Isolating cultural contributions to metacognition

E van der Plas\textsuperscript{1,2,*}, S Zhang\textsuperscript{3}, K Dong, D Bang\textsuperscript{1,4}, J Li\textsuperscript{3,*}, ND Wright\textsuperscript{5,6,*}, SM Fleming\textsuperscript{1,2,7,*}

\textsuperscript{*}Joint senior authors
\textsuperscript{1}Wellcome Centre for Human Neuroimaging, University College London
\textsuperscript{2}Department of Experimental Psychology, University College London
\textsuperscript{3}School of Psychological and Cognitive Sciences, Peking University
\textsuperscript{4}Department of Experimental Psychology, University of Oxford
\textsuperscript{5}Institute of Cognitive Neuroscience, University College London
\textsuperscript{6}Pellegrino Center for Clinical Bioethics, Georgetown University Medical Center
\textsuperscript{7}Max Planck UCL Centre for Computational Psychiatry and Ageing Research, University College London

Email: elisa.plas.18@ucl.ac.uk

Code and data availability: All anonymized behavioral data and analysis scripts supporting the findings of this work are available on GitHub (github.com/elisavanderplas/CulturalMetacognition).

Author Contributions: All authors conceptualized the project. EvdP, KD and SZ collected the data, EvdP analyzed the data, EvdP and SMF wrote the paper with revisions from DB, NDW and JL. All authors approved the manuscript.

Competing Interest Statement: The authors declare no competing interest.

Keywords: Metacognition, culture, confidence, advice-taking, perceptual decision-making

This PDF file includes:

Main Text (7,001 words)
Figures 1 to 4

Supplementary Materials (3,314 words)
Supplementary Figures 1 to 4
ABSTRACT

Some aspects of human metacognition, such as the ability to consciously evaluate our beliefs and decisions, are thought to be culturally acquired. However, direct evidence for this claim is lacking. As an initial step in answering this question, here we examine differences in metacognitive performance between populations matched for occupation (students), income, demographics and general intelligence, but drawn from two distinct cultural milieus (Beijing, China and London, UK). We show that Chinese participants have heightened metacognitive evaluation of perceptual decision-making task performance in comparison with UK participants. These differences manifested in boosts to post-decisional processing following error trials, despite an absence of differences in first-order performance. In a second experiment, we replicate these findings in a new task that replaced post-decision evidence with equivalent social advice. Together, our results are consistent with a proposal that metacognitive capacity is shaped via socio-cultural interactions.
INTRODUCTION

A canonical aspect of human cognition is the ability to reflect and report “I’m sure” or “I’m doubtful” about our perceptions, memories and decisions (Flavell, 1979; Nelson, 1990; Shea et al., 2014). Such self-evaluations, known as explicit metacognition, are thought to facilitate adaptive behavior in two ways: first, by allowing more efficient intrapersonal control, such as prompting further revision when we realize we do not know enough to perform well in an upcoming exam; and second, by facilitating interpersonal communication and collaboration, such as when two football referees pool their confidence about what happened on the pitch (Bahrami et al., 2010). In both of these cases, “better” metacognition, i.e., a tighter coupling between second-order self-evaluations and first-order cognitive or perceptual performance, tends to lead to greater individual and group performance (Bahrami et al., 2010; Bang et al., 2014, 2017; Fusaroli et al., 2012). By leveraging frameworks derived from psychophysics and signal detection theory, it has now become possible to isolate precisely metrics of metacognitive ability in laboratory tasks, for instance the extent to which subjects recognize their mistakes by adjusting their confidence accordingly (Galvin et al., 2003; Maniscalco & Lau, 2012). However, the origin of these high-level, reflective abilities remains poorly understood.

Developmentally, explicit metacognition is thought to be formed between the ages of 3 and 4 (Hembacher & Ghetti, 2014), although implicit precursors have been identified earlier in infancy (Goupil & Kouider, 2016a, 2016b). Intriguingly, explicit metacognition emerges around the same time as the ability to think about the minds of others (Carruthers, 2009; Lockl & Schneider, 2007), suggesting that similar computations may underpin self- and other-evaluation (Fleming & Daw, 2017). A recent theoretical proposal is that aspects of explicit metacognition may be culturally acquired and determined by the extent to which cultures place emphasis on discussing and understanding the mental states of self and other.
(Cleeremans et al., 2020; Heyes et al., 2020; Heyes & Frith, 2014). In other words, just as children learn to understand the meaning of written words from teachers and parents, children who grow up in cultures where working together is the norm may develop a stronger awareness of their own and others’ mental states.

A key implication of this “cultural origins hypothesis” is that metacognition should be subject to cultural variation to the extent that there are cultural differences in social collaboration and integration. For instance, Chinese populations are more likely to pay attention to and conform to others’ opinions than UK or US populations (Korn et al., 2014; Mesoudi et al., 2015; Oeberst & Wu, 2015); are thought to be more interdependent than independent in thinking styles (Singelis, 1994); and be more collectivist in emphasizing harmony with others than Western countries (Hofstede & Hofstede, 2010; Markus & Kitayama, 2010; Weber, 1905). However, whether cultural background similarly affects explicit metacognition remains unknown. Here, by applying recently developed psychophysical tools for isolating and quantifying the capacity for explicit metacognition about simple decisions, we seek to evaluate this hypothesis.

Previous cross-cultural studies of metacognition have focused on quantifying differences in confidence ratings. For example, a typical study might ask subjects general knowledge questions such as “Which one is further north: New York or London?” after which participants indicate their confidence that the decision was correct. Such studies have often found that Chinese populations report higher confidence than US or UK populations (Moore et al., 2018; Yates et al., 1989, 1998). It is important to note, however, that average confidence is only one facet of metacognition, known as metacognitive bias, and can vary independently of metacognitive sensitivity, the ability to discriminate between correct and error trials using confidence ratings (Fleming & Lau, 2014; Maniscalco & Lau, 2012). In other words, a highly confident person may still realize when they are wrong, and rate lower
confidence accordingly—thus demonstrating good metacognitive sensitivity. This capacity for metacognitive sensitivity, rather than idiosyncrasies in metacognitive bias, is also likely to be the key variable for effective collaboration with others (Bahrami et al., 2010; Fusaroli et al., 2012; Bang et al., 2014).

Two previous studies have quantified cross-cultural differences in both metacognitive bias and sensitivity. Yates and colleagues found that, despite a heightened (overconfident) metacognitive bias, metacognitive sensitivity was also higher in Chinese than US populations, as measured by probability judgment discrimination scores (Yates et al., 1989). Another study found heightened metacognitive bias in Chinese people living in Taiwan in comparison to Japanese and American populations, but inconsistent effects on metacognitive sensitivity (Yates et al., 1998). However, in both of these studies, first-order performance (judgment accuracy) was left free to vary across a wide range, and differences in metacognitive sensitivity are known to be potentially confounded by group differences in accuracy (Fleming & Lau, 2014)—people tend to better discriminate between their incorrect and correct decisions when the task at hand is easier. Moreover, both of these studies looked at associations between average confidence and average accuracy collapsed over groups of trials. Much less is known about cultural differences in the computational processes that give rise to fluctuations in confidence. For instance, recent work suggests that confidence is informed by evidence that becomes available after an initial decision has been made (“post-decision evidence”) (Navajas et al., 2016; van den Berg et al., 2016). When post-decision evidence contradicts a past decision, people tend to rate lower confidence; whereas post-decision evidence that confirms a past decision results in higher confidence (Fleming et al., 2018). Given the central role that post-decision processing plays in promoting openness to others’ (conflicting) viewpoints (Rollwage et al., 2018; Schulz et al., 2020), it could be that
cultural norms of harmony and collaboration selectively impact metacognition through shaping the processing of post-decision evidence.

