Title: Do I look like I'm sure?: Partial metacognitive access to the low-level aspects of one's own facial expressions

Running head: Partial metacognition to facial expressions

Authors: Anthony B Ciston*[a,b,c], Carina Forster*[a,b,c,d], Timothy R Brick[e,f], Simone Kühn[g,h], Julius Verrel[i,j], Elisa Filevich[a,b,c,i]

* These authors contributed equally to the work.

Affiliations:

[a] Department of Psychology, Humboldt Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany

[b] Bernstein Center for Computational Neuroscience Berlin, Philippstraße 13 Haus 6, 10115 Berlin, Germany

[c] Berlin School of Mind and Brain, Humboldt-Universität zu Berlin, Luisenstraße 56, 10115 Berlin, Germany

[d] Department of Neurology, Max Planck Institute for Human Cognitive and Brain Sciences, 04103 Leipzig, Germany

[e] Department of Human Development and Family Studies, Pennsylvania State University, 115 HHD Building, University Park, PA, 16802, USA

[f] Institute for Computational and Data Sciences, Pennsylvania State University, 224B Computer Building, University Park, PA, 16802, USA

[g] Lise Meitner Group for Environmental Neuroscience, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany

[h] University Clinic Hamburg-Eppendorf, Clinic and Polyclinic for Psychiatry and Psychotherapy, Martinistraße 52, 20246 Hamburg, Germany

[i] Center for Lifespan Psychology, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany

[j] Institute of Systems Motor Science, University of Lübeck, Ratzeburger Allee 160, 23562 Lübeck

Corresponding author: Elisa Filevich, elisa.filevich@gmail.com

Do I look like I'm sure?: Partial metacognitive access to the low-level aspects of one's own facial expressions

Abstract

As humans we communicate important information through fine nuances in our facial expressions, but because conscious motor representations are noisy, we might not be able to report these fine movements. Here we measured the precision of the explicit metacognitive information that young adults have about their own facial expressions. Participants imitated pictures of themselves making facial expressions and triggered a camera to take a picture of them while doing so. They then rated how well they thought they imitated each expression. We defined metacognitive access to facial expressions as the relationship between objective performance (how well the two pictures matched) and subjective performance ratings. As a group, participants' metacognitive confidence ratings were only about four times less precise than their own similarity ratings. In turn, machine learning analyses revealed that participants' performance ratings were based on idiosyncratic subsets of features. We conclude that metacognitive access to one's own facial expressions is only partial.

Keywords: Metacognition, Facial expressions, Confidence

Introduction

Precise motor planning and execution can occur without the brain having explicit, conscious access to the exact position of our limbs, or the exact degree of contraction of our muscles (Kal et al., 2018; Kleynen et al., 2014; Taylor & Ivry, 2013). For instance, we can simultaneously walk, speak, and gesticulate successfully while concentrating on an argument and not on the movements that enable it, and we are furthermore unable to accurately report the state of each of our muscles. Although explicit access to proprioceptive signals in highly routine tasks like walking or talking may be unnecessary, it might be beneficial in some other cases. For example, it has been suggested (MacIntyre et al., 2014) that metacognitive reasoning plays a central role in developing and improving motor expertise: if an experienced actor has a detailed and sophisticated representation of an ideal facial expression to communicate emotion, they are better able to detect and correct deviations from the ideal, leading in turn to more accurate and consistent performance.

Proprioceptive information about our limbs and their movements is thought to originate primarily from muscle spindles, together with skin receptors, Golgi tendon organs, and joint receptors (Proske & Gandevia, 2012; Sherrington, 1906; Tuthill & Azim, 2018). Artificial vibration of the muscles can lead to activation of the muscle spindles, showing that their activation is sufficient to alter the representation of the body and its position (Goodwin et al., 1972; Lackner, 1988). In addition, position estimates have been found to be more precise following active vs. passive movements, suggesting that efferent motor commands may either affect or inform proprioceptive representations (Craske & Crawshaw, 1975; Fuentes & Bastian, 2010; Gritsenko et al., 2007). Finally, proprioceptive information is combined with visual information, when available, to form a multisensory and integrated representation (Limanowski & Blankenburg, 2016; Ruttle et al., 2018; Sober & Sabes, 2005; van Beers et al., 2002).

Facial expressions present a particularly important yet poorly studied instance of motor control. On the one hand, we communicate a great deal of information with small, nuanced facial movements – on the order of 10 mm or less (Clark Weeden et al., 2001; Coulson et al., 2000). On the other hand, we hardly ever see ourselves while making them. Perhaps apart from actors or public speakers who practice in front of a mirror (or the increased number of video-calls during the 2020 SARS-CoV-2 pandemic), we do not usually have online visual feedback about our facial muscles. Visual feedback information has been shown to be necessary to guide learning in the

imitation of one's own expressions (Cook, Johnston, Heyes 2013), suggesting that facial movements might be poorly represented if they are based on motor information alone. Together, the combination of the high social relevance of small movements in our facial muscles and the general lack of visual information about them raise the interesting question: How accurate and precise is our knowledge about the way we look when we make facial expressions?

Previous studies have focused on related questions. One line of research has quantified metacognitive access to *others'* facial expressions and operationalized metacognitive performance as the precision of participants' representations of uncertainty (Bègue et al., 2019; Chen et al., 2019; Lapate et al., 2020). While our ability to accurately represent both the facial expressions of others and our certainty about them is clearly critical for social interactions, it is equally important to correctly represent and adequately control *one's own* expressions (Shea et al., 2014). In line with this notion, previous research measured how accurately we can report the emotions we feel and express (Gross & John, 1997; Rosenberg & Ekman, 1994; Wagner et al., 2003). While reports of experienced emotions generally match observers' ratings of facial expressions, human volunteers tend to overestimate their expressivity (Barr & Kleck, 1995; Gilovich et al., 1998; Qu et al., 2017). This work has focussed on the emotional content of facial expressions, but it remains possible that our access to the low-level details of our faces and facial expressions is poor. Throughout this work, we will make this distinction between the high-level aspects of a facial expression (namely, the emotion that they communicate) and low-level aspects (the specific shape and constellation of facial features that make up that expression). This distinction can be tied to the motor control literature where, for example, the concept of motor abundance describes the phenomenon that virtually any motor goal can be reached in a multiplicity of ways (Latash, 2012), which need not be consciously controlled. Akin to that concept, we note that any given emotional content could be communicated in a multiplicity of ways. Therefore, we may know *what* we communicate, but not *how* we communicate it. Indeed, participants systematically overestimate the width, but not the length, of their faces (Fuentes, Runa, et al., 2013; Longo & Holmes, 2020; Mora et al., 2018), mimicking what has been described for whole bodies (Fuentes, Longo, et al., 2013) and hands (Longo & Haggard, 2010). More recently, large inter-individual differences have been described in how accurately healthy young adults can represent their own faces (Maister et al., 2020). These previous studies investigated relaxed faces with neutral expressions and captured, in essence, individuals' ability to accurately describe their face, or to discriminate it from the face of another. Importantly, static features of one's face are irrelevant to social interactions, which instead are based on dynamic information. Here, we focussed instead on metacognitive access to how one's face varies when making

different expressions. In a pre-registered experiment, we asked participants to imitate expressions shown in pictures of themselves and to rate their confidence in their own performance. In other words, participants provided a subjective rating about how well they thought they had imitated each expression. We then measured participants' metacognitive access to their own facial expressions as the correspondence between subjective ratings and an objective measure of performance. We calculated the Euclidean distance between landmarks placed automatically on each of the faces in a pair and used this distance as an objective measure of (inverse) performance. We predicted that, if participants had precise metacognitive access to the details of their own facial expressions, we would observe a negative relationship between confidence and distance between two faces. This would imply that participants (correctly) provided low confidence on trials where the two images differed the most. If, on the other hand, participants had no metacognitive access to these low-level details, we expected to find no relationship between subjective and objective measures. The magnitude of the slope parameter between confidence and distance effectively quantifies the precision of confidence judgments.

