



Metacognition of average face perception

Luyan Ji^{1,2} · William G. Hayward¹

Accepted: 21 October 2020 / Published online: 11 November 2020
© The Psychonomic Society, Inc. 2020

Abstract

Individuals have the ability to extract summary statistics from multiple items presented simultaneously. However, it is unclear yet whether we have insight into the process of ensemble coding. The aim of this study was to investigate metacognition about average face perception. Participants saw a group of four faces presented for 2 s or 5 s, and then they were asked to judge whether the following test face was present in the previous set (Experiment 1), or whether the test face was the average of the four member faces (Experiment 2). After each response, participants rated their confidence. Replicating previous findings, there was substantial endorsement for the average face derived from the four member faces in Experiment 1, even though it was not present in the set. When judging faces that had been presented in the set, confidence correlated positively with accuracy, providing evidence for metacognitive awareness of previously studied faces. Importantly, there was a negative confidence-accuracy relationship for judging average faces when duration was 2 s, and a near-zero relationship when duration was 5 s. By contrast, when the average face had to be identified explicitly in Experiment 2, performance was above chance level and there was a positive correlation between confidence and accuracy. These results suggest that people have metacognitive awareness about average face perception when averaging is required explicitly, but they lack insight into the averaging process when member identification is required.

Keywords Face perception · Perceptual categorization and identification · Visual awareness

Introduction

Multiple faces presented at the same time can be integrated to form an ensemble perception (Haberman & Whitney, 2007; Whitney & Yamanashi Leib 2018). Apart from information revealed by each individual face in the group, we can also extract summary statistics from them, for example, the average information from multiple facial identities (de Fockert & Wolfenstein, 2009; Neumann, Schweinberger, & Burton, 2013). However, little is known about our ability to actively monitor and have knowledge of our own averaging

performance. In the current study, we aimed to explore metacognition of average face perception, to shed light on the underlying cognitive mechanisms of multiple face processing.

Several studies have already suggested that we have insight into our face-recognition and face-perception abilities (Arizpe et al., 2019; Bobak, Mileva, & Hancock, 2018; Palermo et al., 2017; Shah, Gaule, Sowden, Bird, & Cook, 2015; Ventura, Livingston, & Shah, 2018). They investigated the correlations between self-report questionnaires on face-recognition ability (for example, 20-item prosopagnosia index and Cambridge Face Memory Questionnaire) and actual face-recognition performance (from Cambridge Face Memory Test, for instance). Positive correlations have generally been found, indicating that individuals do have metacognition about their face-recognition abilities. However, the strength of the correlations varies across different studies and the extent to which people have insight into face-recognition ability is still a topic of debate (Livingston & Shah, 2018).

Metacognition of face identification can also be evaluated by examining how well or confidently an individual considers his or her performance at identifying a face on a trial-by-trial basis. For example, in eyewitness studies, confidence-accuracy (CA) correlations have been computed and have been found to be positive though relatively weak (meta-analyses: $r = .25$, Bothwell, Deffenbacher, & Brigham, 1987; $r = .29$, Sporer, Penrod,

The original online version of this article was revised: In the Electronic Supplementary Material, all the track changes should be accepted.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.3758/s13414-020-02189-7>.

✉ Luyan Ji
luyanji.psy@hotmail.com

¹ Department of Psychology, Faculty of Social Sciences, University of Hong Kong, Hong Kong, China

² Center for Brain and Cognitive Sciences, Department of Psychology, Faculty of Education, Guangzhou University, Guangzhou, China

Read, & Cutler, 1995). The positive CA correlations indicate that witnesses have higher confidence ratings on average for correct responses relative to incorrect ones. Importantly, the CA correlations tend to be higher for participants who make positive identifications than those who make rejection responses (Sporer et al., 1995). Recent studies (Brewer & Wells, 2011; Palmer, Brewer, Weber, & Nagesh, 2013) used a calibration approach and plotted proportion of accurate responses against each confidence level. They also found a positive relationship between confidence and accuracy for positive identifications, suggesting that people have metacognition about face identification.

