Cognition 146 (2016) 377–386

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Individual consistency in the accuracy and distribution of confidence judgments

Joaquín Ais [a,b,1], Ariel Zylberberg [a,c,d,1], Pablo Barttfeld [a,b,e], Mariano Sigman [a,b,*]

[a] Laboratory of Integrative Neuroscience, Physics Department, FCEyN UBA and IFIBA, Conicet, Pabellón 1, Ciudad Universitaria, 1428 Buenos Aires, Argentina
[b] Universidad Torcuato Di Tella, Almirante Juan Saenz Valiente 1010, C1428BIJ Buenos Aires, Argentina
[c] Laboratory of Applied Artificial Intelligence, Computer Science Department, FCEyN UBA, Pabellón 1, Ciudad Universitaria, 1428 Buenos Aires, Argentina
[d] Howard Hughes Medical Institute, Department of Neuroscience and Kavli Institute for Brain Science, Columbia University, New York 10032, NY, USA
[e] Cognitive Neuroimaging Unit, Institut National de la Sante et de la Recherche Medicale, U992, 91191 Gif/Yvette, France

ARTICLE INFO

Article history:
Received 15 April 2014
Revised 18 August 2015
Accepted 11 October 2015
Available online 9 November 2015

Keywords:
Confidence
Perceptual decisions
Metacognitive judgments
Psychophysics

ABSTRACT

We examine which aspects of the confidence distributions – its shape, its bias toward higher or lower values, and its ability to distinguish correct from erred trials – are idiosyncratic of the who (individual specificity), the when (variability across days) and the what (task specificity). Measuring confidence across different sessions of four different perceptual tasks we show that: (1) Confidence distributions are virtually identical when measured in different days for the same subject and the same task, constituting a subjective fingerprint, (2) The capacity of confidence reports to distinguish correct from incorrect responses is only modestly (but significantly) correlated when compared across tasks, (3) Confidence distributions are very similar for tasks that involve different sensory modalities but have similar structure, (4) Confidence accuracy is independent of the mean and width of the confidence distribution, (5) The mean of the confidence distribution (an individual's confidence bias) constitutes the most efficient indicator to infer a subject's identity from confidence reports and (6) Confidence bias measured in simple perceptual decisions correlates with an individual's optimism bias measured with standard questionnaire.

1. Introduction

Every decision we make is earmarked with a confidence label that influences how we learn, how we communicate our decisions to others and when to stop deliberating and commit to a course of action. For instance, a student stops studying for an exam when he thinks he knows enough to get a good grade. To be useful, confidence has to reflect the true likelihood of being correct. In the previous example, the student may not pass the exam if he thinks he knows more than he actually does.

However, research in behavioral economics and cognitive sciences has repeatedly shown that the capacity of confidence to distinguish correct from incorrect knowledge – henceforth referred as the accuracy of confidence – can vary markedly between subjects for a given task. This variability has been related to individual differences in brain structure and function (Barttfeld et al., 2013; Fleming, Weil, Nagy, Dolan, & Rees, 2010). An assumption of this line of research is that there is a shared system for confidence judgements and hence that the accuracy of confidence in different tasks will yield similar scores for a given individual.

Recent studies have investigated this hypothesis. Pleskac and collaborators have shown that a single process can explain confidence in choices made in perceptual (line discrimination and random dot motion) and cognitive (city size inference) tasks (Pleskac & Busemeyer, 2010; Yu, Pleskac, & Zeigenfuse, 2015). Moreover, they show that a model based on a single process to compute confidence in different domains can explain what factors control the degree of correlation between tasks. Song and colleagues found positive correlations in the accuracy of confidence for two tasks that required discrimination of orientation or contrast (Song et al., 2011). McCurdy and colleagues found weaker (but significant) correlations between the accuracy of confidence judgments based on mnemonic and perceptual decisions (McCurdy et al., 2013). Moreover, they found distinct cerebral correlates of confidence accuracy for each task suggesting the existence of functionally segregated confidence systems in the human brain. Baird and colleagues (Baird, Smallwood, Gorgolewski, & Margulies, 2013) showed a non-significant correlation between confidence accuracy for memory and for perception and different connectivity patterns

* Corresponding author at: Laboratory of Integrative Neuroscience, Physics Department, Buenos Aires University, Pabellón I, Ciudad Universitaria (1428), Capital Federal, Buenos Aires, Argentina.
E-mail address: sigman@df.uba.ar (M. Sigman).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.cognition.2015.10.006
0010-0277/

(with the prefrontal cortex as a shared hub) accounting for the variance in the accuracy of mnemonic or perceptual confidence judgments. Hence, the results are somehow mixed suggesting – as was the case in the old literature of intelligence (Spearman, 1904) – that coherence in confidence reports across tasks may be very different, ranging from strong correlations in tasks with similar structures and shared features to moderate or almost negligible correlations in less related tasks.

In this manuscript we sought to examine which aspects of the confidence distributions – its shape, its bias toward higher or lower values, and its ability to distinguish correct from erred decisions – are idiosyncratic of the who (individual specificity) and the what (task specificity).

2. Methods

2.1. Participants

Twenty-three participants (12 female; mean age 24 ± 1.7 years) completed a total of eleven experimental sessions, as detailed in Table 1. All had normal or corrected-to-normal vision. All participants gave written informed consent, and the local ethics committee approved the study.

2.2. Tasks

Each participant performed four different tasks (Fig. 1). All tasks were performed in front of a 19″ CRT computer monitor, at a distance of ∼60 cm. After each choice, subjects reported the degree to which they considered their choice likely to be correct (termed 'confidence') by using a computer mouse to select a point on a scale (−13° to 13° in the horizontal meridian), which ranged from 'guessing' (left) to 'full certainty' (right). Subjects were explicitly asked to use the full range of the scale. No feedback was provided.

(i) Auditory discrimination task (*Aud*)

Two pure tones were presented sequentially, each one lasting 300 ms separated by an inter-stimulus interval of 500 ms. The pitch of the first tone was randomly selected in the range of 300–700 Hz. Subjects had to press keyboard key 1 (2) to indicate that the pitch of the first (second) tone was the highest. The difference in pitch between the first and second tone was adjusted with a Quest procedure to keep accuracy levels at 75% correct (Watson & Pelli, 1983).

(ii) Contrast discrimination (Con)

This forced choice visual discrimination task was adapted from (McCurdy et al., 2013). Each trial started with participants fixating a central red dot (diameter of 0.56°) on a gray background (50 cd/m^2) for 800 ms. On each trial, two circular targets appeared on screen (3°, eccentricity of 6°) for 300 ms, or until the subject responded. One of the targets contained only white noise, and the other a grating of random orientation (spatial frequency of 2 cycles per visual degree) superimposed with white noise. Subjects indicated which target contained the grating. The difficulty of the task was controlled by the contrast of the grating, adjusted with a Quest procedure to keep accuracy at a 75% correct (Watson & Pelli, 1983).

