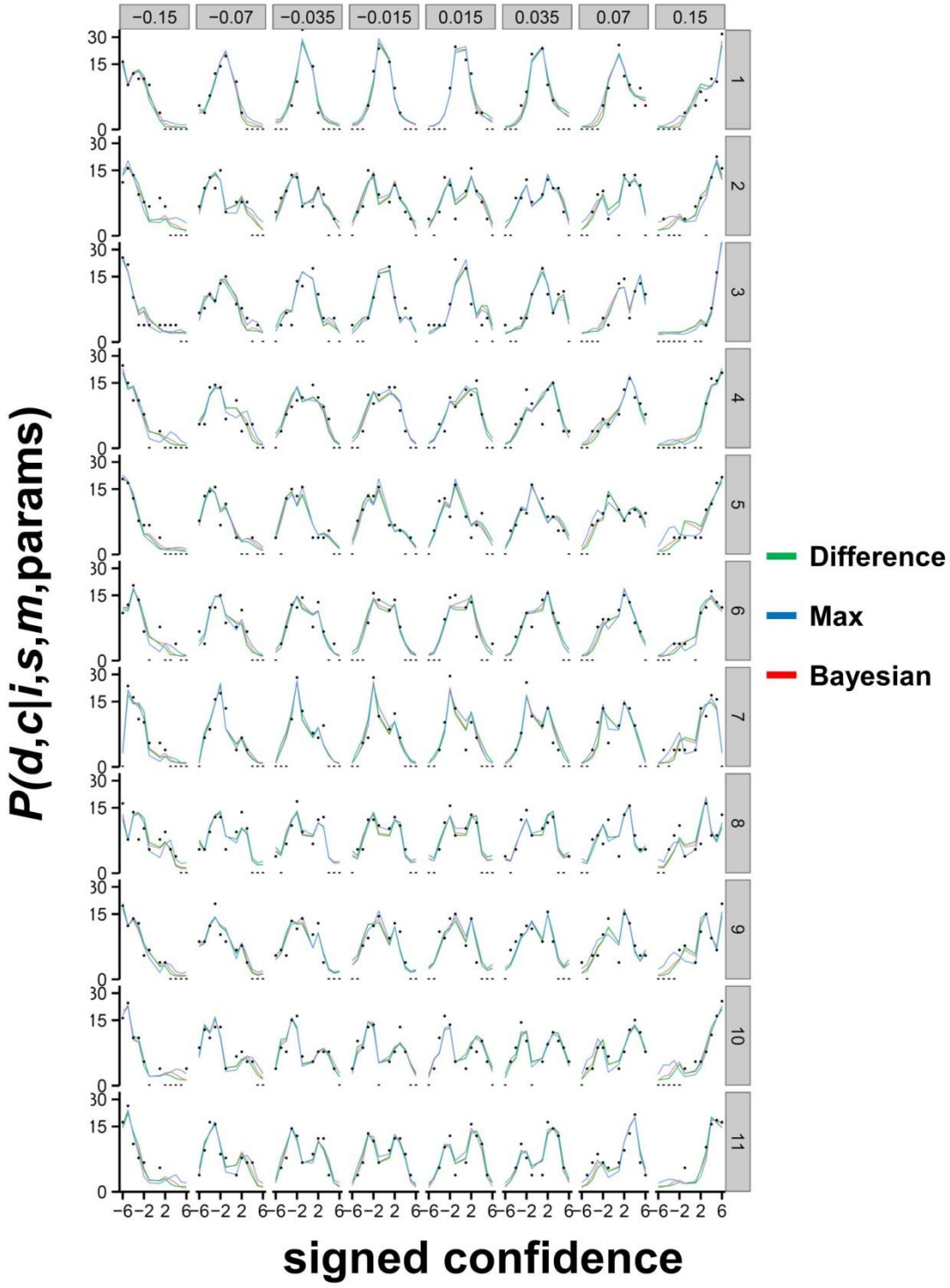
Zarnoth, P., & Sniezek, J. A. (1997). The social influence of confidence in group decision making. *Journal of Experimental Social Psychology*, *33*(4), 345–366.