In eyewitness studies, usually witnesses are asked to identify the previously seen culprit from multiple suspects. However, it is common that a witness may have observed a number of individuals during an episode, and so have to match a suspect against the entire set of individuals. Using the member identification task as described above, researchers have found that if the test face was the average of the previous faces in the set, participants tend to judge this average face to be present, even though it was never visible to them, suggesting an implicit extraction of average identity of a set of faces (de Fockert & Wolfenstein, 2009; Neumann, Schweinberger, & Burton, 2013). The multiple-faces identification task provides an important test of face metacognition: do we have insight into whether a face was actually studied, or is the average of those studied member faces? If the test face was the average face, are individuals confident when they identify the average face as being present, or do they lack the ability to discriminate the average face from studied faces? In the current study, we utilized the member identification task (de Fockert & Wolfenstein, 2009; Haberman & Whitney, 2009) and the modified average identification task, and measured CA relations as an index of metacognition of face identification and average face perception. We computed the within-person Goodman-Kruskal *gamma correlation*, a nonparametric correlation coefficient ranging from -1.0 to 1.0, between confidence and accuracy (Nelson, 1984; Goodman & Kruskal, 1954; Kelly & Metcalfe, 2011; also see Supplementary methods in the [Online Supplementary Material \(OSM\)](#)). If the gamma correlation is significantly above zero, it indicates that people have awareness of their cognitive states, that is, they know what they know and what they don't know. A correlation of zero suggests that people do not have any knowledge of whether their judgments are correct or not. If the correlation is below zero, it indicates that people are more confident about their incorrect judgments than their correct ones. On the other hand, when the average face is required to be identified, if people are certain that the correct average face is present (i.e., confident at the correct response), then it would suggest that people have insight of the averaging process.

In addition, we also investigated the effect of exposure time in metacognition of average face perception. It has been shown that increased exposure time of the target face

improved eyewitness identification accuracy (see meta-analysis by Bornstein, Deffenbacher, Penrod, & McGorty, 2012). Relative to short durations, longer durations of multiple faces in the set also increased the endorsement for a test face that matched one of the target faces; while for the “unseen” average face, the endorsement increased first (e.g., up to 1,600 ms, Neumann, Ng, Rhodes, & Palermo, 2017) and then declined (Li et al., 2016; Neumann et al., 2017). Notwithstanding, the effect of exposure time on CA relations is less clear. Confidence-accuracy correlations of identifying the perpetrator from a lineup have been found to be positively correlated with the exposure time of the target face (Bothwell et al., 1987); metacognition of face identification improved with increased exposure, consistent with an optimality hypothesis that suggests that confidence judgments discriminate correct from incorrect decisions better when information processing becomes more optimal (Bothwell et al., 1987; Deffenbacher, 1980). Some other studies, by contrast, found the opposite effect, in that confidence predicted accuracy better in more difficult conditions when the exposure time was short compared to easier trials with long duration (Memon, Hope, & Bull, 2003; Palmer et al., 2013). There are also studies that found that increasing exposure time did not influence CA correlations (Lindsay, Read, & Sharma, 1998). One possible reason underlying these discrepancies might be that some of the previous studies did not distinguish between different types of responses (e.g., selection vs. rejection). Confidence-accuracy relations have been found to be different between positive selections and negative rejections (Sporer et al., 1995). It would be likely that the effect of exposure time on CA relations was also dependent on the response types, and we analyzed CA relations separately for positive and negative decisions.

In sum, the aim of the current study was to examine metacognition of ensemble coding for face identities and the effect of exposure time on such metacognition. In Experiment 1, we employed an implicit averaging task (de Fockert & Wolfenstein, 2009) where participants were required to identify whether the test face was one of a set of previously presented stimuli. In Experiment 2, we explicitly asked participants to average across the target faces and to identify whether the test face was the average face. A 5-point Likert scale was used to measure confidence levels immediately after each identification response. The four types of test face (*the matching member*, which was presented in the previous set; *the matching average*, which was the average of the four studied faces in the previous set; *the non-matching member*, which was not presented in the set; and *the non-matching average*, which was the average of four unstudied faces) were tested within participants, and two exposure times (2 s, 5 s) were applied between participants. We employed gamma correlations (Goodman & Kruskal, 1954; Nelson, 1984) to measure metacognition of individual and average face identification. We hypothesized that people have metacognition of

identifying member faces (i.e., above-zero correlation), similar to findings in eyewitness studies (Brewer & Wells, 2011; Sporer et al., 1995), and metacognition would be better for positive (“yes” responses) than negative decisions (“no” responses). For average faces, if we observed positive CA correlations in the implicit averaging task, it would suggest that they have good insight into face identification, and they are aware of the differences between average faces and actual studied faces. The positive CA correlations in the explicit averaging task would suggest that participants have metacognition of average face perception. In contrast, if there were zero or even negative CA relations, it would suggest that they did not have knowledge about the averaging process and their metacognition of face identification was even misdirected. Moreover, increasing presentation time of a set of multiple target faces was hypothesized to improve metacognition of identifying whether the following test face was presented before, based on the optimality hypothesis (Bothwell et al., 1987), but the effect of exposure time on metacognition might be dependent on whether participants gave positive or negative identification responses.

