

1 **Title: Limited metacognitive access to one's own facial expressions**

2 **Running head:** Limited metacognition to facial expressions

3 **Authors:** Anthony B Ciston<sup>\*a,b,c</sup>, Carina Forster<sup>\*a,b,c,d</sup>, Timothy R Brice<sup>e,f</sup>, Simone Kühn<sup>g,h</sup>, Julius  
4 Verrel<sup>i,j</sup>, Elisa Filevich<sup>a,b,c,i</sup>

5 \* These authors contributed equally to the work.

6 **Affiliations:**

7 <sup>a</sup> Department of Psychology, Humboldt Universität zu Berlin, Unter den Linden 6, 10099 Berlin,  
8 Germany

9 <sup>b</sup> Bernstein Center for Computational Neuroscience Berlin, Philippstraße 13 Haus 6, 10115  
10 Berlin, Germany

11 <sup>c</sup> Berlin School of Mind and Brain, Humboldt-Universität zu Berlin, Luisenstraße 56, 10115  
12 Berlin, Germany

13 <sup>d</sup> Department of Neurology, Max Planck Institute for Human Cognitive and Brain Sciences,  
14 04103 Leipzig, Germany

15 <sup>e</sup> Department of Human Development and Family Studies, Pennsylvania State University, 115  
16 HHD Building, University Park, PA, 16802, USA

17 <sup>f</sup> Institute for Computational and Data Sciences, Pennsylvania State University, 224B Computer  
18 Building, University Park, PA, 16802, USA

19 <sup>g</sup> Lise Meitner Group for Environmental Neuroscience, Max Planck Institute for Human  
20 Development, Lentzeallee 94, 14195 Berlin, Germany

21 <sup>h</sup> University Clinic Hamburg-Eppendorf, Clinic and Polyclinic for Psychiatry and Psychotherapy,  
22 Martinstraße 52, 20246 Hamburg, Germany

23 <sup>i</sup> Center for Lifespan Psychology, Max Planck Institute for Human Development, Lentzeallee 94,  
24 14195 Berlin, Germany

25 <sup>j</sup> Institute of Systems Motor Science, University of Lübeck, Ratzeburger Allee 160, 23562  
26 Lübeck

27

28 **Corresponding author:** Elisa Filevich, [elisa.filevich@gmail.com](mailto:elisa.filevich@gmail.com)

29 **Word count:**

30 **Abstract:** 148

31 **Introduction, Results and Discussion (excluding figure captions and tables):** 5228

32 **Methods:** 3103

33 **Abstract**

34 As humans we communicate important information through fine nuances in our facial expressions,  
35 but because conscious motor representations are noisy, we might not be able to report these fine  
36 but meaningful movements. Here we measured how much explicit metacognitive information  
37 young adults have about their own facial expressions. Participants imitated pictures of themselves  
38 making facial expressions and triggered a camera to take a picture of them while doing so. They  
39 then rated confidence (how well they thought they imitated each expression). We defined  
40 metacognitive access to facial expressions as the relationship between objective performance  
41 (how well the two pictures matched) and subjective confidence ratings. Metacognitive access to  
42 facial expressions was very poor when we considered all face features indiscriminately. Instead,  
43 machine learning analyses revealed that participants rated confidence based on idiosyncratic  
44 subsets of features. We conclude that metacognitive access to own facial expressions is partial,  
45 and surprisingly limited.

46

## 47 Introduction

48 Precise motor planning and execution can occur without the brain having explicit, conscious  
49 access to the exact position of our limbs, or the exact degree of contraction of our muscles<sup>1-3</sup>. For  
50 instance, we can simultaneously walk, speak, and gesticulate successfully while concentrating on  
51 an argument and not on the movements that enable it, and we are furthermore unable to  
52 accurately report the state of each of our muscles. Although explicit access to proprioceptive  
53 signals in highly routinary tasks like walking or talking may be unnecessary, it might be beneficial  
54 in some other cases. For example, it has been suggested<sup>4</sup> that metacognitive reasoning plays a  
55 central role in developing and improving motor expertise: if an experienced actor has a detailed  
56 and sophisticated representation of an ideal facial expression to communicate emotion, they are  
57 better able to detect and correct deviations from the ideal, leading in turn to more accurate and  
58 consistent performance.

59 Proprioceptive information about our limbs and their movements is thought to originate primarily  
60 from muscle spindles, together with skin receptors, Golgi tendon organs, and joint receptors<sup>5-7</sup>.  
61 Artificial vibration of the muscles can lead to activation of the muscle spindles, showing that their  
62 activation is sufficient to alter the representation of the body and its position<sup>8,9</sup>. In addition, position  
63 estimates have been found to be more precise following active vs. passive movements,  
64 suggesting that efferent motor commands may either affect or inform proprioceptive  
65 representations<sup>10-12</sup>. Finally, proprioceptive information is combined with visual information, when  
66 available, to form a multisensory and integrated representation<sup>13-17</sup>.

67 Facial expressions present a particularly important yet poorly studied instance of motor control.  
68 On the one hand, we communicate a great deal of information with small, nuanced facial  
69 movements (on the order of 10 mm or less<sup>18,19</sup>). On the other hand, we hardly ever see ourselves  
70 while making them. Perhaps with the exception of actors or public speakers who practice in front  
71 of a mirror (or the increased number of video-conferences during the 2020 SARS-CoV-2  
72 pandemic), we do not usually have online visual feedback about our facial muscles. If visual  
73 feedback information is indeed critical to give rise to precise motor representations, facial  
74 movements might be very poorly represented. Together, the combination of the high social  
75 relevance of small movements in our facial muscles and the general lack of visual information  
76 about them raise the interesting question: How much do we know about how we look when we  
77 communicate with others?



78 Previous studies have focused on related questions. One line of research has quantified  
79 metacognitive access to *others'* facial expressions<sup>20–22</sup> and operationalized metacognitive  
80 performance as the precision of participants' representations of uncertainty. While our ability to  
81 accurately represent both the facial expressions of others and our certainty about them is clearly  
82 critical for social interactions, it is equally important to correctly represent and adequately control  
83 *one's own* expressions<sup>23</sup>. In line with this notion, another line of research has aimed at measuring  
84 how accurate the representation of one's own face is (under a neutral facial expression). One  
85 study<sup>24</sup> found that participants showed a systematic bias to underestimate the length of their faces  
86 and slightly overestimate their width, mimicking what has been described for whole bodies<sup>25</sup> and  
87 hands<sup>26</sup>. More recently, large inter-individual differences have been described in how accurately  
88 healthy young adults can represent their own faces<sup>27</sup>. These previous studies investigated relaxed  
89 faces with neutral expressions and captured, in essence, individuals' ability to accurately describe  
90 their face, or to discriminate it from the face of another. Importantly, static features of one's face  
91 are irrelevant to social interactions, which instead are based on dynamic information. Here, we  
92 focussed instead on metacognitive knowledge about how one's face varies when making different  
93 expressions. In a pre-registered experiment, we asked participants to imitate expressions shown  
94 in pictures of themselves and to rate how well they thought they had imitated the expression. We  
95 then measured participants' metacognitive access to their own facial expressions as the  
96 correspondence between subjective ratings and an objective measure of performance.

97 First, participants completed a task to measure their metacognitive access to facial expressions  
98 (Figure 1), consisting of three parts. Briefly, in the first part of the task, participants took pictures  
99 of themselves imitating different cue images done by actors<sup>28</sup> to generate 32 participant-specific  
100 target images. In the second part, participants saw each of the target images on the screen and,  
101 while still looking directly into the digital camera, imitated themselves (Figure 1.B). In both the first  
102 and second parts of the task, participants pressed a keyboard key to trigger the digital camera. In  
103 the second part only, they additionally rated how confident they were in their own performance on  
104 a continuous confidence scale ranging from "Very unsure" to "Very sure". Finally, in the third part  
105 of the task, participants saw the target and response pictures side-by-side and rated them for  
106 similarity on a continuous scale with the same labels as for the confidence rating. We quantified  
107 the distance between each image pair based on landmarks placed automatically on the pictures.

## A. Procedure

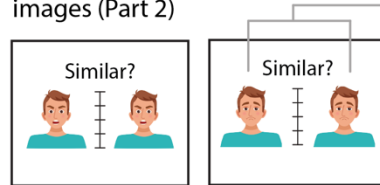
Part 1. Generate participant-specific target images



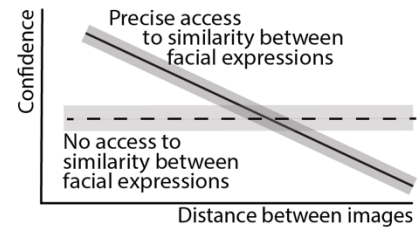
Part 2. Imitate the expressions generated in Part 1 and rate confidence in one's performance



Part 3. Rate similarity between target (Part 1) and response images (Part 2)



## B. Predictions



Distance between image pairs

$$distance = \sqrt{\frac{(x_{target}^i, y_{target}^i)(x_{response}^i, y_{response}^i)}{\sum_{i=1}^{68} (x_{target}^i - x_{response}^i)^2 + (y_{target}^i - y_{response}^i)^2}}$$

108

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

125

126

127

## Results

128

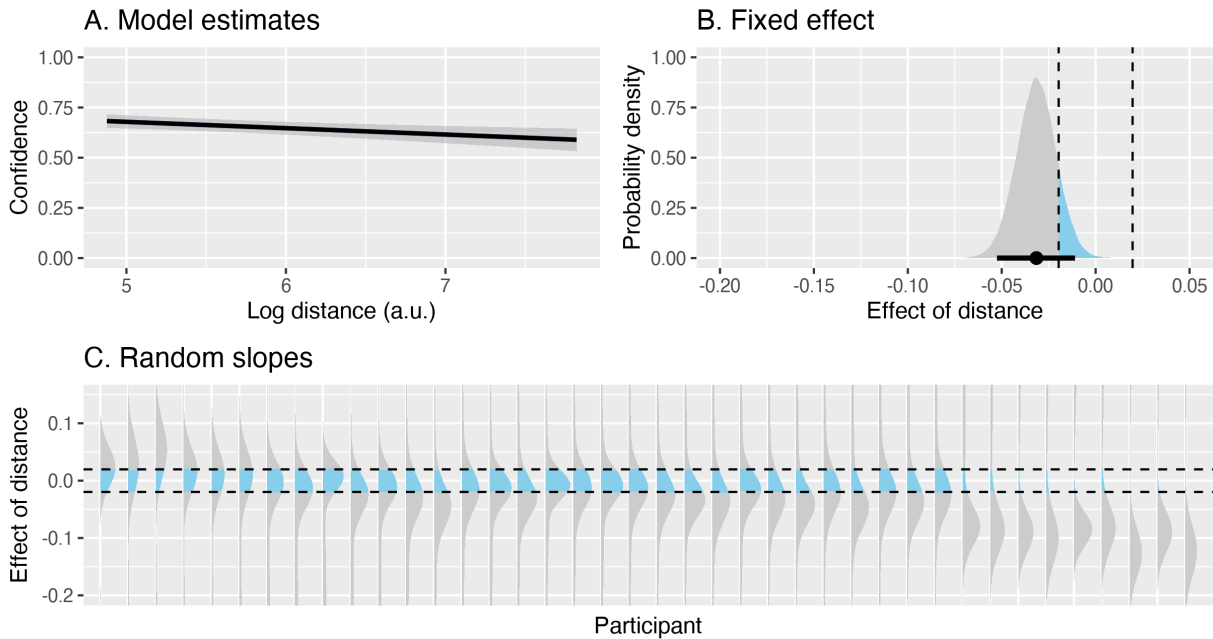
### Confirmatory Analyses

129 The distance between any pair of images is an inverse measure of performance in the task, as  
130 greater distance corresponds to a poorer match between target and response expressions. Thus,  
131 we reasoned that participants with precise metacognitive access to their facial expressions would  
132 have a sharp relationship between the distance between two images and the confidence ratings.  
133 The estimated regression coefficients from a multilevel model of these data should be negative  
134 and clearly different from 0. On the other hand, if a participant had no access to their own  
135 performance, their judgments would bear no relationship to the distance between two images,  
136 and the regression coefficients would be indistinguishable from 0 (Figure 1B, Predictions).

137 To arbitrate between these two possibilities, we first quantified our participants' metacognitive  
138 access to their own facial expressions using a Bayesian linear mixed-effects regression model of  
139 participants' confidence ratings. The model included the log-transformed distances as a fixed  
140 effect (for all 68 landmarks combined), as well as random intercepts for participant and facial  
141 expression. We found that participants' confidence ratings had a small negative relationship to  
142 the distance measured (Figure 2.A,  $M = -0.03 \pm 0.01$ ,  $CI = [-0.05, -0.01]$ ,  $R^2 = 0.21$ , see also  
143 Appendix 1-Figure 1 for the participant-wise data). However, when compared to the null model  
144 without the effect of distance, we found only anecdotal evidence<sup>29</sup> for the relationship between  
145 the two ( $BF_{10} = 2.20$ ). Further, a robustness check revealed that, as expected given the proximity  
146 of the posterior samples to the region of practical equivalence (ROPE, defined following the  
147 default criterion of the region corresponding to a Cohen's  $d$  of 0.1, Figure 2.B), the choice of the  
148 SD of the prior distribution had a strong effect on the  $BF_{10}$ : Widening the prior distribution from  
149 0.4 to 0.7 led to a  $BF_{10} = 1.02$ , and greater SDs also strongly reduced the value of the  $BF_{10}$ .  
150 Together, these results point to no evidence for a relationship between confidence and distance.  
151 For illustration purposes, we plot the participant-wise posterior draws, in relationship to the ROPE  
152 (Figure 2.C).

153

## Confidence ratings



154

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

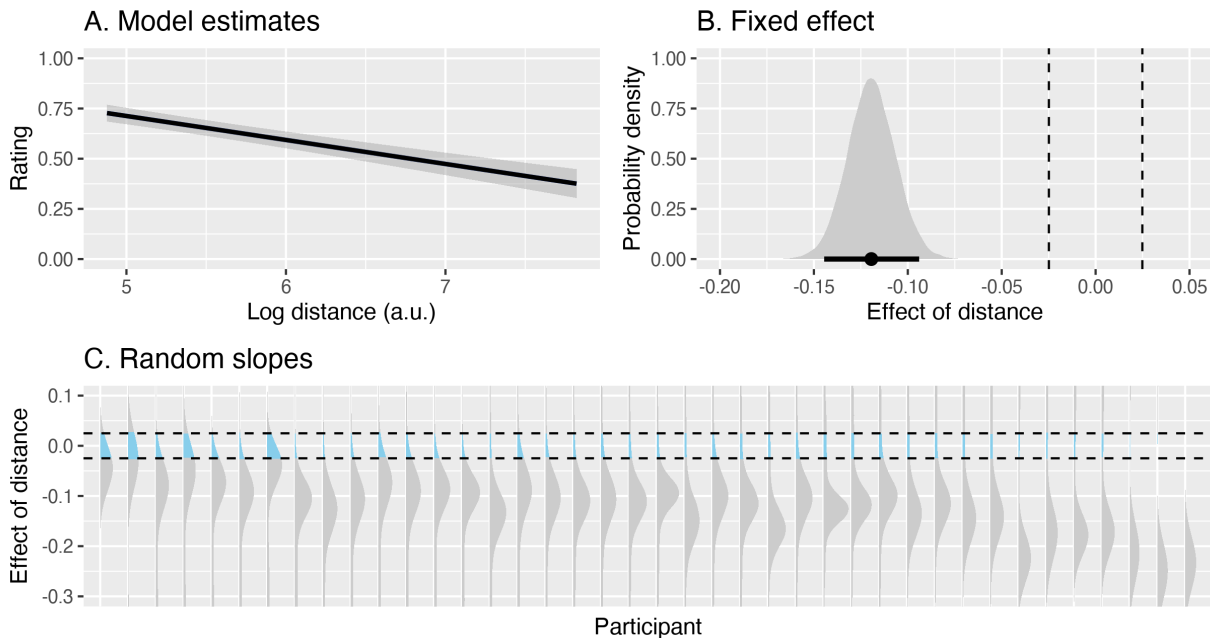
165

166 Then, to quantify the relationship between distance and similarity, we built a regression model of  
167 participants' similarity ratings including, as before, the log-transformed landmark distances as a  
168 fixed effect (for all 68 landmarks combined), as well as random intercepts for participant and facial  
169 expression. Here, similarity ratings did track the distance (Figure 3 and Appendix 1-Figure 2). We  
170 found a clear and, as expected, negative relationship between the two ( $M = -0.12 \pm 0.01$ ,  $CI = [-$   
171  $0.14, -0.09]$ ,  $BF_{10} = 8.01 \times 10^8$ ,  $R^2 = 0.26$ ). This shows that the distance we measured carried  
172 information relevant for similarity ratings and thus the null effect above cannot be simply due to a  
173 poor measure of distance. Additionally, because the same participants rated both confidence and  
174 similarity, the differences between the two ratings cannot be attributed to trivial effects such as a  
175 poor understanding of the confidence scale or task instructions, or simple lack of motivation.