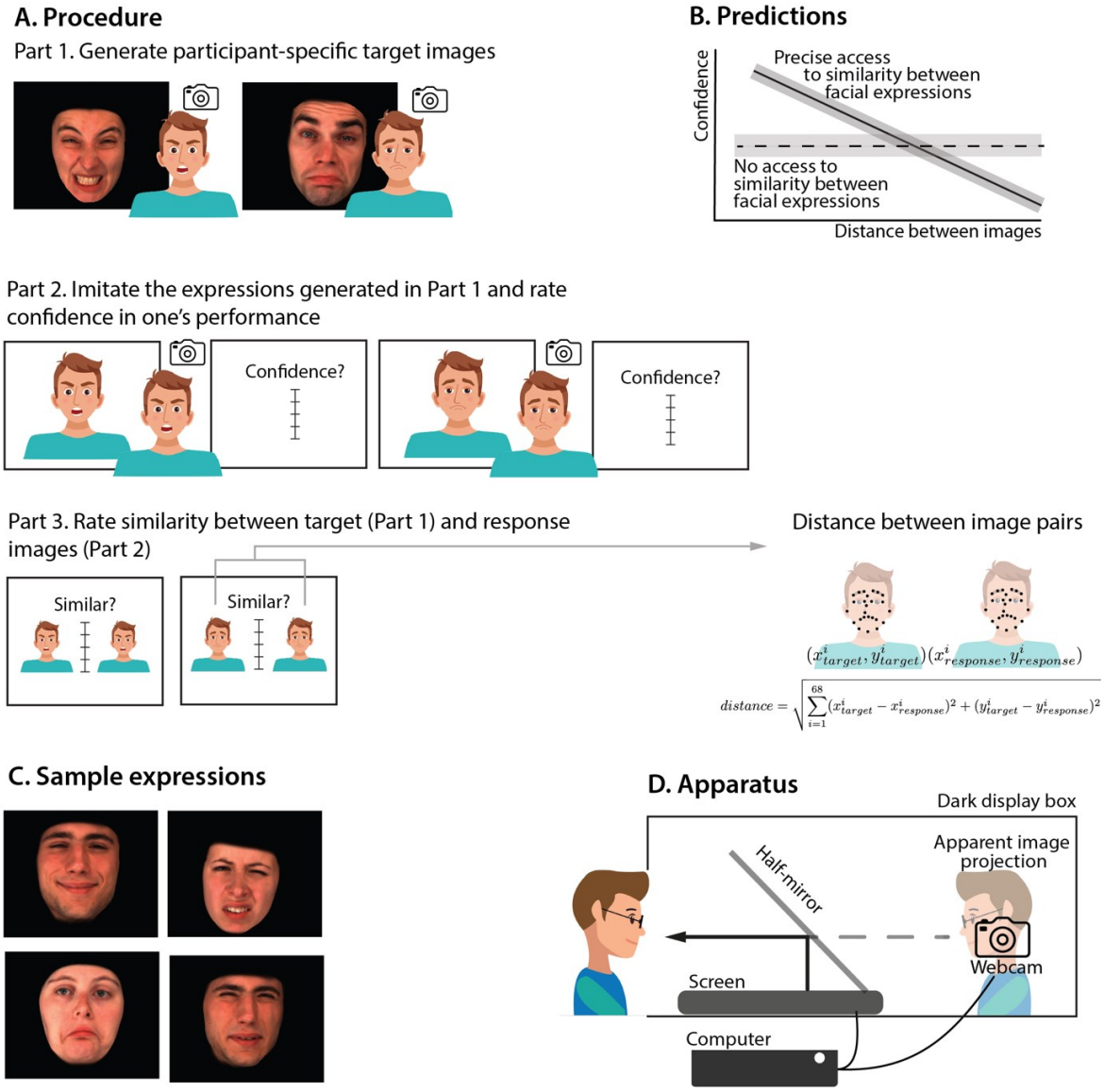

Figure 1: Experimental Design. (A.) Procedure. Cue stimuli were pictures of facial expressions taken from the MPI Small Facial Expression Database (Cunningham et al., 2005). They were performed by actors and represented non-stereotypical expressions (e.g., "You lose the way in a foreign city", see Methods for further details). Participants used these images as cues to produce 32 participant-specific target images. In part 2, each of the 32 target images (of the participants' faces displaying the expression generated in part 1) was shown eight times (256 trials total). Participants reproduced their own expressions shown in the target pictures, pressed a key while holding their expression, and subsequently rated confidence in their own performance. The experiment was self-paced. Squares around the pictures indicate that they were displayed to participants, whereas pictures without a square frame around them represent pictures collected but not shown back to participants. (Expression drawing: Freepik.com) **(B.) Predictions.** The correlation between the two variables indicates the precision of the metacognitive representation. Confidence ratings were expected to be negatively correlated with the distance between two images if participants have metacognitive access to the low-level aspects of their facial expressions (solid line). Confidence ratings were not expected to vary with distance if participants had no metacognitive access to their own facial

expressions (dashed line). **(C.) Sample cue stimuli.** Note that the cue expressions were not unnatural, but hard to label as one of the basic emotions. **(D.) Apparatus.** Participants sat in front of a dark display box and saw the pictures projected from a computer screen reflected on a half-plated mirror (tilted 45°). Behind the mirror, positioned directly in front of participants' gaze, a digital camera took pictures of the participants when they pressed the corresponding key. This way, participants could look simultaneously directly at the to-be-imitated picture and into the camera.

Material and Methods

Participants

Following our pre-registered plan (https://osf.io/pnyw3), 40 healthy participants took part in the study after giving informed consent (21 female, 19 male mean ± SD: 28.2 ± 4.6 years). We based the sample size on pilot data from 12 participants (see SI) and previous studies of motor metacognition from our group. Exclusion criteria were a recent history of psychiatric disease or having a heavy beard, as we reasoned that it would occlude the view of part of the face and placing of the landmarks. The local ethics committee approved all procedures (Nr. 2017-23-R), which conformed to the Declaration of Helsinki.

Apparatus

The experimental setup consisted of a stimulus computer, a digital camera, a screen, and a half-silvered mirror tilted 45° from the vertical (Figure 1.D). Participants saw the image displayed on the screen by the stimulus computer indirectly through its reflection on the half-silvered mirror. Behind the mirror, a digital camera (Fire-i, UniBrain, Athens, Greece) connected to the computer took pictures of the participants' facial expressions. This setup allowed participants to look at the pictures displayed while simultaneously looking directly into the camera. As a result, we obtained pictures of participants looking straight ahead and not downwards at the image, as would have been the case if we had used e.g. a simple laptop computer with a digital camera just above the screen.

Participants sat at approximately 60 cm from the middle-point of the half mirror, which was in turn 45 cm away from the display screen. To reduce head movements, we held participants' torsos loosely in place with an elastic band tied to the chair. Additionally, at the beginning of the experiment, we showed participants the image collected by the camera in real time and asked them not to make large head movements or rotations. While it would have been desirable to further limit whole-head movements using, e.g., a chin rest, we opted against this as it would have made expressions unnatural and, more importantly, because it would have provided a form of

sensory feedback, interfering with the experimental design. We ensured that participants' faces were well-lit and took care that participants did not see any reflections of their own face on the mirror.

Procedure

All experimental tasks were written on MATLAB (R2016b, The Mathworks, Natick, MA), using Psychtoolbox-3 (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997) and ran on MacOS. All tasks were self-paced with no time deadlines. All participants (except for one, due to technical problems) completed all tasks in the same order.

Facial Expressions Task

The facial expressions task consisted of three parts. In the first part (Figure 1.A), participants saw 32 different pictures of four different actors in pseudorandomized order (see the description of *Cue images*, below) and imitated each expression as best they could. Participants pressed a key (the space bar) once they considered that their expression was as close as possible to the actor's expression. We asked participants to try to match the low-level physical features of the face — the curvature of the lips, the elevation of the eyebrows — rather than the emotion conveyed by the expression. Upon pressing the spacebar, the digital camera behind the half-plated mirror took a picture of the participant's facial expression, and a new trial started. On a separate test, we had determined that there was a minimum delay of approximately 80 ms between the time of key press and the time stamp of the image. Accordingly, we included in our instructions to participants to hold the expression in place after they had pressed the key that would trigger the image acquisition.

The 32 pictures of participants generated in this way served as target images for the second part of the paradigm. Here, participants saw the target images and tried to reproduce their own expressions. Once again, we emphasized that the goal was to match the low-level physical features of the face rather than the emotion conveyed. After each trial, participants used a mouse to rate their confidence (on a visual analogue scale) regarding how well they thought that they had imitated their own previous expression. Participants saw each of their 32 target expressions repeated 8 times in random order (256 trials in total). We only revealed that they would have to reproduce their own expressions after the first part of the experiment was complete. Parts 1 and 2 of the experiment took on average approximately 50 minutes. Before starting part 1, participants completed four practice trials where they simply imitated pictures of famous celebrities and took pictures. They did not see the resulting pictures of themselves.

In the third part of the task, participants saw each of the 256 pairs of pictures (target and response) and rated them for similarity on a scale exactly like the one they had used for confidence. This part of the experiment took on average 30 minutes.

Cue images

We used 32 different facial expressions as cue pictures (14 from two different male actors, 18 from three different female actors) which would be used to generate participant-specific target expressions. To prevent participants from producing stereotypical target expressions, we sought pictures representing expressions that could not be unambiguously categorized as one of the basic emotions (Ekman, 1999). We selected pictures from the MPI Small Facial Expression Database (Cunningham et al., 2005), which includes video sequences of expressions based on a method acting protocol in which actors produce non-standard expressions by imagining themselves in a situation described by a brief scenario and reacting accordingly. Example descriptions of expressions include: “Somebody suggests to try something. You hesitate at first, then you agree”, or “You have reached a goal and you are happy to have accomplished it”. Additionally, we selected still images from the video sequence that did not correspond to the peak expression, but instead to an intermediate step. We assume that, as a result, the cue images could not easily be labelled as stereotypical expressions (e.g., “happy”, “sad”) for which participants might have a predefined motor program but are instead the result of an unusual and idiosyncratic combination of gestures. Note that, as the samples in Figure 1.C show, these cue images were not unnatural grimaces and so the paradigm remains ecologically valid. We reasoned that these non-canonical expressions would maximize motor variability, ensuring that confidence ratings could be based only on a true evaluation of trial-by-trial performance and not on a general knowledge of how reproducible a given expression was.

Visual Task

Each participant completed 200 trials of a visual metacognition task (https://github.com/metacoglab/meta_dots). On each trial of this task, two circles enclosing sets of dots appeared for 200 ms on either side of a central fixation cross (each circle with a radius of 5 degrees of visual angle, located along the middle of the screen, with an eccentricity from the vertical midline of 5.5 degrees of visual angle). One of the two circles always contained 50 dots while the other varied in dot number, and the position (left/right) of the circles was randomized on each trial. In a 2-alternative forced-choice (2AFC) task, participants discriminated which of the circles contained more dots by pressing the left or right arrow keys on the keyboard. The

difference in the number of dots was determined by a pair of interleaved 2-down-1-up adaptive staircases aimed at fixing performance at around 71% accuracy. After each response, participants reported their confidence in the accuracy of their own response using the same vertical visual analogue scale that they had used for the two previous tasks rating confidence and similarity for facial expressions.

Before the main visual task, we ran 80 trials of a staircase procedure where participants did only the discrimination task without rating confidence. Here we also included two interleaved 2-down-1-up staircases starting from a difference of 3 and 20 dots respectively. One participant (unintentionally) received feedback about the accuracy of the discrimination task while rating confidence, so we excluded their data from the analysis. The visual task took approximately 20 minutes. Over all participants, we also excluded 2% of the trials where the reaction times to either the discrimination task or the confidence rating were faster than 300 ms or slower than 5 s. We estimated metacognitive efficiency as M-ratio after scaling and binning confidence into four discrete confidence levels based on uniform intervals.

Toronto Alexithymia Scale

At the end of the experiment we collected responses to a computerized version of the Toronto Alexithymia Scale, TAS (Bagby et al., 1994) running on a browser, and the data were stored locally(Lange et al., 2015) (jatos.org). Most participants completed a German version of the scale, except for seven non-German speakers who completed an English version instead. The TAS-20 consists of 20 items that can each be answered on a 5-point Likert scale. We considered three out of the four subscales (Difficulty identifying feelings, Difficulty describing feelings, and Externally-oriented thinking, but excluded the Daydreaming subscale). We calculated Bayes Factors (BF_{10}) for correlations between these covariates and individual slopes from the estimated models using the *BayesFactor* package in R (version 3.6.2).

Data processing and analysis

Following the pre-registered plan, we excluded trials from the facial expressions task at the single participant level if RTs (time between image onset and key press) were above the 95 percentile for that participant. This cut-off was necessary because we noticed that participants sometimes laughed at their own picture or got otherwise distracted. This resulted in seven trials excluded from the entire dataset where the time to take a picture was below 300 ms, and a mean lower threshold of exclusion of 9.43 s (range: 4.0 - 18.0 s).

For each of the pictures taken, we obtained the (x,y) coordinates of landmarks distributed on the face. At pre-registration we planned to estimate the landmark positions using two different toolboxes and choose the best one to estimate distance based on the quality of the relationship to the similarity ratings. Instead, due to technical problems in running one of the toolboxes we opted for the *face-alignment* package (Bulat & Tzimiropoulos, 2017) alone (https://github.com/1adrianb/face-alignment v.1.0.0) together with *scikit-image and pytorch* to extract the landmarks from the faces, running on Python v3 in a Jupyter notebook v5. The face-alignment package automatically places 68 landmarks on the face and excludes the forehead and hairline.

Using MATLAB (R2020a), we computed the distance (in coordinate space) between each pair of target and response images. Using the (x,y) coordinates for all landmarks, we ran a Procrustes rigid alignment of each face in a pair to a standardized set of coordinates. Rather than including all landmarks for the Procrustes alignment, we used three reference points that vary minimally across facial expressions: The outer corners of each eye and a point just below the nose. The transformation allowed for translation, orthogonal rotation, and scaling. Thus, these linear transformations minimized the variance in the distance data that could be accounted for by head rotations and general enlargement or shrinkage due to change in the face position, while also preserving variance resulting from the differences between the expressions. It did not account for other rotations (yaw and pitch), where the relative distance between some face components can change without the facial expression being different. After rigid transformation, we calculated the total distance for each pair of target and response images as the Euclidean distance (the root of the sum of squares, see equation in Figure 1) over all 68 landmarks between the two images. We refer to this measure simply as the distance between two images. Finally, we log-transformed the obtained distances to ensure that the data were normally distributed before fitting the Bayesian mixed models.