Here we sought to provide an initial assessment of whether metacognitive capacity, as measured using performance-controlled laboratory tasks, differs between individuals drawn from distinct Northern European and Chinese cultural milieus. To ensure well matched samples, we compare the profiles of confidence judgments in Chinese and British samples matched for occupation (full-time students at Peking University, PKU, and University College London in the UK), age, gender, income and IQ. We only recruited Chinese/British citizens that had at least one parent that was born and raised in mainland China/Britain and had not lived more than one year abroad. We then leveraged recent methodological advances in metacognition research (Fleming et al., 2012; Fleming & Lau, 2014; Frith, 2012; Yeung & Summerfield, 2012) to disentangle potential effects of cultural background on both first-order and metacognitive processes engaged during the task to examine confidence formation independently of other aspects of task performance. We also asked whether post-decision evidence might differentially modulate confidence across cultural backgrounds. After an initial perceptual decision about the direction of a patch of randomly moving dots (left versus right), participants were shown additional (post-decision) evidence and asked to rate their confidence that the initial decision was correct. Using a calibration procedure, we selected stimuli of similar perceptual strength across individuals and sites to match first-order task difficulty, such that any difference in metacognition between cultures was unrelated to the first-order performance.

To pre-empt our results, in two independent behavioral experiments, we found that Chinese participants had heightened metacognitive sensitivity and post-decisional processing in the absence of differences in first-order perceptual performance, consistent with a hypothesis that cultural variation contributes to metacognition.
METHODS

Experiment 1.

Participants. We recruited N = 83 participants at both Peking University (PKU) in Beijing, China and University College London (UCL) in London, UK (Supplementary Table 1). At both sites the experiment was advertised via an online platform and flyers on campus, from which we recruited participants that were: (1) full-time students at PKU/UCL; (2) Chinese/British citizens; (3) had at least one parent that was born and raised in mainland China/Britain; and (4) had not lived more than one year abroad. All participants had normal or corrected-to-normal vision and no history of neurological or psychiatric illness.

Instructions, advertisements and questionnaires in English were translated to Mandarin Chinese and then back translated by an independent translator. The study was approved by the University College London Ethics Committee (1260/003) and by the Ethics Committee of School of Psychological and Cognitive Science at Peking University. All participants gave written informed consent before taking part in the experiment.

Two participants were excluded from the PKU dataset: one participant did not follow task instructions and one participant performed below our a priori accuracy cut-off threshold (i.e., less than 60% accuracy). Three participants were excluded from the UCL dataset: one participant was found not to have met the recruitment criteria after data collection (not a full-time student), one participant lacked variability in their confidence ratings (881/900 trials were rated as 100% confident) and one participant performed below the accuracy cut-off threshold of 60%. This resulted in the analysis of thirty-nine participants per site (N = 78 participants in total of which 39 female, mean age: 22.63 ± 0.33 years).

To establish that patterns of task performance were consistent with previous literature, we re-analyzed a previous dataset using the same task (Fleming et al., 2018) which was
collected at New York University (NYU). This dataset consisted of N = 25 participants (14 female, mean age: 24.0 ± 0.72 years), although information on the cultural background of the sample was not collected. The NYU recruitment was approved by NYU’s University Committee on Activities Involving Human Subjects and all participants provided written consent before taking part in the experiment.

**Experimental paradigm.** The experiment was programmed in Matlab 2014b (MathWorks) using Psychtoolbox (version 3.0.12) and presented on a desktop monitor at approximately 45 centimeters viewing distance. Stimuli were random dot kinematograms (RDKs): 30 moving dots (0.12° diameter) that appeared in a 7° circular white aperture for 300 milliseconds. The movement of the dots was generated by reploting the dots every three video frames, with a subset moving horizontally to either the left or the right and the remainder moving in a random direction. The subset that moved in the coherent direction was manipulated across conditions as giving rise to weak, medium or strong evidence strength. To ensure that these conditions were perceptually equivalent across participants, we performed a calibration procedure in which we estimated each participants’ psychometric function for a broad range of evidence strength levels and then selected the three evidence strength levels that were associated with three pre-specified levels of accuracy (weak = 60%, medium = 75% and strong = 90%; Supplementary Material 1.2).

On the psychophysical task, participants were shown 900 samples of evidence (RDK stimuli, pre-decision evidence) with variable evidence strength and were asked to judge the direction of dot movement (left or right). Participants indicated their choice by pressing a keyboard button [left: 1; right: 2] within 1,500 ms. After the choice, participants were shown “bonus” post-decision evidence where the dots moved in the same direction but with variable evidence strength (weak, medium, strong). In total, there were thus nine experimental conditions in a 3 (three pre-decision evidence strength levels) x 3 (three post-evidence
strength levels) factorial design (Figure 1a). At the end of every trial, participants were asked to rate their confidence that the initial judgment was correct on a scale ranging from 0 to 100%. Participants indicated their response by selecting a point on the scale with the mouse cursor within 3,000 ms. We implemented a Quadratic Scoring Rule (QSR) to motivate participants to report their confidence as accurately as possible. In particular, participants earned maximum points on a trial if they rated the lowest possible confidence about an incorrect judgment, or if they rated the highest possible confidence about a correct judgment.

Additional measures. After the psychophysical task, we administered three additional surveys: Self-Construal scale (Singelis, 1994), Analysis-Holism scale (Choi et al., 2003), and Culture-Free Intelligence test (Cattell, 1940). One of the authors translated the Analysis-Holism scale and the Culture-Free Intelligence Task to Mandarin Chinese and we used a published Mandarin Chinese translation of the Self-Construal scale (Singelis, 1994). All Mandarin Chinese translations of the questionnaires were back translated by an independent translator to ensure translation quality before the questionnaires were used at PKU. In Supplementary Material 1.1 we report the details of these questionnaires and compare their scores across sites.

Statistics. Group differences were tested with two-tailed independent samples t-tests (assuming equal variances). To assess the effects of our factorial design on accuracy and confidence, we conducted hierarchical mixed-effect regression models using the ‘lme4’ package in R (version 3.3.3) and plotted the behavioral data and the output of the model fits in MATLAB (version R2018a). We obtained the P-values of the regression coefficients using the car package. All models include a random effect at the participant level and all statistics are computed at the group level. We reported type III Wald chi-square tests ($\chi^2$), degrees of freedom (df) for fixed effects, and estimated beta-coefficients ($\beta$) together with their standard errors of the mean (± SEM) and P-values of the associated contrasts.
We investigated the effect of the pre-decision evidence strength (pre) [weak: -0.5, medium: 0, strong: 0.5] across sites [1: PKU, 2: UCL] on trial-by-trial accuracy [0: error, 1: correct] with the following hierarchical mixed-effect logistic regression model:

\[(1) \text{accuracy} \sim \text{site} \times \text{pre} + (1 + \text{pre} | \text{subj})\]

To predict confidence, we used a hierarchical mixed-effect regression model with trial-by-trial confidence (conf) as the dependent variable, and accuracy (acc) [-1: error, 1: correct], z-score of the log response time (RT), pre-decision evidence strength (pre) [weak: -0.5, medium: 0, strong: 0.5], post-evidence strength (post) [weak: -0.5, medium: 0, strong: 0.5], site [1 = PKU, 2 = UCL] and their interactions as predictors:

\[(2) \text{conf} \sim \text{site} \times (\text{acc} + \text{pre} + \text{post} + \text{pre} \times \text{post} + \text{pre} \times \text{acc} + \text{post} \times \text{acc} + \text{pre} \times \text{post} \times \text{acc} + \text{pre} \times \text{acc} + \text{pre} \times \text{post} + \text{pre} \times \text{post} \times \text{acc} + \text{pre} \times \text{post} \times \text{acc} + \text{pre} \times \text{post} \times \text{acc} + \text{RT}) + (1 + \text{acc} + \text{pre} + \text{post} + \text{pre} \times \text{post} + \text{pre} \times \text{acc} + \text{post} \times \text{acc} + \text{pre} \times \text{post} \times \text{acc} \times \text{RT} | \text{subj})\]

After demonstrating that we replicate the results of Fleming et al (2018) in each site separately, we combined the two datasets and included a site interaction term to investigate whether the effects are consistent between PKU and UCL (see Supplementary Material 1.3 for a comparison of all three sites including NYU). To investigate whether the model’s prediction of confidence improved when cross-cultural terms were included, we conducted a Likelihood Ratio Test that assesses the benefit of including interactions with site, here expressed in terms of the Akaike Information Criterion (AIC): \(\Delta \text{AIC} = \text{AIC}_{\text{without site}} - \text{AIC}_{\text{with site}}\), and the Log Likelihood (LL): \(\Delta \text{LL} = \text{LL}_{\text{with site}} - \text{LL}_{\text{without site}}\) with associated P value extracted from a type III Wald chi-square tests (\(\chi^2\)). In addition, we confirmed that simulating data from the summary statistics of the hierarchical regression model in Equation 2 successfully recaptured key features of the actual dataset (Supplementary Material 1.4).