176 We emphasize that an advantage of similarity as compared to confidence ratings is almost trivial,  
177 as participants could see the picture pairs side-by-side to rate similarity, but not confidence.  
178 Hence, we simply take this result as a positive control to ensure that the landmark distances were  
179 at all related to similarity, but make no formal comparisons between the two kinds of ratings.

180

### Similarity ratings



181

182 **Figure 3. The distance between two images captures relevant information. (A.)** Group effects  
183 reflecting the information contained in the distance between two images, namely the relationship between  
184 the similarity ratings provided by participants (when viewing each image pair side-by-side) and distance  
185 between two images. The solid line represents the mean of the posterior draws, and the shaded region  
186 represents the 95% credibility interval. **(B.)** Posterior draws for the group-level fixed effect of distance,  
187 shown in relation to the ROPE, marked with dashed lines. The black horizontal line indicates the mean and  
188 95% HDI. **(C.)** Posterior draws for each participant, shown in relation to the ROPE. Note that the y-axis is  
189 clipped to better display the distributions around the ROPE and therefore excludes the long tails of some  
190 of the distributions. Participants are ordered following the mean slope estimate and might not be aligned  
191 across figures.

192

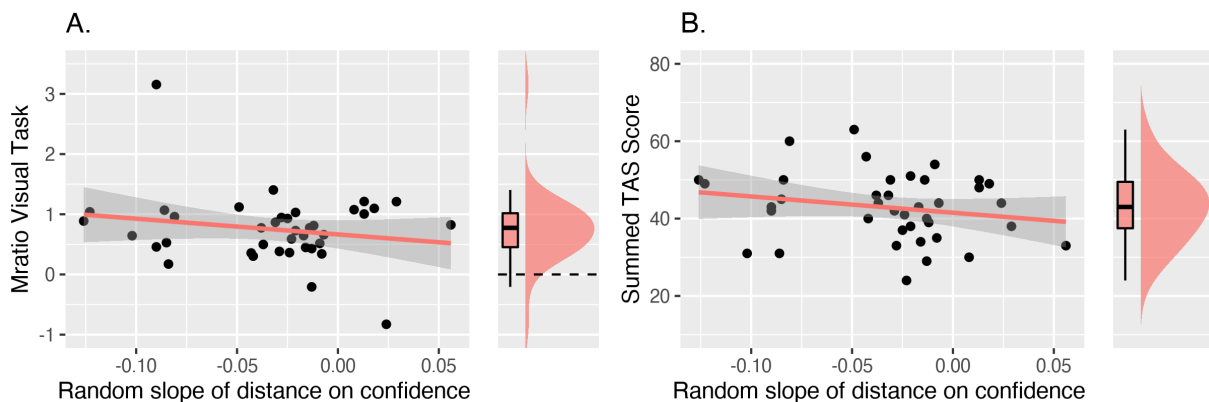
193 Finally, following our pre-registered plan, we explored relationships between the participant-wise  
194 random slopes with Mratio, a measure of visual metacognitive efficiency<sup>30</sup> in a visual task. We  
195 found that visual Mratio was consistently above the chance level of 0 ( $M = 0.75$ ,  $SD = 0.57$ ,  $t(38)$   
196  $= 8.15$ ,  $p < 0.001$ ,  $BF_{10} = 1.54 \times 10^7$ , estimated with a default Cauchy prior) but that it did not

197 correlate with participant-wise effects of distance on confidence (Figure 4.A,  $r = -0.19$ ,  $p = 0.25$ ,  
198  $BF_{10} = 0.64$ , with a default shifted beta prior distribution). While the two measures of metacognitive  
199 access are not strictly comparable (the visual Mratio is controlled for first-order performance but  
200 the individual effects of distance on confidence are not), this analysis shows that poor  
201 metacognitive access to facial expressions cannot be attributed to generally poor domain-general  
202 metacognitive insight<sup>31</sup>.

203 Using Pearson correlations, we also measured potential associations between the inter-individual  
204 differences in metacognitive access to facial expressions and Alexithymia scores, as an indication  
205 of each participant's ability to identify and describe their own feelings. We found no conclusive  
206 evidence for or against any relationships between alexithymia score and the participant-wise  
207 effect of distance on confidence ( $BF_{10} = 0.70$ , Figure 4.B) or on similarity ratings ( $BF_{10} = 0.43$ ).

208

#### Participant-wise metacognitive measures



209

210 **Figure 4: Correlations between participant-wise estimates of metacognitive access to facial**  
211 **expressions and other measures of insight.** Each dot corresponds to one participant's performance  
212 estimate, and the box- and density plots on the right represent the marginal distribution of the corresponding  
213 variable on the y axis. **A. Metacognitive efficiency (Mratio) in a visual task.** Participants' metacognitive  
214 efficiency was significantly better than chance performance (marked with the horizontal dashed line). **B.**  
215 **Alexithymia score (TAS).** We found no evidence for a correlation between metacognitive estimates and  
216 these measures of insight.

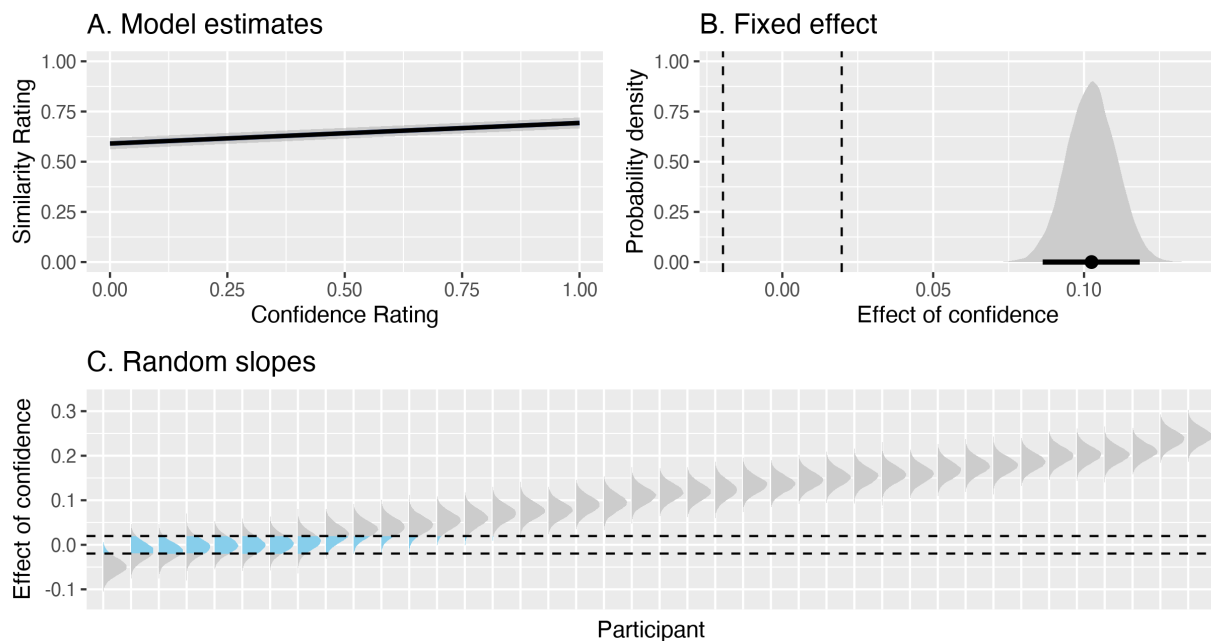
217

#### 218 *Exploratory Analyses*

219 For completeness, we studied the relationship between similarity and confidence ratings. We built  
220 a Bayesian linear regression model of participants' confidence ratings, this time including the

221 similarity ratings as a fixed effect and random intercepts for participant and facial expression. We  
222 found a clear positive relationship between the two ratings ( $M = 0.10 \pm 0.01$ ,  $CI = [0.09, 0.12]$ ,  
223  $BF_{10} = 6.36 \times 10^{31}$ ,  $R^2 = 0.21$ , Figure 5 and Appendix 1-Figure 6). This suggests that participants'  
224 confidence ratings were not random or noisy but rather that they simply did not reflect the low-  
225 level features captured by the distance.  
226

### Similarity and confidence ratings



227  
228 **Figure 5: Similarity ratings vary with confidence ratings. (A.)** Group effects showing the relationship  
229 between the two ratings on image pairs provided by participants (similarity vs. confidence). The solid line  
230 represents the mean of the posterior draws, and the shaded region represents the 95% credibility interval.  
231 **(B.)** Posterior draws for the group-level fixed effect of confidence on similarity, shown in relation to the  
232 ROPE, marked with dashed lines. The black horizontal line indicates the mean and 95% HDI. **(C.)** Posterior  
233 draws for each participant, shown in relation to the ROPE. Participants are ordered following the mean  
234 slope estimate and might not be aligned across figures.

235

236

237 Our results so far suggest that participants' confidence ratings did not reflect performance,  
238 calculated as the Euclidean distance over all landmarks. In a final set of exploratory analyses, we



239 therefore aimed at identifying which pieces of information participants may have taken into  
240 account when rating confidence.

241 The Euclidean distance between image pairs assigns equal weights to the distances of all facial  
242 landmarks and is therefore a relatively naive measure of the difference between expressions, in  
243 that it does not allow for potential differences between landmarks in their contribution to different  
244 individuals' confidence. However, it is in principle possible that participants attended to different  
245 parts of their faces to different degrees and, further, that this differential attention was not  
246 consistent across participants. For example, one participant may have focused almost exclusively  
247 on how well their mouth matched the target image to rate their confidence, and another participant  
248 may have focused exclusively on the eyes and ignored the mouth. While this was against the task  
249 instructions, it remains a possibility that would undermine the strong claim that most participants  
250 did not base their confidence ratings on the landmark distances. To obtain a more fine-grained  
251 and flexible measure of performance we used a simple linear regression machine learning (ML)  
252 model to predict each participant's confidence ratings using a principal component (PC)  
253 decomposition of the distances between corresponding landmarks as features. Building  
254 participant-wise models provided the maximum flexibility in feature weight assignment and was  
255 therefore the harshest test to the conclusion that metacognitive access to facial expressions is  
256 poor. We found that these models could in fact predict confidence ratings (median  $r = 0.26 \pm$   
257  $0.15$ ), suggesting that participants did indeed base their confidence ratings on (specific subsets  
258 of) landmark distances. Further, because confidence is known to correlate negatively with  
259 response times<sup>32,33</sup>, we also asked whether RTs could have served as a proxy for distance. We  
260 found that the landmark distances could be used to build ML models that predicted confidence  
261 ratings above and beyond RT information alone, confirming that participants did use some of the  
262 landmark distance information to rate confidence (see Appendix 1-Figure 4).

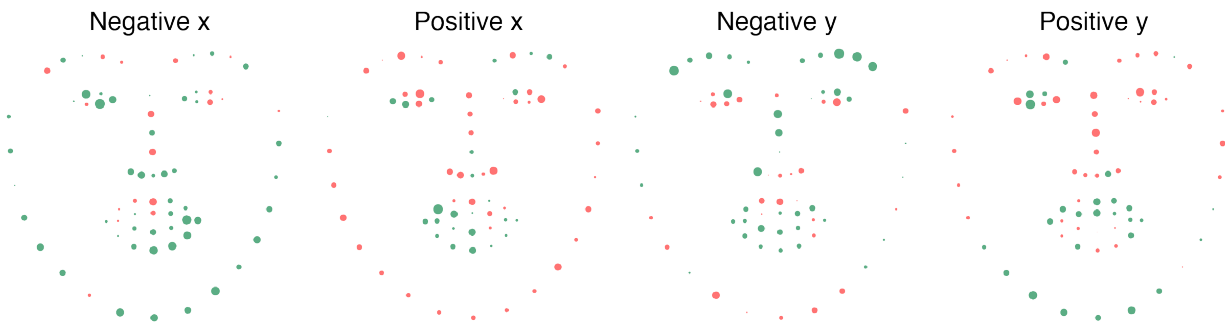
263 To better understand which information participants used to rate their own performance, we  
264 reconstructed the weights of each feature in landmark space (based on the model's weighting of  
265 each principal component and each feature's loading on that component, see Methods). We first  
266 plotted the resulting landmark weights on their corresponding mean locations to explore potential  
267 patterns among participants based on the set of landmarks with the highest weights (both visually  
268 and by considering the median weight over all landmarks); however, we could not identify any  
269 landmarks or features that were consistently prioritized across participants (Figure 6). Individual  
270 participants' ML feature weights can be seen at



271 [https://gitlab.com/elisa.filevich/cistonetal\\_metacognitionoffacialexpressions/](https://gitlab.com/elisa.filevich/cistonetal_metacognitionoffacialexpressions/)). Finally, we  
272 estimated the relationship between the new landmark distance (this time considering the  
273 participant-specific weights) and confidence ratings using, as before, a linear mixed-effects  
274 regression model. In line with the non-zero  $r$  values from the ML models, the reconstructed  
275 distances did in fact show a significant relationship with confidence ratings ( $M = 0.04 \pm 0.004$ ,  $CI$   
276  $= [0.03, 0.04]$ ,  $BF_{10} = 1.34 \times 10^7$ ,  $R^2 = 0.24$ ). Note that the slope estimate is now positive, because  
277 the feature weights must incorporate the negative relationship between landmarks and  
278 confidence, in order to predict confidence ratings. Taken together, the results suggest that  
279 participants were indeed able to base their confidence ratings on the distances between facial  
280 landmarks, but only on a subset of them; and that each participant had access to, or focused on,  
281 different aspects of their facial expressions.

282

Average feature weights across all participants



283

284 **Figure 6: Machine Learning analyses. Average feature weights for participant-wise models of**  
285 **confidence ratings.** Each dot represents the median feature weight for each landmark in models excluding  
286 RTs. Green and red correspond to positive and negative weights, respectively. The size of the dot  
287 corresponds to the relative magnitude of the landmark's approximated weight within the model, and their  
288 positions correspond to a normalized face. Each landmark is split into the four cardinal directions, to yield  
289 four independent features (see Methods for details). We found no consistent pattern over participants where  
290 some features are weighted more strongly than others, see  
291 [https://gitlab.com/elisa.filevich/cistonetal\\_metacognitionoffacialexpressions](https://gitlab.com/elisa.filevich/cistonetal_metacognitionoffacialexpressions) for an interactive table with  
292 participant-wise weights.