Zylberberg, A., Roelfsema, P. R., & Sigman, M. (2014). Variance misperception explains

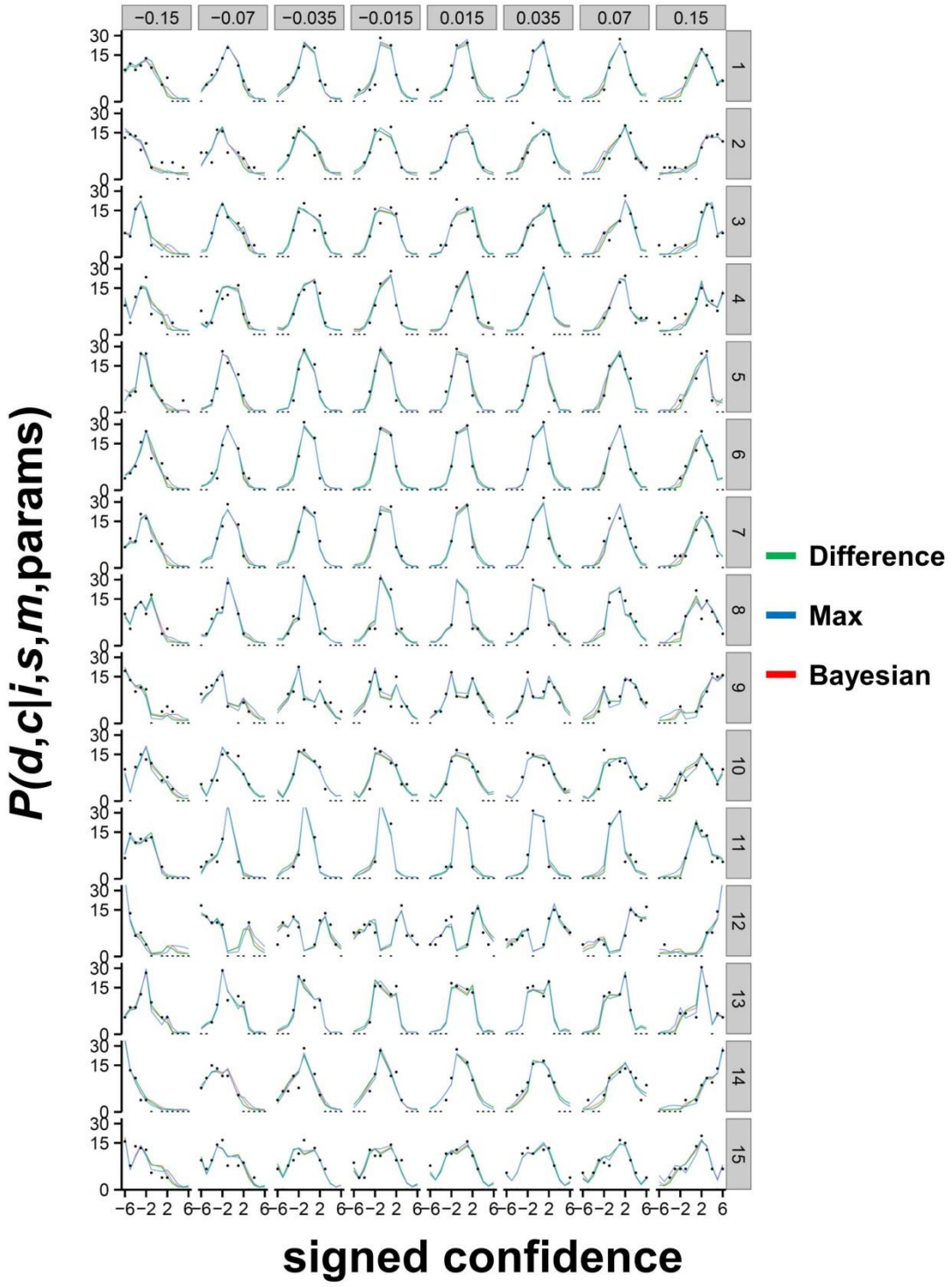
illusions of confidence in simple perceptual decisions. *Consciousness and Cognition*, *27*(1), 246–253.