Experiment 1

In the first experiment, the matching member face that was presented in the set was required to be identified. If people were certain that the average face was present (i.e., confident at incorrect responses, resulting in correlation at zero or even below zero), it would suggest that metacognition of face identification is poor, and people have little insight into the differences between the observed faces and their averages.

Methods

Participants

Forty-eight Chinese students from the University of Hong Kong participated in this experiment for credit, and they were randomly but evenly assigned to the two duration conditions (duration 2 s: 18–25 years, 17 females; duration 5 s: 18–25 years, 15 females). The sample size of 48 (i.e., 24 in each group) was determined a priori, so that the number of participants in each group was consistent with the previous study examining a similar duration effect in a member identification task (Neumann et al., 2017). The participants gave written informed consent prior to the start of the experiment, and reported being right-handed and having normal or corrected-to-normal vision. The study protocol was conducted in accordance with the Declaration of Helsinki and approved by the local ethics committee.

Stimuli

We selected 32 male and 32 female identities from the face images we collected in the lab. All face identities were Chinese and showed neutral expressions with closed mouths. The face images were trimmed to an oval shape and their hair, ears, neck, and other external information were cropped. All images were converted to grayscale, and scaled to the same mean luminance and root-mean-square contrast (Bex & Makous, 2002). Each face image subtended a visual angle $4.47^\circ \times 7.15^\circ$, and was presented against a homogenous black background.

We first chose 32 target sets, each of which consisted of four faces with the same gender but different identities. Since we had only 64 face images in total, every face was present in two different sets, but no two sets had more than one face that overlapped. The average face of each set was generated by morphing together the four individual faces from that set, using FantaMorph 5. The gender of the test face was always the same with the faces in the target set. In the matching member condition, the test face was randomly chosen from the four faces in the target set. In the non-matching member condition, the test face was randomly selected from faces that were not shown in the target set. The matching average test face was the morphed face across the four faces in the target set. The non-matching average test face of one set was randomly selected from the average faces of the other target sets.

Similar to de Fockert and Wolfenstein (2009), we also morphed the original face images with themselves for the member test faces, to create blurred appearances close to those of morphed average faces. In addition, we added Gaussian noises to all the test faces, in order to further reduce the low-level differences between the average and member test faces. The center of the four faces in each set occupied a matrix of $5.37^\circ \times 8.00^\circ$, centered on the screen. The positions of these four individual faces within the set were randomized. The test face was shown on the center of the screen. Figure 1 shows an example of face stimuli used in this study.

Apparatus and procedure

Participants sat 50 cm in front of a 17-in. flat screen with a refresh rate of 60 Hz. Speed of response was not emphasized and feedback was not given. The task was programmed and controlled using PsychoPy (Peirce et al., 2019). The experiment lasted around 20–25 min. A trial began with a fixation cross that appeared at the center of the screen for 500 ms. Then, a set of four faces was presented for 2 s or 5 s, immediately followed by the test face. Participants gave a response about “whether the test face was present or not in the previous four faces,” by binary choices using “left” or “right” buttons (Fig. 1). The response keys were counterbalanced across participants. Immediately after this response, participants rated