Bayesian mixed models

We analysed the data using Bayesian mixed models created in Stan (http://mc-stan.org/) through the *brms* package (Bürkner, 2017, 2018). In all cases, we ran 4 chains with 15,000 iterations, 5,000 burn-in samples each, and no thinning. We checked for convergence by visually examining the MCMC chains and ensured that the scale reduction factor (Rhat) of all models was equal or

close to 1. We considered that ratings might vary across participants both in their mean and in their relationship to the landmark distance, and that different facial expressions might vary in their associated difficulty to both reproduce (leading to greater variability in the landmark distance) and to rate (leading to differences in the ratings). Thus, in all models and unless otherwise stated, we included random slopes for both participants and facial expressions (see the explicit model syntax in Table 1). We extracted the participant-wise random slopes using the *mixedup* package (https://m-clark.github.io/mixedup/).

We followed recommendations (Dienes, 2019) to use heuristics to define prior distributions. We built the prior for the slope between ratings and distance based on the ratio-of-scales heuristic: we found that the range of (log-transformed) distances was approximately 3 a.u. (arbitrary units), whereas the range of confidence ratings is 1 point (minimum: 0). Therefore we used a normal prior centered on 0 with an SD = ⅓ (which corresponds to the ratio between confidence range and distance range) for the slope parameter. To find a prior for the model intercept we followed the logic behind the room-to-move heuristic. Note that raw distances ranged between [131.36 - 2493.78] a.u., hence the expected rating at 0 distance (i.e., perfect performance) can be well approximated by the expected rating at distance = 1, which corresponds to the intercept in a linear model with log-transformed distances. We reasoned that a participant with maximum metacognitive performance would consistently rate their confidence as 1, when the distance between the two images was 0. Because we realistically expect participants to have (at most) less than perfect metacognitive access to their own expressions, we centered the prior at 0.8 with an SD = 0.5. Following a similar logic, we set the prior slope between the two ratings to be centered at 0 with SD = 1, and an intercept of 0 with an SD = ½. For all models, we report the estimate, its associated error mean, the 95% credibility interval (CI), and the BF_{10}, estimated using the *bayestestR* package (Makowski et al., 2020), to compare each model against its null counterpart, containing the same random effects structure but not the fixed effect of interest. We also examined the posterior draws for each participant in relation to the region of practical equivalence (ROPE). We set the ROPE to a default range from -0.1 to 0.1 of a standardized parameter, which corresponds to a negligible effect size (Cohen, 1988; Kruschke & Liddell, 2018). Finally, we estimated R^2 values as implemented by the *brms* package (Gelman et al., 2019).

Table 1: Formulas for the Bayesian mixed models employed

Hypothesis	Model Formula	Corresponding Figures

Participants' confidence in their own performance is inversely related to the distance between two images	confidence ~ logDistance + (1 + logDistance | participantID) + (1 | expressionID)	Figure 2 Figure S6
Participants' similarity ratings are inversely related to the distance between two images	similarity ~ logDistance + (1+ logDistance | participantID) + (1 | expressionID)	Figure 3 Figure S8
Confidence and similarity ratings of the same participant are related	confidence ~ similarity + (1 | participantID) + (1 | expressionID)	Figure 4
Confidence and reaction times are negatively related	confidence ~ RT + (RT | participantID) + (1 | expressionID)	-
Confidence and ML-weighted distances are related	confidence ~ MLweightDist + (1 + MlweightDist | participantID) + (1 | expressionID)	-

We computed metacognitive access to faces using linear regression and estimated the correlation with visual Mratios, deviating from the pre-registered plan. We initially planned to also calculate the area under a type-2 ROC curve (AUROC2) by arbitrarily assuming that first-order performance on the Faces task was at 70% accuracy and by classifying trials with distances above the corresponding threshold as "incorrect". This analysis had the advantage that it would have allowed us to correlate metacognitive performance measured on the same scale for both tasks (Faces and Visual), but we later reasoned that it would make the results less easily interpretable while not adding explanatory power and therefore decided to omit it.

Principal component regression

We used machine learning tools (implemented in Python v3 and *scikit-learn*) to build linear regression models in order to identify predictors of subjective confidence in the landmark information. We analysed each participant separately. We first determined the distance (a 2-component vector) for each landmark distance as the (x,y) coordinate differences between the two images and further decomposed each of the 68 distances into four zero- or positive scalar features (one for each cardinal direction, for a total of 272). This allowed different directions of

movement to be weighted differently by the model. We normalized each feature by dividing it by its median. Then, we applied dimensionality reduction using principal component analysis with a set number of principal components (66, or approximately 90% of the variance from all subjects) in order to avoid multicollinearity among the features. Finally, a least squares linear regression model was trained for each participant using trial-wise leave-one-out cross-validation. The models aimed to predict subjective confidence ratings (or similarity ratings, see Supplementary analyses) using as predictors the values of the principal components derived from the collection of scalar differences for all landmarks together.

The resulting model weights referred to features in principal component space. We translated the model weights back into landmark space (i.e., x,y coordinates of the facial landmarks). To do so, we approximated the weight w of each feature $_f$ using the expression in (1):

$$w_f = \sum_{c=1}^{66} \lambda_{f,c} \times \omega_c \quad (1)$$

Where $\lambda_{f,c}$ is the loading of feature $_f$ on principal component c, and ω_c is the linear regression model's weighting of principal component c.

To reconstruct the distances weighted by the results of each linear regression model, we used expression (2):

$$RSSQ_{weighted} = \sqrt{\sum_{f=1}^{272} w_f \times f^2} \quad (2)$$

Where w_f denotes the weights for each feature *f*, which is in turn the difference between response and target images for each cardinal direction, for a given landmark, if the difference was positive, and 0 otherwise. Note that unlike the case for the Euclidean distance, where distances were forced to be positive and each of them had an effective weight of 1, here we allowed the feature weights to be signed. For those cases where the term under the square root was negative, we calculated the root of the absolute value and then reversed the sign. Note that $RSSQ_{weighted}$ is now better interpreted as a measure of performance, and not distance: because the ML-derived weights already account for the negative relationship between distance and confidence, $RSSQ_{weighted}$ is expected to show a positive relationship to confidence.

We obtained adjusted R2 for each (participant-specific) model values and compared them using a Bayesian Wilcoxon Signed-Rank test (Doorn et al., 2020) as implemented in JASP (JASP Team, 2020) v0.14 with 10,000 MCMC samples and 5 chains, and a default Cauchy prior.

Results

Confirmatory Analyses

The distance between any pair of images is an inverse measure of performance in the task, as greater distance corresponds to a poorer match between target and response expressions. Thus, we reasoned that participants with precise metacognitive access to their facial expressions would have a sharp relationship between the distance between two images and the confidence ratings. The estimated regression coefficients from a multilevel model of these data should be negative and clearly different from zero. On the other hand, if a participant had no access to their own performance, their judgments would bear no relationship to the distance between two images, and the regression coefficients would be indistinguishable from zero (Figure 1B, Predictions).

To arbitrate between these two possibilities, we first quantified our participants' metacognitive access to their own facial expressions using a Bayesian linear mixed-effects regression model of participants' confidence ratings. The model included the log-transformed distances as a fixed effect (for all 68 landmarks combined), as well as random intercepts for participant and facial expression. The random intercepts capture metacognitive bias, or each participant's tendency to rate high confidence, whereas the estimated slope of the model captures a measure akin to metacognitive sensitivity, or the relationship between confidence and performance at the group level (Fleming & Lau, 2014) — note however that these two elements may not be independent (Rausch & Zehetleitner, 2017). We found that participants' confidence ratings had a small negative relationship to the distance measured (Figure 2.A, $M = -0.03 \pm 0.01$, $CI = [-0.05, -0.01]$, $R^2 = 0.21$, see also supplementary Figure S1 for the participant-wise data). However, when compared to the null model without the effect of distance, we found only anecdotal evidence (Jeffreys, 1998) for the relationship between the two ($BF_{10} = 2.20$). Further, a robustness check revealed that, as expected given the proximity of the posterior samples to the region of practical equivalence (ROPE, defined following the default criterion of the region corresponding to a Cohen's d of 0.1, Figure 2.B), the choice of the SD of the prior distribution had a strong effect on the BF_{10}: Widening the prior distribution from 0.4 to 0.7 led to a $BF_{10} = 1.02$, and greater SDs also strongly reduced the value of the BF_{10}. Together, these results point to no evidence for a relationship between confidence and distance. For illustration purposes, we plot the participant-wise posterior draws, in relationship to the ROPE (Figure 2.C).

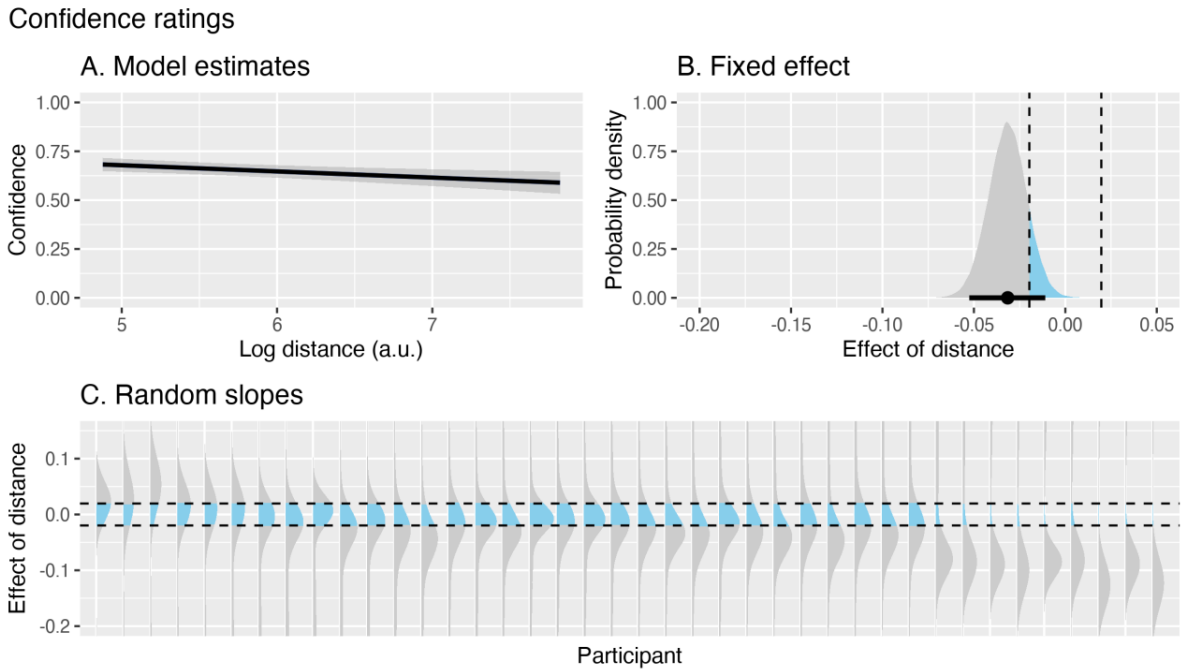

Figure 2. Metacognitive access to facial expressions (A.) Group effects reflecting mean metacognitive access, namely the relationship between confidence ratings and distance between two images (inverse of performance). A small but consistently negative slope suggests that participants had minimal metacognitive access to their own expressions. The solid line represents the mean of the posterior draws, the shaded region represents the 95% credibility interval. **(B.)** Posterior draws for the group-level fixed effect of distance, shown in relation to the ROPE, marked with dashed lines. The black horizontal line indicates the mean and 95% HDI. **(C.)** Posterior draws for each participant, shown in relationship to the ROPE. Note that the y-axis is clipped to better display the distributions around the ROPE and therefore excludes the long tails of some of the distributions. Participants are ordered following the mean slope estimate and might not be aligned across figures.