(iii) Luminance discrimination (Lum)

The task was adapted from Zylberberg, Barttfeld, and Sigman (2012). Each trial started with participants fixating a central red dot (diameter of 0.56°) on a gray background (50 cd/m^2) for 200 ms. Two flickering patches, were presented on the horizontal meridian, centered at ±1.04° from fixation. Each patch was composed of four vertical, spatially adjacent bars (0.14° × 0.56°). The luminance of the bars was updated every 50 ms, sampling from a Gaussian distribution with a standard deviation of 10 cd/m^2. The mean of this distribution equaled the luminance of the background for one of the patches and was set higher for the other (referred as the "target"). Participants pressed key G (H) to indicate that the brighter patch was on the left (right). The trial was aborted if subjects did not respond before 800 ms from the onset of the flickering stimuli. The mean luminance of the target was adjusted online to keep the proportion of correct responses at 75% (Watson & Pelli, 1983).

Table 1
Number of sessions and trials per task and subject.

Task	Sessions	Trials per session
Auditory	3	200
Contrast	3	300
Luminance	2	640
Partial report	3	384

(iv) Partial report (*ParRep*)

This task was adapted from (Graziano, Parra, & Sigman, 2010; Graziano & Sigman, 2008, 2009) where further details of the experiment can be found. Twelve letters (font Time New Roman of height 1.2°) were presented simultaneously for 16 ms. The letters were chosen randomly from the alphabet without repetition. Letters were arranged on a circle around the fixation, at an eccentricity of 5.2°. A red dot (0.1°) on an array of blue dots (with the same configuration as the letters) indicated the position of the target. Participants had to report, using a standard keyboard, the letter in the position cued by the red dot. The time between the offset of the array of letters and the onset of the cue (ISI) was selected pseudo-randomly, with possible values of [24, 71, 129, 200, 306, 506, 753, 1000] ms.

2.3. Data analysis

2.3.1. Metacognitive ability

For each task we measured (i) an individual's ability to correctly discriminate between stimulus alternatives, and (ii) the ability of confidence judgments to discriminate between correct and incorrect responses (Fleming et al., 2010; Maniscalco & Lau, 2012; Persaud, McLeod, & Cowey, 2007), measured for each session as the area under the Receiver Operating Characteristic (ROC) curve. To construct ROC curves, we calculated the parametric function (on the parameter x) of cumulative probabilities $p(confidence < x|correct)$ and $p(confidence < x|incorrect)$. The ROC was calculated independently for each session (Fig. A.1). We determine the accuracy of confidence as the area between the ROC curve and the x-axis, referred as Area Under the Curve (AUC), which ranges from 0 to 1. An AUC of 1 indicates that confidence reports perfectly distinguish correct from incorrect responses, while an AUC ∼0.5 indicates that there's a large overlap in the distribution of confidence reports for correct and incorrect responses (Galvin, Podd, Drga, & Whitmore, 2003).

2.3.2. Kullback–Leibler (KL) distance and hierarchical clustering

We used the Kullback–Leibler divergence to measure similarity between two confidence distributions. Confidence reports from each session were binned in 20 categories of equal length covering the full probability scale. For each subject, we obtained the distribution of confidence ratings for each of the eleven experimental

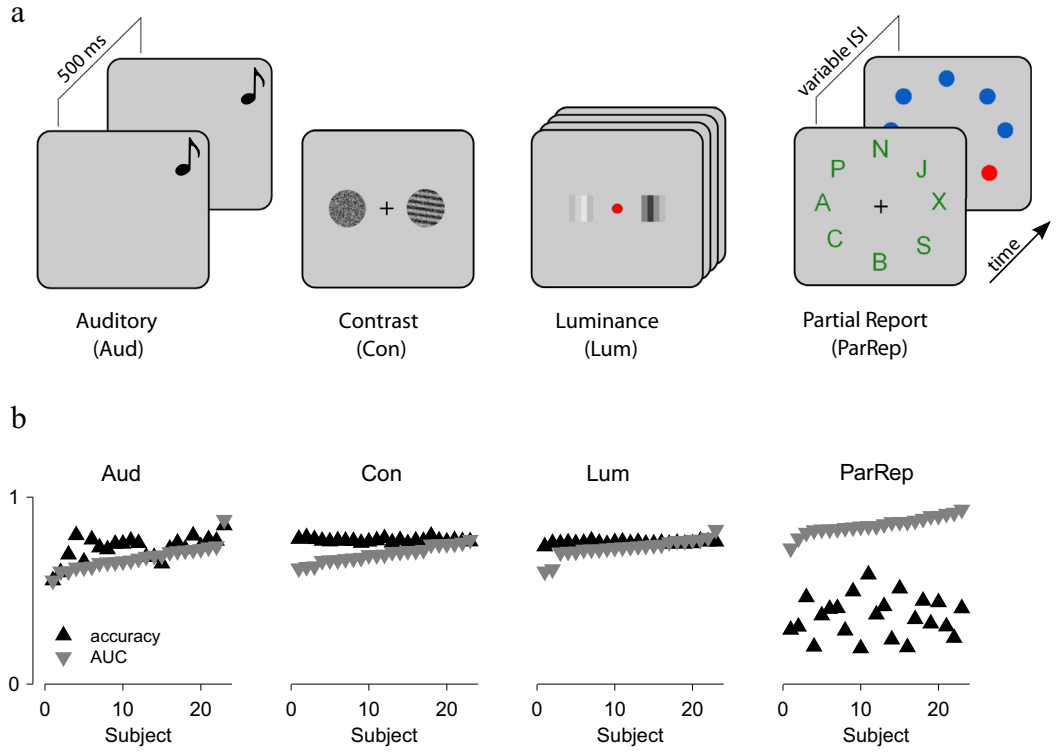

Fig. 1. Schematic depiction of the perceptual tasks and performance metrics. (a) Subjects performed four different tasks. In the Auditory Discrimination Task (*Aud*), subjects were presented with a sequence of two tones, and had to decide which of the two was higher-pitched. In the Contrast Discrimination Task (*Con*), subjects were presented simultaneously with two noisy patches and had to choose the one containing a grating. In the Luminance Discrimination task (*Lum*), subjects had to decide which of two flickering patches had a higher average luminance. In the Partial Report task (*ParRep*), subjects were flashed with an array of letters. At a variable SOA after the offset of the array, a red cue was presented and subjects had to report the identity of the letter that was previously at the position signaled by the cue. (b) The area under the Type-II ROC curve (AUC) was used as a measure of the accuracy of confidence reports. The average proportion of correct responses and the AUC is shown for each subject, sorted by AUC. Except for the partial report task, difficulty was adjusted online with a Quest procedure to maintain the proportion of correct responses at ~75%.

sessions, and computed the Kullback–Leibler divergence for every pair of sessions. We then built a matrix where each entry $M(i,j)$ represents the average over subjects of the within-subject divergence of sessions i and j:

$$M(i,j) = \frac{1}{Ns}\sum_{s=1}^{Ns}\sum_{k=1}^{Nb}\ln\left(\frac{P^{i,s}(k)}{P^{j,s}(k)}\right)P^{i,s}(k)$$

where $Ns = 23$ is the number of subjects, $Nb = 20$ is the number of bins used to discretize the confidence distribution and $P^{i,s}(k)$ is the proportion of confidence responses which fall on bin k for the confidence distribution obtained in session i of subject s. The elements above the main diagonal of the matrix (shown in Fig. 3b) were used to create a hierarchical cluster tree using the minimum of the pairwise distance to determine the distance between clusters – single-linkage clustering (Hastie et al., 2009).