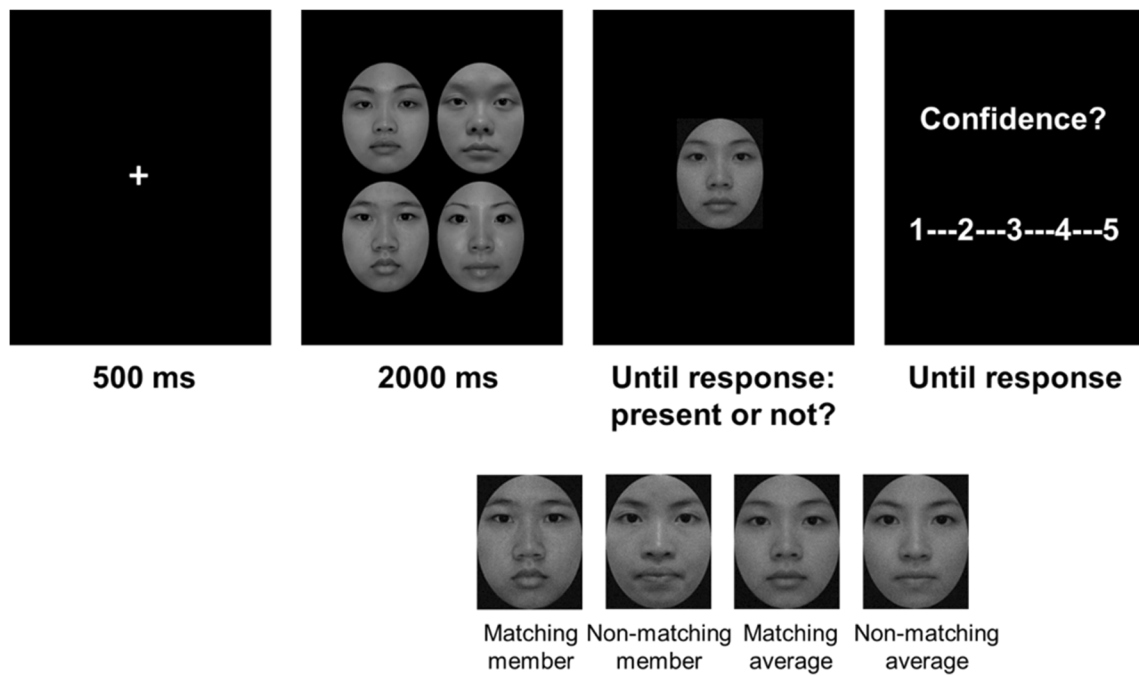


Fig. 1 Procedure of the task. The duration of the set of four faces was 2 s (as shown in the figure) or 5 s. There are four types of test faces: a member of the set, a member of the other set, an average of the four faces in the set,

and an average of the four faces in the other set (from left to right). After judging whether the test face was present or not, participants also rated their confidence about their response

how confident they thought their response was correct on a 5-point Likert scale, where 1 indicates very unconfident and 5 indicates very confident. The next trial started automatically 1,000 ms after the confidence rating.

The face type (member, average) and the match type (match, non-match) were randomized within blocks. The duration of face sets (2 s, 5 s) was a between-subject variable. Participants performed four experimental blocks of 32 trials (eight trials per condition). Each of the 32 face sets was paired once with four kinds of test faces (match member, non-match member, match average, and non-match average). Before the formal task, participants practiced 16 trials with different face stimuli. Practice trials were excluded from all subsequent analyses.

Data analysis

Trials with response times (RTs) exceeding 2.5 SDs above or below the grand mean RT for each participant (overall 3.1% (2 s) and 2.9% (5 s) trials in the two duration conditions) were excluded. This standard cutoff was chosen before running data analyses. The probabilities of endorsing the test face as present and the within-participant gamma correlation between confidence ratings and accuracy (Nelson, 1984; also see Kelly & Metcalfe, 2011) was then analyzed using mixed-design repeated-measures ANOVAs. The within-subject factor was face type (member, average), and the between-subject factor was duration (2 s, 5 s). Given the instructions to participants, the correct answer for judgments of matching member

faces was “yes,” and for the matching average face was “no.” The correct responses for all the non-matching test faces should be “not present.” A Greenhouse-Geisser correction was applied when assumptions of sphericity were violated. A Bonferroni correction was used when multiple comparisons were performed.

Results

Probability of “yes” responses

The probability of “yes” responses in each condition is shown in Table 1. We computed the difference scores between the matching and non-matching conditions for member and average faces separately as in Neumann et al. (2017), to remove the potential response biases. Here, we report the results that

Table 1 Percentages of “yes” responses (mean and standard deviation) in each condition in Experiments 1 and 2

	Experiment 1		Experiment 2	
	2 s	5 s	2 s	5 s
Matching member	.60 (.19)	.64 (.20)	.66 (.15)	.71 (.15)
Non-matching member	.26 (.20)	.16 (.17)	.16 (.10)	.16 (.10)
Matching average	.57 (.17)	.52 (.18)	.80 (.14)	.77 (.14)
Non-matching average	.30 (.19)	.21 (.14)	.37 (.19)	.29 (.12)