To visualize the direction of the effects in Equation 2, we obtained the beta-coefficients of the pre-decision evidence conditions (pre) [weak: -0.5, medium: 0, strong: 0.5] and the post-
evidence conditions (post) [weak: -0.5, medium: 0, strong: 0.5] and their interactions on
certainty for each site separately [1: PKU, 2: UCL] and on error and correct trials
separately:

\[ (3) \quad \text{conf}_{err/corr} \sim pre + post + pre \times post + RT + (1 + pre + post + pre \times post + \]
\[ RT | \text{subj}) \]

Experiment 2

**Participants.** We recruited two new samples of participants at UCL and PKU, using the same
procedure as in Experiment 1. A minimum sample size of N = 53 at each site was defined by
an *a priori* power calculation of the t-test between the impact of post-decision evidence on
certainty in PKU and UCL in Experiment 1 (power = 80%, \( P = 0.05 \), Cohen’s d = 0.54).
Four participants were excluded from the PKU dataset: one participant performed below our
*a priori* accuracy cut-off of 60%; two participants’ calibration data was unusable, and one
participant violated transitivity in performance (i.e., average performance was lower in the
medium evidence condition than in the weak evidence condition). Two participants were
excluded from the UCL dataset: one participant did not believe the social manipulation and
never followed the advice (see ‘Experimental paradigm’), the other participant violated
transitivity. All participants had normal or corrected-to-normal vision and no history of
neurological or psychiatric illness. The study was approved by the University College
London Ethics Committee (1260/003) and by the Ethics Committee of School of
Psychological and Cognitive Science at Peking University. All participants gave written
informed consent before taking part in the experiment.

**Experimental paradigm.** We adapted the task used in Experiment 1. As in the original task,
participants were asked to judge the direction of moving dots (pre-decision evidence) with
varying evidence strength (weak, medium or strong). In addition, we made a number of
changes to the original paradigm. Confidence ratings were made on a confidence scale that
ranged from 100% confidence in the left direction to 100% confidence in the right direction (100%, 80%, 60% left and 60%, 80%, 100% right). Participants were asked to rate their confidence on this scale because, on a randomly selected half of the trials, the same scale was used to display the confidence estimation of a previous participant (‘adviser’) as social post-decision evidence. On the other half of the trials, post-decision evidence was a second RDK stimulus with dots moving in the same direction as pre-decision evidence but with variable evidence strength (weak, medium, strong). Social post-decision evidence was presented below a silhouette with a unique, uninformative background color. Participants were told that, because of the calibration procedure, the performance of the advisers was similar to theirs. In reality, the social advice was obtained from a computational model that made confidence and direction decisions with the same perceptual sensitivity level as the participant. This manipulation allowed us to keep the informativeness of post-decision evidence equal across conditions (social, perceptual) and manipulate the confidence levels of the adviser as a function of three evidence strength levels (with more confident advisers following stronger evidence; Supplementary Material 2.1). Together, this full-factorial design crossed three (pre-decision evidence strength) x three (post-decision evidence strength) x two (social, perceptual post-decision evidence type) within-subject conditions.

Additional measures. In addition to the three questionnaires administered in Experiment 1: the Self-Construal Scale (Singelis, 1994), Cattell Culture Free Intelligence Quotient (Cattell, 1940) and the Analysis Holism Scale (Choi et al., 2003) we also obtained participant’s responses on the Beck Cognitive Insight Scale (BCIS; Beck et al., 2004). This scale was originally developed to measure insight into symptoms within clinical populations but has also been used in non-clinical settings (Fleming et al., 2012). On the BCIS, participants indicated their agreement with statements about the recognition that experienced reality may be different from the objective truth. We were interested in knowing how insight would relate
to differences in post-decision evidence processing on the main task and whether, in light of
the cultural variation hypothesis, we would find cross-cultural differences on the BCIS
(Supplementary Material 2.4, Supplementary Table 1).

Statistics. Statistical inference was conducted similarly to analysis of Experiment 1. As
certainty estimates were given on a different scale in Experiment 2, we first converted
certainty in the dots moving left or right (conf$_{dir}$) to certainty in the chosen direction
[certainty wrong: 0, certainly correct: 1], by subtracting conf$_{dir}$ from 1 when the chosen
direction was left (a = -1), as follows:

$$ (4) \text{ if } (a = -1) \quad \text{confidence} = 1 - \text{conf}_{dir} $$

To index the strength of social post-decision evidence while ignoring the direction of the
advice, we transformed adviser confidence (conf$_{adv}$) on a scale from 100% left to 100%
right. We recoded this variable as ranging from 0-1, such that values < 0.5 indicated greater
adviser confidence in leftward motion and values > 0.5 indicated greater adviser confidence
in rightward motion. We then transformed this signed confidence variable to an unsigned
confidence variable ranging from 0.5 to 1, as follows:

$$ (5) \text{ if } (\text{conf}_{adv} < 0.5) \quad \text{conf}_{adv} = 1 - \text{conf}_{adv} $$

We then binned adviser confidence into three equal quantiles representing the lowest, middle
and highest 33% certainty ratings (conf$_{adv}$) to create 3 levels of social post-decision
evidence [weak: -0.5, medium: 0, strong: 0.5], which we used instead of ‘post’ in Equation 2.
Each individual’s beta coefficient for the main effect of perceptual and social post-decision
evidence (derived from Equation 3) were entered into a robust correlation using the
MATLAB robust correlation toolbox (Pernet et al., 2013).
RESULTS

In Experiment 1, we analyzed the data of N = 78 participants (N = 39 at each site) who were matched in terms of age (MPKU = 22.33 (SE = 0.38), M_UCL = 22.92 (SE = 0.54), independent samples t-test, t_{76} = -0.89, 95% Confidence Interval (CI) = [-1.91, 0.73], P = 0.38), gender (MPKU = 49%, UCL = 51%, t_{76} = -0.22, 95% CI = [-0.25, 0.20], P = 0.82) and annual family income (their parents’ combined gross income before tax, converted from Chinese renminbi (¥) to pounds (£) at 2017 purchasing power parity) relative to the per capita purchasing power parity at the time of recruitment (MPKU = £37,615.38 (SE = 4,535.01) and UCL (M_UCL = £39,381.35 (SE = 3,962.23), t_{75} = -0.29, 95% CI = [-138,52, 10320], P = 0.77). In addition, we administered a non-verbal measure of fluid intelligence which minimizes the influence of verbal fluency, culture and education (Cattell Culture-Free Intelligence test; Cattell, 1940), which showed no differences in intelligence between both sites (MPKU = 102.36 (SE = 1.79), M_UCL = 101.15 (SE = 1.52), t_{73} = 0.51, 95% CI = [-3.55, 5.96], P = 0.61; see Supplementary Table 1 for additional measures).

We next turn to the psychophysical task used in Experiment 1 (Figure 1a). Participants were asked to detect the direction of dot motion in a brief random-dot motion stimulus. The coherence level of random-dot motion was selected from a calibration phase to ensure that accuracy was equal across participants. As a result of the calibration procedure, the accuracy of participants’ initial decisions (first-order performance) was not statistically different between sites (MPKU = 83% (SE = 0.01), M_UCL = 83% (SE = 0.01), independent samples t-test, t_{76} = -0.20, 95% CI = [-0.03, 0.02], P = 0.85). The effect of pre-decision evidence (coherence) level on accuracy, i.e., the slope of the psychometric function, was also similar across sites (Supplementary Material 1.2). Using a hierarchical logistic regression to predict trial-by-trial accuracy, we found that first-order performance was indeed more accurate with stronger evidence (hierarchical linear regression, main effect of pre-decision
evidence: $\chi^2(1) = 363.02$, $P < 2e^{-16}$, $\beta = 2.92$ (SE = 0.15), $z = 19.05$, $P < 2e^{-16}$). As expected, this effect did not interact with site (interaction between site and pre-decision evidence: $\chi^2(1) = 0.94$, $P = 0.33$, $\beta = -0.21$ (SE = 0.21), $z = -0.97$, $P = 0.33$; Figure 1b).