Appendix: Supplementary figures

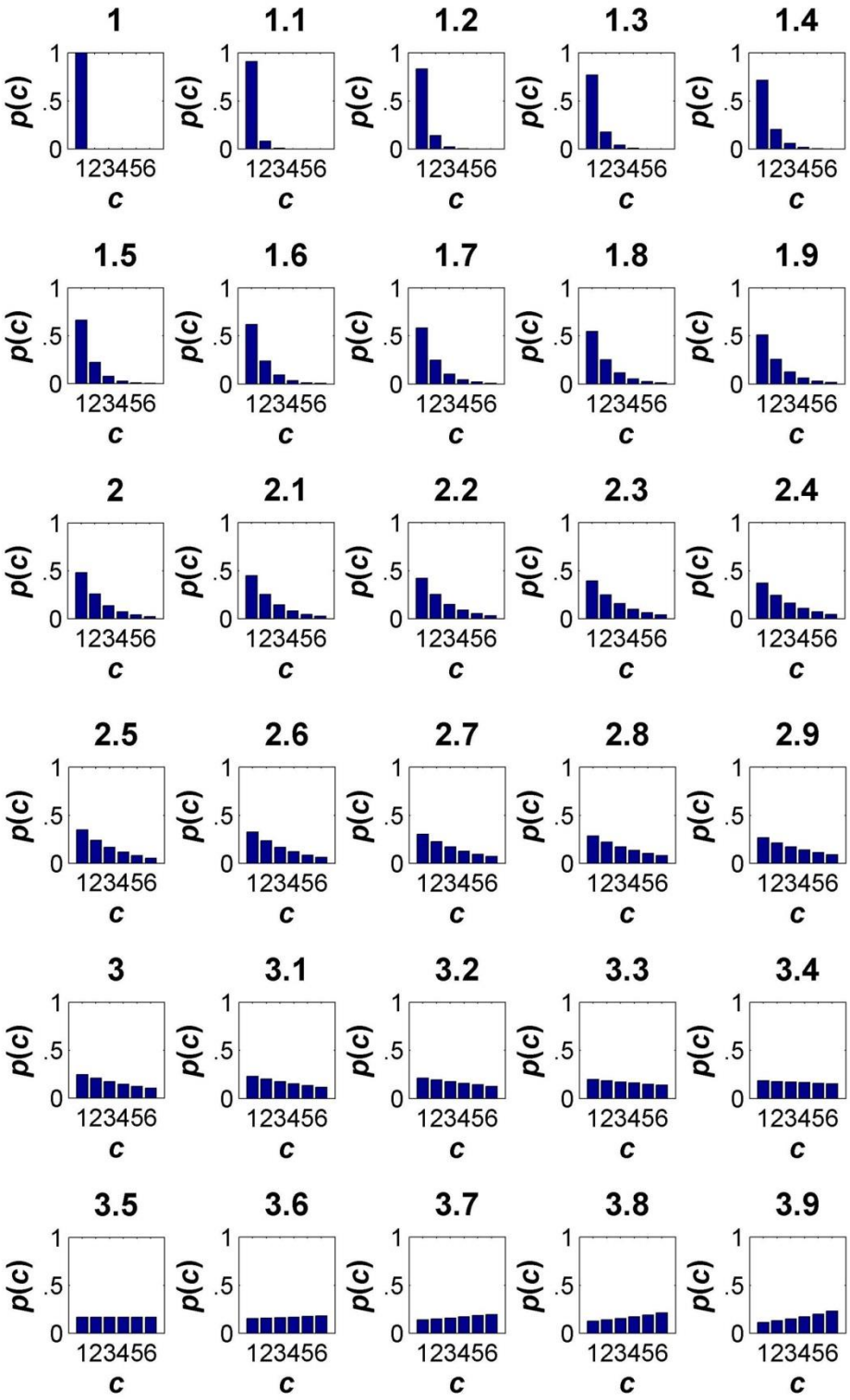
See the following pages.