removed the partially overlapping average faces from the non-matching set, because it is likely that two face sets had one face that overlapped, and the faces that contributed to the non-matching average face on some occasions had an overlap with the faces in the target set. This resulted in the removal of 26.3% and 27.7% trials in the two duration conditions, respectively.¹ The difference scores for the member faces were larger ($M = .41$, $SD = .32$) than those for average faces ($M = .31$, $SD = .22$), $F(1, 46) = 11.88$, $p = .001$, $\eta_G^2 = .03$, which suggests better discrimination when judging the member faces than the average faces (Fig. 2). Consistent with previous studies (de Fockert & Wolfenstein, 2009; Neumann, Schweinberger, & Burton, 2013), the results show that participants could actually discriminate the previously presented member faces from the average of the member faces in the set. When the duration of face sets increased from 2 s to 5 s, the difference scores enlarged numerically for both the member (2 s: $M = .34$, $SD = .31$; 5 s: $M = .47$, $SD = .33$) and average faces (2 s: $M = .29$, $SD = .23$; 5 s: $M = .34$, $SD = .21$), but the improvement did not reach significance, $F(1, 46) = 1.5$, $p = .23$, $\eta_G^2 = .03$, which was slightly different from the previous findings (Neumann et al., 2017). There was no significant interaction between face type and duration, $F(1, 46) = 2.02$, $p = .16$, $\eta_G^2 = .005$.

Confidence-accuracy correlations

We first plotted accuracy against each confidence level for member and average test faces in two duration conditions. As Fig. 3A shows, when the confidence rating increased, accuracy of identifying the member faces increased in both duration conditions. However, for identifying the average faces, accuracy declined with increasing confidence ratings when the duration was 2 s, and the relation between confidence and accuracy was less strong when the duration was 5 s.

Analyses of gamma correlations quantified the CA relations. Similar to above, we report analyses of data from which the partially overlapping non-matching average faces had been trimmed, and as noted above the analyses for the non-trimmed data showed the same pattern. We applied Fisher Z-transformations to the gamma correlations so that they become normally distributed. One participant had a gamma correlation of -1 in one condition, which could not be z-transformed and so was not included in further analyses. The repeated-measures ANOVA was then conducted on the transformed correlations with face type as the within-subject variable and duration as the between-subject variable. The results showed a significant main effect of face type, $F(1, 45) = 22.69$, $p < .0001$, $\eta_G^2 = .24$, and also a significant main effect

¹ We conducted all the reported statistical analyses with the original, untrimmed data set. In all cases, the patterns of statistical significance for main effects and interactions were the same as for the trimmed data.

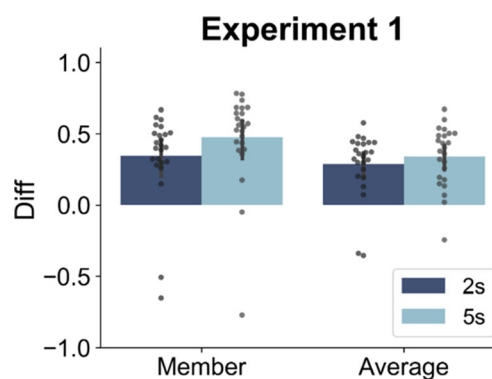


Fig. 2 Difference scores (the difference of the probability of judging “yes” to the test face between the matching and non-matching conditions) when duration was 2 s (dark blue) and 5 s (light blue) for two face types separately in Experiment 1. The error bar represents the 95% confidence interval of the mean, and each dot represents data for each participant

of duration, $F(1, 45) = 7.38$, $p = .009$, $\eta_G^2 = .06$. The gamma correlations were larger for the member faces than the average faces, and they were larger when duration was longer. The interaction between face type and duration was not significant, $F(1, 45) < 1$, $\eta_G^2 = .009$. One-sample t-tests were conducted to compare observed gamma correlations against chance (zero) in each condition. The results showed that gamma correlations were significantly greater than zero for member faces in both 2-s ($M = .66$, $SD = .13$) and 5-s durations ($M = .66$, $SD = .13$), $t(23) = 5.25$, $p < .001$, $d = 1.07$; $t(23) = 5.24$, $p = .001$, $d = 1.07$. For the average face, they were significantly below zero when duration was 2 s, $t(23) = -2.27$, $p = .03$, $d = .46$, and did not differ significantly from zero when duration was 5 s, $t(23) = 1.17$, $p = .25$, $d = .24$ (Fig. 3B).