Then, to quantify the relationship between distance and similarity, we built a regression model of participants' similarity ratings including, as before, the log-transformed landmark distances as a fixed effect (for all 68 landmarks combined), as well as random intercepts for participant and facial expression. Here, similarity ratings did track the distance (Figure 3 and supplementary Figure S2). We found a clear and, as expected, negative relationship between the two ($M = -0.12 \pm 0.01$, $CI = [-0.14, -0.09]$, $BF_{10} = 8.01 \times 10^{8}$, $R^2 = 0.26$). This shows that the distance we measured carried information relevant for similarity ratings and thus the null effect above cannot be simply due to a poor measure of distance. Relatedly, it shows that participants were not consistently poor at reproducing their own facial expressions but that, instead, there was enough variability in their performance that participants recognized in their similarity ratings. Additionally, because the same participants rated both confidence and similarity, the differences between the two ratings cannot

be attributed to trivial effects such as a poor understanding of the confidence scale or task instructions, or simple lack of motivation.

An advantage of similarity as compared to confidence ratings is almost trivial, as participants could see the picture pairs side-by-side to rate similarity, but not confidence. Hence, we interpret this result as suggesting that the landmark distances were indeed related to similarity, but make no statistical comparisons between the two kinds of ratings. Nevertheless, we can use the slope of the relationship between similarity and distance as a plausible maximum for the relationship between confidence and distance. Then, the ratio between mean slope estimates is -0.03 / - 0.12 = 0.25. That is, we found the relationship between confidence and distance to be approximately four times noisier than that between similarity and distance.

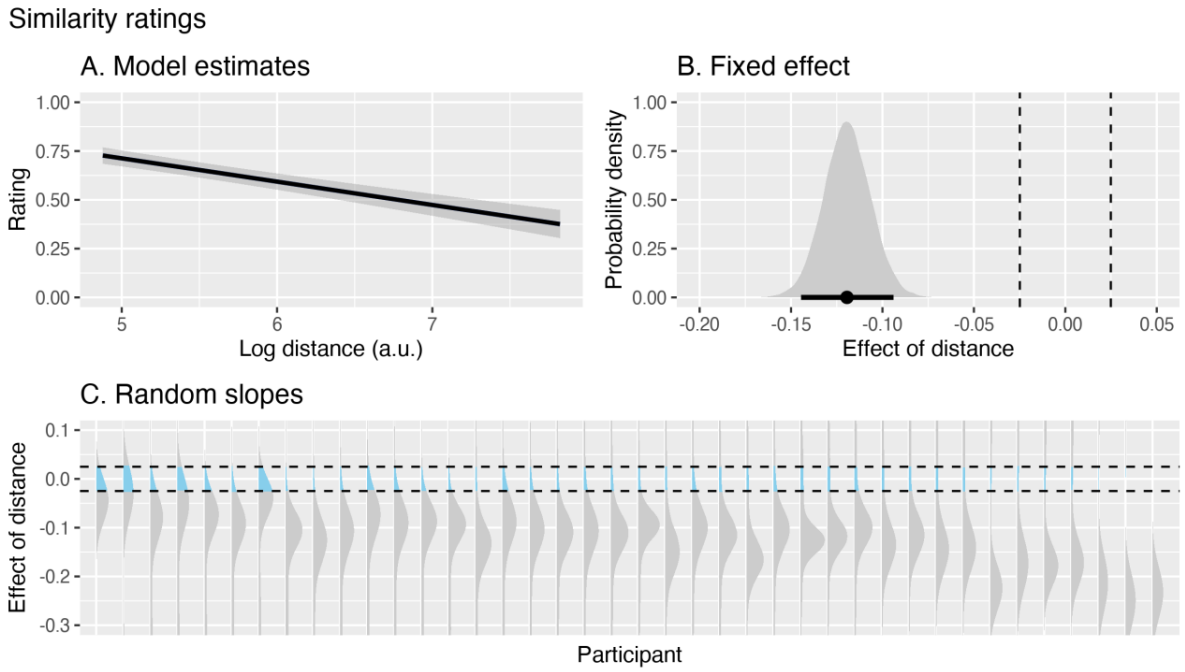

Figure 3. The distance between two images captures relevant information. (A.) Group effects reflecting the information contained in the distance between two images, namely the relationship between the similarity ratings provided by participants (when viewing each image pair side-by-side) and distance between two images. The solid line represents the mean of the posterior draws, and the shaded region represents the 95% credibility interval. **(B.)** Posterior draws for the group-level fixed effect of distance, shown in relation to the ROPE, marked with dashed lines. The black horizontal line indicates the mean and 95% HDI. **(C.)** Posterior draws for each participant, shown in relation to the ROPE. Note that the y-axis is clipped to better display the distributions around the ROPE and therefore excludes the long tails of some of the distributions. Participants are ordered following the mean slope estimate and might not be aligned across figures.

Finally, following our pre-registered plan, we explored relationships between the participant-wise random slopes with Mratio, a measure of visual metacognitive efficiency (Maniscalco & Lau, 2012) in a visual task. We found that visual Mratio was consistently above the chance level of 0 (M= 0.75, SD = 0.57, t(38) = 8.15, p < 0.001, BF_{10} = 1.54x10^7, estimated with a default Cauchy prior) but that it did not correlate with participant-wise effects of distance on confidence (Figure 4.A, r = -0.19, p = 0.25, BF_{10} = 0.64, with a default shifted beta prior distribution). While the two measures of metacognitive access are not strictly comparable (the visual Mratio is controlled for first-order performance but the individual effects of distance on confidence are not), this analysis shows that partial metacognitive access to facial expressions cannot be attributed to generally low domain-general metacognitive insight (Rouault et al., 2018).

Using Pearson correlations, we also measured potential associations between the inter-individual differences in metacognitive access to facial expressions and Alexithymia scores, as an indication of each participant's ability to identify and describe their own feelings. We found no conclusive evidence for or against any relationships between alexithymia score and the participant-wise effect of distance on confidence (*r* = -0.202, *p* = 0.217, BF_{10} = 0.70, Figure 4.B) or on similarity ratings (*r* = -0.108, *p* = 0.513, BF_{10} = 0.43). We also found no association between alexithymia score and visual metacognitive efficiency (*r* = 0.07, *p* = 0.67, BF_{10} = 0.38).

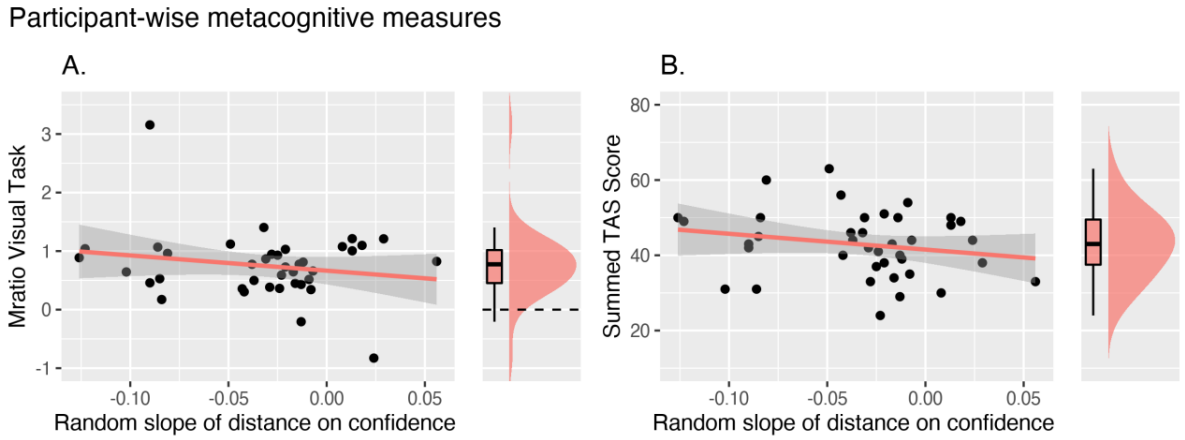

Figure 4: Correlations between participant-wise estimates of metacognitive access to facial expressions and other measures of insight. Each dot corresponds to one participant's performance estimate, and the box- and density plots on the right represent the marginal distribution of the corresponding variable on the y axis. **A. Metacognitive efficiency (Mratio) in a visual task.** Participants' metacognitive

efficiency was significantly better than chance performance (marked with the horizontal dashed line). **B. Alexithymia score (TAS).** We found no evidence for a correlation between metacognitive estimates and these measures of insight.

Exploratory Analyses

For completeness, we studied the relationship between similarity and confidence ratings. We built a Bayesian linear regression model of participants' confidence ratings, this time including the similarity ratings as a fixed effect and random intercepts for participant and facial expression. We found a clear positive relationship between the two ratings (M = 0.10 ± 0.01, CI = [0.09, 0.12], BF_{10} = 6.36 x 10^{31}, R^2 = 0.21, Figure 5 and supplementary Figure S6). This suggests that participants' confidence ratings were not random or noisy but rather that they simply did not reflect the low-level features captured by the distance. Thus, similarity ratings complement the results shown on Figure 3. The trial-wise relationship between confidence and similarity ratings suggests that both ratings corresponded to a valid aspect of the facial expressions. Just only partially to the low-level aspects of facial expressions (as the analysis corresponding to Figure 2 shows).

