**RESPONSE TO REVIEWS**

***EDITOR***

***Your manuscript has now been evaluated by 2 reviewers (note that a third reviewer agreed to review your manuscript, but has been unable to submit their comments; we will forward these to you independently in the event they are submitted). You will see from the reviewers' comments copied below that while they find your work to be both novel and timely, they (and especially Reviewer 1) have raised quite substantial concerns that must be addressed. In light of these comments, we cannot accept the manuscript for publication, but would be very interested in considering a revised version that demonstrates the robustness of your conclusions with additional data, analyses, and simulations.***

***We hope you will find the reviewers' comments useful as you decide how to proceed. If you wish to submit a substantially revised manuscript, please bear in mind that we will be reluctant to approach the referees again in the absence of major revisions.***

We thank the Editor for the opportunity to revise. We have made several major changes to address the Reviewers’ concerns:

* We have restructured the presentation of the results to avoid some of the potential confusion highlighted by the Reviewers and added additional analyses to address issues raised by the Reviewers.
* We have conducted two additional experiments: (A) Experiment 5 tests whether people engage in confidence matching when it is in their financial interest to deviate from the strategy and (B) Experiment 6 tests whether confidence matching may result from mere exposure to another individual’s confidence in the absence of group decision-making. We find (A) that confidence matching persists in the face of financial incentives and (B) that confidence matching is most pronounced in the context of group decision-making.
* We clarify confusion relating to our modelling (the model-comparison issue raised by Reviewer #1). We establish our main claim – that people match their confidence – in six independent experiments. We believe that the behavioural evidence is so strong that model comparison is not needed to support our main claim. The purpose of our modelling (confidence landscapes) is (A) to intuit how confidence matching affects group performance in different social situations and (B) to establish a benchmark for optimal performance against which the empirical groups could be compared – we perform several control analyses to show that this comparison is warranted.
* We believe that readers would be interested in a process-level description of how confidence matching arises. We have therefore added additional analyses in which we show that the behaviour of a simple reinforcement-learning model – which seeks to minimise the distance between its own mean confidence and its estimate of its partner’s mean confidence – is consistent with the observed trial-by-trial behaviour of participants (the serial-dependence results highlighted by Reviewer #1) and other aspects of the behavioural data.
* We previously analysed multiple summary statistics of group members’ confidence in the main text (e.g., mean confidence and confidence distributions). We now focus on convergence in mean confidence; and give results for convergence in the variance of confidence and in confidence distributions in **Supplementary Figure 1** – all of these summary statistics show the same pattern and support our hypothesis. We believe that this simplification has enhanced the readability of the main text as the Reader does not have to keep multiple summary statistics in mind.
* We now complement our use of standard parametric tests (e.g., paired *t*-tests to compare convergence in mean confidence between different conditions) with a non-parametric approach. The added value of this approach is that it allows us to rule out that group members ‘coincidentally’ converged onto the same mean confidence. In Experiment 1, for example, one concern was whether we observed high convergence in the social task compared to the isolated task because group members increased their confidence to dominate the joint decision. We believe that the Reader will find the approach insightful.

Because of these major changes (i.e., additional data, analyses and simulations as requested by the Editor), the current manuscript is longer than the original one (current manuscript: main text is about 3000 words; there are 5 figures; figure captions are about 1100 words; methods text is about 2800 words). In the event that the Editor and the Reviewers are satisfied with our changes, we are happy to discuss with the Editor how to best shorten the manuscript to conform to NHB guidelines, or possibly, if the Editor prefers this option, to change the manuscript from a Letter to a Research Article.

We address the Reviewers’ concerns below.

***Reviewer #1***

***The manuscript is the report of four experiments on group decision making carried out in Iran and in the UK. The issue at hand is how two observers engaged in the same task can increase their overall performance if they are both able to communicate their percept and some confidence rating of their percept. The authors argue that in this scenario, observers adopt a strategy, called "confidence matching", that consists in matching both the mean and the variance of their confidence judgments. The study is timely, following the fascinating work by a subgroup of the authors (Bahrami et al., 2010, Science). The manuscript itself is well-written. However, I am not convinced by the authors' interpretation that their participants adopt a confidence matching strategy, and therefore I am doubtful about the exact value of this report. My doubts have three major aspects.***

We thank the Reviewer for positive assessment of our work. We have conducted additional analyses to address the Reviewer’s concerns – all were consistent with our hypothesis.