We were interested in particular in the extent to which participants had metacognitive awareness of whether an average face had been shown in the set or not. A matching average face contained information about the studied faces, while a non-matching average did not, although both were actually not studied. Thus, we analyzed metacognitive awareness for matching and non-matching averages separately (Fig. 4). As can be seen from Fig. 4A, for both 2-s and 5-s durations, participants show a negative relationship between confidence and accuracy for matching averages. In other words, when accuracy of identifying the matching average faces decreased, the confidence rating actually increased. Participants were more confident about their incorrect endorsement for the matching average faces in Experiment 1. This pattern is relatively consistent across the two duration conditions. On the other hand, the CA relation for the non-matching average faces showed a different pattern. When duration was 2 s, the slope seemed flat; while when duration increased to 5 s, there was a positive relationship between confidence and accuracy for identifying the non-matching average faces. The statistical

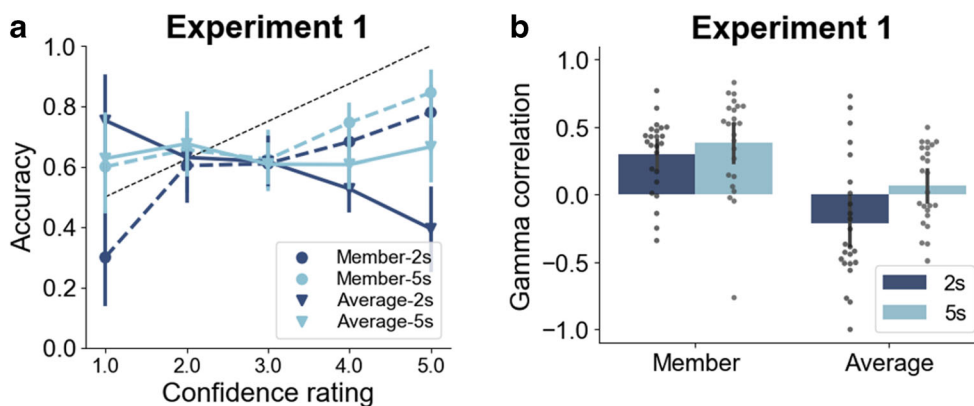


Fig. 3 (a) Accuracy against each confidence level for member (dashed line) and average test face (solid line) when duration was 2 s (dark blue) and 5 s (light blue) in Experiment 1. The error bar represents the 95% confidence interval of the mean. In each figure, a dotted black line has been drawn for reference purposes, showing an idealized function of chance performance of lowest confidence and perfect performance at highest confidence. The line is drawn in such a way that the lowest

confidence rating corresponds to accuracy at chance level and the highest confidence rating corresponds to perfect accuracy (accuracy equals to one). (b) Gamma correlations when duration was 2 s (dark blue) and 5 s (light blue), in each face type condition in Experiment 1. The error bar represents the 95% confidence interval of the mean, and each dot represents data for each participant

analyses of gamma correlations confirmed the above observations (see OSM, and also Fig. 4B).

Previous studies have shown that the confidence-accuracy relations differed between positive and negative decisions. Here, we also computed the gamma correlations for “yes” (positive) and “no” (negative) responses separately (Fig. 5). To be noted, the correct responses for the average faces in the member identification task used in Experiment 1 were always “no” responses, no matter whether in the matching or the non-matching conditions, and the “yes” response were always incorrect. Therefore, we could not measure gamma correlations for the positive and negative decisions separately for the average faces. We decided to collapse the two face types, and conducted the repeated-measures ANOVA on the Fisher Z-transformed gamma correlations with response type and

duration as two factors. The results revealed a significant main effect of response type, $F(1, 46) = 30.34, p < .0001, \eta_G^2 = .16$. The main effect of duration, as well as the interaction between duration and response type, was not significant, $F(1, 46) < 1, \eta_G^2 = .34; F(1, 46) = 1.45, p = .23, \eta_G^2 = .009$. Gamma correlations for the positive decisions ($M = .40, SD = .31$) were significantly larger than the negative decisions ($M = .14, SD = .30$), and both of them were above zero under two duration conditions, $t_s > 2.00, p_s < .057$.

Discussion

Using the member identification task (de Fockert & Wolfenstein, 2009; Li et al., 2016), we replicated the previous