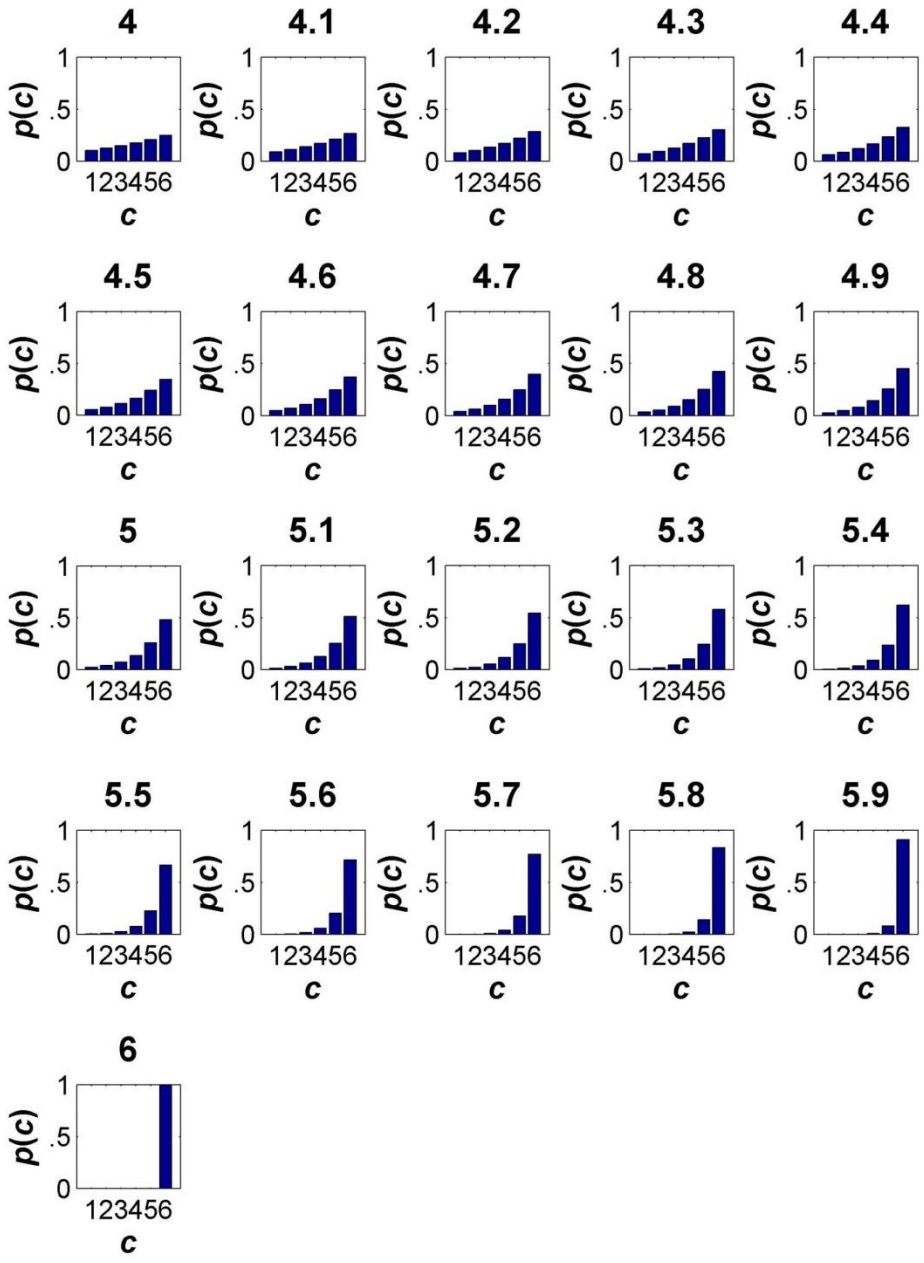


Supplementary Figure 1. Chapter 2: Empirical and fitted distributions over signed confidence given signed contrast for each participant in Experiment 1. The lines show the fitted models, and the points show the data. Each row gives the complete responses for one participant. Each column gives the responses for a given signed contrast. The axis has been square-root transformed, in order to emphasize differences in low probabilities.