***First, if the participants match their confidence distributions, we should see some gradual confidence convergence over time. The serial dependence analysis (line 86 and Figure 2B) goes in that direction but only provides a very local analysis (time lag of 1 trial). Actually, the outcome of this 1-trial back analysis goes against the authors' hypothesis. This is because if indeed "participants were more likely to report high confidence after their partner had reported high confidence" (line 89), then the two participants will preserve their difference in confidence space and never converge to a common mean value. So a local time serial dependence analysis goes against confidence matching.***

As the Reviewer says, serial dependence in confidence (what we now call ‘interdependence in confidence’) does not necessarily imply confidence matching: participants could increase and decrease their trial-by-trial confidence in synchrony while keeping a fixed distance in their mean confidence. As such, at first sight, this result shows that confidence is subject to social influences but it does not provide direct evidence for confidence matching.

To address the Reviewer’s concern, we simulated the behaviour of a simple reinforcement-learning model which seeks to minimise the distance between its own mean confidence and its estimate of its partner’s mean confidence (*L232-L249*). The model makes two testable predictions (**Figure 5**). First, a participant’s confidence on a given trial should be influenced by the partner’s confidence on recent trials. Second, while the magnitudes of their of mean confidence may vary (e.g., around 2 at the start and around 5 at the end), the difference in their mean confidence should be more stable. In this way, the dynamics of confidence matching may be thought of as a dance: the group members occupy different positions in confidence space but stay close to each other. Our empirical trial-by-trial data are in line with these predictions (**Figure 5**). While it is true that interdependence in confidence does not imply confidence matching, it seems that the reverse is not true: a simple trial-by-trial model of confidence matching implies interdependence in confidence.

We have re-organised the presentation of our results so that we first focus on results which are more obviously related to our hypothesis, such as the convergence in mean confidence. We present the trial-by-trial results towards the end of manuscript where – after having established the general behavioural effect in six independent experiments – it seems most appropriate to consider its underlying mechanisms.

***Second, I find the design of Experiment 4 very interesting. But this experiment offers some data that seem to go against the authors' hypothesis. Given the definition of measure delta\_m (line 188), we should expect this measure to be positive if participants are engaged in confidence matching. Even though this analysis is not explicitly presented (why?), one can estimate the distribution of delta\_m by marginalizing confidence out of Figure 4B, and this distribution clearly straddles the origin. So an experimental design where accuracy and confidence was adjusted for each participant does not provide strong evidence for confidence matching.***

We thank the reviewer for drawing our attention to this analysis. We did report this analysis in the previous version, and found that the $Δm$ values were significantly positive (*L199-L202* in original version), but it was ‘buried’ in the reporting of the regression results relating confidence matching ($Δm$) to changes in joint accuracy ($Δa\_{joint}$). We now say in a separate sentence after introducing $Δm$: “*We note that the* $Δm$ *values pooled across participants and conditions were positive (t(151) = 5.066, p < .001, one-sample, null: 0), providing further evidence for our hypothesis*.” (*L194-L195*)

***Third, it seems that there are several models that could account for the sub-optimal performance of the observers and one cannot help asking why do the authors emphasize confidence matching. The modelling effort is interesting, but the authors face the curse of overfitting their data with a model that has too many degrees of freedom. This is the reason they propose to adopt maximum entropy distributions to avoid having to estimate all the confidence thresholds. But then again, how can we assess the quality of the fit of this model, given that no other model is being presented?***

With reference to the Reviewer’s concern about overfitting, we would like to clarify that we only fitted *one* parameter to our data: a participant’s sensory noise. The subjective mapping was *determined* by a participant’s response distribution observed across stimuli – a method used effectively in other studies (Aitchison et al., *PLoS Computational Biology*, 2015; Sanders et al., *Neuron*, 2016). The reason why we can claim to have obtained good fits is that the model captures empirical responses (decisions and confidence) when analysed *separately* for each stimulus (see **Supplementary Figures 5-6**) even though the subject mapping was determined using the response distribution observed *across* stimuli. We now explicitly make this point in the **Methods** (*L430-433*).