**Figure 1. Experiment 1. Task design and matched first-order performance.** a, Participants made judgments about the direction (left versus right) of random dot motion. After seeing this pre-decision evidence, participants were shown additional post-decision evidence in the same direction as the pre-decision evidence but of potentially differing strength. Finally, they were asked to rate their confidence of their initial decision being correct on a scale from 0% to 100%, with percentages indicating probability of being correct. b, Choice accuracy was matched between sites (n.s.) and higher following stronger pre-decision evidence levels ($P < 0.001$, $N = 39$ participants at each site). Error bars represent group mean ± SEM.

Having shown that we matched choice accuracy (first-order performance) across sites, our next question was whether confidence ratings varied as a function of the strength of confirming or disconfirming post-decision evidence (weak, medium or strong) that each participant received (Figure 2a). Participants were instructed that the new evidence would always move in the same direction as the initial evidence and that they could use both pieces of evidence to rate their confidence about their initial response on a scale from 0 to 100%.

We crossed three levels of pre-decision evidence strength with three levels of post-decision
evidence strength to create a fully factorial 3 (pre-decision evidence strength) x 3 (post-decision evidence strength) factorial design (Figure 1a).

Across both sites, we replicated key patterns of confidence modulation reported previously using this task (Fleming et al., 2018): stronger post-decision evidence after an incorrect choice led to lower confidence (as participants could use the new evidence to realize that they were wrong), whereas stronger post-decision evidence after a correct choice led to higher confidence (as participants could use the new evidence to confirm that they were correct; Figure 2a and Supplementary Material 1.3).

We next tested whether a hierarchical regression model better predicted trial-by-trial confidence when the predictor variables (pre- and post-decision evidence levels, accuracy, standardized log response time (RT) and their interactions) were allowed to vary across sites. A Likelihood Ratio Test indicated that this was indeed the case (log likelihood (LL): ΔLL = 11 and Akaike Information criteria (AIC): ΔAIC = 5, χ²(9) = 23.38, \( P = 0.005 \); Supplementary Material 1.4), suggesting a significant role for cultural differences in affecting the construction of confidence. In addition, we replicated previous findings of higher average confidence ratings in Chinese participants (\( M_{PKU} = 85\% \) (SE = 0.01), \( M_{UCL} = 80\% \) (SE = 0.01), independent samples t-test, \( t_{76} = 2.32, 95\% \) CI = [0.01, 0.08], \( P = 0.02 \)), driven by PKU subjects tending to use higher confidence ratings on correct trials (Figure 2a).

The variance of confidence ratings was not different between sites (\( M_{PKU} = 85\% \) (SE = 0.01), \( M_{UCL} = 80\% \) (SE = 0.01), independent samples t-test, \( t_{76} = 1.35, 95\% \) CI = [0.004, 0.02], \( P = 0.18 \)).

We next asked how cultural background modulated the impact of new evidence on confidence by testing which predictor variables interacted with site. We found that post-decision evidence had a higher impact on confidence in the PKU dataset than in the UCL
dataset (hierarchical linear regression, interaction of post-decision evidence x site: \( \chi^2(1) = 6.89, P = 0.009, \beta = 0.05 \) (SE = 0.02). This effect was most evident on error trials, as shown by the steeper slope in the PKU dataset (Figure 2a). Indeed, when we fitted a hierarchical regression model on error trials only, the impact of post-decision evidence on confidence was significantly higher in the PKU dataset than in the UCL dataset (interaction between site x post-decision evidence on error trials: \( \chi^2(1) = 4.85, P = 0.03, \beta = 0.08 \) (SE = 0.04) but not on correct trials: \( \chi^2(1) = 2.40, P = 0.12, \beta = 0.02 \) (SE = 0.02); Figure 2b). However, the three-way interaction between post-decision evidence, accuracy and site did not reach statistical significance when tested within a single hierarchical regression model (\( \chi^2(1) = 2.23, P = 0.14, \beta = -0.03 \) (SE = 0.02), \( t_{74.04} = -1.49, P = 0.13 \), suggesting an enhanced susceptibility to new evidence in the PKU sample that was more pronounced on, but not necessarily restricted to, error trials.

Figure 2. Behavioral results for Experiment 1. a, Confidence as a function of post-decision evidence strength on error trials (red) and correct trials (blue) for each pre-decision evidence level. Shaded error bars represent group mean ± SEM. N = 39 at each site. b, Impact of post-decision evidence on confidence indicated as standardized beta-coefficients from a hierarchical mixed-effect regression
model on error trials (red) and correct trials (blue) at each site. Error bars represent group mean ± SEM, * $P < 0.05$. 
In summary, in Experiment 1 we found enhanced susceptibility to post-decision evidence in PKU participants compared with UCL participants, providing initial support for a heightened metacognitive evaluation of performance. Importantly, since first-order performance was matched between sites, these results support a hypothesis that metacognitive processes are liable to cultural influence.

In order to replicate and extend our results we conducted Experiment 2. Two new samples of N = 53 PKU participants (25 females, $M_{\text{age}} = 21.91$ (SE = 0.46) and N = 53 UCL participants (29 females, $M_{\text{age}} = 22.49$ (SE = 0.41), again with similar age ($t_{104} = -0.95$, 95% CI = [-1.81, 0.64], $P = 0.34$), gender (MPKU = 47%, MUCL = 55%, $t_{104} = -0.77$, 95% CI = [-0.27, 0.12], $P = 0.44$), Culture Free Intelligence Quotient (MPKU = 99.21 (SE = 1.41), MUCL = 102.00 (SE = 1.46), $t_{102} = -1.37$, 95% CI = [-6.82, 1.24], $P = 0.17$) and annual family income (MPKU = £41,373.58 (SE = 5,454.69) and UCL (MUCL = £56,988.89 (SE = 13,766.63), $t_{102} = -1.05$, 95% CI = [-45060, 13830], $P = 0.30$) were recruited. In Experiment 2 (but not in Experiment 1) we also included a measure of cognitive insight as quantified using the Beck Cognitive Insight Scale (BCIS; Beck et al., 2004). The BCIS includes questions about a person’s ability to recognize that objective reality may be different from what one subjectively feels to be true. In light of the findings of enhanced metacognition in PKU participants in Experiment 1, we hypothesized that PKU participants would report having greater insight than UCL participants. This hypothesis was confirmed by the questionnaire data, with PKU participants having higher average BCIS scores than UCL participants (MPKU = 40.26 (SE = 0.49); MUCL = 20.96 (SE = 0.82), independent samples t-test, $t_{104} = 20.08$, 95% CI = [17.40, 21.21], $P < 2.2e^{-16}$; see Supplementary Material 1.1. for other questionnaire measures and a comparison with Experiment 1).

In Experiment 2, participants again made a binary perceptual discrimination (left versus right random dot motion) based on pre-decision evidence of varying strength (weak,
medium or strong). Half of the trials were similar to those in Experiment 1. In the other half
of trials, the perceptual post-decision evidence was replaced by the confidence and direction
judgment provided by an anonymous previous participant (‘adviser’). This manipulation
allowed us to assess whether cultural differences in post-decision processing would
generalize across different domains (perceptual, social). In practice, we generated adviser
choices from a model that mimicked the perceptual sensitivity of the participant. The
stimulus that we presented to the simulated adviser was that trial’s perceptual post-decision
evidence level, i.e., the evidence strength that would have been presented to the participant in
the equivalent perceptual condition (with the same dot direction as the participant’s pre-
decision evidence yet with potentially variable strength). As a result of this, adviser accuracy
and confidence levels were contingent on the perceptual post-decision evidence strength on
any particular trial, which was counterbalanced with respect to the pre-decision evidence
strength just as for the perceptual condition. Participants were paired with a new adviser on
every trial and were told that all advisers had the same accuracy in detecting the motion
direction as themselves due to completion of an identical calibration procedure. One
participant reported not to believe the social manipulation and was excluded from further
analyses (see Methods).