293

## 294 **Discussion**

295 We asked how much we know about how our faces look when we make expressions. We  
296 quantified young, healthy adults' metacognitive access to the low-level details of their own facial  
297 expressions. We emphasized to participants that we were focused on the specific shape of the  
298 face and activation of the muscles, not on the emotion that the expression conveyed. Surprisingly,  
299 our results suggest that participants were only very poorly able to consistently base their  
300 confidence ratings on the complete set of facial features. A priori, this can be interpreted in two  
301 (non-exclusive) ways: Participants' confidence ratings may not have strongly relied on the  
302 distance between a pair of images because they truly had little or no metacognitive access to their  
303 own facial expressions. Alternatively, our measured distance based on the whole set of landmarks  
304 may have been a very noisy or even invalid measure of performance. In turn, this alternative  
305 explanation would mean that it would be invalid to quantify metacognitive access as we did. To  
306 ensure that the second alternative could not fully explain our results, we quantified the relationship  
307 between ratings of similarity (provided by the participants themselves while viewing image pairs  
308 side-by-side) and distance (based on the whole set of landmarks, combined with equal weights).  
309 Here, we did find a clear relationship between the two, suggesting that the distance between  
310 image pairs does carry information that is — to some extent — relevant for similarity. This result  
311 also shows that a poor relationship between confidence and distance cannot be attributed simply  
312 to poor use or understanding of the confidence scale. It is important to emphasize that we draw  
313 no conclusions from the direct comparison of the strengths of the association between distance  
314 and the two kinds of ratings (namely confidence and similarity), as it would not be a valid  
315 comparison. Participants had no visual information about the expression they were making when  
316 rating confidence, whereas they could do careful comparisons of image pairs using all available  
317 visual information to rate similarity. Instead, we make separate inferences based solely on the  
318 estimation of the effect size and reliability for each of the associations, and the comparison  
319 between each full model including the effect of interest and its null counterpart. Simply put, the  
320 analysis of the relationships between confidence and distance suggests that participants could  
321 access their performance only poorly. On the other hand, the analysis of the relationships  
322 between similarity and distance suggests that we measured performance adequately.

323 Beyond the group-level effects, we found variation between individuals. We aimed at explaining  
324 this variation by exploring correlations between these individual estimates of the relationship  
325 between distance and confidence and other measures of insight, namely visual metacognitive

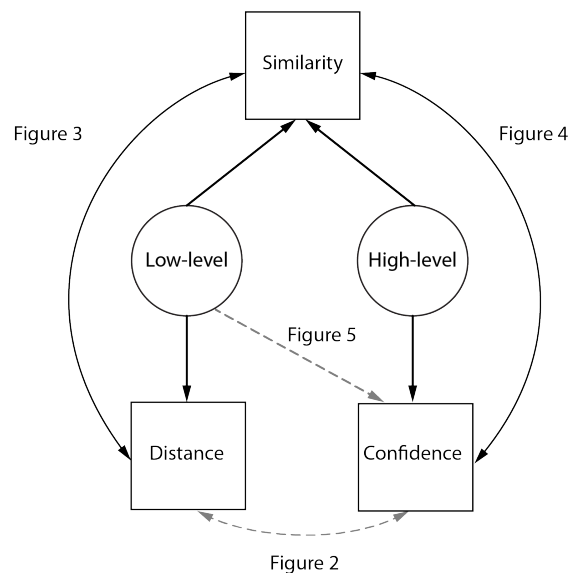
326 efficiency and alexithymia score. No conclusive relationships emerged that could explain the  
327 variations between individuals.

328 Further, in another exploratory analysis, we considered that the summary distance measure could  
329 not discriminate between landmarks that heavily informed participants' confidence ratings and  
330 those that were ignored. In other words, confidence ratings may have depended on performance  
331 defined by a subset of landmarks, which may not have been the same for all participants. To  
332 examine this possibility, we built linear regression ML models on confidence ratings that included  
333 the differences for each landmark as individual features (each of them separated into the four  
334 cardinal directions). This analysis revealed that the models built for all participants could predict  
335 confidence from the combined features (and could do so with better accuracy than the models  
336 relying solely on reaction times, which we expected to be predictive of confidence based on  
337 previous literature<sup>32,33</sup>). This result suggests that participants' confidence ratings do indeed carry  
338 information about the landmark distance between target and response expressions. But, unlike  
339 what the linear regression analyses assumed, not all landmarks contribute equally. In fact, some  
340 landmarks contributed in a way that was contrary to what was expected (i.e. larger distances were  
341 associated with higher confidence). Further, the contributions from each landmark were not  
342 consistent between participants. In sum, because some variability in facial expressions did not  
343 appear to inform confidence ratings, we argue that these findings show that there is a disconnect  
344 between participants' ability to control their faces (through their low level features) and their  
345 assessment of performance. While some aspects of participants' facial expressions led  
346 (idiosyncratically) to higher confidence ratings, these ratings were not indicative of performance.

347 If it is indeed the case that young, healthy volunteers have only partial access to their own facial  
348 expressions, the obvious question arises: How do we communicate effectively in society?  
349 Drawing from previous literature, we assume that each facial expression carries both low-level  
350 information (the specific degree of contraction of each muscle and consequent location of the  
351 landmarks) and high-level information (the emotion conveyed) and that these two bits of  
352 information are not necessarily correlated. We note that the effects we observed here are valid  
353 for the low-level features which we asked participants to concentrate on, but they may not  
354 extrapolate to the high-level features of facial expressions.

355 In fact, we suggest a simple model (Figure 7) consistent with our results where these two aspects  
356 are dissociated. We obtained the distance using an algorithm that, we assume, has no access to

357 high-level information. Similarity ratings, on the other hand, were made by human observers (the  
358 study participants) and therefore were based on both the low-level features (by design, in line with  
359 our instructions) and high-level emotional information that is automatically processed<sup>34</sup>, as we  
360 discussed above. On the basis of our results, we contend that confidence ratings may be based  
361 chiefly on high-level information, as they can only poorly incorporate low-level information. Then,  
362 the shared (high-level) information between similarity and confidence ratings explains why they  
363 correlate and the dissociation between low- and high-level information, together with their unequal  
364 contribution to different ratings, explains why confidence and distance are in turn dissociated.



365

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

376

377 The distinction between metacognitive access to high- and low-level features of facial expressions  
378 is compatible with previous literature. It has been shown that the brain regions involved in  
379 assigning confidence to the accuracy of purely perceptual decisions (the thickness of a horizontal  
380 bar presented above-fixation) were different from those assigning confidence to decisions about

381 emotional faces<sup>20</sup>. Two recent studies presented participants with two conditions with more  
382 closely matched stimuli. In the first one, two groups of participants underwent one of two kinds of  
383 perceptual learning<sup>21</sup>. One group trained to discriminate between two faces based either on their  
384 identity (high-level features) and the other group trained to discriminate the contrast between two  
385 faces (low-level features). The results showed that, while there was perceptual learning (first-  
386 order performance remained stable despite increased task difficulty) in both groups,  
387 metacognitive accuracy improved for the low-, but not high-level features training group. The  
388 authors argued for a dissociation between metacognitive access to these two levels and for a  
389 dual-stage model of metacognition whereby perceptual learning reduces noise in the  
390 representations for low- (but not high-) level facial features. A second study used a causal  
391 intervention<sup>22</sup> to show that continuous theta-burst suppression to the lateral prefrontal cortex led  
392 to a decrease in metacognitive performance in a task that relied on the low-level aspects of faces  
393 (discriminating between the orientation of two faces) but not one that relied on high-level aspects  
394 (discriminating the expression they communicated). Together, these results support a distinction  
395 between metacognitive access to high- and low-level features of *seen* faces (i.e., others' faces).  
396 We extend these results and suggest that this distinction may also apply to the case of one's own  
397 face, even when not seen.

398 Facial muscles appear to lack muscle spindles<sup>35-38</sup>, which are the main sensors for skeletal  
399 muscle stretching<sup>5-7</sup>. Instead, other mechanoreceptors have been suggested to replace muscle  
400 spindles in their transduction of electric signals elicited by facial muscles<sup>39</sup>. In contrast to what we  
401 described for facial muscles, young, healthy participants have above-chance and precise  
402 metacognitive access to movements that are controlled by skeletal muscles<sup>40</sup>. Moreover, unlike  
403 the case of metacognition of facial expressions, measures of metacognitive performance in motor  
404 control do partially correlate with those from a visual task<sup>41</sup>. Speculatively, at least two factors  
405 may explain these discrepancies. First, different stretch receptors may lead to different kinds of  
406 representations that may be differentially accessible to metacognitive monitoring. Second, visual  
407 feedback during development and motor learning might play an important role. Extensive motor  
408 learning and concomitant visual information for limbs that are in the field of view may shape and  
409 lead to sharper conscious representations in a way that is not possible for facial expressions.

410

411

412 *Relationship to other metacognitive tasks*

413 Many of the recent studies measuring metacognitive performance have capitalized on a relatively  
414 rigid operationalization of metacognition that quantifies metacognitive performance as the  
415 relationship between subjective confidence ratings (the second-order task) and objective  
416 performance in a 2AFC (the first-order task), and especially in whether a participant is able to  
417 assign high confidence exclusively to correct trials<sup>42</sup>. Unlike most experiments on metacognition,  
418 where experimenters can very easily control the (often visual) stimuli that they present to  
419 participants, the study of motor metacognition requires participants to make a movement in the  
420 first place, thereby adding another task to the standard operationalization. Participants make a  
421 movement (zero-order), then make a (first-order) judgment about it, and finally provide a (second-  
422 order) subjective confidence rating. Examples of a zero-order task include moving a finger at a  
423 given pace<sup>40</sup> or throwing a ball to hit a target<sup>41</sup>. A different approach, which we took here, consists  
424 in operationalizing the metacognitive judgment not as confidence in accuracy of a binary choice,  
425 but instead as a judgment of performance<sup>43–45</sup>. While both operationalizations may be valid, it is  
426 important to note the differences between them to prevent assuming unwarranted relationships:  
427 The first approach, borrowed from paradigms developed for perceptual tasks, makes a very clear  
428 distinction between three different tasks with, in principle, independent performance levels. In a  
429 ball-throwing task, a person could miss a target often (poor zero-order performance), be good at  
430 discriminating whether the movement they made would hit the target or not (high first-order  
431 performance), but assign high and low confidence equally often to correct and incorrect  
432 discrimination trials (low second-order performance). This sharp distinction between three  
433 cognitive levels is elegant and makes metacognitive motor tasks directly comparable to  
434 perceptual ones. On the other hand, the comparison may not be as straightforward as it appears  
435 to be<sup>46</sup>. It has been argued that this rigid operationalization ignores a distinctive feature of  
436 (sensori)motor performance monitoring: In making a movement, we must monitor our  
437 performance in relationship to the intended goal, which includes not only perceptual uncertainty  
438 but also motor noise and skill<sup>43,47</sup>. Thus, the approach of asking participants to rate their own  
439 performance allowed us to measure metacognitive access as the relationship between true  
440 performance and the (arguably) ecologically relevant estimate of subjective performance.

441

442

443 *Introspective vs. extrospective access*

444 These results contribute with an interesting case to the question of introspective privilege. A  
445 classic view has argued that introspection has privileged first-person access to — and is thus the  
446 ultimate authority on — mental and emotional states<sup>48</sup>. In the motor domain, this would mean that  
447 the agents always have the most precise representation of their movement. This makes intuitive  
448 sense, as a precise representation of an ongoing movement is presumably a prerequisite for fine  
449 and efficient motor control and execution, as well as for the emergence of a sense of agency<sup>49,50</sup>.  
450 On the other hand, a reading of the empirical literature does not provide a clear answer, perhaps  
451 due to the diversity of motor paradigms examined. Some studies have shown that precise access  
452 to movements is not always available at an explicit representational level. Participants failed to  
453 report large corrections to their ongoing movements<sup>51</sup>, and explicit instructions about how to solve  
454 a visuomotor rotation task can in fact be detrimental for performance, because explicit control is  
455 not a substitute for implicit corrections, which occur without participants' awareness<sup>52</sup>. Healthy  
456 participants also appear to have poor access to their own eye movements and a poor (i.e., noisy)  
457 representation of their own bodies that can be easily affected by visual cues<sup>13,14</sup>. On the other  
458 hand, almost directly contradicting the results above, other studies have shown that metacognitive  
459 representations of movements are as precise as those of exteroceptive signals<sup>40</sup> and that explicit  
460 instructions can sometimes be indeed beneficial for performance by leading to quicker adaptation  
461 times and shorter after-effects, as compared to no explicit instructions<sup>53</sup>. To understand these  
462 discrepancies, it may be helpful to measure metacognitive access systematically across different  
463 muscle effectors and motor and metacognitive tasks. By examining healthy participants' explicit  
464 knowledge of their own facial expressions, then, we explored another — and in our view very  
465 important — instance of motor control. We suggest that, perhaps just like eye movements, some  
466 parts of motor control might be opaque to explicit introspective access. This contributes to the  
467 body of literature questioning the privileges that introspective access has been argued to have as  
468 a matter of principle and levels the balance of epistemic access towards the complementary  
469 notion of extrospection<sup>48,54</sup>.

470 *Limitations*

471 One important limitation in our analyses is related to one basic assumption of our approach. In  
472 our exploratory analyses, we found a clear relationship between confidence and similarity ratings  
473 at the single-participant level. We explicitly relied on the distance estimated by the algorithms as



474 the ‘true’ measure of performance. We argue that this assumption is valid for two main reasons.  
475 First, we specifically instructed participants to focus on these low-level aspects. Second, we found  
476 very similar results using two completely different algorithms to place facial landmarks (see SI),  
477 suggesting that this measure of distance captures true differences in facial features and does not  
478 depend heavily on the idiosyncrasies of the algorithm. However, it could be argued that similarity  
479 ratings are in fact a better, truer measure of performance because they reflect how similarly two  
480 faces are perceived by a person (either a judge or the very same participant) in an ecologically  
481 valid setting. Against this intuition, we argue that similarity ratings could have been subject to the  
482 same biases and heuristics that confidence may have relied on. As a very simplistic example, a  
483 given participant could have consistently rated positive expressions with higher confidence and  
484 similarity than negative expressions, leading to a relationship between the two kinds of ratings  
485 that needn’t be explained by metacognitive access. We note, however, that this alternative  
486 analysis of the data, based on different assumptions, would have led to the cardinally opposite  
487 conclusion that participants *do* have precise metacognitive access to their own expressions.

488 A second limitation has to do with the predictive power of our statistical models. Despite robust  
489 effects in the Bayesian mixed models, a significant amount of variability is left unexplained (see  
490 SI). Better measures of distance, more precise motion tracking technologies (like infrared  
491 reflectors placed on the face), or different analysis methods may have reduced this unexplained  
492 variance. Additionally, we note that our analyses are based on static images, namely the  
493 endpoints of otherwise dynamic expressions. But, important information is conveyed in the  
494 dynamic pattern of facial expressions<sup>55–57</sup>, and a future direction of this work might be to relate  
495 confidence to dynamic aspects of facial expressions instead.

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

501 It could be argued that the use of non-canonical expressions limits the ecological validity of our  
502 paradigm. However, we note that in this study we were interested in studying a potential  
503 disconnect between (zero-order) motor control and (second-order) metacognitive access to it.  
504 Canonical expressions, where a highly trained and stereotypical set of movements correspond,



505 one-to-one, to a specific expression, confound motor control with emotional content and would  
506 not have allowed us to make any inferences about which kind of information participants were  
507 accessing to make their judgments. For instance, had we asked participants to make a  
508 stereotypical “happy” expression and then rated confidence, we would not have been able to  
509 determine whether their confidence judgments were well calibrated with the emotional state they  
510 recreated, the highly-trained motor program, or the end state of the target expression. In short,  
511 canonical expressions would have carried with them a set of confounds that our paradigm  
512 avoided.

## 513 **Conclusion**

514 Our analyses suggest that healthy young volunteers were only able to estimate their performance  
515 in producing non-stereotypical facial expressions based on partial information. This indicates that  
516 we not only do not have metacognitive access to the low-level details of our facial expressions,  
517 but also suggest that we cannot access them, even when explicitly asked to do so under  
518 experimental conditions. This is surprising, we argue, because it sets facial movements apart  
519 from other body movements (namely those of arms and fingers), for which, as previous studies  
520 have shown, we do have precise metacognitive access to lower-level motor information, even  
521 when this information is decoupled from the motor goal. We speculate that this distinction might  
522 be related to the lack of concurrent visual information during social interactions, but our  
523 speculation will need to be examined in future studies.