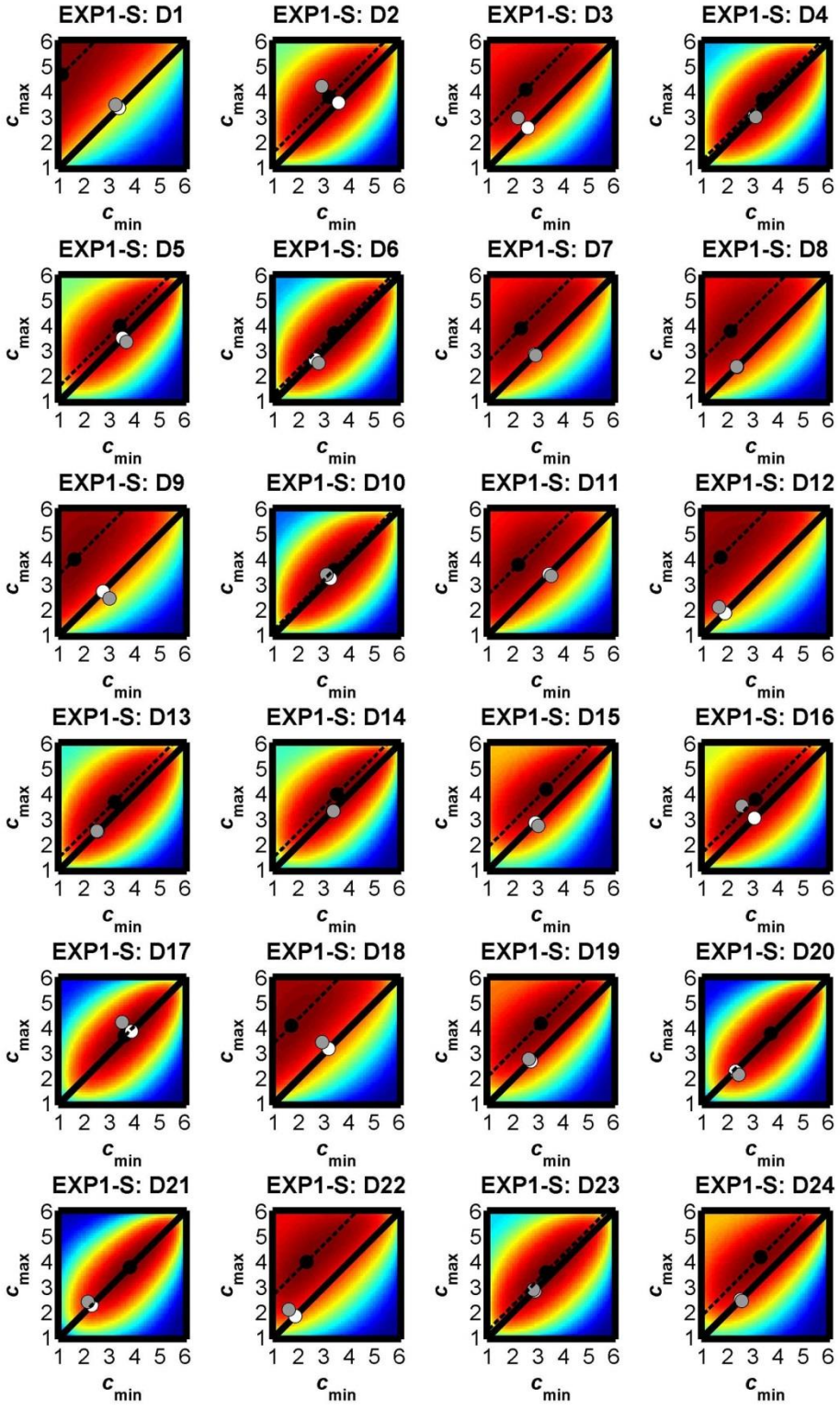


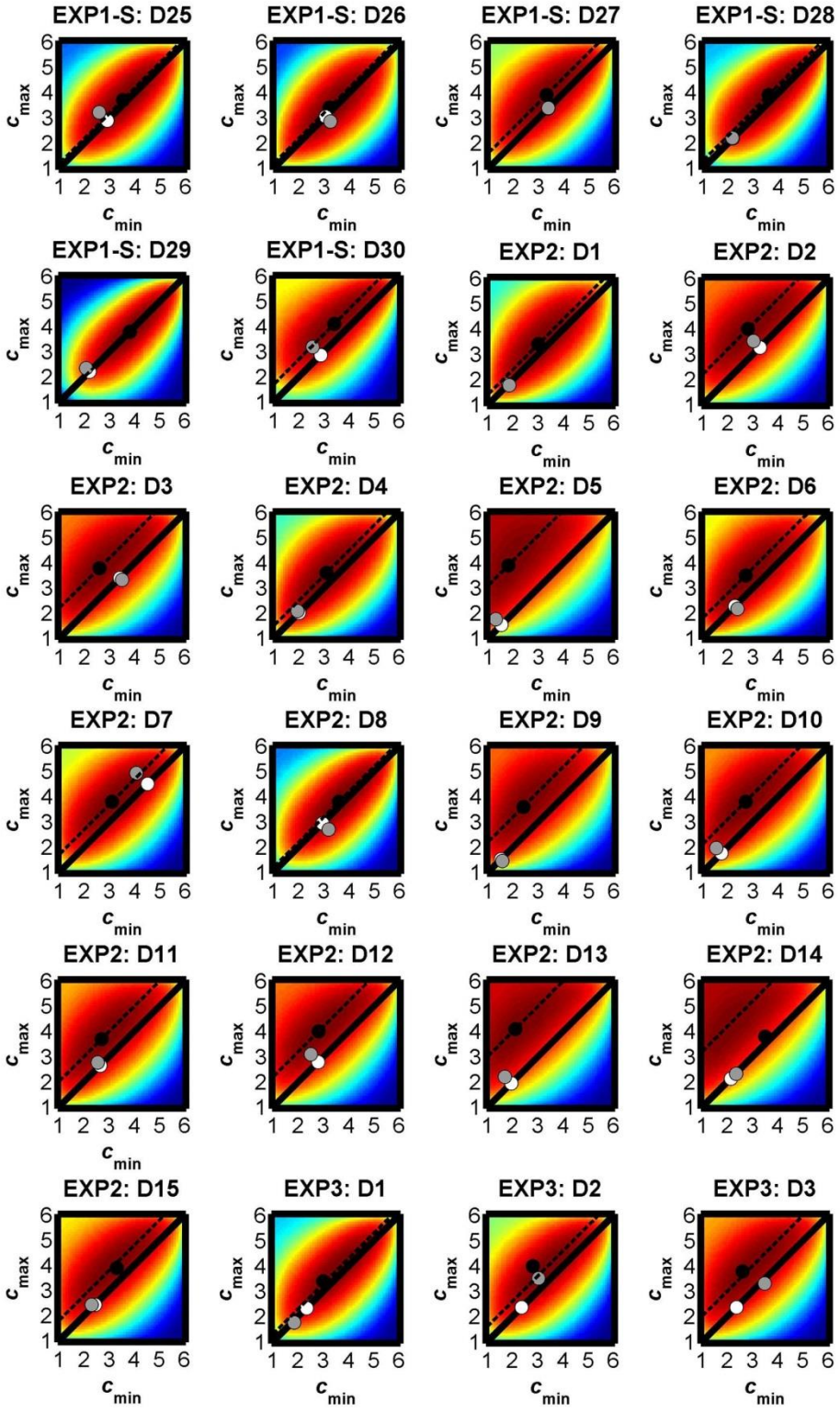
Supplementary Figure 2. Chapter 2: Empirical and fitted distributions over signed confidence given signed contrast for each participant in Experiment 2. The lines show the fitted models, and the points show the data. Each row gives the complete responses for one participant. Each column gives the responses for a particular signed contrast value. The axis has been square-root transformed, in order to emphasize differences in low probabilities.