We defined social post-decision evidence strength as the adviser’s confidence rating
binned into three levels (low, medium, high), creating a fully factorial 3 (pre-decision
evidence strength) x 3 (post-decision evidence strength) x 2 (post-decision evidence type)
design (Figure 3a and Supplementary Material 2.1). We again ensured that first-order
performance was matched across participants and across both post-decision evidence types
(Figure 3b and Supplementary Material 2.2). Accordingly, there was also no difference in
average confidence across sites (M_PKU = 82% (SE = 0.01), M_UCL = 79% (SE = 0.01),
independent samples t-test, $t_{104} = 1.64$, 95% CI = [-0.01, 0.06], $P = 0.10$).
Figure 3. Task design and first-order performance in Experiment 2. a, Participants were asked to make judgments about the direction (left, right) of random dot motion stimuli. Afterwards participants were either shown perceptual post-decision evidence or what an anonymous ‘adviser’ had decided on the same trial (social post-decision evidence, which was generated from a computational model). At the end of each trial, participants were asked to rate their confidence that the initial decision was correct on a scale from 100% left-stimulus to 100% right-stimulus. b, Choice accuracy was matched between sites (n.s.) and higher following stronger pre-decision evidence levels (*P < 0.001, N = 53 at each site). Error bars represent group mean ± SEM.

In the perceptual condition, we replicated our findings from Experiment 1 that PKU participants, in comparison with UCL participants, show heightened metacognitive evaluation in the processing of post-decision evidence. Specifically, perceptual post-decision evidence had a higher impact on confidence in the PKU dataset than in the UCL dataset (hierarchical linear regression, interaction perceptual post-decision evidence x site: $\chi^2(1) = 10.39$, $P = 0.001$, $\beta = 0.06$ (SE = 0.02), Figure 4a). This effect was again most evident on error trials, which in Experiment 2 led to a significant three-way interaction (hierarchical linear regression, interaction perceptual post-decision evidence x accuracy x site: $\chi^2(1) = 7.07$, $P = 0.008$, $\beta = -0.05$ (SE = 0.02)).
We next asked whether these differences in metacognition between cultural backgrounds would generalize to a situation in which post-decision evidence is presented as social advice. In the social condition of Experiment 2, we calculated how often participants changed their mind towards the direction suggested by the adviser on trials in which the participant and adviser disagreed. This tendency to change one’s mind and comply with the adviser was higher in PKU participants than in UCL participants (M_{PKU} = 17.9%, M_{UCL} = 12.6%, independent samples t-test, t_{104} = 2.21, 95% CI = [0.005, 0.10], P = 0.03). In keeping with a metacognitive advantage in PKU participants, this effect was restricted to trials on which the participant was wrong (and accordingly, the adviser correct; M_{PKU} = 33.8%, M_{UCL} = 24.1%, independent samples t-test, t_{104} = 2.59, 95% CI = [0.02, 0.17], P = 0.01), and was not seen on trials in which the participant was correct (and the adviser wrong; M_{PKU} = 8.3%, M_{UCL} = 6.5%, independent samples t-test, t_{104} = 0.92, 95% CI = [-0.02, 0.06], P = 0.36). This result suggests that the cross-cultural asymmetries in post-decision processing identified using perceptual stimuli generalize to cases in which new evidence is presented as social advice.

To further examine the drivers of cultural differences in advice-taking, we computed the impact (beta coefficient) of adviser confidence [low, medium, high] on participants’ confidence levels using a hierarchical mixed-effects model. Similar to the cross-cultural differences in perceptual post-decision evidence processing reported in Experiments 1 and 2, advice had a greater impact on the confidence ratings of PKU participants compared to UCL participants (hierarchical linear regression, interaction between social post-decision evidence x site: \( \chi^2(1) = 8.38, P = 0.004, \beta = 0.04 \) (SE = 0.02). As expected from the previous analyses, this asymmetry in the impact of adviser confidence was most evident on trials where the participant made an error (hierarchical linear regression, interaction social post-decision evidence x initial choice accuracy x site: \( \chi^2(1) = 10.56, P = 0.001, \beta = -0.05 \) (SE = 0.02), see
Figure 4a), consistent with a hypothesis of cultural differences in the metacognitive evaluation of performance.

At both sites, social post-decision evidence had a lower impact on confidence than perceptual post-decision evidence (hierarchical linear regression, interaction evidence types x post-decision evidence strength: $\chi^2(1) = 77.34, P < 2.2e^{-16}, \beta = 0.06$ (SE = 0.007). However, an enhanced susceptibility to post-decision evidence in PKU compared with UCL participants was found irrespective of whether the evidence was social or perceptual (no three-way interaction between evidence type, post-decision evidence and site: $\chi^2(1) = 3.35, P = 0.07, \beta = -0.02$ (SE = 0.01).

The similar manner in which social and perceptual post-decision evidence was processed suggests a domain-general component of post-decision evidence processing (Rouault, McWilliams, et al., 2018). In line with the pattern of confidence reports obtained in the perceptual version of the task, participants across both sites reported higher confidence after receiving more confident confirming advice and lower confidence after receiving more confident disconfirming advice (hierarchical linear regression, interaction-effect of social post-decision evidence and accuracy: $\chi^2(1) = 93.18, P = 2.2e^{-16}, \beta = 0.08$ (SE = 0.01; Supplementary Material 2.3). To further investigate this putative domain-generality, we next asked whether the impact of perceptual and social post-decision evidence was similar for any given individual. Figure 4b shows that this was the case: the impact of these two evidence types were positively correlated among both PKU participants (robust correlation, $r = 0.45, 95\% \ CI = [0.19, 0.64], P = 0.0006$) and UCL participants (robust correlation, $r = 0.39, 95\% \ CI = [0.13, 0.64], P = 0.004$), suggesting that participants who are more likely to integrate new perceptual evidence to update their confidence are also more likely to make use of social advice.
**Figure. 4. Post-decision evidence processing across domains.** a, Impact of perceptual and social post-decision evidence on confidence on error trials (red) and correct trials (blue) across sites and experiments. The coefficients from Experiment 1 (Figure 2b) are replotted for comparison. b, Standardized beta-coefficients for the impact of perceptual and social post-decision evidence on confidence for each participant from a hierarchical mixed-effect regression model standardized within each site. Error bars represent the group means ± SEM, *** $P < 0.001$, ** $P < 0.01$ and * $P < 0.05$. 
DISCUSSION

Across two behavioral experiments we show that Chinese participants were more susceptible to post-decision evidence than UK participants. In particular, Chinese participants changed their minds more after errors than their British counterparts, consistent with enhanced metacognitive evaluation of performance facilitated by adaptive post-decision processing. Using a psychophysical task that enabled the separation of first-order and metacognitive processes in simple perceptual decisions, our data supports a proposal that metacognition is sensitive to socio-cultural variation. Strikingly, these differences in confidence were found specifically on error trials, suggesting that cultural background may shape a metacognitive faculty to evaluate one’s own performance.

Our results are consistent with the recent theoretical proposal that explicit metacognition, the ability to self-evaluate one’s perceptions, memories and decisions, is subject to cultural variation (Heyes et al., 2020). The routes by which these differences emerge, and their stability over time, remains to be determined. One possibility is that the extent to which a culture places emphasis on the group over the individual may lead to more willingness to question and doubt one’s own decisions. For instance, in more collectivist societies it might be adaptive for all agents to communicate more accurate beliefs (Bang et al., 2017; Mahmoodi et al., 2015) allowing the correction of errors before they escalate to a societal level. In contrast, in more individualistic societies, cultivating distorted metacognition for one’s own ends (e.g., an overconfident style) may be prioritized. It also remains unclear as to what aspects of self-evaluative processing are affected by culture. In previous studies using related tasks within cultures, a distinction has been drawn between brain areas that are sensitive to post-decision evidence (in posterior medial frontal cortex) and those in more anterior frontal regions that mediate a mapping between private and public aspects of confidence (Bang et al., 2017, 2020; Fleming et al., 2018; Gherman & Philiastides,
Either or both of these levels of processing may plausibly be affected by culture and contribute to the current results.