524

## 525 **Material and Methods**

### 526 *Participants*

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

534

535 *Apparatus*

536 The experimental setup consisted of a stimulus computer, a digital camera, a screen, and a half-  
537 silvered mirror tilted 45° from the vertical (Figure 8). Participants saw the image displayed on the  
538 screen by the stimulus computer indirectly through its reflection on the half-silvered mirror. Behind  
539 the mirror, a digital camera (Fire-i, UniBrain, Athens, Greece) connected to the computer took  
540 pictures of the participants' facial expressions. This setup allowed participants to look at the  
541 pictures displayed while simultaneously looking directly into the camera. As a result, we obtained  
542 pictures of participants looking straight ahead and not downwards at the image, as would have  
543 been the case if we had used e.g. a simple laptop computer with a digital camera just above the  
544 screen.

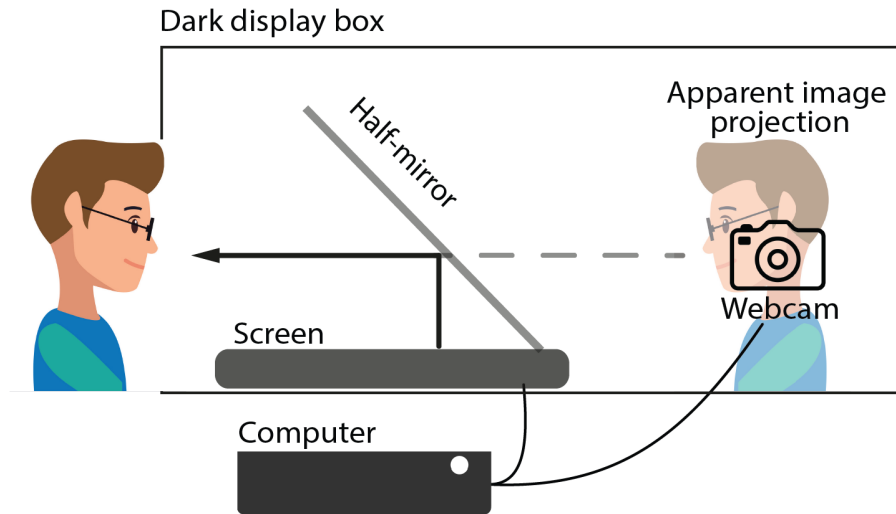
545 Participants sat at approximately 60 cm from the middle-point of the half mirror, which was in turn  
546 45 cm away from the display screen. In order to reduce head movements, we held participants'  
547 torsos loosely in place with an elastic band tied to the chair. Additionally, at the beginning of the  
548 experiment, we showed participants the image collected by the camera in real time and asked  
549 them not to make large head movements or rotations. While it would have been desirable to  
550 further limit whole-head movements using, e.g., a chin rest, we opted against this as it would have  
551 made expressions unnatural and, more importantly, because it would have provided a form of  
552 sensory feedback, interfering with the experimental design. We ensured that participants' faces  
553 were well-lit and took care that participants did not see any reflections of their own face on the  
554 mirror.

555

556

557

## Apparatus



558

559 **Figure 8. Experimental Apparatus.** Participants sat in front of a dark display box and saw the pictures  
560 projected from a computer screen reflected on a half-plated mirror (tilted 45°). Behind the mirror, positioned  
561 directly in front of participants' gaze, a digital camera took pictures of the participants when they pressed  
562 the corresponding key. This way, participants could look simultaneously directly at the to-be-imitated picture  
563 and into the camera.

564

565

566 *Procedure*

567 All experimental tasks were written on MATLAB (R2016b, The Mathworks, Natick, MA), using  
568 Psychtoolbox-3<sup>58–60</sup> and ran on MacOS. All tasks were self-paced with no time deadlines. All  
569 participants (except for one, due to technical problems) completed all tasks in the same order.

570 *Facial Expressions Task*

571 The facial expressions task consisted of three parts. In the first part (Figure 1.A), participants saw  
572 32 different pictures of four different actors in pseudorandomized order (see the description of  
573 *Cue images*, below) and imitated each expression as best they could. Participants pressed a key  
574 (the space bar) once they considered that their expression was as close as possible to the actor's  
575 expression. We asked participants to try to match the low-level physical features of the face —  
576 the curvature of the lips, the elevation of the eyebrows — rather than the emotion conveyed by  
577 the expression. Upon pressing the spacebar, the digital camera behind the half-plated mirror took  
578 a picture of the participant's facial expression, and a new trial started. On a separate test, we had  
579 determined that there was a minimum delay of approximately 80 ms between the time of key  
580 press and the time stamp of the image. Accordingly, we included in our instructions to participants  
581 to hold the expression in place after they had pressed the key that would trigger the image  
582 acquisition.

583 The 32 pictures of participants generated in this way served as target images for the second part  
584 of the paradigm. Here, participants saw the target images and tried to reproduce their own  
585 expressions. Once again, we emphasized that the goal was to match the low-level physical  
586 features of the face rather than the emotion conveyed. After each trial, participants used a mouse  
587 to rate their confidence (on a visual analog scale) regarding how well they thought that they had  
588 imitated their own previous expression. Participants saw each of their 32 target expressions  
589 repeated 8 times in random order (256 trials in total). We only revealed that they would have to  
590 reproduce their own expressions after the first part of the experiment was complete. Parts 1 and  
591 2 of the experiment took on average approximately 50 minutes. Before starting part 1, participants  
592 completed four practice trials where they simply imitated pictures of famous celebrities and took  
593 pictures. They did not see the resulting pictures of themselves.

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

### 597 *Cue images*

598 We used 32 different facial expressions as cue pictures (14 from two different male actors, 18  
599 from three different female actors) which would be used to generate participant-specific target  
600 expressions. To prevent participants from producing stereotypical target expressions, we sought  
601 pictures representing expressions that could not be unambiguously categorized as one of the  
602 basic emotions<sup>61</sup>. We selected pictures from the MPI Small Facial Expression Database<sup>28</sup>, which  
603 includes video sequences of expressions based on a method acting protocol in which actors  
604 produce non-standard expressions by imagining themselves in a situation described by a brief  
605 scenario and reacting accordingly. Example descriptions of expressions include: “Somebody  
606 suggests to try something. You hesitate at first, then you agree”, or “You have reached a goal and  
607 you are happy to have accomplished it”. Additionally, we selected still images from the video  
608 sequence that did not correspond to the peak expression, but instead to an intermediate step. As  
609 a result, the cue images could not easily be labeled as stereotypical expressions (e.g., “happy”,  
610 “sad”) for which participants might have a predefined motor program but could instead be  
611 assumed to be the result of an unusual and idiosyncratic combination of gestures. Note that, as  
612 the samples in Figure 1.C show, these cue images were not unnatural grimaces and so the  
613 paradigm remains ecologically valid. We reasoned that these non-canonical expressions would  
614 maximize motor variability, ensuring that confidence ratings could be based only on a true  
615 evaluation of trial-by-trial performance and not on a general knowledge of how reproducible a  
616 given expression was.

### 617 *Visual Task*

618 Each participant completed 200 trials of a visual metacognition task  
619 ([https://github.com/metacoglab/meta\\_dots](https://github.com/metacoglab/meta_dots)). On each trial of this task, two circles enclosing sets  
620 of dots appeared for 200 ms on either side of a central fixation cross (each circle with a radius of  
621 5 degrees of visual angle, located along the middle of the screen, with an eccentricity from the  
622 vertical midline of 5.5 degrees of visual angle). One of the two circles always contained 50 dots  
623 while the other varied in dot number, and the position (left/right) of the circles was randomized on  
624 each trial. In a 2-alternative forced-choice (2AFC) task, participants discriminated which of the

625 circles contained more dots by pressing the left or right arrow keys on the keyboard. The  
626 difference in the number of dots was determined by a pair of interleaved 2-down-1-up adaptive  
627 staircases aimed at fixing performance at around 71% accuracy. After each response, participants  
628 reported their confidence in the accuracy of their own response using the same vertical visual  
629 analog scale that they had used for the two previous tasks rating confidence and similarity for  
630 facial expressions.

631 Before the main visual task, we ran 80 trials of a staircase procedure where participants did only  
632 the discrimination task without rating confidence. Here we also included two interleaved 2-down-  
633 1-up staircases starting from a difference of 3 and 20 dots respectively. One participant  
634 (unintentionally) received feedback about the accuracy of the discrimination task while rating  
635 confidence, so we excluded their data from the analysis. The visual task took approximately 20  
636 minutes. Over all participants, we also excluded 2% of the trials where the reaction times to either  
637 the discrimination task or the confidence rating were faster than 300 ms or slower than 5 s. We  
638 estimated metacognitive efficiency as  $M_{ratio}^{30}$  after scaling and binning confidence into four  
639 discrete confidence levels based on uniform intervals.

#### 640 *Toronto Alexithymia Scale*

641 At the end of the experiment we collected responses to a computerized version of the Toronto  
642 Alexithymia Scale (TAS<sup>62</sup>) running on a browser, and the data were stored locally<sup>63</sup> (jatos.org).  
643 Most participants completed a German version of the scale, except for seven non-German  
644 speakers who completed an English version instead. The TAS-20 consists of 20 items that can  
645 each be answered on a 5-point Likert scale. We considered three out of the four subscales  
646 (Difficulty identifying feelings, Difficulty describing feelings, and Externally-oriented thinking, but  
647 excluded the Daydreaming subscale). We calculated Bayes Factors ( $BF_{10}$ ) for correlations  
648 between these covariates and individual slopes from the estimated models using the *BayesFactor*  
649 package<sup>64</sup> in R (version 3.6.2).

#### 650 *Data processing and analysis*

651 Following the pre-registered plan, we excluded trials from the facial expressions task at the single  
652 participant level if RTs (time between image onset and key press) were above the 95 percentile  
653 for that participant. This cutoff was necessary because we noticed that participants sometimes  
654 laughed at their own picture or got otherwise distracted. This resulted in seven trials excluded

655 from the entire dataset where the time to take a picture was below 300 ms, and a mean lower  
656 threshold of exclusion of 9.43 s (range: 4.0 - 18.0 s).

657 For each of the pictures taken, we obtained the x,y coordinates of landmarks distributed on the  
658 face. In our pre-registered plan we stated that we would estimate the landmark positions using  
659 two different toolboxes and choose the best one to estimate distance based on the quality of the  
660 relationship to the similarity ratings. Instead, due to technical problems in running one of the  
661 toolboxes we opted for the Face Alignment package<sup>65</sup> alone ([https://github.com/1adrianb/face-](https://github.com/1adrianb/face-alignment)  
662 [alignment](#) v.1.0.0), a fully automated deep-learning based face alignment network (FAN) that  
663 places landmarks on the pictures. We used the *face-alignment* package together with *scikit-image*  
664 *and pytorch* to extract the landmarks from the faces, running on Python v3 in a Jupyter notebook  
665 v5. The face-alignment package automatically places 68 landmarks on the face and excludes the  
666 forehead and hairline.

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

### 681 *Bayesian mixed models*

682 In our central analysis we computed metacognitive access to facial expressions as the  
683 relationship between confidence ratings and performance. We take the distance as an inverse  
684 measure of performance: if a response image closely matches the target image, the distance  
685 between them will be small. Furthermore, a strong negative relationship between confidence



686 ratings and distance will indicate that participants had metacognitive access to their own facial  
687 expressions, as they (correctly) provided low ratings in trials where the two images differed the  
688 most. Conversely, no relationship between confidence and distance would indicate that  
689 participants had no metacognitive access to their own expressions.

690 Because finding no relationships between variables was a plausible outcome from our analyses,  
691 we used Bayesian statistics that, unlike frequentist statistics, provide evidence for the null  
692 hypotheses. We analyzed the data using Bayesian mixed models created in Stan ([http://mc-  
693 stan.org/](http://mc-stan.org/)) through the *brms* package<sup>66,67</sup>. In all cases, we ran 4 chains with 15,000 iterations,  
694 5,000 burn-in samples each, and no thinning. We checked for convergence by visually examining  
695 the MCMC chains and ensured that the scale reduction factor (Rhat) of all models was equal or  
696 close to 1. We considered that ratings might vary across participants both in their mean and in  
697 their relationship to the landmark distance, and that different facial expressions might vary in their  
698 associated difficulty to both reproduce (leading to greater variability in the landmark distance) and  
699 to rate (leading to differences in the ratings). Thus, in all models and unless otherwise stated, we  
700 included random slopes for both participants and facial expressions (see the explicit model syntax  
701 in Table 1). We extracted the participant-wise random slopes using the *mixedup* package  
702 (<https://m-clark.github.io/mixedup/>).

703 Because, to the best of our knowledge, there was no existing data to inform our priors, we followed  
704 recommendations<sup>68</sup> to use heuristics to define prior distributions. We built the prior for the slope  
705 between ratings and distance based on the ratio-of-scales heuristic: we found that the range of  
706 (log-transformed) distances was approximately 3 a.u. (arbitrary units), whereas the range of  
707 confidence ratings is 1 point (minimum: 0). Therefore we used a normal prior centered on 0 with  
708 an  $SD = \frac{1}{3}$  (which corresponds to the ratio between confidence range and distance range) for the  
709 slope parameter. To find a prior for the model intercept we followed the logic behind the room-to-  
710 move heuristic. Note that raw distances ranged between [131.36 - 2493.78] a.u., hence the  
711 expected rating at 0 distance (i.e., perfect performance) can be well approximated by the  
712 expected rating at distance = 1, which corresponds to the intercept in a linear model with log-  
713 transformed distances. We reasoned that a participant with maximum metacognitive performance  
714 would consistently rate their confidence as 1, when the distance between the two images was 0.  
715 Because we realistically expect participants to have (at most) less than perfect metacognitive  
716 access to their own expressions, we centered the prior at 0.8 with an  $SD = 0.5$ . Following a similar  
717 logic, we set the prior slope between the two ratings to be centered at 0 with  $SD = 1$ , and an



718 intercept of 0 with an SD =  $\frac{1}{2}$ . For all models, we report the estimate, its associated error mean,  
 719 the 95% credibility interval (CI), and the  $BF_{10}$ , estimated using the *bayestestR* package<sup>69</sup>, to  
 720 compare each model against its null counterpart, containing the same random effects structure  
 721 but not the fixed effect of interest. We also include the posterior draws for each participant in  
 722 relation to the region of practical equivalence (ROPE). We set the ROPE to a default range from  
 723 -0.1 to 0.1 of a standardized parameter, which corresponds to a negligible effect size<sup>70,71</sup>. Finally,  
 724 we estimated  $R^2$  values as implemented by the *brms* package<sup>72</sup>.

725

726 **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 Appendix 1- Figure 5
The (mean) similarity ratings are inversely related to the distance between two images	meanSimilarity ~ logDistance + (1 + logDistance   participantID) + (1   expressionID)	Figure 3 Appendix 1- Figure 7
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)	-

727

728

729 We computed metacognitive access to faces using only linear regression and estimated the  
 730 correlation with visual Mratios, deviating from the pre-registered plan. We initially planned to also  
 731 calculate the area under a type-2 ROC curve (AUROC2) by arbitrarily assuming that first-order

732 performance on the Faces task was at 70% accuracy and by classifying trials with distances  
733 above the corresponding threshold as “incorrect”. This analysis had the advantage that it would  
734 have allowed us to correlate metacognitive performance measured on the same scale for both  
735 tasks (Faces and Visual), but we later reasoned that it would make the results less easily  
736 interpretable while not adding explanatory power and therefore decided to omit it.

### 737 *Machine learning models*

738 Using Python v3, and *scikit-learn*, we created a separate model for each subject wherein, first,  
739 each landmark distance was determined by (x,y) coordinate differences between the two images.  
740 We further decomposed the differences into four zero- or positive features (one for each cardinal  
741 direction). This allowed different directions of movement to be weighted differently by the model.  
742 We normalized each feature by dividing it by its median. Then, we applied dimensionality  
743 reduction using principal component analysis with a set number of principal components (66, or  
744 approximately 90% of the variance from all subjects) in order to avoid multicollinearity among the  
745 features. Finally, a least squares linear regression model was trained for each participant using  
746 trial-wise leave-one-out cross-validation.

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

$$750 \quad w_f = \sum_{c=1}^{66} \lambda_{f,c} \times \omega_c \quad (1)$$

751 Where  $\lambda_{f,c}$  is the loading of feature  $f$  on principal component  $C$ , and  $\omega_c$  is the ML model’s  
752 weighting of principal component  $C$ .