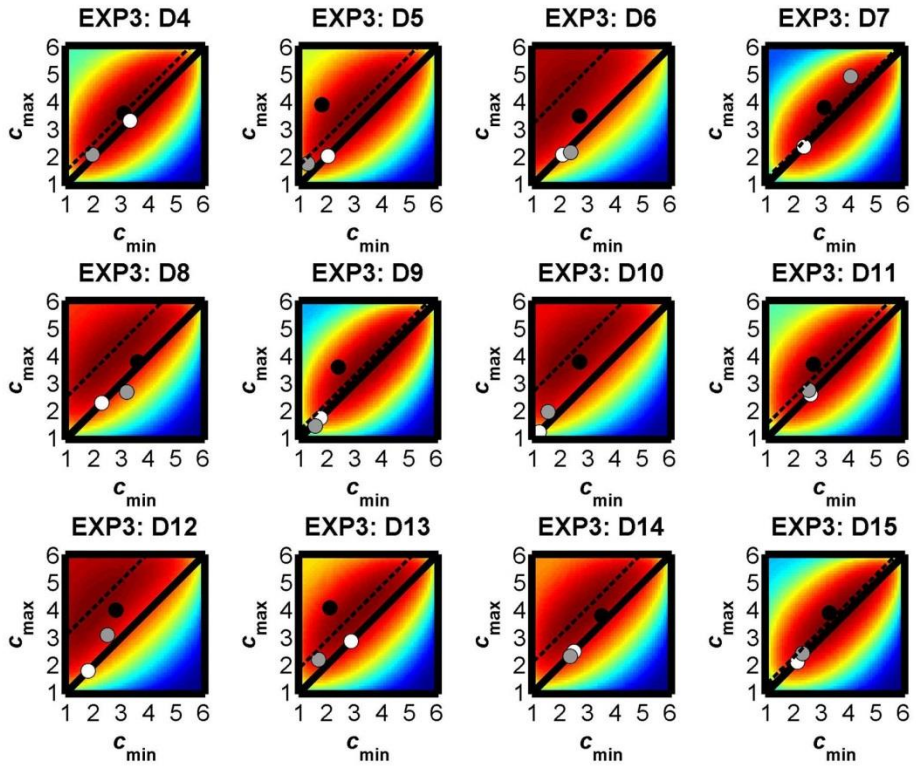

1
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
2
2.1
2.2
2.3
2.4
2.5
2.6
2.7
2.8
2.9
3
3.1
3.2
3.3
3.4
3.5
3.6
3.7
3.8
3.9
p(c)
1
.5
0
123456
c



Supplementary Figure 3. Chapter 4: Set of maximum entropy distributions. The maximum entropy distributions are shown as confidence distributions; transforming these into response distributions for our analysis is straightforward. The x-axis shows the confidence levels, c. The y-axis shows the proportion of times that each confidence level was selected, $p(c)$.