2.3.3. Features that characterize the distribution of confidence reports

We also used three scalar metrics to characterize the similarity of the confidence distribution across sessions: the mean confidence rating (μ), the standard deviation (σ), and a multimodality index (*mi*). The latter measures the tendency of confidence reports (obtained from a single session) to be organized around multiple peaks as opposed to a single peak. We used Hartigan's dip test statistic, as this non-parametric test is consistent for testing any unimodal against any multimodal distribution (Hartigan & Hartigan, 1985). A larger *mi* indicates higher probability of rejecting the hypothesis of unimodality. Significance was tested ranking the empirical dip value against 5000 random samples of uniform distributions of sample size equal to the real data (Hartigan & Hartigan, 1985).

2.3.4. Decoding a subject's identity from her confidence distributions

We studied whether a classifier can identify the subject identity based on different features of the distribution of confidence, with a multinomial logistic regression model. In multinomial regression, each class (here subject) has its own linear discriminant function f^c:

$$f^c = \beta_0^c + \beta_1^c \cdot ses + \beta_2^c \cdot X \quad \text{for } c = 1,\ldots,nsuj$$

where *ses* is the session number (as *dummy* variable) and the βs are the fitted coefficients. We performed three different regressions where X was either the mean (μ), standard deviation (σ) or multimodality index(*mi*) of the confidence distribution. The probability that the data from a session (*ses* and X) corresponds to subject s follows the softmax distribution:

$$p(s) = \frac{\exp(f^s)}{\sum_k \exp(f^k)}$$

For classification, the predicted subject for each session is chosen as the subject class with the highest probability. The accuracy of the decoder is determined by the percentage of sessions that were correctly classified.

We quantified the classification efficacy of each parameter by comparing it against values obtained from a distribution of surrogated data. We generated a matrix $A_p(sub,ses)$ such that each entry (*sub*,*ses*) has the value of the parameter (p) for subject (*sub*) and session (*ses*). We compared the values obtained from the classification with surrogates of the data in which we random shuffled the

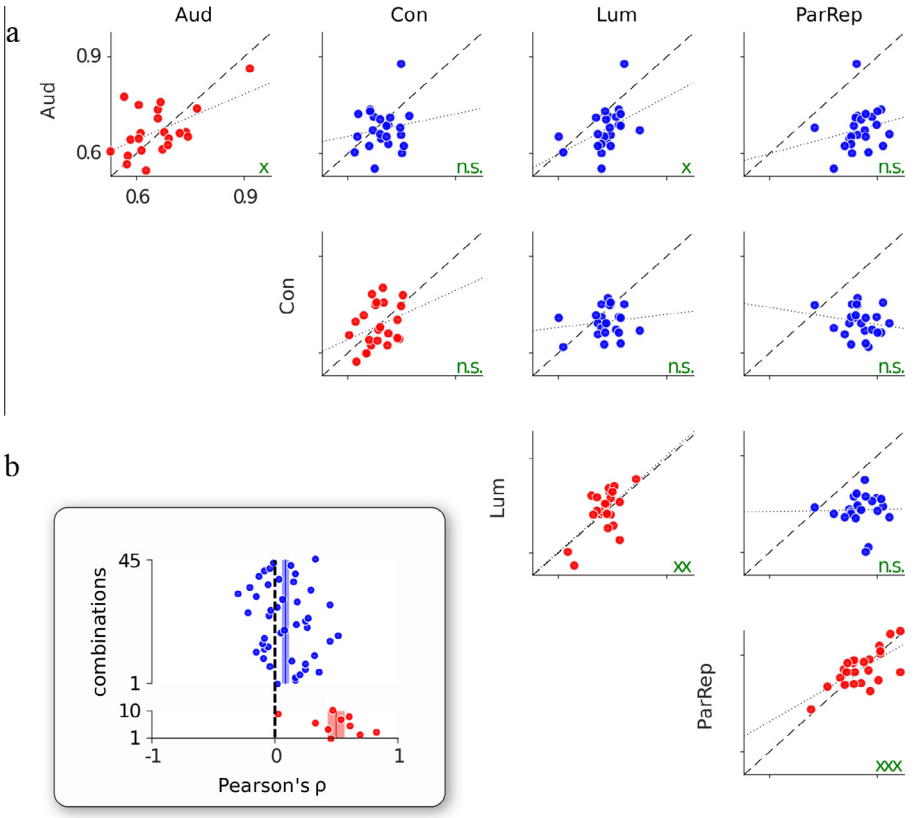

Fig. 2. Correlation of confidence accuracy within and across tasks. (a) Each dot represents for a single subject, the AUC value for a pair of tasks indexed by row and column. Elements of the main diagonal (same perceptual task, in red) were constructed by plotting the AUC values of the last two sessions against each other. For off-diagonal elements (different perceptual tasks, in blue), we first averaged AUC values across sessions for each experiment before plotting them against each other. Inside each graph, the dotted line represents the best fitting straight line, and the dashed line represents the identity. The significance of the Pearson's correlation coefficient is indicated in each graph: n.s.: not significant; x: p-value < 0.05; xx: p-value < 0.005; xxx: p-value < 0.0005. (b) Pearson's correlation coefficients for the AUC values extracted from every pair of sessions. Correlation coefficients computed from sessions of the same (different) task are shown in red (blue). Colored vertical lines indicate the mean value of the correlation, and shaded areas indicate s.e.m. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

elements on each column of the matrix $A_p(sub,ses)$. This results in a permutation of the subject identity for each session. This procedure allows us to obtain a measure of the variance of the distribution in surrogated data that can be used to assign a probability to the null hypothesis. We did so by comparing the measured classification accuracy (in the empiric data) to the distribution of classification accuracy obtained from 10,000 different runs of surrogated data.

To assess whether μ was a better predictor than AUC, we implemented a nonparametric bootstrap algorithm (Efron & Tibshirani, 1994). We generated a bootstrap distribution based on 1000 bootstrap samples. For each sample, we repeated the regression analysis including only a subset of participants, which were obtained as a random sample with replacement from the pool of all subjects. For each bootstrap sample, we computed the classification accuracy for regressors μ and AUC. We considered μ to be a better predictor than AUC if it was so for at least 95% of the bootstrap samples.

2.3.5. Life Orientation Test (LOT-R)

Participants completed the Life Orientation Test (LOT-R), a test devised to measure individual differences in generalized optimism/pessimism (Scheier, Carver, & Bridges, 1994), in its Spanish version (Perczek, Carver, Price, & Pozo-Kaderman, 2000). The test includes statements like "I'm always optimistic about my future" and "I rarely count on good things happening to me". Subjects indicate the extent to which they agree with each statement selecting one out of five alternatives: "I agree a lot", "I agree a little", "I neither agree nor disagree", "I disagree a little", "I disagree a lot". Each response is given a score from 1 to 5, with 5 being the most compatible with an optimistic trait. Participants responded to nine statements, six of which were relevant (three 'fillers'). Final scores range from 30 (maximal optimism) to 6 (maximum pessimism).