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

$$754 \quad RSSQ_{weighted} = \sqrt{\sum_{f=1}^{272} w_f \times f^2} \quad (2)$$

755 Where  $w_f$  denotes the weights for each feature  $f$ , which is in turn the difference between response  
756 and target images for each cardinal direction, for a given landmark, if the difference was positive,  
757 and 0 otherwise. The 272 features result from decomposing 68 landmarks into the four cardinal  
758 directions. Note that unlike the case for the Euclidean distance, where distances were forced to

759 be positive and each of them had an effective weight of 1, here we allowed the feature weights to  
760 be signed. For those cases where the term under the square root was negative, we calculated  
761 the root of the absolute value and then reversed the sign. Note that  $RSSQ_{weighted}$  is now better  
762 interpreted as a measure of performance, and not distance: because the ML-derived weights  
763 already account for the negative relationship between distance and confidence,  $RSSQ_{weighted}$  is  
764 expected to show a positive relationship to confidence.

765 We obtained adjusted R2 for each (participant-specific) model values and compared them using  
766 a Bayesian Wilcoxon Signed-Rank test<sup>73</sup> as implemented in JASP<sup>74</sup> v0.14 with 10,000 MCMC  
767 samples and 5 chains, and a default Cauchy prior.

768

769

770 **Acknowledgements**

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

779 **Competing Interests**

780 The authors declare no competing interests.

781 **Data and Code Availability**

782 Raw data (excluding images from participants and any other personally identifiable information)  
783 along with reproducible analysis scripts are available under  
784 [https://gitlab.com/elisa.filevich/cistonetal\\_metacognitionoffacialexpressions](https://gitlab.com/elisa.filevich/cistonetal_metacognitionoffacialexpressions).

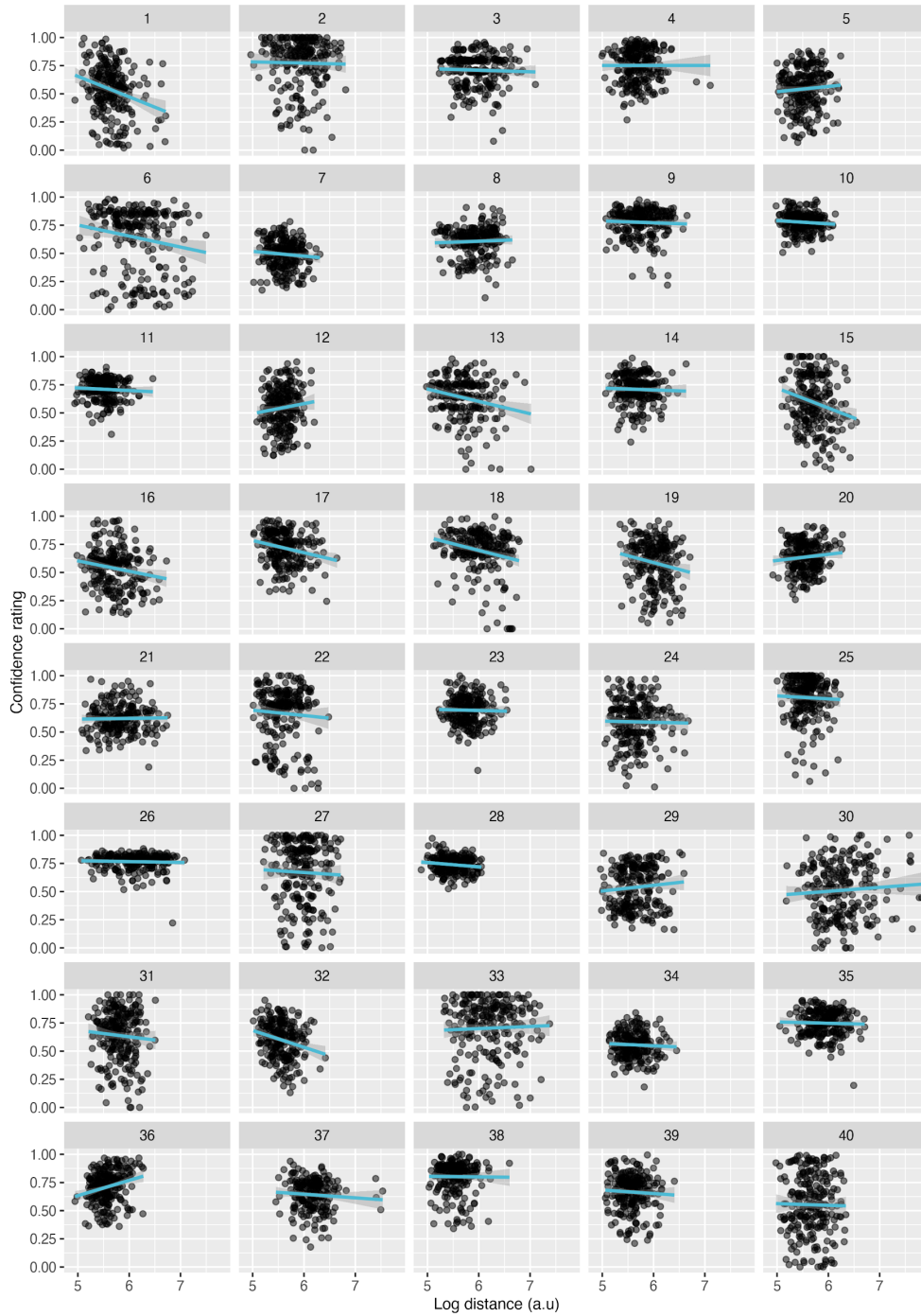
785

786 **Appendix 1**

787 **Supplementary Information**

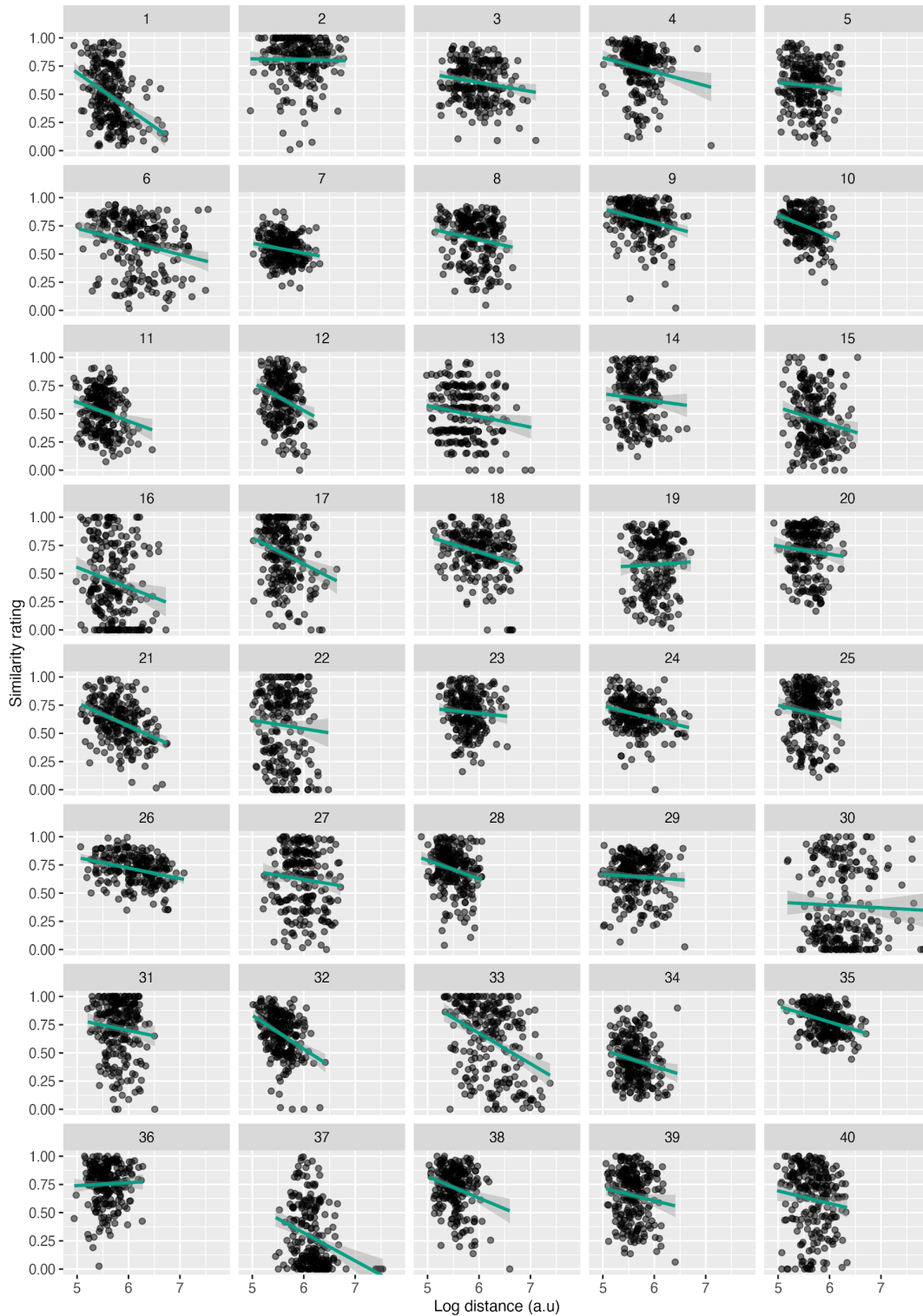
788 Appendix 1-Figures 1-3 show the single-trial data (and linear regressions at the single-participant

789 level) for the data reported in the main text.



790

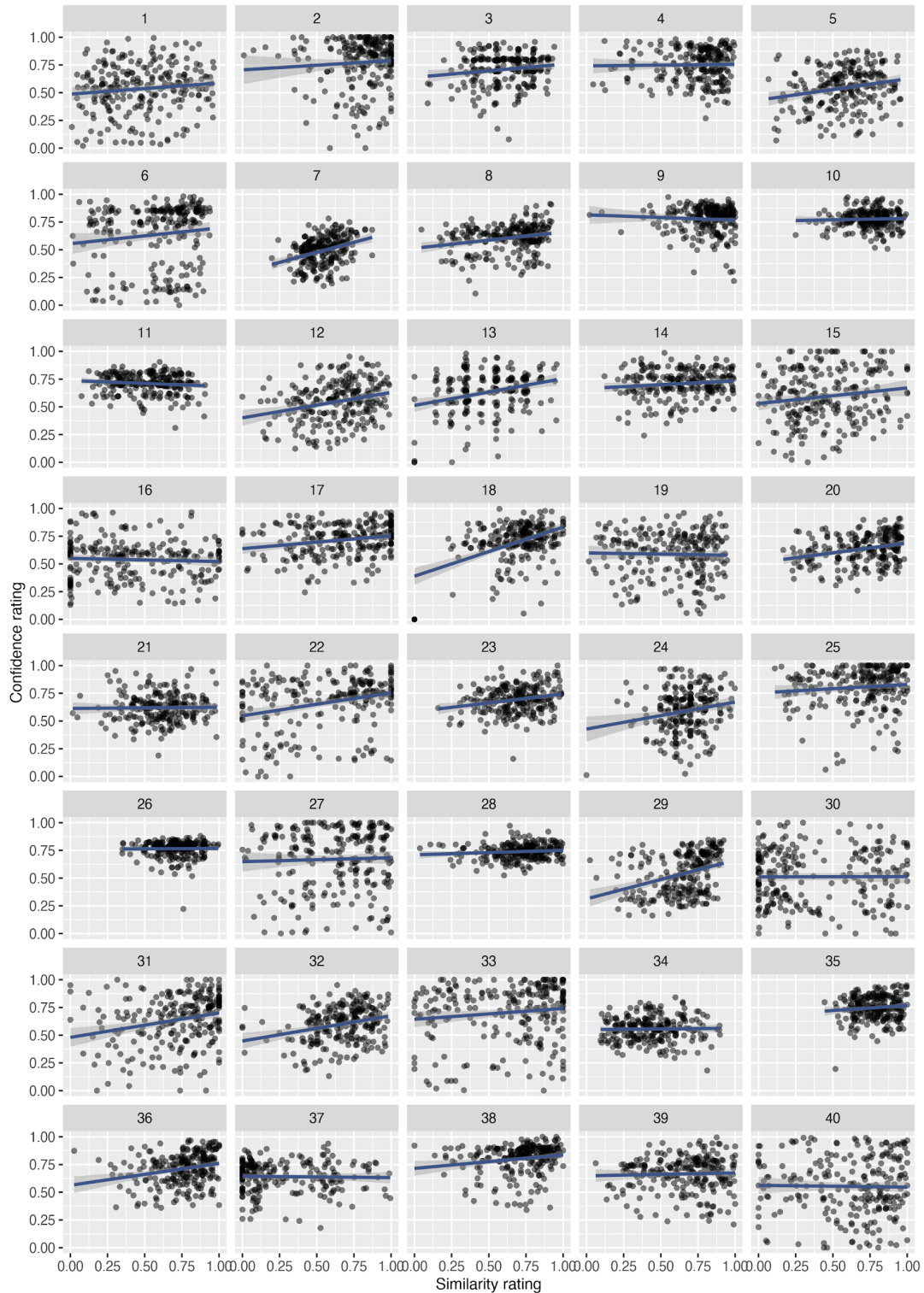
791 **Appendix 1-Figure 1: Linear regressions of confidence ratings as a function of distance (at**  
792 **the single-participant level, for Experiment 2).** Note that all statistical inferences are made on the  
793 **basis of Bayesian linear regressions, and this plot is for illustrative purposes only.**



794



795 **Appendix 1-Figure 2: Linear regressions of similarity ratings as a function of distance** (at  
796 the single-participant level, for Experiment 2). Note that all statistical inferences are made on the  
797 basis of Bayesian linear regressions, and this plot is for illustrative purposes only.



798

799 **Appendix 1-Figure 3: Linear regressions of confidence vs. similarity ratings** (at the single-  
800 participant level, for Experiment 2). Note that all statistical inferences are made on the basis of  
801 Bayesian linear regressions, and this plot is for illustrative purposes only.

802

803

## 804 **Supplementary analyses**

### 805 *Machine Learning - Effects of RT*

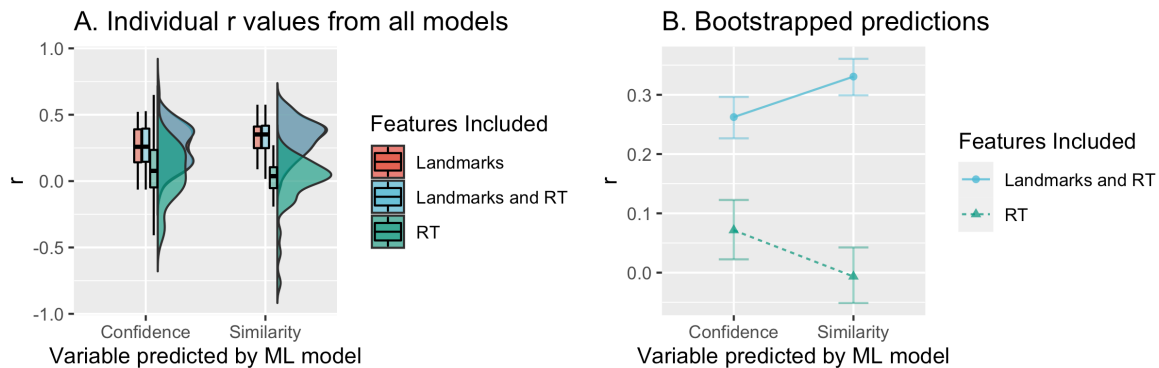
806 Because confidence is known to correlate negatively with response times<sup>75,76</sup> (RT), we first  
807 explored a potential relationship between the two and asked whether RTs could have served as  
808 a proxy for performance. We ran a Bayesian linear regression model of participants' confidence  
809 ratings including the RT as a fixed effect and random intercepts for participant and facial  
810 expression, as well as a per-participant random slope for RT. We based the prior distribution for  
811 this analysis on previous data from our group<sup>3</sup>, and set a wide prior for the intercept centered at  
812 around confidence = 0.8 with SD = 0.5, and a prior for the slope centered on 0 with an SD = 0.20,  
813 which roughly corresponds to the ratio-of-scales. We confirmed that there was a small but  
814 consistent effect of RT on confidence ( $M = -0.01 \pm 0.00$ ,  $CI = [-0.02, -0.00]$ ,  $BF_{10} = 5.66 \times 10^{39}$ ,  $R^2$   
815 = 0.20).