The differences between cultural milieus in susceptibility to new evidence reported here complement and extend previous findings that Chinese populations are more affected by social influence than German and British populations (Korn et al., 2014; Mesoudi et al., 2015). In particular, we suggest that such differences in susceptibility to new evidence may partly be explained by heightened metacognition, rather than normative social compliance. In other words, recognizing the potential for error may prompt a search for corrective information from our peers (Schulz et al., 2020). Notably, Chinese participants were more susceptible to both social and perceptual forms of post-decision evidence, and such effects were most prominent on trials where mistakes had been made. Together, these findings suggest that the informativeness of the evidence—rather than mere social compliance—underpinned the cultural differences observed in the current study.

As perceptual post-decision evidence always disconfirmed a previous decision after errors (i.e., was always helpful), an alternative explanation of these findings is that Chinese participants simply processed disconfirming evidence to a greater extent than UK participants—in other words, they were less prone to confirmation bias (Kappes et al., 2020; Talluri et al., 2018). However, additional analyses of the social task data nuance this interpretation. The social task allowed us to distinguish between cases of disagreement when advice was correct (‘good advice’) as well as when advice was wrong (‘bad advice’). Notably, both Chinese and UK participants were equally susceptible to bad advice that agreed with their wrong decision (suggesting similar susceptibility to confirmatory social information) and to bad advice that disagreed with their correct decision (suggesting similar susceptibility to social disagreement). Instead, differences between cultural backgrounds selectively manifested in a heightened susceptibility of Chinese participants to ‘good’ advice,
even when it disagreed with their decision (Supplementary Material 2.3). This finding suggests that Chinese participants had heightened metacognitive evaluation of their performance, allowing them to selectively follow the advice when it is most beneficial.

Another line of evidence supporting a metacognitive explanation of our findings between sites is an association between our index of metacognitive processing (the tendency to specifically process new evidence on error trials) and an independent measure of cognitive insight (BCIS; Beck et al., 2004). Chinese participants had substantially higher baseline levels of self-reported cognitive insight than UK participants (Supplementary Material 1.1). In addition, inter-individual differences in cognitive insight, but not differences in sociocultural flexibility (as measured with the self-construal scale; Choi et al., 2007), predicted the degree of metacognitive processing in the sample as a whole (Supplementary Material 2.4).

In Experiment 2, we were also able to evaluate the domain-general nature of the cultural difference. On half of the trials post-decision evidence was perceptual, whereas on the other half it was presented as social advice. Differences between sites in post-decisional processing were similar across the social and perceptual forms of post-decision evidence, and the impact of both types of evidence was correlated across participants. Despite this similarity, participants at both sites adjusted their confidence levels to a lesser degree in response to social compared to perceptual evidence (Figure 4a), a difference that may have been due to the model generating simulated advisers with generally lower confidence levels than the participants (see Supplementary Material 2.1 for further discussion). Whether social and perceptual evidence have a similar impact on post-decision processing when advisers’ confidence is matched to that of the participant could be investigated in future experiments.
This study aimed at a robust and replicated assessment—using new, sensitive and specific methods that provide an in-depth analysis of individuals’ metacognitive processes—to compare two closely-matched samples drawn from distinct cultural milieus (for which a priori evidence suggested cross-cultural differences) and so provide evidence for or against an important hypothesis regarding human metacognition. We do not claim that either China’s or any other state or region’s culture is monolithic, or that our samples are representative of all 1,398 million Chinese or 67 million UK citizens, and instead we chose two well-matched subgroups. The strengths of such a tightly controlled, robust and replicated approach to explore a specific hypothesis can be complemented by future work using other approaches, which can, for example, look across broader groups of samples drawn from other ages, different socio-economic backgrounds, different levels of education (including adaptations to semi-literate populations) and other regions (within Northern Europe, within China and globally). Combining diverse types of study—both tightly controlled studies and those testing greater generalizability (Tiokhin et al., 2019)—will likely provide greater advances in understanding of human cognition and its cultural contributions than either types of study alone.

In summary, across two behavioral experiments we demonstrate that Chinese participants show heightened metacognitive evaluations of performance in comparison with UK participants. These differences manifested in boosts to post-decisional processing following error trials, in the absence of differences in first-order performance. This pattern was also obtained in a new task where post-decision evidence was replaced with equivalent social advice, suggesting that socio-cultural background shapes a domain-general tendency to evaluate and reflect on previous decisions.
REFERENCES


SUPPLEMENTARY MATERIAL

Experiment 1

1.1 Demographic and questionnaire measures

**CFIT.** We administered the Cattell Culture-Free Intelligence Test (CFIT), a non-verbal measure of individual’s fluid intelligence that minimizes the influence of verbal fluency, culture and education (Cattell, 1940). Participants were asked to complete visual puzzles by selecting one of four multiple choice options, which each pertained to four subtasks and had different instructions. The experimenter explained the instructions and signaled a pre-defined time limit for each subtask. Raw accuracy scores were converted to IQ scores following a standard coding table.

**BCIS.** In Experiment 2 we also administered the Beck Cognitive Insight Scale (BCIS) which evaluates a person’s tendency to reflect on their inner experiences (‘self-reflectiveness’ subscale) and the ability to critically reconsider inner experiences based on counterevidence (‘self-confidence’ subscale; Beck et al., 2004). Participants rated their agreement with fifteen items on a scale from 0 (‘do not agree at all’) to 3 (‘agree completely’), from which we computed a main composite score following a standard coding procedure. As reported in the manuscript, we found that main composite BCIS and both subscales were notably higher in PKU participants than in UCL participants (see Supplementary Table 1). Our reported scores on the BCIS were comparable with the scores of a control group in a large-scale clinical study conducted in India (Jacob et al., 2019). In line with our findings, Jacob et al proposed that BCIS may differ between collectivist versus individualist cultures but did not test this empirically.

In addition to the CFIT, BCIS and standard demographics (age, gender and family income) we also administered two questionnaires that have been developed to assess cultural differences in thinking styles.
AHS. The Analysis-Holism scale (AHS) measures individuals’ analytical versus holistic thinking tendency (Choi, Koo & Choi, 2007). People that think more analytically focus more on objects instead of on the whole, and usually desire one true answer instead of accepting that multiple dissimilar or even opposing truths can be valid at the same time. A total of 24 items (4 subscales) are rated from 1 (‘strongly disagree’) to 7 (‘strongly agree’). Following a standard coding procedure, we reverse-coded seven items and summed the resulting scores.

SCS. The Self-Construal Scale (SCS) measures the strength of individuals’ inter-dependent and independent self-construal (Singelis, 1994), i.e., how important people think that maintaining harmony within their social group is. A total of 24 items (corresponding to either the independent or inter-dependent subscale) were rated from 1 (‘strongly disagree’) to 7 (‘strongly agree’). We used an available translation from Huang et al. (2007) which was back-translated by an independent Chinese-English speaker.

Supplementary Table 1 | Demographical and trait differences between PKU and UCL datasets in Experiment 1 and 2. Group mean ± standard error from the mean. Income in pounds (£) is given relative to the purchase power parity (PPP) ratio between UK and China at the time of recruitment (ratio 1:1.71). Mean composite score ± standard error from the mean is given for the CFIT (Cattell Culture-Free Intelligence Test), AHS (Analysis-Holism scale), SCS (Self-Construal Scale), main composite BCIS (Beck Cognitive Insight Scale), and the self-reflectiveness and self-confidence subscales of the BCIS. *P < 0.05, ** P < 0.01, *** P < 0.001 for the independent samples t-test contrast between sites (assuming equal variance), bold values indicate a significant group difference. N/A means not administered.
Before the main task, participants were shown 240 random dot motion stimuli and had to judge the direction of the movement (left, right) without making a confidence estimation. The coherence of dot movement was manipulated across six coherence levels: 3%, 8%, 12%, 24%, 48% and 100%. Participants heard auditory feedback that signaled the accuracy of their judgment (high-pitched tone signaled a correct judgment and low-pitched tone signaled an error judgment). For every participant, a cumulative normal psychometric function was fitted to the data and the three coherence levels that resulted in 60%, 75% and 90% accuracy were used in the main task. In Supplementary Figure 1a we plot the likelihood of participants’ rightward judgement across each level of rightward motion coherence (ranging from 100% coherence left to 100% coherence right). Performance during the calibration phase was 76.6% correct (SE = 0.01) in the PKU sample, 75.7% correct (SE = 0.01) in the UCL sample and 73.5% correct (SE = 0.01) in the NYU sample. Using independent samples t-tests, we show that performance on the calibration phase was not different between PKU and UCL (t_{66} = 0.86, P = 0.39) or between UCL and NYU (t_{62} = 1.54, P = 0.13), but that it was higher in PKU than NYU (t_{62} = 2.40, P = 0.02, uncorrected). These performance levels were successfully reflected in the evidence strength levels that the participants received on the main task. The coherence
levels of the weak, medium and strong evidence levels were \([0.08, 0.21, 0.40]\) for PKU (average: \(M = 0.22\) (SE = 0.01)), \([0.10, 0.26, 0.46]\) for UCL (average: \(M = 0.27\) (SE = 0.02)) and \([0.13, 0.34, 0.56]\) for NYU (average: \(M = 0.34\) (SE = 0.03). As a result of this, first-order performance in the main experiment was matched across sites (Figure 1b in the main manuscript, Supplementary Figure 1b).