3. Results

Each participant performed four different perceptual tasks (Fig. 1a). The four tasks span a set of typical parameters in perceptual experiments: (a) visual vs. auditory discrimination, (b) reaction-time vs. fixed duration, (c) fixed or freely-varying performance, (d) decisions between events separated in time or in space, (e) binary or multiple-choices and (f) decisions based on symbolic or analogical stimuli. All tasks had a common structure where participants made a perceptual decision (choice) followed by a judgment about the confidence in their choice. Participants reported confidence in a Likert scale (Likert, 1932) by clicking on a horizontal line, which span the range from complete guessing on the left to absolute certainty on the right. Each participant completed between two and three sessions of each task, for a total of eleven sessions (days). The proportion of correct responses for each subject, experiment, and session is indicated in Table A.1.

3.1. Metacognitive ability within and across tasks

We used the area under the Type-II ROC curve (AUC) to measure the accuracy of confidence (Galvin et al., 2003). The AUC curve

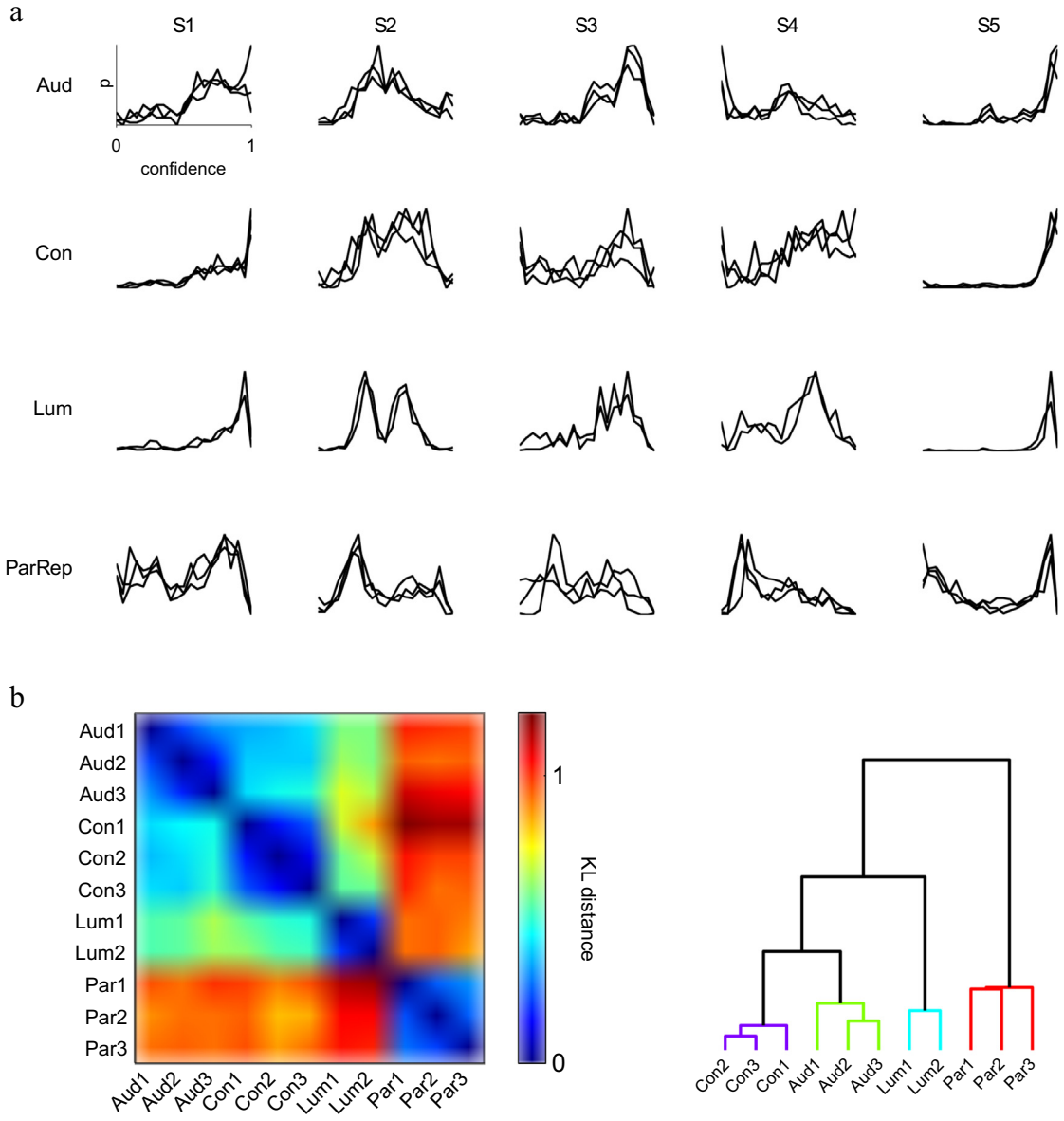

Fig. 3. Similarity of confidence reports across sessions. (a) Distribution of confidence reports in the four tasks for a sample of five different subjects. Distributions from different sessions are plotted on top of each other. To construct the distribution of responses, reports were classified in twenty bins of equal size covering the full scale. The same data is shown for every subject in Supplementary Fig. A.4. (b) Pairwise distance between every pair of sessions (left). Distance was defined as the average across subjects of the inter-subject KL divergence of the distribution of confidence reports (see Section 2). The matrix of pairwise distances was hierarchically clustered (right), revealing four clear clusters where sessions from the same task grouped together.

measures the degree to which confidence reports distinguish correct from incorrect responses (see Section 2). AUC values were above 0.5 for every subject and experiment, indicating that all subjects were reliably above chance in assigning confidence judgments relative to correct and error trials (Fig. 1b).

To determine whether the individual variability in AUC in one task is predictive of AUC in other tasks, we computed the Pearson's correlation coefficient across all pairs of tasks (Fig. 2a). Correlation coefficients of the distribution of individual AUC measured in different sessions of the same task consistently showed high positive values (Fig. 2a, diagonal, red dots). The AUC values obtained from different tasks typically resulted in weak but positive correlations (Fig. 2a, in blue).

To investigate whether there is a consistent within-subject correlation of AUC, we computed the correlation for every pair of sessions. The average correlation coefficient was significantly positive both within (one-sided T-value = 7.24, df = 9, $p < 5 \times 10^{-5}$) and across (one-sided T-value = 2.97, df = 44, $p < 0.005$) tasks (Fig. 2b). Correlations coefficients were significantly higher when obtained from sessions that belonged to the same experiment (one-sided T-value = 5.96, df = 53, $p < 10^{-6}$).