816 To evaluate whether the landmarks informed confidence ratings above and beyond RT, we  
817 compared the resulting individual  $r$  values from the ML models (including both RTs and the  $x$ ,  $y$   
818 positions of the landmarks) to those of a ML model including only RTs as their single feature  
819 (Appendix 1-Figure 4). A non-parametric ANOVA computed with the *ez* package for R revealed  
820 an interaction effect ( $p = 0.001$ ) on the  $r$  values between the variable predicted (confidence or  
821 similarity) and the features included in the model. Note that we do not interpret the main effect of  
822 the number of features included, as these are known to inflate the  $r$  values. Instead, we focus on  
823 the interaction effect. In particular, the interaction revealed higher  $r$  values for the models of  
824 similarity that included both landmarks and RT as compared to confidence (Wilcoxon signed rank  
825 test,  $p = 0.015$ ), but lower  $r$  values for models including RTs only (Wilcoxon signed rank test,  $p =$   
826 0.068). This pattern of results is consistent with landmarks being predictive of confidence ratings,  
827 above and beyond RTs To understand the contribution of RTs relative to the other features, we  
828 obtained the rank of importance of RTs within the ML model. We found that RTs varied in  
829 importance with each participant, but ranged between the 5th and the 100th percentile (Mean =



830 71.15, Median = 93.41), suggesting that, for some participants, RTs were the most reliable piece  
831 of information for confidence ratings, even if the variance explained by them was very low.

832



833

834 **Appendix 1-Figure 4: r values resulting from the linear regression models built using ML. A.**  
835 Distributions of individual r values (summarized with boxplots and violin plots) for models on confidence or  
836 similarity ratings, using different sets of features. **B. Bootstrapped predictions from a non-parametric**  
837 **ANOVA** for models of confidence and similarity built using RTs alone or also including landmark  
838 information.

839

## 840 **Supplementary Pilot Experiment**

841 Prior to pre-registering and collecting the data reported in the main text, we collected a smaller  
842 dataset as a pilot. Because there are some important differences in the experimental details, we  
843 report the methods and results as supporting information, that serve as a conceptual replication.

## 844 **Supplementary Methods - Pilot Experiment**

845 The methods for the pilot experiment were largely similar to those of the main experiment. We  
846 only describe here the differences between the two.

### 847 *Participants*

848 Thirteen healthy participants took part in the experiment after giving informed consent (seven  
849 female, mean  $\pm$  SD: 24  $\pm$  3 years). One participant was excluded from the analysis because four  
850 external judges agreed (see below) that there was no variability in their facial expressions.  
851 Participants had no recent history of psychiatric disease. The local ethics committee approved all  
852 procedures, which conformed to the Declaration of Helsinki.

853 *Apparatus*

854 Behind the mirror, a digital camera (Logitech HD C310) connected to the computer captured  
855 images of the participants' facial expressions. The apparatus was similar to the one described in  
856 the main text, with some minor differences. Unlike in the main experiment, where the screen  
857 rested on top of the stimulus box and projected downwards, the screen lay on the table for the  
858 pilot experiment and projected upwards. From the point of view of the participants, this did not  
859 change the visual display.

860 *Procedure*

861 The task was programmed on GNU Octave and displayed stimuli using Psychtoolbox-3<sup>58-60</sup>, and  
862 ran on a Linux Debian (Gnome 3.4.2) operating system. The task consisted of two parts (not  
863 three). Participants saw 30 (not 32) different photos of four different actors and imitated each  
864 expression as best they could. The images were presented in one of five possible pre-defined  
865 random orders to each participant. As in the main experiment, participants first generated 30  
866 participant-specific pictures that then served as target images for the second part of the paradigm.  
867 After each trial, participants rated their confidence (on a scale from 1 to 6) regarding how well  
868 they thought that they had imitated their own previous expression. To make the task intuitive, we  
869 kept the mapping of the scale consistent with the German education system, where the best grade  
870 is a 1.0. We then reversed the ratings for further analyses, so that a rating of 6 corresponds to  
871 the highest confidence. In all cases, we recorded each picture taken, the response time (RT,  
872 measured as the time between image onset and key press) and participants' confidence ratings.  
873 Participants saw each of their 30 target expressions repeated 8 times in random order, for a total  
874 of 240 trials. We only revealed that they would have to reproduce their own expressions after the  
875 first part of the experiment was complete. On average, the experiment took approximately 50  
876 minutes.

877 *Data Processing and Analysis*

878 We first used the Face Modeling GUI<sup>78</sup> to manually position 99 landmarks on their corresponding  
879 locations on a small subset of images (3-5) of each participant. The Face Modeling GUI then uses  
880 the location of these landmarks to automatically find their optimal locations in the remaining  
881 images. After the automatic fit, the landmarks in each of the images were corrected manually. In  
882 this way, we reduced the dimensionality of each of the 240 response images along with the 30  
883 target images for each of the participants to 99 pairs of (x,y) coordinates. We then did the same

884 Procrustes rigid-alignment as described in the main text, with 5 reference points instead of 3 (the  
885 inner and outer corners of each eye and a point just below the nose). We did not use a mean  
886 reference face, but instead minimized the distance of each response picture to its corresponding  
887 target picture.

#### 888 *Similarity ratings by external judges*

889 Unlike what was the case in the main experiment, here four independent judges (student research  
890 assistants) rated the image pairs for similarity on a scale from 1 to 6, exactly like the one the  
891 participants had used.

#### 892 *Data processing and analysis*

893 Here as well we followed recommendations<sup>68</sup> to use heuristics to define prior distributions. We  
894 built the prior for the slope based on the ratio-of-scales heuristic: we found that the range of (log-  
895 transformed) distances was approximately 4.93 a.u. (arbitrary units), whereas the maximum  
896 possible range of ratings is 5 points (maximum: 6, minimum: 1). The ratio between the two is  
897 approximately 1, so we used a normal prior centered on 0 with an SD = 1 for the slope parameter.  
898 To find a prior for the model intercept (the expected rating at 0 distance, i.e., perfect performance),  
899 we followed the logic behind the room-to-move heuristic. We reasoned that a participant with  
900 maximum metacognitive performance would consistently rate their confidence as 6, when the  
901 distance between the two images was 0. Because we realistically expect participants to have (at  
902 most) less than perfect metacognitive access to their own expressions, we centered the prior at  
903 4 with an SD = 3.

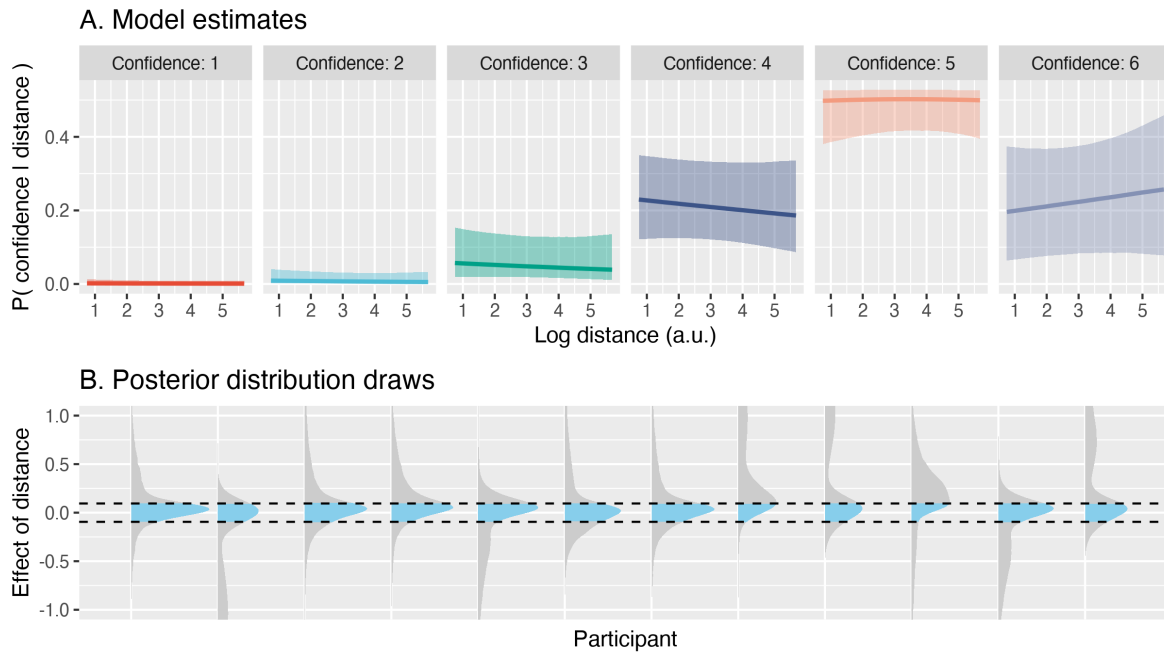
904

#### 905 **Supplementary Results - Pilot Experiment**

906 Because ratings were not on a visual analog scale but instead on a Likert scale, we first quantified  
907 our participants' metacognitive access to their own facial expressions using an ordinal Bayesian  
908 mixed-effects regression model of participants' confidence ratings. The model included the log-  
909 transformed landmark distances as a fixed effect (for all 99 landmarks combined) as well as  
910 random intercepts for participant and facial expression (See Appendix 1-Table 1). The estimated  
911 ordinal regression coefficient was indistinguishable from 0 ( $M = 0.04 \pm 0.07$ ,  $CI = [-0.10, 0.16]$ )  
912 and the evidence ratio favoured the (point) null hypothesis of no relationship between confidence  
913 and distance ( $BF_{10} = 0.082$ ). This is illustrated by the flat probability profiles for each rating shown

914 in Appendix 1-Figure 5.A: while there were differences in the overall probability of each confidence  
915 rating (e.g. a rating of 5 occurring more often than others), the probability of a participant providing  
916 a given confidence rating was similar over all landmarks distances (see also Appendix 1-Figure  
917 6 for the single-participant data).

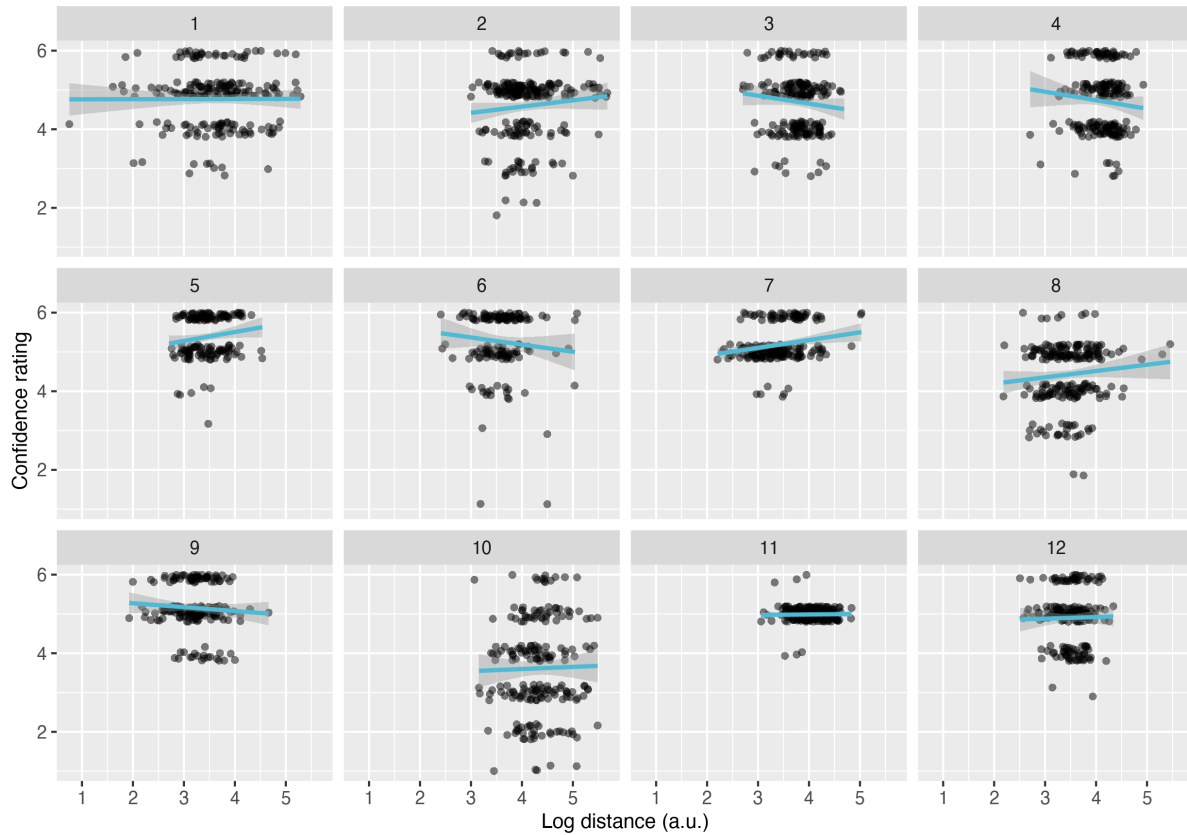
918



919

920 **Appendix 1-Figure 5. No evidence for metacognitive access to facial expressions (A.)** Group effects  
921 reflecting mean metacognitive access, namely the relationship between confidence ratings and distance  
922 between two images (inverse of performance). While different confidence ratings appear at different  
923 frequencies in the data, they do not vary with distance as would be expected if participants had  
924 metacognitive access to their own expressions. Solid lines represent the mean of the posterior draws, the  
925 shaded regions represent the 95% credibility interval. **(B.)** Posterior draws for each subject, shown in  
926 relation to the ROPE. Note that the y-axis is clipped to better display the distributions around the ROPE  
927 and therefore excludes the long tails of some of the distributions.

928



929

930 **Appendix 1-Figure 6: Linear regressions of confidence ratings as a function of distance** (at the  
931 single-participant level, for Experiment 1). Note that all statistical inferences are made on the basis of  
932 Bayesian ordinal regressions, and this plot is for illustrative purposes only.

933

934 That there is no observable relationship between the combined landmark distances and  
935 participants' confidence ratings suggests, at face value, that participants did not have access to  
936 the details of their face. However, other alternative explanations must be considered. First, it is  
937 possible that the landmark distance measure, which is essentially the result of an algorithm  
938 placing landmarks based on pixel information plus some rigid transformations, may not capture  
939 enough information relevant for the similarity of two faces. If this were true, there should also be  
940 no relationship between the landmark distance and the similarity ratings provided by external  
941 judges looking at each image pair side by side. In fact, this was not the case. To evaluate this  
942 possibility we used a Bayesian linear mixed-effects regression model on the mean of four judges.  
943 The model included the same fixed and random effects factors as in the mixed ordinal model  
944 above (namely, the log-transformed distance as a fixed effect, intercepts for participant and  
945 expression as random effects, and a by-participant random slope for the fixed effect). However,

946 unlike in the mixed-effects regression model on participants' confidence ratings, we did find a  
947 consistent negative relationship between the distance and the similarity ratings ( $M = -0.54 \pm 0.06$ ,  
948  $CI = [-0.67, -0.42]$ ,  $BF_{10} = 71551.85$ ). That is, unlike the confidence ratings, the similarity ratings  
949 did show a consistent and (as expected) negative relationship to the distance (Appendix 1-Figure  
950 7.B and Appendix 1-Figure 8). This suggests that the distance did carry some information about  
951 face similarity meaningful to human observers. For illustration purposes only, we repeated the  
952 analysis between similarity ratings and distance but this time rounded the mean ratings and ran  
953 an ordinal model (Appendix 1-Figure 7.B). We do not make any statistical inferences from this  
954 analysis but use it only to illustrate the differences between the probability profiles of the ratings  
955 that vary with distance and those who do not (Appendix 1-Figure 5.A).