**Supplementary Figure 1. First-order performance across sites.** a, Probability of choosing right ‘\(P(\text{right})\)’ on the calibration task as a function of six coherence levels multiplied by their direction (dir: 100% left = -1, 100% right = 1). b, Fitted cumulative normal psychometric function (red) and behavioral data (blue) of the probability of choosing rightward direction on the main task as a function of the three coherence levels (1 = weak, 2 = medium, 3 = strong) multiplied by their direction (dir: left = -1, right: 1). Solid lines represent the predictions from the model, dots represent the group mean ± standard error.
1.3 Combined analysis of NYU, PKU and UCL datasets

PKU participants reported higher confidence on correct compared to error trials (main effect of accuracy: $\chi^2(1) = 261.77, P < 2.2e^{-16}, \beta = 0.25$ (SE = 0.02), and reported higher confidence after seeing stronger pre-decision evidence (main effect of pre-decision evidence: $\chi^2(1) = 5.30, P = 0.02, \beta = 0.03$ (SE = 0.01). The direction in which post-decision evidence influenced confidence was dependent on the accuracy of the initial choice (interaction of post-decision evidence x accuracy: $\chi^2(1) = 298.99, P < 2.2e^{-16}, \beta = 0.25$ (SE = 0.01). Specifically, receiving stronger disconfirming post-decision evidence on error trials decreased confidence (as participants could use the new evidence to realize that they were wrong), whilst receiving stronger confirming post-decision evidence on correct trials increased confidence (as participants could use the new evidence to confirm that they were correct). This V-shaped pattern is illustrated in Supplementary Figure 2a, b (blue lines indicate confidence about correct choices and red lines indicate confidence about incorrect choices). Post-decision evidence decreased confidence on error trials more than it increased confidence on correct trials, as indicated by a negative main effect of post-decision evidence on confidence (main effect of post-evidence: $\chi^2(1) = 112.02, P < 2.2e^{-16}, \beta = -0.15$ (SE = 0.01).

UCL participants also reported higher confidence on correct compared to error trials (main effect of accuracy: $\chi^2(1) = 230.39, P < 2.2e^{-16}, \beta = 0.23$ (SE = 0.02). Again, there was an interaction between post-decision evidence and the accuracy of the initial judgement (interaction of post-decision evidence x accuracy: $\chi^2(1) = 237.11, P < 2.2e^{-16}, \beta = 0.21$ (SE = 0.01). In addition, a negative main effect of post-decision evidence on confidence shows that post-decision evidence decreased confidence more on error trials than it increased confidence on correct trials (main effect of post-decision evidence: $\chi^2(1) = 40.29, P < 2.2e^{-10}, \beta = -0.09$ (SE = 0.01).
We next report how these effects interacted with site when also introducing the NYU dataset (setting PKU as a baseline in the regressions). In line with the site interactions described in Experiment 1, we find that the impact of post-decision evidence varied across sites (interaction post-decision evidence x site: $\chi^2(2) = 6.66, P = 0.04$). Contrasts show that this effect is mainly driven by a higher impact of post-decision evidence on confidence ratings in the PKU dataset than in the UCL dataset (contrast post-decision evidence PKU and UCL: $\beta = 0.06, SE = 0.02, t_{94.17} = 2.58, P = 0.01$). The contrast between PKU and NYU was in the same direction but did not reach significance (contrast post-decision evidence PKU and NYU: $\beta = 0.03, SE = 0.02, t_{100.74} = 1.02, P = 0.31$). As shown in Supplementary Figure 2a, b the negative slope on error trials (red line) is steeper in PKU than in UCL or NYU. The three-way interaction between post-decision evidence, accuracy and site was not significant (interaction accuracy, post-decision evidence and site: $\chi^2(2) = 3.54, P = 0.17$). PKU participants were marginally more susceptible to post-decision evidence on error trials than NYU participants ($\beta = -0.04, SE = 0.02, t_{101.49} = -1.72, P = 0.09$; Supplementary Figure 2c).

In addition to the hypothesized cultural differences in post-decision evidence processing, we also found cross-cultural differences in the impact of pre-decision evidence. The impact of pre-decision evidence varied across sites (interaction pre-decision evidence x site: $\chi^2(2) = 6.69, P = 0.04$). Contrasts reveal that pre-decision evidence had a lower impact on confidence in the PKU dataset than in the UCL dataset (contrast pre-decision evidence PKU and UCL: $\beta = -0.05, SE = 0.02, t_{92.37} = -2.56, P = 0.01$), the contrast between PKU and NYU is in the same direction but did not reach significance (contrast pre-decision evidence PKU and NYU: $\beta = -0.03, SE = 0.02, t_{98.84} = -1.40, P = 0.16$). In particular, this effect was restricted to error trials (3-way interaction accuracy, pre-decision evidence and site: $\chi^2(2) = 9.18, P = 0.01$). The impact of pre-decision evidence on error trials was lower in PKU than in UCL (contrast pre-decision evidence x accuracy: $\beta = 0.05, SE = 0.02, t_{95.22} = 2.80, P = 0.006$) and also lower in PKU than
in NYU (contrast pre-decision evidence x accuracy : $\beta = 0.05$, $SE = 0.02$, $t_{101.96} = 2.23$, $P = 0.03$; Supplementary Figure 2c).

**Supplementary Figure 2. Behavioral results from Experiment 1 across PKU, UCL and NYU datasets.**

**a.** Confidence as a function of post-decision evidence strength on error trials (red) and correct trials (blue) for each pre-decision evidence level. As we excluded the NYU dataset from the main analyses, this dataset is shown with dashed grey lines. Shaded error bars represent group mean ± SEM. $N = 25$ at NYU and $N = 39$ at UCL and PKU. **b.** Impact of pre-decision evidence level and post-decision evidence level on confidence as simulated from the beta-coefficients of the main hierarchical regression model reported in the main text (Equation 2). **c.** Impact of pre-decision evidence on confidence indicated as standardized beta-coefficients from a hierarchical regression model on error trials (red) and correct trials (blue) at each site. **d.** Impact of post-decision evidence on confidence indicated as standardized beta-coefficients from a hierarchical regression model on error trials (red) and correct trials (blue) at each site. Error bars represent group mean ± SEM.
1.4 Regression model comparison

We conducted a Likelihood Ratio Test to assess whether allowing for site interactions in our main hierarchical regression model (Equation 2) was justified, by comparing the likelihood of Equation 2 with and without site interactions using the ‘anova’-function in R (Winter, 2013). The validity of Equation 2 was also assessed by simulating data from its summary statistics and confirming that the simulated data recaptured key features of the actual datasets (Supplementary Figure 2b).