The main purpose of our modelling (confidence landscapes) was not to provide evidence for confidence matching but (A) to quantify how group performance varies with differences in mean confidence and expertise (sensory noise) and (B) to establish an optimal benchmark against which we could compare the empirical group performance. Any confidence model would arrive at the same conclusion: the optimal relationship between group members’ mean confidence depends on their relative expertise. To avoid confusion about the purpose of our modelling, we clarify in the main text that the confidence landscapes shown in **Figure 3** are based on *simulated* data (*L148*) and explicitly state the aforementioned purpose of our modelling in the **Methods** (*L383-386*).

We used maximum entropy distributions to create the confidence landscapes but *not* for fitting participant’ sensory noise. There are many distributions that can give rise to the same mean; it therefore seemed sensible to limit the simulations to one family of distributions. We chose maximum entropy distributions because they carry maximum information about the internal estimate of the evidence strength for a given mean confidence. We note that we, when comparing empirical group performance with that expected under the optimal solution (maximum in a confidence landscape), take into account the fact that the optimal expectation was computed using maximum entropy distributions and not the observed response distributions (*L85-L112* in **Supplementary Materials**)**.**

***Minor comments:***

***- If observers match the distribution over confidence reports, they will in particular match both the mean and the variance of the distribution. So why, in the abstract and throughout the text, do the authors say that "[the observers] match their mean confidence and confidence distributions" rather than either "match their confidence distributions" or "match the mean and variance of their confidence distributions"?***

We thank the reviewer for drawing our attention to this inconsistency. We now simply say ‘match their confidence’. We want to stress that the fact that we can observe convergence in confidence distributions does not necessarily imply that we can detect significant effects when analysing the mean and the variance of confidence; however, the fact that we do detect significant effects for these two moments (always for the mean and sometimes for the variance) corroborates our main claim.

***- It is not clear what is the added benefit of running the experiment both in Iran and the UK, given that no inter-individual difference analysis is being presented.***

The running of experiments in multiple countries reflects the international nature of our team. We have removed the reference to different cultural contexts in the Abstract so as not to detract focus from the main story. We did not observe significant differences between countries.

***Reviewer #3***

***In this study by Bang and colleagues, participants were tested both in pairs (social condition) and alone (isolated condition), using a joint decision/confidence measure. When partners disagreed in the social condition, the joint decision was the decision made by the partner who expressed higher confidence. The results show that pair members were more alike in their mean confidence, and in the distributions of their confidence reports in the social condition than in the isolated condition. This confidence 'matching' effect was replicated when participants completed twice as many trials in the social condition (Experiment 2), and when they used a continuous confidence scale (Experiment 3). Simulations of the data from these experiments using a signal detection model indicated that confidence matching is suboptimal in general, but can be relatively helpful in the case of dyads whose confidence ratings are poorly calibrated. Experiment 4, in which participants were paired with "virtual" partners that were simulated using the stimulus detection model confirmed that confidence matching had a negative effect on the accuracy of joint decisions when the partner was well-calibrated, and a positive effect when the partner was poorly-calibrated.***

***In general, this work makes a novel contribution to the important issue of how subjective confidence impacts group decision-making. The experiments are well designed, especially the clever 'virtual opponent' experiment and clearly presented. The work makes exemplary use of simple computational models to define and quantify confidence. The use of carefully chosen models and statistical analysis allow the authors to define the optimal behaviour under a limited set of assumptions and to compare that to individual and group performance. This revealed that performance failed to match optimal models, showing a significant bias of participants toward matching their partner's confidence ratings. The fact that a social equality bias can impair group decision-making is a significant issue that may be of general relevance to understanding human social behaviour.***

We thank the Reviewer for the positive assessment of our work. We have conducted two additional experiments to address the Reviewer’s concerns: (A) Experiment 5 introduces financial incentives (as per optimality issue) and (B) Experiment 6 asks whether confidence matching can result from mere exposure to another person’s confidence (as per specificity issue). As we describe below, the results of the two experiments show (A) that confidence matching persists even when participants have financial incentives to deviate and (B) that confidence matching is most pronounced in the context of group decision-making.