3.2. Confidence distributions as individuals' fingerprints

The shape of confidence distributions differed markedly across participants: some distributions were bimodal, others were packed in a high confidence mode, others a single and wide bell-shaped distribution with close to zero mean. In contrast with this broad variability across participants, the shape of the confidence distributions was extremely consistent across sessions of the same task for the same participant. In fact, confidence distributions for individual participants were almost overlapping for different sessions of

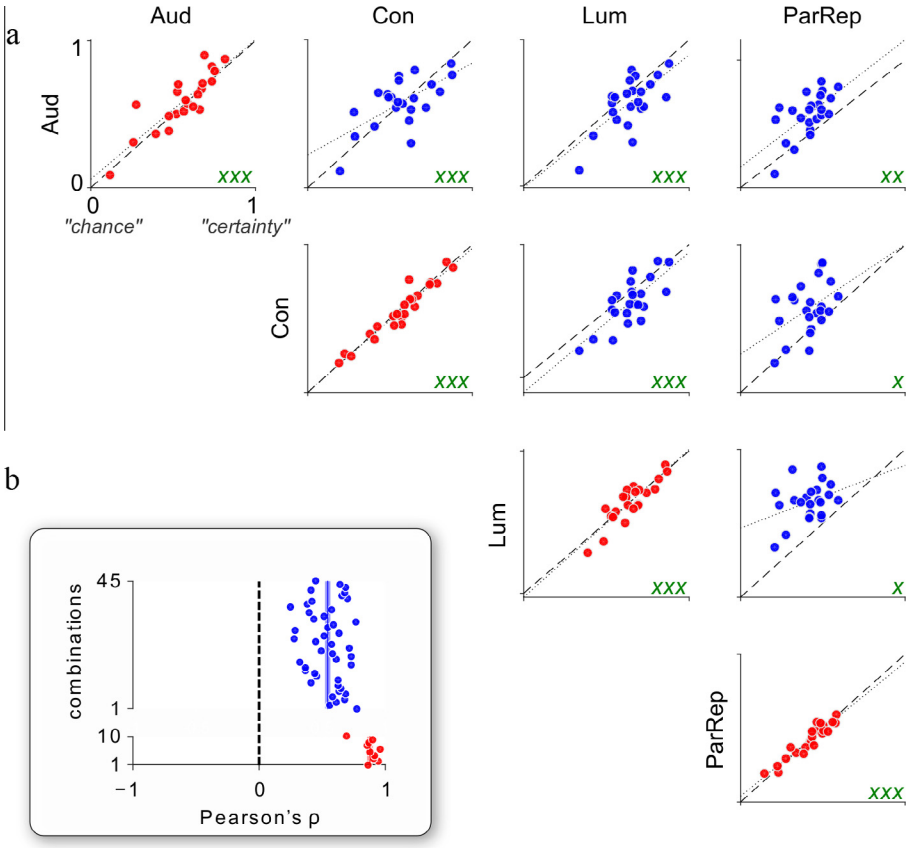

Fig. 4. Correlation of average confidence within and across tasks. Same as Fig. 2, but using the average of confidence (μ) instead of the AUC.

the same task (Fig. 3a). Across different tasks, confidence distribution showed larger variability (Fig. 3a). To formalize these observations, we computed the Kullback–Leibler divergence (KLdiv) to measure the distance between confidence distributions. For each subject, we computed the KLdiv for every pair of sessions (a total of 11^2 comparisons). After averaging across subjects, we obtained a matrix $M(i,j)$ which measures the average distance between sessions i and j (Fig. 3b, left). The visualization of similarity matrix confirms that confidence distributions have very similar shapes for all participants when measured in different sessions of the same task. It also shows that the shape of confidence distributions for the contrast and auditory discrimination task were very similar, less so to the luminance discrimination task. The partial report task (as could be expected due to large differences in performance) shows instead much higher scores of dissimilarity. These observations were confirmed by a hierarchical cluster analysis, which revealed four clear clusters (coded in different colors in Fig. 3b, right), each grouping together the sessions that belonged to the same task. The four clusters further organize into two clusters, with the partial report task separated from the rest. Hence, the clustering analysis reveals that the similarity in the usage of the confidence scale is supramodal, as we failed to observe a clear separation between auditory and visual tasks.

3.3. Confidence bias within and across tasks

The previous analysis highlights the similarity of confidence distributions across sessions. To study which features of the distributions are more consistent across sessions and tasks, we extracted from each distribution of confidence, the mean (μ; referred as 'bias'), standard deviation (σ), and an index of the multimodality of the distribution (*mi*). The *mi* represents the degree to which the distribution of confidence in one session is multimodal as opposed to unimodal, measured by the statistic for Hartigan's dip test for unimodality (Section 2).

We observed very strong correlations in the bias (μ) across sessions of same task (Fig. 4a). The correlation coefficient between sessions of different experiments was also significantly positive for every comparison (Fig. 4a). This indicates that an individual's confidence bias is consistent across tasks. Across the population, the mean correlation coefficient was highly significant both when the session belonged to the same ($p < 10^{-6}$, one sided T-value = 38, df = 9) and different ($p < 10^{-6}$, one sided T-value = 25.9, df = 44) tasks (Fig. 4b). In Figs. A.2 and A.3 we show the same analysis for the standard deviation and multimodality index of the confidence distribution, which were also highly significant both within and across tasks. Table A.2 shows the value of each index for every subject, experiment and session.

Together, these results indicate that the features that characterize the shape of the confidence distributions are a more robust fingerprint of a participant than the confidence accuracy as measured by the AUC. We formalize this assertion with a multinomial logistic regression analysis used to map each session to an individual subject. In different regression analyses we used μ, σ, *mi*, or AUC as independent regressors. Classification accuracy was 16.2%, 10.3%, 10.7%, and 7.9% for regressors μ, σ, *mi*, or AUC respectively. All these values are above chance level of 4.35%. However, the statistical power of these parameters was very different. Classification power for the three parameters of the distribution was highly significant ($p < 10^{-3}$) while classification power for the AUC was only marginally significant ($p = 0.053$). A nonparametric bootstrap test (Efron & Tibshirani, 1994) revealed that the classification accuracy

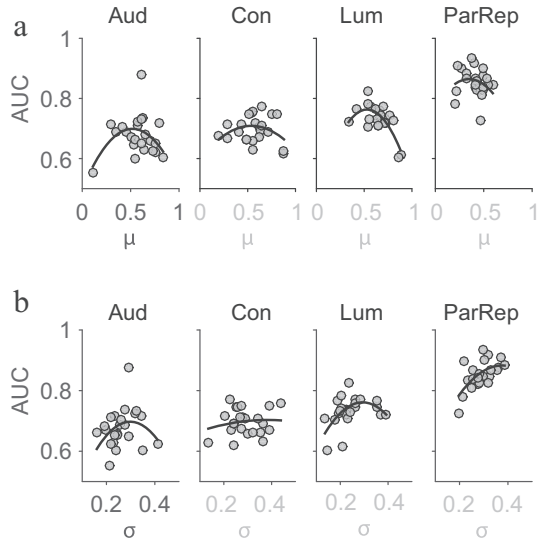

Fig. 5. Relation between the distribution and accuracy of confidence reports. Relation between the shapes of the confidence distributions characterized by μ and σ, and the AUC values. Each point represents a different participant (with μ, σ and AUC averaged across sessions of the same task), and the solid line represents best fits to second order polynomials.