956 As in the main experiment, here we also found that distance was related to similarity ratings.  
957 Neither the procedure to estimate distance nor the similarity ratings were identical between the  
958 two experiments (two different algorithms placed 68 or 99 landmarks respectively; and either the  
959 participants themselves or external judges rated similarity), which validated our measure of  
960 distance by showing that it does not depend on idiosyncratic properties of the algorithm or the  
961 rating process.

962

963

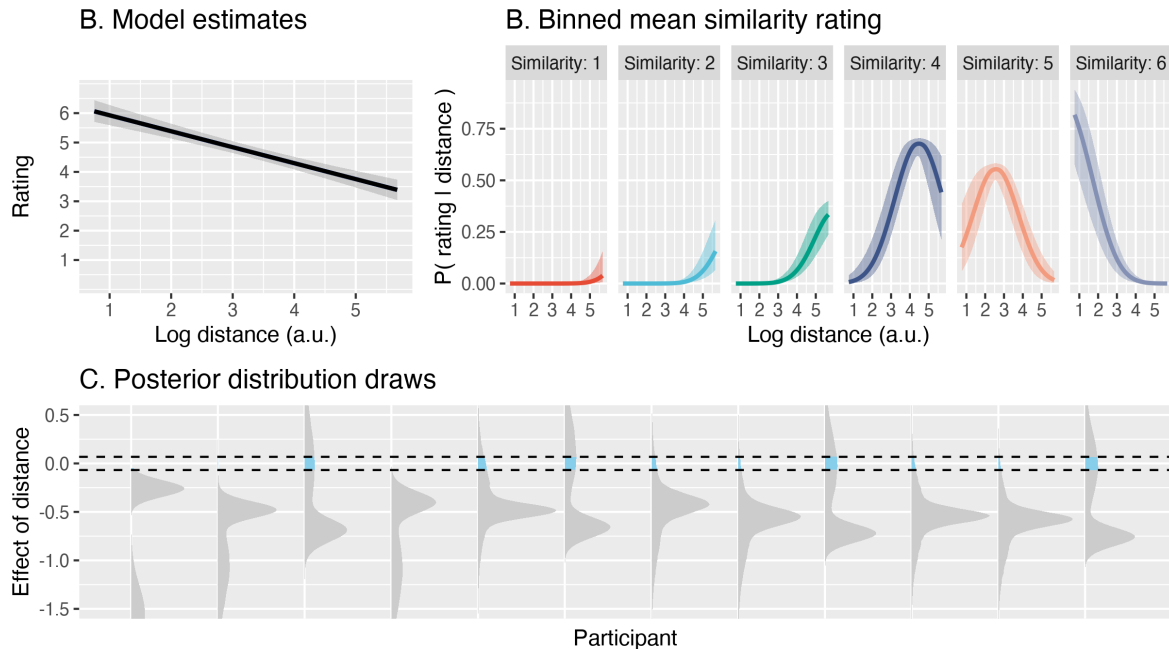
964

965

966

967

968



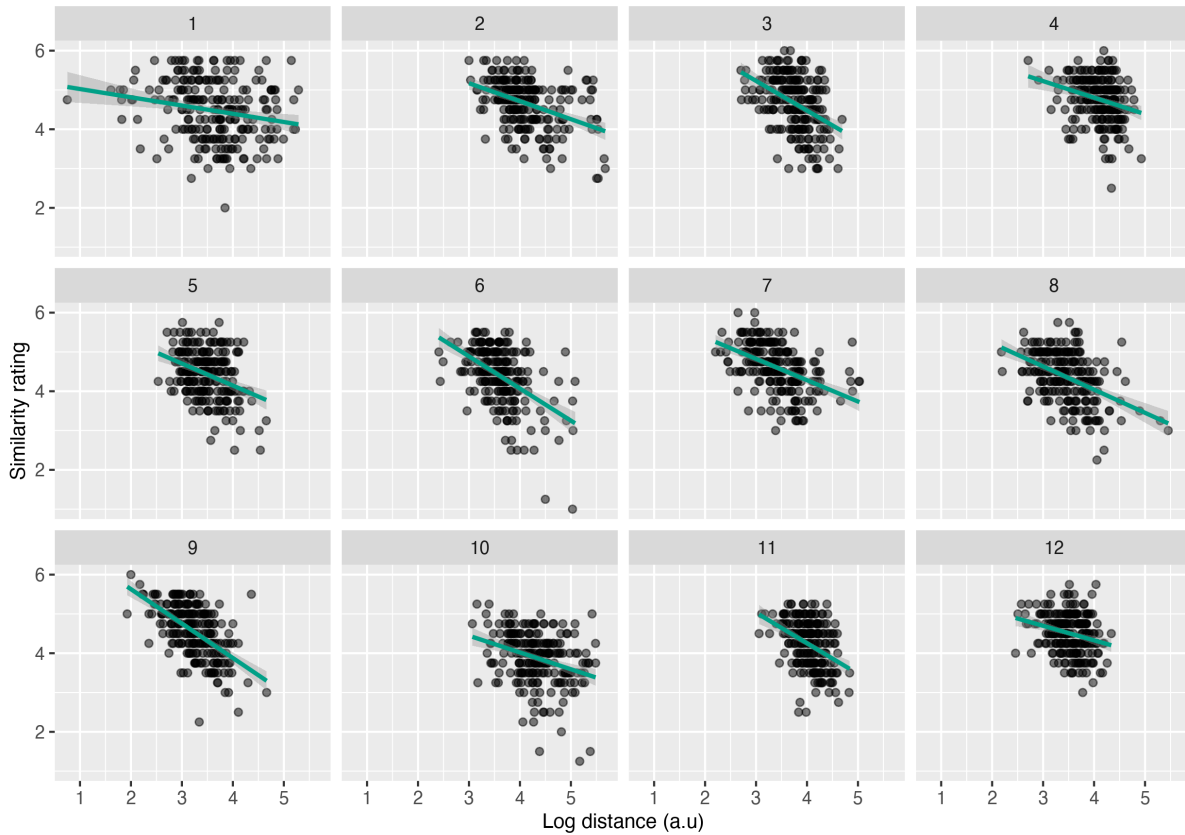
969

970 **Appendix 1-Figure 7. The distance between two images captures relevant information. (A.)** Group  
971 effects reflecting the information contained in the distance between two images, namely the relationship  
972 between the mean similarity ratings of four judges (who viewed each image pair side-by-side) and distance  
973 between two images. There is a clear relationship between mean similarity and distance, suggesting that  
974 distance contains meaningful variability. **(B.)** An ordinal version of the model shown in (A.) presented only  
975 to illustrate the contrast to Appendix 1-Figure 5. For both panels (A.) and (B.), solid lines represent the  
976 mean of the posterior draws, and the shaded regions represent the 95% credibility interval. **(C.)** Posterior  
977 draws for each subject, shown in relationship to the region of practical equivalence (ROPE). Note that the  
978 y-axis is clipped to better display the distributions around the ROPE and therefore excludes the long tails  
979 of some of the distributions.

980

981





982

983 **Appendix 1-Figure 8: Linear regressions of mean similarity ratings as a function of distance.** The y  
984 axis represents the mean of all four judges, and each panel represents a single participant, from the pilot  
985 experiment).

986

987 Importantly, we note that the relationships shown in Appendix 1-Figure 7, panels B. and C. and  
988 Appendix 1-Figure 8 are the result of taking the mean of four judges. Thus, this significant  
989 relationship might be accounted for by a Wisdom of the crowds effect, whereby the mean of the  
990 estimates of many individuals is better than any single individual's estimate<sup>79</sup>. To evaluate this  
991 possibility, we ran Bayesian ordinal mixed regressions for the similarity ratings of each individual  
992 judge. In all cases, we found that the estimates were negative, and clearly different from 0 (all  
993 mean slope estimates  $< -0.53$ , all  $BF_{10} > 554$ . See Appendix 1-Table 1 and Appendix 1-Figure 9  
994 for the model predictions and single-participant data, respectively).

995

996

997

998

999

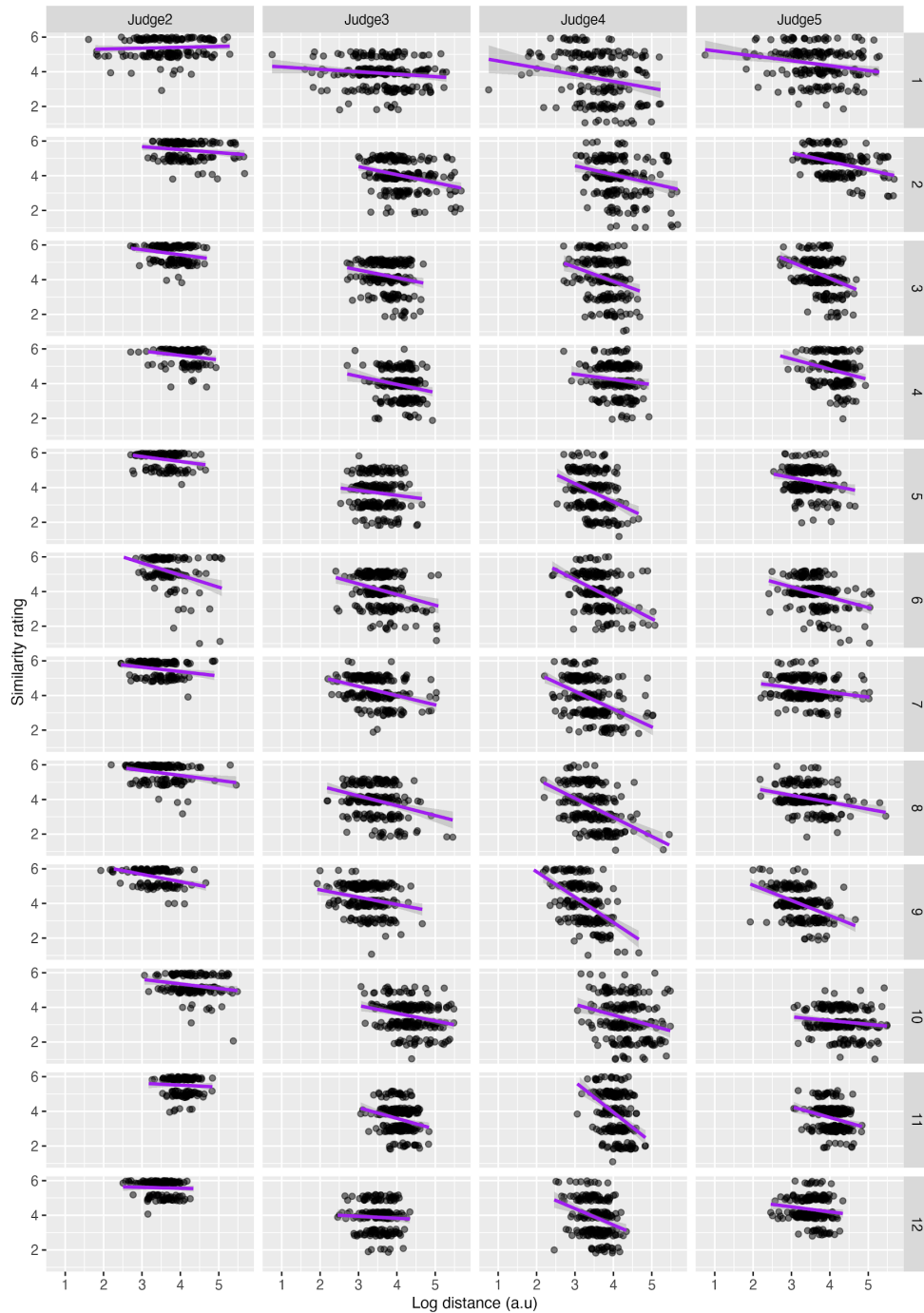
1000 **Appendix 1-Table 1: Bayesian ordinal model estimates for the effect of distance on**  
1001 **similarity.** Each row contains the estimates for a single judge (and all participants in the pilot  
1002 experiment) and includes the mean, standard deviation, 95% credibility interval and  $BF_{10}$  relative  
1003 to the null model.

Judge	Effect of distance on similarity rating ( $M \pm SD$ )	95% CI	$BF_{10}$
1	-0.53 $\pm$ 0.11	[-0.75, -0.32]	554.91
2	-0.55 $\pm$ 0.07	[-0.70, -0.41]	15175.95
3	-0.89 $\pm$ 0.11	[-1.11, -0.67]	57346.06
4	-0.73 $\pm$ 0.11	[-0.95, -0.51]	7608.01

1004

1005

1006



1007

1008 **Appendix 1-Figure 9: Linear regressions of similarity ratings from each judge as a function**  
1009 **of distance.** Each panel represents a single judge (columns) and participant (rows) from the pilot  
1010 experiment.

1011

1012

1013

## 1014 **Brief Discussion - Pilot Experiment**

1015 Briefly, these results suggest that participants did not have access to the low-level details of their  
1016 own facial expressions. This could not be explained by any of the several alternatives we  
1017 explored: neither lack of variability in performance or a poor benchmark measure (similarity  
1018 ratings from external judges did show a clear relationship to the landmark distances, Appendix 1-  
1019 Figure 7) nor the fact that the confidence ratings were from a single person (individual judges'  
1020 similarity ratings also showed the same clear relationship, Appendix 1-Table 1) proved to be  
1021 sufficient to explain the apparent null relationship between confidence ratings and distance.

1022 Despite these controls, alternative explanations remain in principle possible, which we  
1023 incorporated when designing the experiment reported in the main text. First, participants provided  
1024 their confidence ratings on a Likert scale from 1-6. Perhaps, a continuous scale would have given  
1025 them the opportunity to provide more nuanced and precise ratings. Second, metacognitive ability  
1026 — in both the visual<sup>80</sup> and the motor<sup>41</sup> domains — is known to vary in the normal population.  
1027 Perhaps, due to mere chance, participants with poor general metacognitive access to their own  
1028 facial expressions were overrepresented in the relatively small sample of 12 participants. Hence,  
1029 to exclude the possibility that our conclusions in this pilot experiment resulted from a small (and  
1030 potentially biased) sample of 12 participants, we tested a larger sample. Third, we considered the  
1031 possibility that the differences we observed in this pilot experiment between the relationships of  
1032 distance and confidence and similarity ratings could be attributed to differences in metacognitive  
1033 traits between groups of individuals. We therefore did not recruit external judges but asked the  
1034 same participants to rate their own performance in the image pairs.

1035 **References**

- 1036 1. Kal, E., Prosée, R., Winters, M. & Kamp, J. van der. Does implicit motor learning lead to  
1037 greater automatization of motor skills compared to explicit motor learning? A systematic  
1038 review. *PLOS ONE* **13**, e0203591 (2018).
- 1039 2. Kleynen, M. *et al.* Using a Delphi technique to seek consensus regarding definitions,  
1040 descriptions and classification of terms related to implicit and explicit forms of motor  
1041 learning. *PloS One* **9**, e100227 (2014).
- 1042 3. Taylor, J. & Ivry, R. Implicit and Explicit Processes in Motor Learning. 63–87 (2013)  
1043 doi:10.7551/mitpress/9780262018555.003.0003.
- 1044 4. MacIntyre, T., Igou, E. R., Campbell, M. J., Moran, A. P. & Matthews, J. Metacognition and  
1045 action: a new pathway to understanding social and cognitive aspects of expertise in sport.  
1046 *Front. Psychol.* **5**, (2014).
- 1047 5. Proske, U. & Gandevia, S. C. The Proprioceptive Senses: Their Roles in Signaling Body  
1048 Shape, Body Position and Movement, and Muscle Force. *Physiol. Rev.* **92**, 1651–1697  
1049 (2012).
- 1050 6. Sherrington, C. S. *The integrative action of the nervous system.* (Scribner, 1906).
- 1051 7. Tuthill, J. C. & Azim, E. Proprioception. *Curr. Biol.* **28**, R194–R203 (2018).
- 1052 8. Goodwin, G. M., McCloskey, D. I. & Matthews, P. B. The contribution of muscle afferents to  
1053 kinaesthesia shown by vibration induced illusions of movement and by the effects of  
1054 paralysing joint afferents. *Brain J. Neurol.* **95**, 705–748 (1972).
- 1055 9. Lackner, J. R. SOME PROPRIOCEPTIVE INFLUENCES ON THE PERCEPTUAL  
1056 REPRESENTATION OF BODY SHAPE AND ORIENTATION. *Brain* **111**, 281–297 (1988).
- 1057 10. Craske, B. & Crawshaw, M. Shifts in kinesthesia through time and after active and passive  
1058 movement. *Percept. Mot. Skills* **40**, 755–761 (1975).