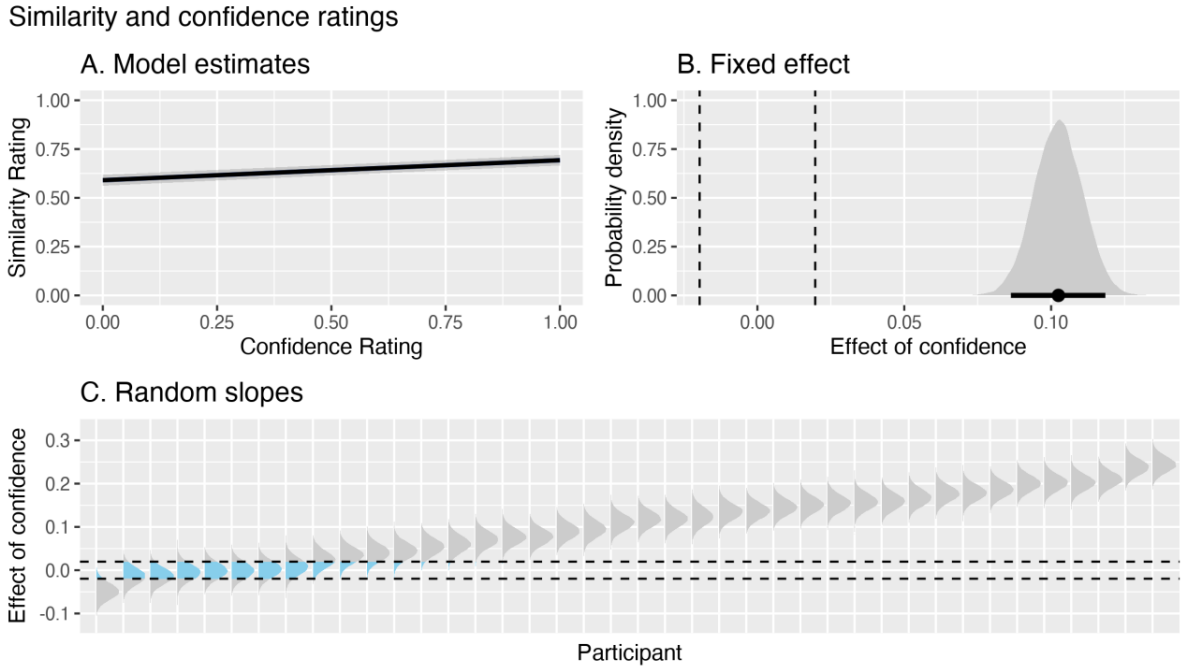

Figure 5: Similarity ratings vary with confidence ratings. (A.) Group effects showing the relationship between the two ratings on image pairs provided by participants (similarity vs. confidence). The solid line represents the mean of the posterior draws, and the shaded region represents the 95% credibility interval.

(B.) Posterior draws for the group-level fixed effect of confidence on similarity, shown in relation to the ROPE, marked with dashed lines. The black horizontal line indicates the mean and 95% HDI. **(C.)** Posterior draws for each participant, shown in relation to the ROPE. Participants are ordered following the mean slope estimate and might not be aligned across figures.

Our results so far suggest that participants' confidence ratings only partially reflected performance, calculated as the Euclidean distance over all landmarks. In a final set of exploratory analyses, we therefore aimed at identifying which pieces of information participants may have informed confidence ratings.

The Euclidean distance between image pairs assigns equal weights to the distances of all facial landmarks and is therefore a relatively naive measure of the difference between expressions, in that it does not allow for potential differences between landmarks in their contribution to different individuals' confidence. However, it is in principle possible that participants attended to different parts of their faces to different degrees and, further, that this differential attention was not consistent across participants. For example, one participant may have focused almost exclusively on how well their mouth matched the target image to rate their confidence, and another participant may have focused exclusively on the eyes and ignored the mouth. While this was against the task instructions, it remains a possibility that would undermine the strong claim that most participants did not base their confidence ratings on the landmark distances. To obtain a more fine-grained and flexible measure of performance we used machine learning (ML) algorithms to build a simple linear regression model to predict each participant's confidence ratings using a principal component (PC) decomposition of the distances between corresponding landmarks as features. Building participant-wise models provided the maximum flexibility in feature weight assignment and was therefore the harshest test to the conclusion that metacognitive access to facial expressions is partial. We found that these models could in fact predict confidence ratings (median $r = 0.26 \pm 0.15$), suggesting that participants did indeed base their confidence ratings on (specific subsets of) landmark distances. Further, because confidence is known to correlate negatively with response times, we also asked whether RTs could have served as a proxy for distance. We found that the landmark distances could be used to build ML models that predicted confidence ratings above and beyond RT information alone, confirming that participants did use some of the landmark distance information to rate confidence (see supplementary Figure S4).

To better understand which information participants used to rate their own performance, we reconstructed the weights of each feature in landmark space (based on the model's weighting of each principal component and each feature's loading on that component, see Methods). We first plotted the resulting landmark weights on their corresponding mean locations to explore potential patterns among participants based on the set of landmarks with the highest weights (both visually and by considering the median weight over all landmarks); however, we could not identify any landmarks or features that were consistently prioritized across participants (Figure 6). Individual participants' ML feature weights can be seen at https://metamotorlab.filevich.com/onlineInfo_papers/cistonEtAl_2021/table2D.html). Finally, we estimated the relationship between the new landmark distance (this time considering the participant-specific weights) and confidence ratings using, as before, a linear mixed-effects regression model. In line with the non-zero r values from the ML models, the reconstructed distances showed a significant relationship with confidence ratings ($M = 0.04 \pm 0.004$, $CI = [0.03, 0.04]$, $BF_{10} = 1.34 \times 10^7$, $R^2 = 0.24$). Note that the slope estimate is now positive, because the feature weights must incorporate the negative relationship between landmarks and confidence, in order to predict confidence ratings. This result is expected, as it is merely a transformation from principal components space to landmarks space, and we provide it here only to offer a result that is more intuitive to interpret. Taken together, the results suggest that participants were indeed able to base their confidence ratings on the distances between facial landmarks, but only on a subset of them; and that each participant had access to, or focused on, different aspects of their facial expressions.

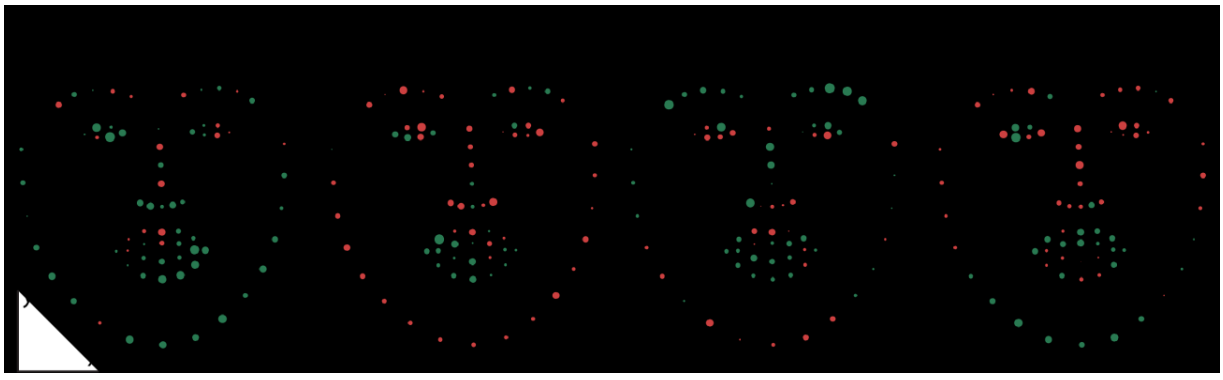

Figure 6: Machine Learning analyses. Average feature weights for participant-wise models of confidence ratings. Each dot represents the median feature weight for each landmark in models excluding RTs. Green and red correspond to positive and negative weights, respectively. The size of the dot corresponds to the relative magnitude of the landmark's approximated weight within the model, and their

positions correspond to a normalized face. Each landmark is split into the four cardinal directions, to yield four independent features (see Methods for details). We found no consistent pattern over participants where some features are weighted more strongly than others, see https://gitlab.com/elisa.filevich/cistonetal_metacognitionoffacialexpressions for an interactive table with participant-wise weights.

Discussion

We asked how precisely we can describe how our faces look when we make expressions. We quantified young, healthy adults' metacognitive access to the low-level details of their own facial expressions. We emphasized to participants that we were focused on the specific shape of the face and activation of the muscles, not on the emotion that the expression conveyed. We found a negative, but weak, relationship between subjective confidence and distance. A priori, this can be interpreted in two (non-exclusive) ways: Participants' confidence ratings may not have strongly relied on the distance between a pair of images because they truly had little or no metacognitive access to their own facial expressions. Alternatively, our measured distance based on the whole set of landmarks may have been a very noisy or even invalid measure of performance. In turn, this alternative explanation would mean that it would be invalid to quantify metacognitive access as we did. To ensure that the second alternative could not fully explain our results, we quantified the relationship between ratings of similarity (provided by the participants themselves while viewing image pairs side-by-side) and distance (based on the whole set of landmarks, combined with equal weights). The maximum theoretical slope of the relationship between a rating and the distance between image pairs is given by the ratio between the range of possible ratings and the maximum distance between two corresponding images. However, we expected the empirical relationships between ratings and distance to be lower than this theoretical maximum. We reasoned that the magnitude of the relationship between similarity and distance effectively quantifies the empirical maximum for this paradigm, as it accounts for noise in the estimation of distance (which in turn includes the resolution of the images, errors in the landmark placement, and limitations of the rigid transformations) as well as noise in the use of the rating scale. We found that the slope of the relationship between similarity ratings and distance was approximately four times greater than that of the relationship between confidence and distance. This result suggests, first, that the relationship between confidence and distance reveals sources of noise beyond those due to a poor use or understanding of the confidence scale, or to generally poor performance in the movement task. More specifically, if participants had not understood the instructions, or chosen not to follow them when providing confidence ratings, then presumably the same would have been the case for similarity ratings. It would require an additional assumption to interpret these data as evidence that participants could have, but chose not to, attend to the low-level aspects of their facial expressions when rating confidence and that they then changed their behaviour when rating similarity. Instead, a more parsimonious explanation is that participants found it hard to access the low-level details of their expressions when they did not have visual access to them. Second, the analysis of similarity ratings suggests that

participants were not simply generally poor at the self-imitation task. Instead, there was enough variability in performance across trials that informed participants' similarity ratings. In other words, the clear relationship between distance and similarity indicates that there was enough meaningful variability in both variables included in the analysis.

It is important to emphasize that a statistical difference between the strengths of the association between distance and the two kinds of ratings (namely confidence and similarity), is expected, but also trivial. Participants had no visual information about the expression they were making when rating confidence, whereas they could do careful comparisons of image pairs using all available visual information to rate similarity. We therefore did not compute a statistical comparison. Instead, we make separate inferences based solely on the estimation of the effect size and reliability for each of the associations, and the comparison between each full model including the effect of interest and its null counterpart. Simply put, the analysis of the relationships between confidence and distance suggests that participants only had partial access to their own performance. On the other hand, the analysis of the relationships between similarity and distance suggests that we measured performance adequately. A numerical comparison between the two allows us to interpret the magnitude of the relationship between confidence and distance, by providing a range of values that this relationship could plausibly take.