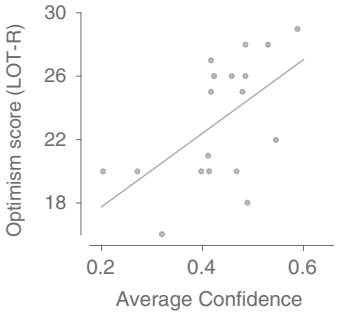

Fig. 6. Confidence in a perceptual choice indexes traits of optimism. The average confidence obtained after aggregating all trials is correlated with an index of optimism versus pessimism obtained from LOT-R. Each dot represents a different participant, and the continuous line is the least-squares fit. Pearson's correlation ($\rho = 0.56$) was significantly positive ($p = 0.007$).

for μ was significantly higher than for AUC ($p < 0.02$, $N = 1{,}000$ bootstrap samples). The classification capacity of these parameters was synergic which was revealed by the fact that the accuracy of the decoder increased significantly to 37.1% with a multinomial regression that includes all four features simultaneously.

We wished to ascertain whether the shape of the confidence distribution constraints the capacity of confidence judgments to distinguish between correct and incorrect choices. For example, in the limit case in which a participant is completely certain of her choices (a distribution with $\mu = 1$ and $\sigma = 0$), confidence judgments won't be able to distinguish correct from erred responses. Yet, the relation between the distribution and accuracy of confidence reports may not be significant for typical values of these observables. We investigated the relation between the mean and standard deviation of the confidence distribution and the AUC (Fig. 5). First, we fit a second order polynomial to the relation between AUC and μ (Fig. 5a). This relation tended to be concave (indicating that for extreme cases of very high or very low confidence there is a slight decrease in the accuracy of the confidence reports). However, the explanatory power of these fits was very low (the average R^2 for the data shown in Fig. 5a is 0.22) and did not reach significance indicating that the subjective choice of whether to report confidence in higher or lower average values (which is persistent and reliable across sessions and experiments, as described above) has a very weak effect on the accuracy of the confidence system. Similarly, an analysis of the linear and quadratic relations between AUC and σ (Fig. 5b) did not reach significance indicating that subjects can report confidence within narrower regions of the confidence scale without affecting its ability to distinguish between correct and incorrect responses.

3.4. Confidence bias is correlated with an individual's bias toward optimism versus pessimism

The previous analysis indicates that the mean of the distribution (μ) is the most reproducible feature of how an individual expresses confidence. It's conceivable that μ only reflects that subjects use the scale differently, i.e. that the same point on the scale represents different measures of correct probability for different subjects. Alternatively, μ could be an index of a subjects tendency to be underconfident or overconfident, or, more generally, optimistic or pessimistic.

We tested the degree to which the average confidence indexes a subject's generalized bias toward optimism versus pessimism. A few months after the completion of the perceptual experiments, we asked participants to complete the Life Orientation Test (LOT-R), a test devised to measure individual differences in generalized optimism/pessimism (Scheier et al., 1994). Eighteen of the twenty-three participants completed the test. The bias toward optimism/pessimism as measured with the LOT-R was strongly correlated with the average confidence computed from the psychophysics experiments ~1 year earlier (Fig. 6).

4. Discussion

Our aim was to examine which aspects of the confidence distributions are characteristic of the *who* (individual specificity) and the *what* (task specificity). To this aim we conducted a large-scale psychophysical experiment in which 23 subjects performed 11 sessions for a total of about 100.000 trials.

The summary of the main results of this study is:

(1) Most aspects of confidence distribution (AUC, mean, shape estimators, variance) are highly reproducible across sessions for the same individual in the same context.
(2) The precision of confidence reports to distinguish correct and incorrect (AUC) trials is only modestly correlated when compared across tasks.
(3) The distributions of confidence are – as said above in (1) – almost identical for different sessions of a given individual and task. Across tasks the distributions of confidence show several important changes and regularities. First, similarity of these distributions is not affected by modality. In fact, confidence distributions where very similar for the contrast and auditory discrimination tasks which are based on different modalities but have very similar task structures. The task that showed greater dissimilarity is the Partial Report Paradigm. The space of parameters is too large to infer which of the many factors that vary between tasks is the critical one. It is natural to think that a main factor explaining why confidence distributions for the partial report task were very different from the rest is accuracy, which is close to 35% percent on this task (where chance level is ~3%) while in other tasks it was held close to 75% with chance levels of 50%. However, an interesting observation is that spikes of very high (almost perfect) confidence where much more frequent

in this low-accuracy task (Fig. 3). Two possible explanations of this observation are that the partial report task is discrete and based on symbols and not analog quantities, and that the task has perceptual analogs of what psychologists have referred as contrary, misleading or deceptive problems (Lichtenstein, Fischhoff, & Phillips, 1977; May & Scholz, 1986). In the partial report task the subjects see an array of letters and then are required to answer which letter was presented in a given location. We demonstrated that spatial transposition of a given shape is a frequent unconscious operation which results in subjects being very confident about a response which is perfectly correct within a flawed scheme of reasoning, a sort of visual illusion (Graziano & Sigman, 2009). The flaw is of course unnoticed to subject's introspection and subjects are hence highly confident as they would also be in the wrong assessment of a visual illusion. This presents an intrinsic methodological difficulty for the study of universal aspects of confidence judgments (Klayman, Soll, González-Vallejo, & Barlas, 1999).

(4) Despite the fact that there were large variations in the shape of confidence distributions across tasks, the mean of this distribution was highly indicative of a subject's identity. A regression analyses showed that the average confidence was a more efficient indicator of a subject's identity than the capacity of confidence to distinguish correct from incorrect decisions. In fact every single correlation of the average confidence across different sessions (measured in different days and performing different tasks) was positive.

4.1. Accuracy of confidence

Many behavioral and theoretical studies have addressed previously under which circumstances confidence judgments are accurate or inaccurate with some discrepancy on the mechanisms and in the conclusions. Tversky and Kahneman have argued that confidence judgments are often inaccurate because they rely on heuristics and simplifications that yield systematic errors (Griffin & Tversky, 1992; Tversky & Kahneman, 1974). It has also been suggested that confidence judgments may not be precise due to unbiased random variation or noise in the decision process (Erev, Wallsten, & Budescu, 1994). Gigerenzer, Hoffrage, and Kleinbolting (1991) criticized some of these conclusions arguing that inaccuracies in confidence judgments may be due to selective sampling of questions by the experimenter. In support of this hypothesis, Juslin, Winman, and Olsson (2000) analyzed a large set of studies (with random sampling) and showed that the average difference between confidence and accuracy was basically indistinguishable from zero.

Previous studies have also measured the accuracy of confidence judgments using – as we do in this study – a signal detection approach (Fleming & Lau, 2014; Galvin et al., 2003; Zylberberg, Roelfsema, & Sigman, 2014). Pleskac and colleagues developed a 2-stage dynamic signal detection (2DSD) theory which assumes that evidence can continue to accumulate after the choice. By increasing the amount of evidence that they collect during the second stage of 2DSD (after choice), decision makers can increase the resolution of their confidence reports (Pleskac & Busemeyer, 2010; Yu et al., 2015).

An important question to this general aim is to understand the reliability of the accuracy of confidence across difference sessions and tasks. In our study the accuracy of confidence reports to distinguish correct and incorrect trials is only modestly correlated when compared across tasks. The "half-empty or half-full glass" can be described in a more clarifying manner. For each pair of sessions of different tasks the correlation is weak. However, when all possible pairs of sessions (for different tasks) are taken together, the distribution of r values of the distribution is highly significantly shifted toward positive values, revealing a more likely tendency toward positively correlated measures.