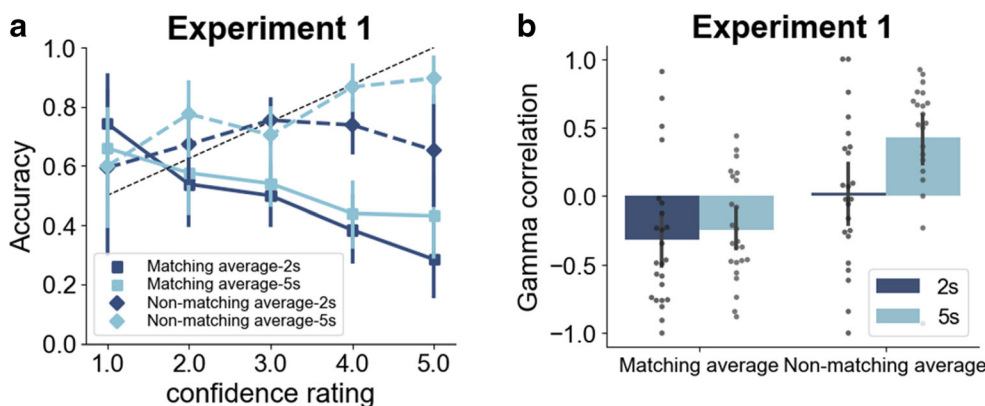


Fig. 4 (a) Accuracy against each confidence level for matching average (solid line) and non-matching average faces (dashed line) when duration was 2 s (dark blue) and 5 s (light blue) in Experiment 1. The error bar represents the 95% confidence interval of the mean. In each figure, a dotted black line has been drawn for reference purposes, showing an idealized function of chance performance of lowest confidence and perfect performance at highest confidence. The line is drawn in such a way

that the lowest confidence rating corresponds to accuracy at chance level and the highest confidence rating corresponds to perfect accuracy (accuracy equals to one). (b) Gamma correlations when duration was 2 s (dark blue) and 5 s (light blue) for matching average and non-matching average faces in Experiment 1. The error bar represents the 95% confidence interval of the mean, and each dot represents data for each participant

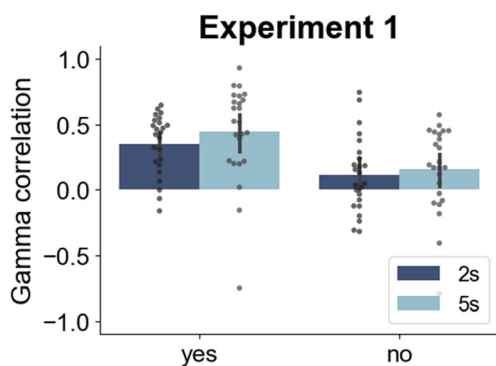


Fig. 5 Gamma correlations for “yes” and “no” responses separately, when duration was 2 s (darker blue) and 5 s (lighter blue) in Experiment 1. The two face types were combined. The error bar represents the 95% confidence interval of the mean, and each dot represents data for each participant

findings that participants incorrectly identified the average face (morphed from the previous four member faces) as the member face. The gamma correlations between confidence and accuracy for average faces had large variance across participants (Figures 3 & 4), especially when duration was short. We observed a negative CA relation for average faces when duration was 2s, that is participants were more confident about the erroneous decisions compared to the correct ones. When we looked at the CA relation for matching average and non-matching average faces separately (Figure 4), it is clear that the negative slope was mainly driven by the negative correlation when the average face was matching (i.e., the average of the member faces). Participants actually had high confidence when they incorrectly endorsed the matching average face, which suggests that they were not aware of the mean identity representation and had poor insight into the differences between the average faces and the studied faces. Relative to 2s, for 5s-display duration participants show clear evidence for sensitivity to the accumulation of information about non-matching averages, and they show higher rates of accuracy at judging that non-matching averages were not studied as their confidence grows (Figure 4, also see Supplementary materials). However, importantly, the CA relations did not differ from zero under 5s duration condition, confirming the lack of insight to implicit averaging. On the other hand, people had metacognition of identifying individual faces (e.g., Arizpe et al., 2019; Palmer et al., 2013), as shown by the positive CA relation for member faces. The overall gamma correlation scores which combined two face types were also positive, confirming the insight to identifying individual faces even when the test faces include average face distractors which were perceptually similar to the target faces. To be noted, the overall gamma correlations were not impacted by duration, which was different from the findings that gamma correlations increased with increased duration when two face types were analyzed separately. Thus, we should be cautious about the conclusion of the duration effect on CA relations,

especially for identifying individual faces. Moreover, consistent with the previous studies (Sporer et al., 1995), we found that CA relations were higher for positive decisions (judging “yes”) than negative ones (judging “no”); however, duration seems not to modulate this distinction.