- 1059 11. Fuentes, C. T. & Bastian, A. J. Where is your arm? Variations in proprioception across  
1060 space and tasks. *J. Neurophysiol.* **103**, 164–171 (2010).
- 1061 12. Gritsenko, V., Krouchev, N. I. & Kalaska, J. F. Afferent input, efference copy, signal noise,  
1062 and biases in perception of joint angle during active versus passive elbow movements. *J.*  
1063 *Neurophysiol.* **98**, 1140–1154 (2007).
- 1064 13. Limanowski, J. & Blankenburg, F. Integration of Visual and Proprioceptive Limb Position  
1065 Information in Human Posterior Parietal, Premotor, and Extrastriate Cortex. *J. Neurosci.* **36**,  
1066 2582–2589 (2016).
- 1067 14. Ruttle, J. E., Hart, B. M. & Henriques, D. Y. P. The fast contribution of visual-  
1068 proprioceptive discrepancy to reach aftereffects and proprioceptive recalibration. *PLOS*  
1069 *ONE* **13**, e0200621 (2018).
- 1070 15. Sober, S. J. & Sabes, P. N. Flexible strategies for sensory integration during motor planning.  
1071 *Nat. Neurosci.* **8**, 490–497 (2005).
- 1072 16. van Beers, R. J., Wolpert, D. M. & Haggard, P. When Feeling Is More Important Than  
1073 Seeing in Sensorimotor Adaptation. *Curr. Biol.* **12**, 834–837 (2002).
- 1074 17. van Beers, R. J., Sittig, A. C. & van der Gon Denier, J. J. How humans combine  
1075 simultaneous proprioceptive and visual position information. *Exp. Brain Res.* **111**, 253–261  
1076 (1996).
- 1077 18. Clark Weeden, J., Trotman, C.-A. & Faraway, J. J. Three Dimensional Analysis of Facial  
1078 Movement in Normal Adults: Influence of Sex and Facial Shape. *Angle Orthod.* **71**, 132–140  
1079 (2001).
- 1080 19. Coulson, S. E., Croxson, G. R. & Gilleard, W. L. Quantification of the Three-Dimensional  
1081 Displacement of Normal Facial Movement. *Ann. Otol. Rhinol. Laryngol.* **109**, 478–483  
1082 (2000).

- 1083 20. Bègue, I. *et al.* Confidence of emotion expression recognition recruits brain regions outside  
1084 the face perception network. *Soc. Cogn. Affect. Neurosci.* **14**, 81–95 (2019).
- 1085 21. Chen, B., Mundy, M. & Tsuchiya, N. Metacognitive Accuracy Improves With the Perceptual  
1086 Learning of a Low- but Not High-Level Face Property. *Front. Psychol.* **10**, 1712 (2019).
- 1087 22. Lapate, R. C., Samaha, J., Rokers, B., Postle, B. R. & Davidson, R. J. Perceptual  
1088 metacognition of human faces is causally supported by function of the lateral prefrontal  
1089 cortex. *Commun. Biol.* **3**, 1–10 (2020).
- 1090 23. Shea, N. *et al.* Supra-personal cognitive control and metacognition. *Trends Cogn. Sci.* **18**,  
1091 186–193 (2014).
- 1092 24. Fuentes, C. T., Runa, C., Blanco, X. A., Orvalho, V. & Haggard, P. Does My Face FIT?: A  
1093 Face Image Task Reveals Structure and Distortions of Facial Feature Representation. *PLoS*  
1094 *ONE* **8**, e76805 (2013).
- 1095 25. Fuentes, C. T., Longo, M. R. & Haggard, P. Body image distortions in healthy adults. *Acta*  
1096 *Psychol. (Amst.)* **144**, 344–351 (2013).
- 1097 26. Longo, M. R. & Haggard, P. An implicit body representation underlying human position  
1098 sense. *Proc. Natl. Acad. Sci.* **107**, 11727–11732 (2010).
- 1099 27. Maister, L., De Beukelaer, S., Longo, M. & Tsakiris, M. *The Self in the Mind's Eye: Reverse-*  
1100 *correlating one's self reveals how psychological beliefs and attitudes shape our body-image.*  
1101 <https://osf.io/f2b36> (2020) doi:10.31234/osf.io/f2b36.
- 1102 28. Cunningham, D. W., Kleiner, M., Wallraven, C. & Bühlhoff, H. H. Manipulating Video  
1103 Sequences to Determine the Components of Conversational Facial Expressions. *ACM*  
1104 *Trans Appl Percept* **2**, 251–269 (2005).
- 1105 29. Jeffreys, H. *The Theory of Probability*. (OUP Oxford, 1998).
- 1106 30. Maniscalco, B. & Lau, H. A signal detection theoretic approach for estimating metacognitive  
1107 sensitivity from confidence ratings. *Conscious. Cogn.* **21**, 422–430 (2012).



- 1108 31. Rouault, M., McWilliams, A., Allen, M. G. & Fleming, S. M. Human metacognition across  
1109 domains: insights from individual differences and neuroimaging. *Personal. Neurosci.* **1**,  
1110 (2018).
- 1111 32. Rahnev, D. *et al.* The Confidence Database. *Nat. Hum. Behav.* **4**, 317–325 (2020).
- 1112 33. Vickers, D. & Packer, J. Effects of alternating set for speed or accuracy on response time,  
1113 accuracy and confidence in a unidimensional discrimination task. *Acta Psychol. (Amst.)* **50**,  
1114 179–197 (1982).
- 1115 34. LeDoux, J. & Bemporad, J. R. The emotional brain. *J. Am. Acad. Psychoanal.* **25**, 525–528  
1116 (1997).
- 1117 35. Stål, P., Eriksson, P.-O., Eriksson, A. & Thornell, L.-E. Enzyme-histochemical differences in  
1118 fibre-type between the human major and minor zygomatic and the first dorsal interosseus  
1119 muscles. *Arch. Oral Biol.* **32**, 833–841 (1987).
- 1120 36. Stål, P., Eriksson, P.-O., Eriksson, A. & Thornell, L.-E. Enzyme-histochemical and  
1121 morphological characteristics of muscle fibre types in the human buccinator and orbicularis  
1122 oris. *Arch. Oral Biol.* **35**, 449–458 (1990).
- 1123 37. Goodmurphy, C. W. & Ovalle, W. K. Morphological study of two human facial muscles:  
1124 orbicularis oculi and corrugator supercilii. *Clin. Anat. N. Y. N* **12**, 1–11 (1999).
- 1125 38. Happak, W., Burggasser, G., Liu, J., Gruber, H. & Freilinger, G. Anatomy and Histology of  
1126 the Mimic Muscles and the Supplying Facial Nerve. in *The Facial Nerve* (eds. Stennert, E.  
1127 R., Kreutzberg, G. W., Michel, O. & Jungehülsing, M.) 85–86 (Springer, 1994).  
1128 doi:10.1007/978-3-642-85090-5\_23.
- 1129 39. Cobo, J. L., Abbate, F., de Vicente, J. C., Cobo, J. & Vega, J. A. Searching for  
1130 proprioceptors in human facial muscles. *Neurosci. Lett.* **640**, 1–5 (2017).
- 1131 40. Charles, L., Chardin, C. & Haggard, P. Evidence for metacognitive bias in perception of  
1132 voluntary action. *Cognition* **194**, 104041 (2020).

- 1133 41. Arbuzova, P. *et al.* Measuring Metacognition of Direct and Indirect Parameters of Voluntary  
1134 Movement. *bioRxiv* 2020.05.14.092189 (2020) doi:10.1101/2020.05.14.092189.
- 1135 42. Fleming, S. M. & Lau, H. C. How to measure metacognition. *Front. Hum. Neurosci.* **8**, 443  
1136 (2014).
- 1137 43. Locke, S. M., Mamassian, P. & Landy, M. S. Performance monitoring for sensorimotor  
1138 confidence: A visuomotor tracking study. *Cognition* 104396 (2020)  
1139 doi:10.1016/j.cognition.2020.104396.
- 1140 44. McIntosh, R. D., Fowler, E. A., Lyu, T. & Della Sala, S. Wise up: Clarifying the role of  
1141 metacognition in the Dunning-Kruger effect. *J. Exp. Psychol. Gen.* **148**, 1882–1897 (2019).
- 1142 45. Mole, C. D., Jersakova, R., Kountouriotis, G. K., Moulin, C. J. & Wilkie, R. M. Metacognitive  
1143 judgements of perceptual-motor steering performance: *Q. J. Exp. Psychol.* (2018)  
1144 doi:10.1177/1747021817737496.
- 1145 46. Chambon, V., Filevich, E. & Haggard, P. What is the Human Sense of Agency, and is it  
1146 Metacognitive? in *The Cognitive Neuroscience of Metacognition* (eds. Fleming, S. M. &  
1147 Frith, C. D.) 321–342 (Springer Berlin Heidelberg, 2014).
- 1148 47. Froemer, R., Nassar, M. R., Stuermer, B., Sommer, W. & Yeung, N. I knew that! Confidence  
1149 in outcome prediction and its impact on feedback processing and learning. *BioRxiv* 442822  
1150 (2018).
- 1151 48. Pauen, M. *Die Natur des Geistes*. (S. Fischer Verlag, 2016).
- 1152 49. Marcel, A. J. Agency and Self-Awareness: Issues in Philosophy and Psychology. (2003).
- 1153 50. Metcalfe, J. & Greene, M. J. Metacognition of agency. *J. Exp. Psychol. Gen.* **136**, 184–199  
1154 (2007).
- 1155 51. Fournieret, P. & Jeannerod, M. Limited conscious monitoring of motor performance in  
1156 normal subjects. *Neuropsychologia* **36**, 1133–1140 (1998).
- 1157 52. Mazzoni, P. & Krakauer, J. W. An Implicit Plan Overrides an Explicit Strategy during

- 1158 Visuomotor Adaptation. *J. Neurosci.* **26**, 3642–3645 (2006).
- 1159 53. Malone, L. A. & Bastian, A. J. Thinking About Walking: Effects of Conscious Correction  
1160 Versus Distraction on Locomotor Adaptation. *J. Neurophysiol.* **103**, 1954–1962 (2010).
- 1161 54. Pauen, M. The Functional Mapping Hypothesis. *Topoi* **36**, 107–118 (2017).
- 1162 55. Chiovetto, E., Curio, C., Endres, D. & Giese, M. Perceptual integration of kinematic  
1163 components in the recognition of emotional facial expressions. *J. Vis.* **18**, 13 (2018).
- 1164 56. Dobs, K., Bühlhoff, I. & Schultz, J. Use and Usefulness of Dynamic Face Stimuli for Face  
1165 Perception Studies—a Review of Behavioral Findings and Methodology. *Front. Psychol.* **9**,  
1166 (2018).
- 1167 57. Krumhuber, E. G., Skora, L., Küster, D. & Fou, L. A Review of Dynamic Datasets for Facial  
1168 Expression Research: *Emot. Rev.* (2016) doi:10.1177/1754073916670022.
- 1169 58. Brainard, D. H. The Psychophysics Toolbox. *Spat. Vis.* **10**, 433–436 (1997).
- 1170 59. Kleiner, M. *et al.* What's new in Psychtoolbox-3. *Perception* **36**, 1–1 (2007).
- 1171 60. Pelli, D. G. The VideoToolbox software for visual psychophysics: transforming numbers into  
1172 movies. *Spat. Vis.* **10**, 437–442 (1997).
- 1173 61. Ekman, P. Basic emotions. *Handb. Cogn. Emot.* **98**, 16 (1999).
- 1174 62. Bagby, R. M., Parker, J. D. A. & Taylor, G. J. The twenty-item Toronto Alexithymia scale—I.  
1175 Item selection and cross-validation of the factor structure. *J. Psychosom. Res.* **38**, 23–32  
1176 (1994).
- 1177 63. Lange, K., Kühn, S. & Filevich, E. "Just Another Tool for Online Studies" (JATOS): An Easy  
1178 Solution for Setup and Management of Web Servers Supporting Online Studies. *PLoS ONE*  
1179 **10**, e0130834 (2015).
- 1180 64. Morey, R. D., Rouder, J. N. & Jamil, T. BayesFactor: Computation of Bayes Factors for  
1181 common designs. R package version 0.9. 12-4.2. *Comput. Softw.* Retrieved [https://CRAN.R-](https://CRAN.R-project.org/package=BayesFactor)

- 1182        *Proj. Orgpackage BayesFactor* (2018).
- 1183    65. Bulat, A. & Tzimiropoulos, G. How far are we from solving the 2D & 3D Face Alignment  
1184        problem? (and a dataset of 230,000 3D facial landmarks). *2017 IEEE Int. Conf. Comput.*  
1185        *Vis. ICCV* 1021–1030 (2017) doi:10.1109/ICCV.2017.116.
- 1186    66. Bürkner, P.-C. Advanced Bayesian Multilevel Modeling with the R Package brms. *R J.* **10**,  
1187        395–411 (2018).
- 1188    67. Bürkner, P.-C. brms: An R Package for Bayesian Multilevel Models Using Stan. *J. Stat.*  
1189        *Softw.* **80**, 1–28 (2017).
- 1190    68. Dienes, Z. How Do I Know What My Theory Predicts? *Adv. Methods Pract. Psychol. Sci.* **2**,  
1191        364–377 (2019).
- 1192    69. Makowski, D. *et al.* *bayestestR: Understand and Describe Bayesian Models and Posterior*  
1193        *Distributions.* (2020).
- 1194    70. Kruschke, J. K. & Liddell, T. M. The Bayesian New Statistics: Hypothesis testing, estimation,  
1195        meta-analysis, and power analysis from a Bayesian perspective. *Psychon. Bull. Rev.* **25**,  
1196        178–206 (2018).
- 1197    71. Cohen, J. *Statistical power analysis for the behavioral sciences.* (L. Erlbaum Associates,  
1198        1988).
- 1199    72. Gelman, A., Goodrich, B., Gabry, J. & Vehtari, A. R-squared for Bayesian Regression  
1200        Models. *Am. Stat.* **73**, 307–309 (2019).
- 1201    73. Doorn, J. van, Ly, A., Marsman, M. & Wagenmakers, E.-J. Bayesian rank-based hypothesis  
1202        testing for the rank sum test, the signed rank test, and Spearman's  $\rho$ . *J. Appl. Stat.* **47**,  
1203        2984–3006 (2020).
- 1204    74. JASP Team. JASP (Version 0.14)[Computer software]. *JASP - Free and User-Friendly*  
1205        *Statistical Software* <https://jasp-stats.org/faq/how-do-i-cite-jasp/> (2020).

- 1206 75. Rahnev, D. et al. The Confidence Database. <https://osf.io/h8tju> (2019)  
1207 doi:10.31234/osf.io/h8tju.
- 1208 76. Vickers, D. & Packer, J. Effects of alternating set for speed or accuracy on response time,  
1209 accuracy and confidence in a unidimensional discrimination task. *Acta Psychol. (Amst.)* 50,  
1210 179–197 (1982).
- 1211 77. Response-Related Signals Increase Confidence But Not Metacognitive Performance I  
1212 eNeuro. <https://www.eneuro.org/content/7/3/ENEURO.0326-19.2020>.
- 1213 78. Brick, T. R., Braun, J., Harrill, C. & Yu, M. Face Modeling GUI, Version 0.2 $\beta$ .” Software for  
1214 facial expression analysis and stimulus synthesis. (2013).
- 1215 79. Surowiecki, J. *The wisdom of crowds*. (Anchor, 2005).
- 1216 80. Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J. & Rees, G. Relating Introspective  
1217 Accuracy to Individual Differences in Brain Structure. *Science* 329, 1541–1543 (2010).  
1218