4.2. Idiosyncrasies in the expression of confidence

In contrast to the modest reliability observed in the accuracy of confidence, most aspects of confidence distribution (AUC, mean, shape estimators, variance) are highly reproducible across sessions for the same individual in the same context. The distribution of confidence in a given task constitutes a subjective fingerprint. Confidence distributions have a very high capacity to classify a subject's identity.

More generally, our work conveys the idea that confidence is not the mere report of an internal probability but that, instead, it is expressed in an idiosyncratic manner. For some individuals collapsed in a single mode, for others being expressed in two different categories. There does not seem to be a universal language to express probabilities. This is in line with many studies investigating how people convey uncertainty with words such as "possible", "likely" "doubtful". It has been generally found that each individuals uses different expressions to describe identical situations (Wallsten & Budescu, 1995; Wallsten, Budescu, Rapoport, Zwick, & Forsyth, 1986). In fact, Budescu and colleagues have shown that they can reduce communication errors by a "translation" device that standardizes an individual's linguistic idiosyncrasy to map uncertainty (Karelitz & Budescu, 2004). It also resonates with previous findings showing that individual differences in policies to express confidence may be cultural, distinguishing groups of people who express (verbally and numerically) continuous notions of probabilities from those who express it in a more categorical (all-or-none) fashion (Phillips & Wright, 1977). Phillips and Wright have extended this idea beyond the domain of decision confidence to argue that there are cultural and individual differences in the way that people think and conceptualize probabilities and uncertainty (Wright et al., 1978). Individual persistence in the policy to express confidence has also been reported in a seminal study by Adams and Adams (Adams & Adams, 1961) who reported a grossly overconfident calibration curve of a schizophrenic who believed he was Jesus Christ (Lichtenstein et al., 1977). Swets, Tanner, Wilson, and Birdsall (1961) also observed that the calibration curves of each of their observers were widely different.

It is important to emphasize that in the present study subjects report in a non-calibrated scale, and thus we cannot make precise claims about whether a participant is over or under confident. Still, we observed a significantly positive correlation between confidence bias in the simple perceptual decisions we studied, and the general optimism bias measured with standard questionnaire (Scheier et al., 1994). Accordingly, while confidence bias is a highly reproducible trait of how confidence is reported (Stankov & Crawford, 1996), previous research has also identified that being over or under confident and generally optimistic or pessimist is a stable individual trait (Plomin et al., 1992). In the optimism/pessimism classification, there seems to be a bias toward optimism, typically expressed in people believing that they will live longer than average, underestimating divorce chances, health problems and risks such as car accidents (Weinstein, 1980). Pessimism is typical in patients with depressive symptoms indicating that specific populations may have a bias (the causality of this correlation is not known) toward a specific policy of confidence (Drevets et al., 1997). These findings are consistent with our observation that the confidence bias is the most stable individual trait of confidence distributions.

4.3. Questions for future research

Some questions derive naturally from the present study, which could be addressed in future studies:

(1) One natural question that derives from this observation is which aspects of people's personality are related to persistent differences in confidence distributions, preserved across days and tasks. As mentioned above, depression is associated with under-confidence (Drevets et al., 1997) but it is likely that more subtle manifestations of personality, brain function and structure, may predict average scores of confidence in the same way that some cerebral structures (most reliably the frontal cortex (Barttfeld et al., 2013; Fleming et al., 2010; Lau & Rosenthal, 2011)) have been shown to index the capacity of confidence judgments to distinguish correct from incorrect decisions.

(2) A second question is what are the benefits and costs of typically reporting (and presumably feeling) within different ranges of the continuum of confidence. This of course relates to risk policies in decision making and becomes critical when the outcomes of a decision are associated with different values. Extreme overconfidence can lead to an underestimation of risk which can be harmful in the evaluation of future events (Lovallo & Kahneman, 2003); for instance, driving under the effect of alcohol when being overconfident of one's own driving abilities. On the other hand moderate optimism can promote exploration and leads (or is a cause of) better health (Scheier & Carver, 1987; Taylor & Brown, 1988). Our study shows that, contrary to what intuition may have suggested, the ability of confidence reports to distinguish correct from incorrect decisions is largely independent from the confidence bias. The fact that some subjects are responding with very high scores (and then seemingly saturating and loosing precision in a narrow range of confidence) leads to a re-scaling of confidence without losing precision.

(3) A third question which remains to be addressed in future studies involves the link between confidence bias, accuracy and over and under confidence. In behavioral economics, confidence studies have focused on individual tendencies to be over- or under-confident, and the conditions that promote these two modes of operation (Griffin & Tversky, 1992). Only recently, neuroscientist begun to investigate the neural correlates of confidence judgments (Kepecs, Uchida, Zariwala, & Mainen, 2008; Kiani & Shadlen, 2009; Komura, Nikkuni, Hirashima, Uetake, & Miyamoto, 2013; Middlebrooks & Sommer, 2012). The approach taken by the majority of neuroscience studies is to examine the structural and functional cerebral correlates that index confidence accuracy (Baird et al., 2013; Fleming et al., 2010; McCurdy et al., 2013). Focus on confidence accuracy has somehow shadowed one of the initial aims of confidence research, its calibration to an objective norm and its tendency to be shifted toward overall higher or lower values of confidence. Here we performed a step toward bridging this gap, addressing the relation within individuals and tasks, between confidence accuracy and bias.

(4) Few studies explored whether the way in which confidence is reported has an influence on the calibration and resolution of confidence (Overgaard & Sandberg, 2012). Klayman and colleagues reported that the tendency toward overconfidence is stronger when participants report confidence using subjective confidence intervals instead of probabilities (Klayman et al., 1999). Tunney and colleagues (Tunney & Shanks, 2003) (Tunney, 2005) reported that the resolution of confidence is higher when confidence is reported with a dichotomic scale than when using a continuous scale. In contrast, Dienes (2007) found very small differences in the resolution of confidence for six different confidence scales (with a slightly worst resolution when confidence was reported with numerical categories). Confidence judgments may also differ depending on the time at which they are solicited. For instance, higher overconfidence is observed if confidence is reported after each decision rather than after the test is completed (Gigerenzer et al., 1991). These studies are somehow orthogonal to what was investigated here, focusing on a single task and inquiring about different ways of conveying confidence. More studies are needed to clarify whether confidence bias is a truly stable individual property that will surface regardless of how confidence is measured.

Acknowledgments

This study was funded by CONICET and UBACYT. M.S. is sponsored by a scholar award of the James McDonnell Foundation.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cognition.2015.10.006.

References

Adams, J. K., & Adams, P. A. (1961). Realism of confidence judgments. *Psychological Review, 68*(1), 33.

Baird, B., Smallwood, J., Gorgolewski, K. J., & Margulies, D. S. (2013). Medial and lateral networks in anterior prefrontal cortex support metacognitive ability for memory and perception. *The Journal of Neuroscience, 33*(42), 16657–16665.

Barttfeld, P., Wicker, B., McAleer, P., Belin, P., Cojan, Y., Graziano, M., ... Sigman, M. (2013). Distinct patterns of functional brain connectivity correlate with objective performance and subjective beliefs. *Proceedings of the National Academy of Sciences, 110*(28), 11577–11582.