We also ran a series of exploratory analyses. First, to exploit inter-individual variability, we estimated the correlations between individual estimates of the relationship between distance and confidence and other measures of insight, namely visual metacognitive efficiency estimates and alexithymia scores. No conclusive relationships emerged that could explain the variations between individuals. Further, in another exploratory analysis, we considered that the summary distance measure could not discriminate between landmarks that heavily informed participants' confidence ratings and those that were ignored. In other words, confidence ratings may have depended on performance defined by a subset of landmarks, which may not have been the same for all participants, or indeed for all trials of a given participant. To examine this possibility, we built linear regression models on confidence ratings that included the differences for each landmark as individual features (each of them separated into the four cardinal directions). This analysis revealed that the models built for all participants could predict confidence from the combined features — and could do so with better accuracy than the models relying solely on reaction times, which we expected to be predictive of confidence based on previous literature (Rahnev et al., 2020; Vickers & Packer, 1982). This result suggests that participants' confidence ratings do indeed carry information about the landmark distance between target and response

expressions. But, unlike what the linear regression analyses assumed, not all landmarks contribute equally: The contributions from each landmark were not consistent between participants and, in fact, some landmarks contributed in a way that was contrary to what was expected (i.e. larger distances were associated with higher confidence). In sum, while some aspects of participants' facial expressions led (idiosyncratically) to higher confidence ratings, these ratings were not indicative of performance. Critically, we note that a portion of the variability in facial expressions did not inform confidence ratings. This implies that participants were able to control specific aspects of their faces without having metacognitive access to this control. On the basis of these findings, we argue that there is a disconnect between participants' ability to control their faces (through their low-level features) and their assessment of performance. The (four times) greater amount of variance in distance captured by similarity ratings as compared to confidence ratings supports this interpretation and argues against the simpler alternative interpretation, that the measured distance that included all landmarks was too coarse or inadequate because it ignored idiosyncrasies in facial landmarks relevant for performance.

To rate confidence, participants may have used heuristics or proxies for performance, like perceived difficulty of the target expression. Further research may be necessary to study this or other potential strategies that participants may have used to solve the task.

If it is indeed the case that young, healthy volunteers have only partial access to their own facial expressions, the obvious question arises: Does this affect our ability to communicate effectively in society? Drawing from previous literature, we assume that each facial expression carries both low-level information (the specific degree of contraction of each muscle and consequent location of the landmarks) and high-level information (the emotion conveyed) and that these two bits of information are not necessarily correlated. Previous work has addressed healthy adults' awareness of their emotions and consequent expressions. Overall, these studies revealed that while participants often overestimate their expressivity, their access is not zero (Barr & Kleck, 1995; Gilovich et al., 1998; Gross & John, 1997; Qu et al., 2017; Rosenberg & Ekman, 1994; Wagner et al., 2003). Hence, the effects we observed here are valid for the low-level features which we asked participants to concentrate on, but they may not extrapolate to the high-level features of facial expressions, namely the expression they communicate.

To bring our results together with those of previous literature, we put forward a simple model (Figure 7). In our study, we measured distance using an algorithm that, we assume, has no access to high-level information. Similarity ratings, on the other hand, were made by human observers (the study participants) and therefore were based on both the low-level features (by design, in

line with our instructions) and high-level emotional information that is automatically processed (LeDoux & Bemporad, 1997), as we discussed above. On the basis of our results, we contend that confidence ratings may be based chiefly on high-level information, as they only partially incorporate low-level information. Then, the shared (high-level) information between similarity and confidence ratings explains the correlation between the two. Finally, the dissociation between low- and high-level information, together with their unequal contribution to different ratings, explains why confidence and distance are in turn dissociated.

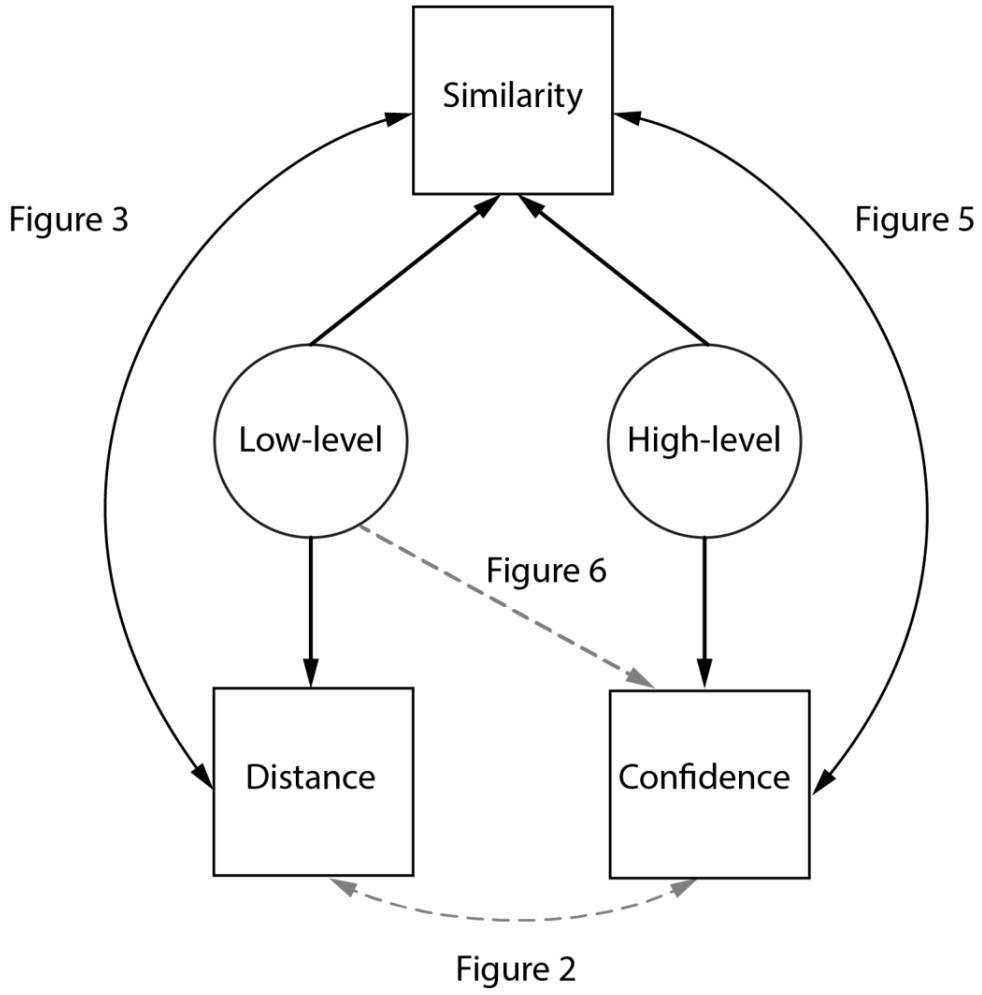

Figure 7: Suggested model for metacognitive access to facial expressions. We consider that each facial expression carries both low-level and high-level information (here depicted as circles because they are akin to latent variables in a structural equation model, whereas the measured variables of Distance and Confidence are depicted as squares). We also consider that the distance we measured is solely based on low-level information that the algorithm has access to. Thus, this simple suggested model (where confidence has accurate access to high-level but only partial access to low-level information, and where similarity ratings by human judges are informed by both low- and high-level aspects of each image) is sufficient to explain both, on the one hand, the relationships that we observed between distance and

similarity and between similarity and confidence, and on the other hand, the dissociations we found between confidence and distance.

The distinction between metacognitive access to high- and low-level features of facial expressions is compatible with previous literature. First, brain regions involved in assigning confidence to the accuracy of purely perceptual decisions (the thickness of a horizontal bar presented above-fixation) differ from those assigning confidence to decisions about emotional faces (Bègue et al., 2019). Second, continuous theta-burst suppression to the lateral prefrontal cortex led to a decrease in metacognitive performance in a task that relied on the low-level aspects of faces (discriminating between the orientation of two faces) but not one that relied on high-level aspects (discriminating the expression they communicated) (Lapate et al., 2020). Together, these results support a distinction between metacognitive access to high- and low-level features of *seen* faces (i.e., others' faces). We extend these results and suggest that this distinction may also apply to the case of one's own face, even when not seen.

Facial muscles appear to lack muscle spindles (Goodmurphy & Ovalle, 1999; Happak et al., 1994; Rinn, 1984; Stål et al., 1987, 1990), which are the main sensors for skeletal muscle stretching (Proske & Gandevia, 2012; Sherrington, 1906; Tuthill & Azim, 2018). Instead, other mechanoreceptors have been suggested to replace muscle spindles in their transduction of electric signals elicited by facial muscles (Cobo et al., 2017). In contrast to what we described for facial muscles, young, healthy participants have above-chance and precise metacognitive access to movements that are controlled by skeletal muscles (Charles et al., 2020). Moreover, unlike the case of metacognition of facial expressions, measures of metacognitive performance in motor control do partially correlate with those from a visual task (Arbuzova et al., 2021). Speculatively, at least two factors may explain these discrepancies. First, different stretch receptors may lead to different kinds of representations that may be differentially accessible to metacognitive monitoring. Second, visual feedback during development and motor learning might play an important role. Extensive motor learning and concomitant visual information for limbs that are in the field of view may shape and lead to sharper conscious representations in a way that is not possible for facial expressions.

Relationship to other metacognitive tasks

Many of the recent studies measuring metacognitive performance have capitalized on a relatively rigid operationalization of metacognition that quantifies metacognitive performance as the relationship between subjective confidence ratings (the second-order task) and objective performance in a 2AFC (the first-order task), and especially in whether a participant is able to assign high confidence exclusively to correct trials (Fleming & Lau, 2014). Unlike most experiments on metacognition, where experimenters can very easily control the (often visual) stimuli that they present to participants, the study of motor metacognition requires participants to make a movement in the first place, thereby adding another task to the standard operationalization. Participants make a movement (zero-order), then make a (first-order) judgment about it, and finally provide a (second-order) subjective confidence rating. Examples of a zero-order task include moving a finger at a given pace (Charles et al., 2020) or throwing a ball to hit a target (Arbuzova et al., 2021). A different approach, which we took here, consists in operationalizing the metacognitive judgment not as confidence in accuracy of a binary choice, but instead as a judgment of performance (Locke et al., 2020; McIntosh et al., 2019; Mole et al., 2018). While both operationalizations may be valid, it is important to note the differences between them to prevent assuming unwarranted relationships: The first approach, borrowed from paradigms developed for perceptual tasks, makes a very clear distinction between three different tasks with, in principle, independent performance levels. In a ball-throwing task, a person could miss a target often (poor zero-order performance), be good at discriminating whether the movement they made would hit the target or not (high first-order performance), but assign high and low confidence equally often to correct and incorrect discrimination trials (low second-order performance). This sharp distinction between three cognitive levels is elegant and makes metacognitive motor tasks directly comparable to perceptual ones. To test metacognitive access to the low-level details of facial expressions, as we did here, but using a 2AFC task, future studies could require participants to produce one expression and then discriminate between two images of themselves (one corresponding to the current expression, and one corresponding to a previous trial) to decide which of them they produced in the current trial. However, the comparison between motor and perceptual tasks may not be as straightforward as it appears to be (Chambon et al., 2014). It has been argued that this rigid operationalization ignores a distinctive feature of (sensori)motor performance monitoring: In making a movement, we must monitor our performance in relationship to the intended goal, which includes not only perceptual uncertainty but also motor noise and skill (Froemer et al., 2018; Locke et al., 2020). Thus, the approach of asking participants to rate their own performance allowed us to measure metacognitive access

as the relationship between true performance and the (arguably) ecologically relevant estimate of subjective performance.