Experiment 2

2.1. Simulating advisers

Adviser’s responses (a_{adv}) were simulated under a signal detection theoretic model. We computed the perceptual sensitivity levels (d’) that an adviser who had experienced the same calibration procedure as subjects should show for each level of simulated evidence strength. This was obtained by transforming the target probability correct values (P_{adv}) used in calibration:

\begin{align*}
(1) \quad & P_{adv} = [0.6, 0.75, 0.9] \\
(2) \quad & d' = 2 \cdot \text{norminv}(P_{adv}) = [0.507, 1.3491, 2.563]
\end{align*}

From d’ we could calculate samples of evidence experienced by the adviser on each trial (), sampled from a normal distribution (~N) with mean determined by the perceptual sensitivity on a given trial (s; [weak: 1, medium: 2, strong: 3]), sign dependent on the true direction of the dots (dir, indicated as [left: -1, right: 1]) and a standard deviation of 1:

\begin{align*}
(3) \quad & x_{dir} \sim N\left(dir \cdot \frac{d'(s)}{2}, 1\right)
\end{align*}

The adviser reported rightward movement (a = 1) if exceeded an internal decision criterion which we assumed to be unbiased [m = 0]:
(4) \( if(x_{dir} > m) \)

\[ a_{adv} = 1 \]

\[ else \]

\[ a_{adv} = -1 \]

In addition to generating the choices of the adviser \((a_{adv})\), we used the same signal detection theory model to generate trial-by-trial adviser confidence levels. Due to an error in this model that mis-specified the mean and variance during inference, advisers were generally less confident than most participants. Despite this general tendency towards under-confidence, adviser confidence levels mimicked key features of human confidence levels: advisers were generally more confident about correct decisions (Supplementary Figure 3a) and less confident about wrong decisions (Supplementary Figure 3b). Furthermore, adviser confidence on correct trials was lowest in the weak post-decision evidence condition (57%, SE = 0.01), higher in the medium post-decision evidence condition (67%, SE = 0.01) and highest in the strong post-decision evidence condition (76%, SE = 0.02; Supplementary Figure 3a).

To avoid relying on the model when analyzing data, we decided to bin adviser confidence into three levels based on the 33% interquartile cumulative distribution and entered this as social post-decision evidence [-0.5: weak, 0: medium, 0.5: strong] in all regression analyses.
Supplementary Figure 3. Confidence levels of advisers. a, The probability density distributions of adviser confidence levels on correct trials. The left y-axis represents the probability density estimate along the three post-decision evidence levels [weak, medium, strong]. The right y-axis represents advisers’ choice accuracy for each level of post-decision evidence. Error bars represent group mean ± SEM. b, The probability density distributions of adviser confidence levels on error trials. The y-axis represents the probability density estimate along three post-decision evidence levels (weak, medium, strong).

2.2 First-order performance in Experiment 2

Using a hierarchical logistic regression on trial-by-trial accuracy in Experiment 2, we confirmed that choice accuracy was higher when participants had seen stronger pre-decision evidence (main effect pre-decision evidence: $\chi^2(1) = 484.85, P < 2e^{-16}, \beta = 2.97$ (SE = 0.14)). As per our calibration, this effect did not interact with site (no interaction-effect pre-decision evidence and site: $\chi^2(1) = 0.003, P = 0.96, \beta = -0.01$ (SE = 0.19) nor post-decision evidence type (no interaction-effect pre-decision evidence level and post-decision evidence type: $\chi^2(1) = 0.0003, P = 0.99, \beta = -0.01$; SE = 0.19; Figure 3b in the main manuscript).
2.3 Cultural differences in social advice-taking

A key difference between social and perceptual evidence is that perceptual post-decision evidence is always in the correct direction, whereas the advisers could sometimes be wrong. As a result, there were four possible trial scenarios in the social condition: (1) The participant was correct, and the adviser agreed (‘good’ agreement); (2) The participant as wrong and the adviser disagreed (‘good’ disagreement); (3) The participant was correct, yet the adviser disagreed (‘bad’ disagreement); (4) The participant was wrong, yet the adviser agreed (‘bad’ agreement).

To facilitate exploratory analyses of social post-decision evidence, we transformed participants’ confidence in the chosen direction to confidence in the objectively correct direction (ranging from higher confidence in the incorrect direction to higher confidence in the correct direction) as the dependent variable in an extended hierarchical regression model (Equation 3), to which we introduced agreement between the participant and adviser [disagree: -1, agree: 1] as an additional predictor variable. Confidence in the objectively correct direction was higher on disagree trials than on agree trials (main effect of agreement: $\chi^2(1) = 24.54, P = 7.28 \times 10^{-7}, \beta = -0.07$ (SE = 0.01). This effect is explained by a two-way interaction with accuracy, indicating that ‘good’ agreement increased participants’ confidence in the objectively correct direction, yet to a smaller extent than ‘bad’ agreement decreased participants’ confidence in the objectively correct direction (interaction effect of agreement x accuracy: $\chi^2(1) = 82.74, P = 2.2 \times 10^{-6}, \beta = 0.26$ (SE = 0.03).

When we allow the predictors to interact with site, we find that more confidently disagreeing advisers had a more pronounced impact on PKU participants than UCL participants (interaction agreement x social post-decision evidence level x site: $\chi^2(1) = 5.53, P = 0.02, \beta = 0.06$ (SE = 0.03). PKU participants were especially more susceptible to strongly disagreeing advisers when
their initial decision was, in fact, wrong (interaction accuracy x agreement x confidence adviser x site: $\chi^2(1) = 5.55, P = 0.02, \beta = -0.11$ (SE = 0.05). This effect is shown in Supplementary Figure 4a with the consistently steeper upwards sloping dark red lines in PKU than in UCL; and in Supplementary Figure 4b with heightened impact of ‘good’ disagreement in PKU than in UCL (compare the dark red dots across sites), but a similar impact of ‘bad’ disagreement across sites (compare the light blue dots across sites). In sum, these results show that PKU participants were more influenced by social post-decision evidence than UCL participants, but only when the advice was useful.

**Supplementary Figure 4. Cross-cultural differences in social advice-taking.** a, Linear fit of the association between the participant’s confidence in the objectively correct direction and the advisers’ binned confidence level in the objectively correct direction on correct (blue) and error trials (red) across sites, and as a function of adviser accuracy (good advice in darker colours, bad advice in lighter colors). b, The impact of advisers’ binned confidence level in the objectively correct direction on the participant’s confidence in the objectively correct direction on error (red) and correct trials (blue) across sites, and as a function of adviser accuracy (good advice in darker colors, bad advice in lighter colors), indicated as standardized beta-coefficients from a hierarchical mixed-effect regression model. As expected, the impact coefficients of good advice resemble the impact coefficients of perceptual post-decision evidence (as shown in Figure 4a in the main manuscript).
2.4 Correlations between regression model beta coefficients and questionnaire scores

In a final analysis we explored whether individual differences on the questionnaires were associated with beta coefficients on the main task. We computed each participant’s interaction coefficient between accuracy x social/perceptual post-decision evidence on confidence in Experiment 2 and correlated this measure with main composite self-construal scores (SCS; Singelis, 1994), IQ (Cattell, 1940) and cognitive insight (BCIS; Beck et al., 2004).

Post-decisional processing did not correlate with main composite independent construal (perceptual condition: Pearson’s $r = 0.03$, $P = 0.73$; social condition: Pearson’s $r = 0.09$, $P = 0.35$). In line with recent findings that metacognitive sensitivity and IQ are independent constructs (Rouault et al., 2018) we also do not find any association between post-decisional processing and IQ (perceptual condition: Pearson’s $r = 0.09$, $P = 0.39$; social condition: Pearson’s $r = 0.10$, $P = 0.32$). The interaction between post-decision evidence and accuracy positively correlated with BCIS scores (perceptual condition: Pearson’s $r = 0.23$, $P = 0.02$; social condition: Pearson’s $r = 0.34$, $P = 4.44e^{-04}$). These effects were driven by a positive correlation between post-decisional processing and both the BCIS self-certainty subscale (perceptual condition: Pearson’s $r = 0.18$, $P = 0.07$; social condition: Pearson’s $r = 0.35$, $P = 2.51e^{-04}$) and self-reflectiveness subscale (perceptual condition: Pearson’s $r = 0.23$, $P = 0.02$; social condition: Pearson’s $r = 0.27$, $P = 0.005$). Perhaps because of a lack of power to investigate individual differences in a sample of $N = 53$ participants, this correlation did not reach significance when both sites were analyzed separately (PKU perceptual: Pearson’s $r = 0.21$, $P = 0.14$, social: Pearson’s $r = 0.07$, $P = 0.60$; UCL perceptual: Pearson’s $r = 0.0005$, $P = 0.10$, social: Pearson’s $r = 0.22$, $P = 0.11$).
SUPPLEMENTARY REFERENCES