Dienes, Z. (2007). Subjective measures of unconscious knowledge. *Progress in Brain Research, 168*, 49–269.

Drevets, W. C., Price, J. L., Simpson, J. R., Todd, R. D., Reich, T., Vannier, M., & Raichle, M. E. (1997). Subgenual prefrontal cortex abnormalities in mood disorders. *Nature, 386*(6627), 824–827.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC Press.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review, 101*(3), 519–527.

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

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

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

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

Graziano, M., & Sigman, M. (2008). The dynamics of sensory buffers: Geometric, spatial, and experience-dependent shaping of iconic memory. *Journal of Vision*, 8 (5), 9 1–13.

Graziano, M., Parra, L. C., & Sigman, M. (2010). *Neurophysiology of perceived confidence*. Paper presented at the engineering in medicine and biology society (EMBC), 2010 Annual International Conference of the IEEE.

Graziano, M., & Sigman, M. (2009). The spatial and temporal construction of confidence in the visual scene. *PLoS ONE, 4*(3), e4909.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology, 24*(3), 411–435.

Hartigan, J. A., & Hartigan, P. M. (1985). The dip test of unimodality. *The Annals of Statistics, 13*(1), 70–84.

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., & Tibshirani, R. (2009). *The elements of statistical learning* (Vol. 2). Springer.

Juslin, P., Winman, A., & Olsson, H. (2000). Naive empiricism and dogmatism in confidence research: A critical examination of the hard–easy effect. *Psychological Review, 107*(2), 384.
Karelitz, T. M., & Budescu, D. V. (2004). You say" probable" and I say" likely": Improving interpersonal communication with verbal probability phrases. *Journal of Experimental Psychology: Applied, 10*(1), 25.
Kepecs, A., Uchida, N., Zariwala, H. A., & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature, 455*(7210), 227–231.
Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science, 324*(5928), 759.
Klayman, J., Soll, J. B., González-Vallejo, C., & Barlas, S. (1999). Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior and Human Decision Processes, 79*(3), 216–247.
Komura, Y., Nikkuni, A., Hirashima, N., Uetake, T., & Miyamoto, A. (2013). Responses of pulvinar neurons reflect a subject's confidence in visual categorization. *Nature Neuroscience, 16*(6), 749–755.
Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences, 15*(8), 365–373.
Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1977). *Calibration of probabilities: The state of the art*. Springer.
Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140*, 1–55.
Lovallo, D., & Kahneman, D. (2003). Delusions of success. How optimism undermines executives' decisions. *Harvard Business Review, 81*(7), 56–63. 117.
Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition, 21*(1), 422–430.
May, R. S., & Scholz, R. W. (1986). *Overconfidence as a result of incomplete and wrong knowledge*. New York, NY: Peter Lang Publishing.
McCurdy, L. Y., Maniscalco, B., Metcalfe, J., Liu, K. Y., de Lange, F. P., & Lau, H. (2013). Anatomical coupling between distinct metacognitive systems for memory and visual perception. *The Journal of Neuroscience, 33*(5), 1897–1906.
Middlebrooks, P. G., & Sommer, M. A. (2012). Neuronal correlates of metacognition in primate frontal cortex. *Neuron, 75*(3), 517–530.
Overgaard, M., & Sandberg, K. (2012). Kinds of access: Different methods for report reveal different kinds of metacognitive access. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*(1594), 1287–1296.
Perczek, R., Carver, C. S., Price, A. A., & Pozo-Kaderman, C. (2000). Coping, mood, and aspects of personality in Spanish translation and evidence of convergence with English versions. *Journal of Personality Assessment, 74*(1), 63–87.
Persaud, N., McLeod, P., & Cowey, A. (2007). Post-decision wagering objectively measures awareness. *Nature Neuroscience, 10*(2), 257–261.
Phillips, L. D., & Wright, C. (1977). *Cultural differences in viewing uncertainty and assessing probabilities*. Springer.
Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review, 117*(3), 864–901.
Plomin, R., Scheier, M. F., Bergeman, C. S., Pedersen, N. L., Nesselroade, J. R., & McClearn, G. E. (1992). Optimism, pessimism and mental health: A twin/adoption analysis. *Personality and Individual Differences, 13*(8), 921–930.
Scheier, M. F., & Carver, C. S. (1987). Dispositional optimism and physical well-being: The influence of generalized outcome expectancies on health. *Journal of Personality, 55*(2), 169–210.
Scheier, M. F., Carver, C. S., & Bridges, M. W. (1994). Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): A reevaluation of the Life Orientation Test. *Journal of Personality and Social Psychology, 67*(6), 1063–1078.
Song, C., Kanai, R., Fleming, S. M., Weil, R. S., Schwarzkopf, D. S., & Rees, G. (2011). Relating inter-individual differences in metacognitive performance on different perceptual tasks. *Consciousness and Cognition, 20*(4), 1787–1792.
Spearman, C. (1904). *The abilities of man*. New York.
Stankov, L., & Crawford, J. D. (1996). Confidence judgments in studies of individual differences. *Personality and Individual Differences, 21*(6), 971–986.
Swets, J. A., Tanner, J., Wilson, P., & Birdsall, T. G. (1961). *Psychological Review, 68*(5), 301–340 (September).
Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin, 103*(2), 193–210.
Tunney, R. J. (2005). Sources of confidence judgments in implicit cognition. *Psychonomic Bulletin & Review, 12*(2), 367–373.
Tunney, R. J., & Shanks, D. R. (2003). Subjective measures of awareness and implicit cognition. *Memory & Cognition, 31*(7), 1060–1071.
Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131.
Wallsten, T. S., & Budescu, D. V. (1995). A review of human linguistic probability processing: General principles and empirical evidence. *The Knowledge Engineering Review, 10*(01), 43–62.
Wallsten, T. S., Budescu, D. V., Rapoport, A., Zwick, R., & Forsyth, B. (1986). Measuring the vague meanings of probability terms. *Journal of Experimental Psychology: General, 115*(4), 348.
Watson, A. B., & Pelli, D. G. (1983). QUEST: A Bayesian adaptive psychometric method. *Attention, Perception and Psychophysics, 33*(2), 113–120.
Weinstein, N. D. (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology, 39*(5), 806.
Wright, G. N., Phillips, L. D., Whalley, P. C., Choo, G. T., Ng, K.-O., Tan, I., & Wisudha, A. (1978). Cultural differences in probabilistic thinking. *Journal of Cross-Cultural Psychology, 9*(3), 285–299.
Yu, S., Pleskac, T. J., & Zeigenfuse, M. D. (2015). Dynamics of postdecisional processing of confidence (vol 144, pg 489, 2015). *Journal of Experimental Psychology – General, 144*(3). 638-638.
Zylberberg, A., Barttfeld, P., & Sigman, M. (2012). The construction of confidence in a perceptual decision. *Frontiers in Integrative Neuroscience, 6*(79).
Zylberberg, A., Roelfsema, P. R., & Sigman, M. (2014). Variance misperception explains illusions of confidence in simple perceptual decisions. *Consciousness and Cognition, 27*, 246–253.