Limitations

Our conclusion relies on two main assumptions. First, we assume that participants followed our instructions to focus on the low-level facial features both while making and while rating facial expressions. While, as we argued above, this assumption is supported by the similarity ratings, it would lead to different conclusions if our assumption were incorrect. Second, our conclusions are only valid if the distance estimated by the algorithms is valid as a 'true' measure of performance on each trial, which we assumed. We argue that this assumption is valid for two main reasons. First, we specifically instructed participants to focus on the low-level aspects of their facial expressions. Second, we found very similar results using two completely different algorithms to place facial landmarks (see Supplementary Information), suggesting that this measure of distance captures true differences in facial features and does not depend heavily on the idiosyncrasies of the algorithm. However, it could be argued that similarity ratings are in fact a better, truer measure of performance because they reflect how similarly two faces are perceived by a person (either a judge or the very same participant) in an ecologically valid setting. Against this intuition, we argue that similarity ratings could have been subject to the same biases and heuristics that confidence may have relied on. As a very simplistic example, a given participant could have consistently rated positive expressions with higher confidence and similarity than negative expressions, leading to a relationship between the two kinds of ratings that need not be explained by metacognitive access. We note, however, that this alternative analysis of the data, based on different assumptions, would have led to the cardinally opposite conclusion that participants *do* have precise metacognitive access to their own expressions.

A second limitation has to do with the predictive power of our statistical models. Despite robust effects in the Bayesian mixed models, a significant amount of variability is left unexplained (see SI). Better measures of distance, more precise motion tracking technologies (like infrared reflectors placed on the face), or different analysis methods may have reduced this unexplained variance. Additionally, we note that our analyses are based on static images, namely the endpoints of otherwise dynamic expressions. But important information is conveyed in the dynamic pattern of facial expressions (Chiovetto et al., 2018; Dobs et al., 2018; Krumhuber et al.,

2016), and a future direction of this work might be to relate confidence to dynamic aspects of facial expressions instead.

Finally, while the exploratory machine learning analyses allowed us to identify potential aspects of the face that participants attended to while ignoring others, we might have failed to detect any true effects where the relationship between confidence and distance differed between expressions, or relationships that changed significantly over the course of the experimental session.

It could be argued that the use of non-canonical expressions limits the ecological validity of our paradigm. However, we note that in this study we were interested in studying a potential disconnect between (zero-order) motor control and (second-order) metacognitive access to it. Canonical expressions, where a highly trained and stereotypical set of movements correspond, one-to-one, to a specific expression, confound motor control with emotional content and would not have allowed us to make any inferences about which kind of information participants were accessing to make their judgments. For instance, had we asked participants to make a stereotypical “happy” expression and then rated confidence, we would not have been able to determine whether their confidence judgments were well calibrated with the emotional state they recreated, the highly-trained motor program, or the end state of the target expression. In short, canonical expressions would have carried with them a set of confounds that our paradigm avoided. However, we speculate that, had we chosen to use standard, instead of non-canonical expressions as cues, confidence ratings of emotions would have been even more poorly attuned to the low-level details of the expressions. If there is indeed a disconnect between low- and high-level aspects of facial expressions, using expressions that could have been well imitated by focusing on the high-level emotional content alone would have made it even harder to report on the low-level details.

Conclusion

Our analyses suggest that healthy young volunteers were only able to estimate their performance in producing non-stereotypical facial expressions based on partial information. This is surprising, we argue, because it sets facial movements apart from other body movements (namely those of arms and fingers), for which, as previous studies have shown, we do have precise metacognitive access to lower-level motor information, even when this information is decoupled from the motor

782 goal. We speculate that this distinction might be related to the lack of concurrent visual information
783 during social interactions, but our speculation will need to be examined in future studies.

784

785

786

Acknowledgements

We thank student assistants for help in data collection in Experiment 1, and Manuel Zellhöfer for help in programming the experimental paradigm. We thank Soledad Galli for assistance with the ML models and Nathan Faivre for comments on an earlier version of this manuscript. ABC, CF and EF were supported by a Freigeist Fellowship to EF from the Volkswagen Foundation (grant number 91620). This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 337619223 / RTG2386 and the Max-Planck Society. The funders had no role in the conceptualization, design, data collection, analysis, decision to publish, or preparation of the manuscript.

Competing Interests

The authors declare no competing interests.

Data and Code Availability

Raw data (excluding images from participants and any other personally identifiable information) along with reproducible analysis scripts are available under https://gitlab.com/elisa.filevich/cistonetal_metacognitionoffacialexpressions.

References

Arbuzova, P., Peters, C., Röd, L., Koß, C., Maurer, H., Maurer, L. K., Müller, H., Verrel, J., & Filevich, E. (2021). Measuring metacognition of direct and indirect parameters of voluntary movement. *Journal of Experimental Psychology: General*, No Pagination Specified-No Pagination Specified. https://doi.org/10.1037/xge0000892

Bagby, R. M., Parker, J. D. A., & Taylor, G. J. (1994). The twenty-item Toronto Alexithymia scale—I. Item selection and cross-validation of the factor structure. *Journal of Psychosomatic Research*, *38*(1), 23–32. https://doi.org/10.1016/0022-3999(94)90005-1

Barr, C. L., & Kleck, R. E. (1995). Self-other perception of the intensity of facial expressions of emotion: Do we know what we show? *Journal of Personality and Social Psychology*, *68*(4), 608–618. https://doi.org/10.1037//0022-3514.68.4.608

Bègue, I., Vaessen, M., Hofmeister, J., Pereira, M., Schwartz, S., & Vuilleumier, P. (2019). Confidence of emotion expression recognition recruits brain regions outside the face perception network. *Social Cognitive and Affective Neuroscience*, *14*(1), 81–95. https://doi.org/10.1093/scan/nsy102

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436.

Bulat, A., & Tzimiropoulos, G. (2017). How far are we from solving the 2D & 3D Face Alignment problem? (And a dataset of 230,000 3D facial landmarks). *2017 IEEE International Conference on Computer Vision (ICCV)*, 1021–1030. https://doi.org/10.1109/ICCV.2017.116

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal*, *10*(1), 395–411.

Chambon, V., Filevich, E., & Haggard, P. (2014). What is the Human Sense of Agency, and is it Metacognitive? In S. M. Fleming & C. D. Frith (Eds.), *The Cognitive Neuroscience of Metacognition* (pp. 321–342). Springer Berlin Heidelberg. http://link.springer.com/chapter/10.1007/978-3-642-45190-4_14

Charles, L., Chardin, C., & Haggard, P. (2020). Evidence for metacognitive bias in perception of voluntary action. *Cognition*, *194*, 104041. https://doi.org/10.1016/j.cognition.2019.104041

Chen, B., Mundy, M., & Tsuchiya, N. (2019). Metacognitive Accuracy Improves With the Perceptual Learning of a Low- but Not High-Level Face Property. *Frontiers in Psychology*, *10*, 1712. https://doi.org/10.3389/fpsyg.2019.01712

Chiovetto, E., Curio, C., Endres, D., & Giese, M. (2018). Perceptual integration of kinematic components in the recognition of emotional facial expressions. *Journal of Vision*, *18*(4), 13. https://doi.org/10.1167/18.4.13

Clark Weeden, J., Trotman, C.-A., & Faraway, J. J. (2001). Three Dimensional Analysis of Facial Movement in Normal Adults: Influence of Sex and Facial Shape. *The Angle Orthodontist*, *71*(2), 132–140. https://doi.org/10.1043/0003-3219(2001)071<0132:TDAOFM>2.0.CO;2

Cobo, J. L., Abbate, F., de Vicente, J. C., Cobo, J., & Vega, J. A. (2017). Searching for proprioceptors in human facial muscles. *Neuroscience Letters*, *640*, 1–5. https://doi.org/10.1016/j.neulet.2017.01.016

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. L. Erlbaum Associates.

Coulson, S. E., Croxson, G. R., & Gilleard, W. L. (2000). Quantification of the Three-Dimensional Displacement of Normal Facial Movement. *Annals of Otology, Rhinology & Laryngology*, *109*(5), 478–483. https://doi.org/10.1177/000348940010900507

Craske, B., & Crawshaw, M. (1975). Shifts in kinesthesis through time and after active and passive movement. *Perceptual and Motor Skills*, *40*(3), 755–761.

Cunningham, D. W., Kleiner, M., Wallraven, C., & Bülthoff, H. H. (2005). Manipulating Video Sequences to Determine the Components of Conversational Facial Expressions. *ACM Trans. Appl. Percept.*, *2*(3), 251–269. https://doi.org/10.1145/1077399.1077404

Dienes, Z. (2019). How Do I Know What My Theory Predicts? *Advances in Methods and Practices in Psychological Science*, *2*(4), 364–377. https://doi.org/10.1177/2515245919876960

Dobs, K., Bülthoff, I., & Schultz, J. (2018). Use and Usefulness of Dynamic Face Stimuli for Face Perception Studies—A Review of Behavioral Findings and Methodology. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.01355

Doorn, J. van, Ly, A., Marsman, M., & Wagenmakers, E.-J. (2020). Bayesian rank-based hypothesis testing for the rank sum test, the signed rank test, and Spearman's ρ. *Journal of Applied Statistics*, *47*(16), 2984–3006. https://doi.org/10.1080/02664763.2019.1709053

Ekman, P. (1999). Basic emotions. In *Handbook of cognition and emotion* (pp. 45–60). John Wiley & Sons Ltd. https://doi.org/10.1002/0470013494.ch3

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

Froemer, R., Nassar, M. R., Stuermer, B., Sommer, W., & Yeung, N. (2018). I knew that! Confidence in outcome prediction and its impact on feedback processing and learning. *BioRxiv*, 442822.

Fuentes, C. T., & Bastian, A. J. (2010). Where is your arm? Variations in proprioception across space and tasks. *Journal of Neurophysiology*, *103*(1), 164–171. https://doi.org/10.1152/jn.00494.2009

Fuentes, C. T., Longo, M. R., & Haggard, P. (2013). Body image distortions in healthy adults. *Acta Psychologica*, *144*(2), 344–351.

Fuentes, C. T., Runa, C., Blanco, X. A., Orvalho, V., & Haggard, P. (2013). Does My Face FIT?: A Face Image Task Reveals Structure and Distortions of Facial Feature Representation. *PLoS ONE*, *8*(10), e76805. https://doi.org/10.1371/journal.pone.0076805

Gelman, A., Goodrich, B., Gabry, J., & Vehtari, A. (2019). R-squared for Bayesian Regression Models. *The American Statistician*, *73*(3), 307–309. https://doi.org/10.1080/00031305.2018.1549100

Gilovich, T., Savitsky, K., & Medvec, V. H. (1998). The illusion of transparency: Biased assessments of others' ability to read one's emotional states. *Journal of Personality and Social Psychology*, *75*(2), 332–346. https://doi.org/10.1037//0022-3514.75.2.332

Goodmurphy, C. W., & Ovalle, W. K. (1999). Morphological study of two human facial muscles: Orbicularis oculi and corrugator supercilii. *Clinical Anatomy (New York, N.Y.)*, *12*(1), 1–11. https://doi.org/10.1002/(SICI)1098-2353(1999)12:1<1::AID-CA1>3.0.CO;2-J

Goodwin, G. M., McCloskey, D. I., & Matthews, P. B. (1972). The contribution of muscle afferents to kinaesthesia shown by vibration induced illusions of movement and by the effects of paralysing joint afferents. *Brain: A Journal of Neurology*, *95*(4), 705–748.