EXP1-S: D1
EXP1-S: D2
EXP1-S: D3
EXP1-S: D4
EXP1-S: D5
EXP1-S: D6
EXP1-S: D7
EXP1-S: D8
EXP1-S: D9
EXP1-S: D10
EXP1-S: D11
EXP1-S: D12
EXP1-S: D13
EXP1-S: D14
EXP1-S: D15
EXP1-S: D16
EXP1-S: D17
EXP1-S: D18
EXP1-S: D19
EXP1-S: D20
EXP1-S: D21
EXP1-S: D22
EXP1-S: D23
EXP1-S: D24
c_max
c_min
1 2 3 4 5 6


EXP1-S: D25
EXP1-S: D26
EXP1-S: D27
EXP1-S: D28
EXP1-S: D29
EXP1-S: D30
EXP2: D1
EXP2: D2
EXP2: D3
EXP2: D4
EXP2: D5
EXP2: D6
EXP2: D7
EXP2: D8
EXP2: D9
EXP2: D10
EXP2: D11
EXP2: D12
EXP2: D13
EXP2: D14
EXP2: D15
EXP3: D1
EXP3: D2
EXP3: D3
c_{max}
c_{min}
1 2 3 4 5 6



Supplementary Figure 4. Chapter 4: Confidence landscapes for Experiments 1 to 3. Each subplot corresponds to a confidence landscape for one dyad. The axes show the mean confidence of the worse, $c_{\min}$ (x-axis), and the better dyad member, $c_{\max}$ (y-axis) – with this division determined by comparing their fitted levels of sensory noise. Each mean indexes a response distribution constrained to have maximum entropy but associated with a specific mean confidence (see Supplementary Figure 3). The colour of each coordinate indicates the joint accuracy expected under a particular pair of response distributions, with expected joint accuracy normalised to the range 0-1 (blue-red) separately for each dyad. The grey dot indexes the joint accuracy expected under maximum entropy distributions associated with the same mean confidence as dyad members' observed response distributions, a_{maxent}. The white dot indexes the joint accuracy expected under confidence matching, a_{match} (a pair of maximum entropy distributions associated with the mean of the dyad members' mean confidence, $\text{mean}(c_1, c_2)$). The black dot indexes the joint accuracy expected under the optimal strategy, a_{opt} (the pair of maximum entropy distributions that maximises the expected joint accuracy).

Appendix: Supplementary tables

See the following pages.

data	mean beta weights (*p*-value)						
	intercept	k_t	k_{t-1}	c_self_{t-1}	$c\_partner_{t-1}$	a_self_{t-1}	$a\_partner_{t-1}$
EXP1-I	.24 (<.001)	.47 (<.001)	–.01 (.149)	.10 (<.001)	.00 (.848)	.01 (.005)	–.00 (.419)
EXP1-S	.21 (<.001)	.49 (<.001)	–.02 (.118)	.07 (<.001)	.06 (<.001)	.01 (<.001)	–.01 (.002)
EXP2	.20 (<.001)	.38 (<.001)	–.02 (.031)	.09 (<.001)	.06 (<.001)	.01 (<.001)	–.00 (.324)
EXP3	.16 (<.001)	.44 (<.001)	–.04 (<.001)	.14 (<.001)	.05 (<.001)	.01 (.039)	–.01 (.113)
EXP4-S	.19 (<.001)	.47 (<.001)	–.06 (<.001)	.18 (<.001)	.11 (<.001)	.01 (.014)	–.01 (<.001)
ALCL	.24 (<.001)	.47 (<.001)	–.01 (.397)	.12 (<.001)	.01 (.492)	.01 (.093)	–.00 (.593)
ALCH	.30 (<.001)	.45 (<.001)	–.03 (.007)	.15 (<.001)	.03 (.003)	.01 (.008)	–.01 (.018)
AHCL	.18 (<.001)	.47 (<.001)	–.02 (.176)	.13 (<.001)	.04 (.011)	.01 (.042)	–.01 (.149)
AHCH	.25 (<.001)	.48 (<.001)	–.02 (.197)	.15 (<.001)	.03 (.052)	.01 (.009)	–.00 (.234)

Supplementary Table 1. Chapter 4: Coefficients and *p*-values from linear trial-by-trial regression model of confidence reports in Experiments 1 to 4. The coefficients encode how participants' confidence report on trial t was influenced by the level of contrast added to the visual target (k) on trial t and $t-1$, the participant's confidence report (c_self) on trial $t-1$, the partner's confidence report (*c_partner*) on trial $t-1$, the accuracy of the participant's decision (a_self) on trial $t-1$, and the accuracy of the partner's decision ($a\_partner$) on trial $t-1$. We normalised confidence to the range 0-1, normalised the level of contrast added to the visual target to the range 0-1, and coded response accuracy as -1 when wrong and 1 when correct. We calculated significance by testing (one-sample t-test) whether the coefficients pooled across participants were significantly different from 0.