***I would caution the authors' to be careful with the use of the term 'optimal'. I am not sure the sense in which they are using it in p. 2, Fig. 1 is correct.***

We realise that our use of the term was underspecified. We now state that what is ‘optimal’ in our social task is for group members to report their confidence in a way that maximises the probability that the group makes the correct decision (*L61-L62* and**Figure 1**).

***I also see a tension in how to explain the departure from optimality and would suggest the authors address that. There are broadly three options: performance reflects mere 'heuristics', ad hoc short cuts lacking a normative basis, (2) performance may be under-constrained (optimality is not well defined) or (3) performance is constrained by additional variables not being considered, such as priors derived from natural experience or additional cost/benefit factors. The paper does an admirable job of dancing amongst these possibilities, but it might have been nice to see a stronger position taken. Regardless, the paper elegantly demonstrates that the computational approach has "legs" in facing difficult terrain. It may serve as an important anchor point for future studies to isolate and quantify "social cost functions", "social optimality", "social priors", etc.***

We have written an additional paragraph (*L289-L301*) where we discuss these explanations of the observed departure from optimality and how the ‘social’ hypotheses could be tested in future research using our framework. We have also conducted Experiment 5 (*L205-L217*) where participants had monetary incentives to report their confidence in a way that would maximise the probability that the group makes the correct decision – thus addressing the issue of optimality not being well-defined in our task.

***Specific points:***

***(1) The issue of gender interaction with confidence is present but not well laid out. First, information about the groups of participants is scarce. E.g. no mention of how many participants actually were female, or at least I couldn't find that information.***

We now give gender information in the **Methods** (*L311-L313*). In Experiments 1 to 3, all participants were male (*N* = 90). In Experiments 4-6, participants were male and female (male: *N* = 36; female: *N* = 46).

***Given that confidence judgements are said to vary with gender, and the likelihood that group decision-making varies according to whether the members of the group have on-going relationships, the authors should indicate, for each experiment, the gender balance of participants, and whether they were recruited in pairs.***

We now state in the **Methods** (*L313-L315*) whether participants were recruited in pairs (i.e., they knew each beforehand; Experiments 1 to 3) or individually (Experiments 4 to 6). We note that the interactions in Experiments 4 to 6 were anonymous.

***How do any gender effects compare to previous studies by the same group?***

We have not previously published gender effects. In fact, our earlier papers (e.g., Bahrami et al., *Science*, 2010; Mahmoodi et al., *PNAS*, 2015) only tested male participants. The current study is therefore a significant extension; the consistency of our main findings across single-gender (Experiments 1 to 3) and mixed-gender groups (Experiments 4 to 6) indicates that they are not gender-specific but general.

***The introduction of the issue of gender effects on social norming in the discussion is underdeveloped.***

We decided to remove the text on gender effects on social norming so as to make way for the additional data, analyses and simulations.

***(2) It is not clear whether the experimental design is sufficient to establish whether confidence matching and equality bias are phenomena associated specifically with group or social decision-making or rather with some component thereof. In experiment 4, performance in the "social" condition was compared with performance in an "isolated" condition, participants neither saw the confidence reports of other participants, nor contributed to joint decisions. Therefore it is possible in principle that the differences between these conditions were due, at least in some measure, to mere exposure to the confidence reports of another participant.***

We agree with the Reviewer that this is an important question and therefore conducted Experiment 6 (*L218-L231*) where we compared the social task with a task where participants *observed* their partner’s response after having made their own response but where no joint decision was selected. There were some markers of confidence matching in the observe condition (e.g., correlation in the magnitudes of group members’ mean confidence), but these markers were *significantly* weaker than in the social condition. We conclude that confidence matching is most pronounced in group contexts where individual confidence has consequences for the group outcome.