Gritsenko, V., Krouchev, N. I., & Kalaska, J. F. (2007). Afferent input, efference copy, signal noise, and biases in perception of joint angle during active versus passive elbow movements. *Journal of Neurophysiology*, *98*(3), 1140–1154.

Gross, J. J., & John, O. P. (1997). Revealing feelings: Facets of emotional expressivity in self-reports, peer ratings, and behavior. *Journal of Personality and Social Psychology*, *72*(2), 435–448. https://doi.org/10.1037//0022-3514.72.2.435

Happak, W., Burggasser, G., Liu, J., Gruber, H., & Freilinger, G. (1994). Anatomy and Histology of the Mimic Muscles and the Supplying Facial Nerve. In E. R. Stennert, G. W. Kreutzberg, O. Michel, & M. Jungehülsing (Eds.), *The Facial Nerve* (pp. 85–86). Springer. https://doi.org/10.1007/978-3-642-85090-5_23

JASP Team. (2020). *JASP (Version 0.14)[Computer software].* JASP - Free and User-Friendly Statistical Software. https://jasp-stats.org/faq/how-do-i-cite-jasp/

Jeffreys, H. (1998). *The Theory of Probability*. OUP Oxford.

Kal, E., Prosée, R., Winters, M., & Kamp, J. van der. (2018). Does implicit motor learning lead to greater automatization of motor skills compared to explicit motor learning? A systematic review. *PLOS ONE*, *13*(9), e0203591. https://doi.org/10.1371/journal.pone.0203591

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3. *Perception*, *36*(14), 1–1.

Kleynen, M., Braun, S. M., Bleijlevens, M. H., Lexis, M. A., Rasquin, S. M., Halfens, J., Wilson, M. R., Beurskens, A. J., & Masters, R. S. W. (2014). Using a Delphi technique to seek consensus regarding definitions, descriptions and classification of terms related to implicit and explicit forms of motor learning. *PloS One*, *9*(6), e100227. https://doi.org/10.1371/journal.pone.0100227

Krumhuber, E. G., Skora, L., Küster, D., & Fou, L. (2016). A Review of Dynamic Datasets for Facial Expression Research: *Emotion Review*. https://doi.org/10.1177/1754073916670022

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, *25*(1), 178–206. https://doi.org/10.3758/s13423-016-1221-4

Lackner, J. R. (1988). SOME PROPRIOCEPTIVE INFLUENCES ON THE PERCEPTUAL REPRESENTATION OF BODY SHAPE AND ORIENTATION. *Brain*, *111*(2), 281–297. https://doi.org/10.1093/brain/111.2.281

Lange, K., Kühn, S., & Filevich, E. (2015). "Just Another Tool for Online Studies" (JATOS): An Easy Solution for Setup and Management of Web Servers Supporting Online Studies. *PLoS ONE*, *10*(6), e0130834. https://doi.org/10.1371/journal.pone.0130834

Lapate, R. C., Samaha, J., Rokers, B., Postle, B. R., & Davidson, R. J. (2020). Perceptual metacognition of human faces is causally supported by function of the lateral prefrontal cortex. *Communications Biology*, *3*(1), 1–10. https://doi.org/10.1038/s42003-020-1049-3

Latash, M. L. (2012). The bliss (not the problem) of motor abundance (not redundancy). *Experimental Brain Research*, *217*(1), 1–5. https://doi.org/10.1007/s00221-012-3000-4

LeDoux, J., & Bemporad, J. R. (1997). The emotional brain. *Journal of the American Academy of Psychoanalysis*, *25*(3), 525–528.

Limanowski, J., & Blankenburg, F. (2016). Integration of Visual and Proprioceptive Limb Position Information in Human Posterior Parietal, Premotor, and Extrastriate Cortex. *Journal of Neuroscience*, *36*(9), 2582–2589. https://doi.org/10.1523/JNEUROSCI.3987-15.2016

940 Locke, S. M., Mamassian, P., & Landy, M. S. (2020). Performance monitoring for sensorimotor
941 confidence: A visuomotor tracking study. *Cognition*, 104396.
942 https://doi.org/10.1016/j.cognition.2020.104396
943 Longo, M. R., & Haggard, P. (2010). An implicit body representation underlying human position
944 sense. *Proceedings of the National Academy of Sciences*, *107*(26), 11727–11732.
945 https://doi.org/10.1073/pnas.1003483107
946 Longo, M. R., & Holmes, M. (2020). Distorted perceptual face maps. *Acta Psychologica*, *208*,
947 103128. https://doi.org/10.1016/j.actpsy.2020.103128
948 MacIntyre, T., Igou, E. R., Campbell, M. J., Moran, A. P., & Matthews, J. (2014). Metacognition
949 and action: A new pathway to understanding social and cognitive aspects of expertise in
950 sport. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.01155
951 Maister, L., De Beukelaer, S., Longo, M., & Tsakiris, M. (2020). *The Self in the Mind's Eye:*
952 *Reverse-correlating one's self reveals how psychological beliefs and attitudes shape our*
953 *body-image* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/f2b36
954 Makowski, D., Lüdecke, D., Ben-Shachar, M. S., Wilson, M. D., Bürkner, P.-C., Mahr, T.,
955 Singmann, H., & Gronau, Q. F. (2020). *bayestestR: Understand and Describe Bayesian*
956 *Models and Posterior Distributions* (0.7.2) [Computer software]. https://CRAN.R-
957 project.org/package=bayestestR
958 Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating
959 metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1),
960 422–430. https://doi.org/10.1016/j.concog.2011.09.021
961 McIntosh, R. D., Fowler, E. A., Lyu, T., & Della Sala, S. (2019). Wise up: Clarifying the role of
962 metacognition in the Dunning-Kruger effect. *Journal of Experimental Psychology:*
963 *General*, *148*(11), 1882–1897. https://doi.org/10.1037/xge0000579

Mole, C. D., Jersakova, R., Kountouriotis, G. K., Moulin, C. J., & Wilkie, R. M. (2018). Metacognitive judgements of perceptual-motor steering performance: *Quarterly Journal of Experimental Psychology*. https://doi.org/10.1177/1747021817737496

Mora, L., Cowie, D., Banissy, M. J., & Cocchini, G. (2018). My true face: Unmasking one's own face representation. *Acta Psychologica*, *191*, 63–68.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442.

Proske, U., & Gandevia, S. C. (2012). The Proprioceptive Senses: Their Roles in Signaling Body Shape, Body Position and Movement, and Muscle Force. *Physiological Reviews*, *92*(4), 1651–1697. https://doi.org/10.1152/physrev.00048.2011

Qu, F., Yan, W.-J., Chen, Y.-H., Li, K., Zhang, H., & Fu, X. (2017). "You Should Have Seen the Look on Your Face…": Self-awareness of Facial Expressions. *Frontiers in Psychology*, *8*, 832. https://doi.org/10.3389/fpsyg.2017.00832

Rahnev, D., Desender, K., Lee, A. L. F., Adler, W. T., Aguilar-Lleyda, D., Akdoğan, B., Arbuzova, P., Atlas, L. Y., Balcı, F., Bang, J. W., Bègue, I., Birney, D. P., Brady, T. F., Calder-Travis, J., Chetverikov, A., Clark, T. K., Davranche, K., Denison, R. N., Dildine, T. C., … Zylberberg, A. (2020). The Confidence Database. *Nature Human Behaviour*, *4*(3), 317–325. https://doi.org/10.1038/s41562-019-0813-1

Rausch, M., & Zehetleitner, M. (2017). Should metacognition be measured by logistic regression? *Consciousness and Cognition*, *49*, 291–312. https://doi.org/10.1016/j.concog.2017.02.007

Rinn, W. E. (1984). The neuropsychology of facial expression: A review of the neurological and psychological mechanisms for producing facial expressions. *Psychological Bulletin*, *95*(1), 52–77. https://doi.org/10.1037/0033-2909.95.1.52

Rosenberg, E. L., & Ekman, P. (1994). Coherence between expressive and experiential systems in emotion. *Cognition and Emotion*, *8*(3), 201–229. https://doi.org/10.1080/02699939408408938

Rouault, M., McWilliams, A., Allen, M. G., & Fleming, S. M. (2018). Human metacognition across domains: Insights from individual differences and neuroimaging. *Personality Neuroscience*, *1*. https://doi.org/10.1017/pen.2018.16

Ruttle, J. E., Hart, B. M. 't, & Henriques, D. Y. P. (2018). The fast contribution of visual-proprioceptive discrepancy to reach aftereffects and proprioceptive recalibration. *PLOS ONE*, *13*(7), e0200621. https://doi.org/10.1371/journal.pone.0200621

Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., & Frith, C. D. (2014). Supra-personal cognitive control and metacognition. *Trends in Cognitive Sciences*, *18*(4), 186–193. https://doi.org/10.1016/j.tics.2014.01.006

Sherrington, C. S. (1906). *The integrative action of the nervous system*. Scribner.

Sober, S. J., & Sabes, P. N. (2005). Flexible strategies for sensory integration during motor planning. *Nature Neuroscience*, *8*(4), 490–497. https://doi.org/10.1038/nn1427

Stål, P., Eriksson, P.-O., Eriksson, A., & Thornell, L.-E. (1987). Enzyme-histochemical differences in fibre-type between the human major and minor zygomatic and the first dorsal interosseus muscles. *Archives of Oral Biology*, *32*(11), 833–841. https://doi.org/10.1016/0003-9969(87)90011-2

Stål, P., Eriksson, P.-O., Eriksson, A., & Thornell, L.-E. (1990). Enzyme-histochemical and morphological characteristics of muscle fibre types in the human buccinator and orbicularis oris. *Archives of Oral Biology*, *35*(6), 449–458. https://doi.org/10.1016/0003-9969(90)90208-R

Taylor, J., & Ivry, R. (2013). *Implicit and Explicit Processes in Motor Learning*. 63–87. https://doi.org/10.7551/mitpress/9780262018555.003.0003

Tuthill, J. C., & Azim, E. (2018). Proprioception. *Current Biology*, *28*(5), R194–R203. https://doi.org/10.1016/j.cub.2018.01.064

van Beers, R. J., Wolpert, D. M., & Haggard, P. (2002). When Feeling Is More Important Than Seeing in Sensorimotor Adaptation. *Current Biology*, *12*(10), 834–837. https://doi.org/10.1016/S0960-9822(02)00836-9

Vickers, D., & Packer, J. (1982). Effects of alternating set for speed or accuracy on response time, accuracy and confidence in a unidimensional discrimination task. *Acta Psychologica*, *50*(2), 179–197. https://doi.org/10.1016/0001-6918(82)90006-3

Wagner, A. W., Roemer, L., Orsillo, S. M., & Litz, B. T. (2003). Emotional experiencing in women with posttraumatic stress disorder: Congruence between facial expressivity and self-report. *Journal of Traumatic Stress*, *16*(1), 67–75. https://doi.org/10.1023/A:1022015528894