data	mean beta weights (*p*-value)					
	k_t	k_{t-1}	c_self_{t-1}	$c\_partner_{t-1}$	a_self_{t-1}	$a\_partner_{t-1}$
EXP1-I	3.54 (<.001)	–.19 (.019)	.87 (<.001)	–.07 (.285)	.10 (.001)	–.04 (.132)
EXP1-S	3.75 (<.001)	–.15 (.100)	.61 (<.001)	.50 (<.001)	.12 (<.001)	–.09 (.004)
EXP2	3.21 (<.001)	–.22 (.017)	.82 (<.001)	.55 (<.001)	.12 (<.001)	–.03 (.296)
EXP3	3.68 (<.001)	–.40 (<.001)	.25 (<.001)	.10 (<.001)	.07 (.007)	–.06 (.027)
EXP4-S	3.44 (<.001)	–.48 (<.001)	1.51 (<.001)	.85 (<.001)	.06 (.005)	–.10 (<.001)
ALCL	3.71 (<.001)	–.09 (.395)	1.13 (<.001)	.09 (.197)	.06 (.125)	–.02 (.462)
ALCH	3.55 (<.001)	–.27 (.002)	1.28 (<.001)	.23 (.002)	.11 (.002)	–.07 (.023)
AHCL	3.64 (<.001)	–.21 (.103)	1.10 (<.001)	.42 (.007)	.08 (.012)	–.04 (.146)
AHCH	3.63 (<.001)	–.17 (.147)	1.18 (<.001)	.20 (.062)	.09 (.006)	–.04 (.227)

Supplementary Table 2. Chapter 4: Coefficients and *p*-values from ordered probit trial-by-trial regression model of confidence reports in Experiments 1 to 4. The coefficients encode how participants' confidence report on trial t was influenced by the level of contrast added to the visual target (k) on trial t and $t-1$, the participant's confidence report (c_self) on trial $t-1$, the partner's confidence report (c_partner) on trial $t-1$, the accuracy of the participant's decision (a_self) on trial $t-1$, and the accuracy of the partner's decision ($a\_partner$) on trial $t-1$. We normalised the level of contrast added to the visual target to the range 0-1, and coded response accuracy as -1 when wrong and 1 when correct. For Experiment 3, the continuous confidence values were discretised to the range 1 to 6 in steps of 1 by applying a set of linearly spaced thresholds. We calculated significance by testing (one-sample t-test) whether the coefficients pooled across participants were significantly different from 0. For simplicity, we do not show the intercepts (thresholds) associated with each confidence level.

Model	beta weights (*p*-value)			
	intercept	*Δm*	*cal*	*Δm*cal*
full model	.01 (.080)	.01 (.052)	–.00 (.786)	.04 (<.001)
well-calibrated	.01 (.121)	–.01 (.041)	–.00 (.786)	.04 (<.001)
poorly calibrated	.01 (.324)	.03 (<.001)	–.00 (.786)	.04 (<.001)

Supplementary Table 3. Chapter 4: Coefficients and *p*-values from robust linear regression testing for the effect of confidence matching on joint accuracy in Experiment 4. We used robust linear regression to estimate the relative influence of confidence matching (Δm), calibration (cal), and their interaction ($\Delta m * cal$) on a measure of joint accuracy (Δm) that had been normalised using participants' baseline data (Δa_{dyad}) – for details on how the latter quantity was computed see Section 4.4.1.4: Joint accuracy expected prior to interaction. Our measure of confidence matching, $\Delta m = |c_{\mathrm{sbj}}^{\mathrm{isolated}} - c_{\mathrm{agent}}^{\mathrm{social}}| - |c_{\mathrm{sbj}}^{\mathrm{social}} - c_{\mathrm{agent}}^{\mathrm{social}}|$, tells us whether the absolute difference between a participant's mean confidence and that of their respective partner was larger prior to ($\Delta m > 0$: $|c_{\mathrm{sbj}}^{\mathrm{isolated}} - c_{\mathrm{agent}}^{\mathrm{social}}| > |c_{\mathrm{sbj}}^{\mathrm{social}} - c_{\mathrm{agent}}^{\mathrm{social}}|$) or during interaction ($\Delta m < 0$: $|c_{\mathrm{sbj}}^{\mathrm{isolated}} - c_{\mathrm{agent}}^{\mathrm{social}}| < |c_{\mathrm{sbj}}^{\mathrm{social}} - c_{\mathrm{agent}}^{\mathrm{social}}|$). For the full-model, cal was a dummy contrast, with well-calibrated = -.5 and poorly calibrated = .5. For the simple-effect model for well-calibrated dyad members, well-calibrated = 0 and poorly calibrated = 1. For the simple-effect model for poorly calibrated agents, well-calibrated = -1 and poorly calibrated = 0. We found the same results (both direction of results and significance) using a linear mixed-effects model with a random intercept for each participant.