***(3) The authors propose that variation across single trials is normally assumed to arise from sensory noise, but that here it arises due to social interaction. What amount of this variation is due to slow 'social learning' over the course of the experiment? This could be tested by breaking down sessions into shorter blocks for analysis.***

We thank the Reviewer for recognising the relevance of our findings for current models of ‘metacognition’ – that is, the ability to evaluate ongoing performance. We present two (albeit related) social-learning effects: (a) convergence in group members’ mean confidence and (b) coupling between dyad members’ trial-by-trial confidence (i.e., interdependence in confidence discussed above). Our argument about *systematic* variation in confidence across single trials pertains to the ‘trial-by-trial coupling’ and not ‘mean convergence’. Breaking sessions into shorter blocks does not seem to be the best way to assess the impact of such history effects on metacognition – the trial-by-trial coupling is present throughout each experiment (and would thus be present for every short block) and is by definition very fast (and thus best analysed in a trial-by-trial analysis).

To get a handle on the impact of trial-by-trial coupling on metacognition, we used our learning model to simulate data under different degrees of social learning (i.e., different degrees of adaptation) and applied a standard measure of metacognitive ability to the data (meta-d’ first proposed by Maniscalco & Lau, *Consciousness &* Cognition, 2012). Briefly, this measure identifies the level of sensory noise that corrupts ‘first-order’ decision performance (i.e., d’) and then asks how much more noise is needed to account for ‘second-order’ confidence performance (i.e., meta-d’). The standard interpretation is that such additional noise reflects a noisy read-out of the first-order information and thus lower metacognitive ability. In **Supplementary Figure 9C** (referred to on *L284*), we now show that higher degrees of social learning correlate with higher differences between d’ and meta-d’. This analysis provides – in signal-to-noise units – a quantitative intuition for the relationship between variation in confidence across single trials and social learning. Importantly, by design, there is *no* noisy read-out in our learning model – only the mapping function is changing. It would therefore be wrong to attribute a *higher* difference between d’ and meta-d’ to *lower* metacognitive ability. In a sense, such differences may actually reflect *higher* metacognitive ability as the mapping function is adaptive to the contingencies of the local context.

***The claim that subjects have 'insight' into their confidence matching (line 203) is hard to defend against the alternative that they simply adapt, without any 'insight', to the social situation. Perhaps more detailed analyses of time courses could reveal whether sudden 'insights' were made.***

We thank the Reviewer for drawing our attention to this issue. We found in Experiment 4 that participants engaged more in confidence matching when it was good for the group – in the ‘poorly calibrated’ scenarios – than when it is bad for the group – in the ‘well-calibrated’ scenarios (*L348-351* in unrevised version). However, since these differences were weak, and for the sake of space, we removed the text discussing this analysis and the issue of ‘insight’.

***(4) 397-310: Why should accuracy be more discordant in the social condition? Larger differences then what, in the private condition? Was this the case no matter what the order of the blocks was?***

We did not have a prior expectation that the differences between dyad members’ accuracy should be larger in the social condition than in the isolated condition in Experiment 1. It is beyond the scope of the current paper to unpack this effect (but there is an extensive literature on social loafing in social psychology). We note, however, that the effect was not strong and it did not depend on the condition order (we now show in **Supplementary Figure 2** results for Experiment 1 split by the order in which dyads performed the two conditions). In response to the Reviewer’s second question (“Larger differences than what?”), we have rephrased the text to clarify that we were indeed reporting on differences between the social and the isolated condition (*L90-94*).

***(5) More care could have been used in preparing the figures for the supplemental material - they are a bit hard to read, due to missing colour bars, etc.***

We did not add colour bars for individual confidence landscapes (**Supplementary Figure 7**) as the values (joint accuracies) in each landscape had been normalised to the range 0-1. However, to facilitate interpretation of the landscapes, we now specify the range of ‘raw’ values shown in a given landscape and the fitted levels of sensory noise for each group member. The poor figure resolution arose when creating the PDF version of the manuscript for submission; the original figures are of high resolution.