Experiment 2

In Experiment 1, participants were not explicitly asked to make judgments about the average faces but they implicitly did so. However, previous research suggests that participants can explicitly extract the average information from faces when asked to do so (e.g., Haberman & Whitney, 2009; Ji, Chen, Loeyes, & Pourtois, 2018; Ji, Rossi, & Pourtois, 2018). Beyond the well-known ensemble perception for faces, would participants be able to monitor their internal processes when making judgments about the average faces? In Experiment 2, we ran a new task in which participants were explicitly asked about the presence of the average face, in order to further explore metacognition of explicit average perception. We reasoned that they may have sensitivity to the average face when specifically asked. If people are certain that the matching average face is present and the non-matching average is not (i.e., confident at the correct response, resulting in a positive CA correlation), then it would suggest that people have insight of the averaging process. In Experiment 2, we also examined the effect of duration on CA relations and whether they differ between positive and negative decisions.

Methods

Participants

A new group of 48 Chinese students from the University of Hong Kong participated in this experiment for credit or cash payments. They were randomly assigned to the two duration conditions (duration 2 s: 18–21 years, 17 females; duration 5 s: 18–23 years, 20 females). The sample size of 48 (i.e., 24 in each group) was determined a priori, and was chosen to be the same as in Experiment 1. The participants gave written informed consent prior to the start of the experiment, and reported being right-handed and having normal or corrected-to-normal vision. The study protocol was conducted in accordance with the Declaration of Helsinki and approved by the local ethics committee.

Stimuli

Apart from the 64 face images used in Experiment 1, we collected 64 new face stimuli in the lab, resulting in a total of 64 female and 64 male faces. All face identities were Chinese and showed neutral expressions with closed mouth.

Similar to Experiment 1, the face images were cropped to an oval shape and converted to grayscale. Each face image subtended a visual angle $4.47^\circ \times 7.15^\circ$, and was presented against a homogenous black background. The face sets and the test faces were created similarly to Experiment 1, with the only difference being that every face was present in only one face set in Experiment 2. In other words, the face sets did not have any overlap in face images.

Apparatus and procedure

Participants sat 50 cm in front of a 17-in. flat screen with a refresh rate of 75 Hz. Speed of response was not emphasized and feedback was not given, as in Experiment 1. The task was programmed and controlled using PsychoPy (Peirce et al., 2019). The procedure was similar to that in Experiment 1, except that the two different face types (member, average) were shown in separate blocks in Experiment 2. In the member-block, the participants were required to judge “whether or not the test face was one of the member faces in the previous set you have just seen.” In the average-block, the participants were asked “whether the test face was the average of the faces you have just seen in the set.” They were reminded that the test face was always an average face in the average-block, and the non-matching average face was the average of four different faces that were not presented earlier in the trial. The order of the two blocks and the response keys (“left” or “right” buttons) were counterbalanced across participants. Each of the 32 face sets was paired once with the matching and non-matching targets in each block. The match type (match, non-match) was randomized within blocks, and the duration of face sets (2s, 5s) was a between-subject variable. Before the formal task, participants practiced eight trials in each block with different face stimuli. Practice trials were excluded from all subsequent analyses.

Data analysis

Trials with RTs exceeding 2.5 SDs above or below the grand mean RT for each participant (overall 2.8%, and 2.8% trials in two duration conditions) were excluded. This standard cutoff was chosen before running data analyses. Here, we considered the “yes” judgments to the matching member and the matching average faces as correct, and the “no” responses to the non-matching member and average face as correct responses in the analysis. Similar to Experiment 1, the probabilities of endorsing the test face to be matched (i.e., the “yes” judgments) and the within-participant gamma correlation between confidence ratings and accuracy were computed and analyzed using repeated-measures ANOVAs. The within-subject factors were face type (member, average) and the between-subject factor was duration (2 s, 5 s). One participant gave the highest confidence rating (level 5) for all member

faces under 2-s duration, thus gamma correlations could not be computed for this participant in this condition. A Greenhouse-Geisser correction was applied when assumptions of sphericity were violated. A Bonferroni correction was used when multiple comparisons were performed.

Results

Probability of “yes” responses

The probability of “yes” responses in each condition is shown in Table 1. We also computed the difference scores between the matching and non-matching conditions for member and average faces separately (Fig. 6), in order to remove potential response biases. Similar to Experiment 1, the difference scores for the member faces were larger ($M = .52$, $SD = .16$) than those for average faces ($M = .45$, $SD = .21$), $F(1, 46) = 5.93$, $p = .02$, $\eta_G^2 = .03$, suggesting better discrimination of the member than of the average faces. When the duration of face sets increased from 2 s to 5 s, the difference scores did not change significantly (2 s: $M = .46$, $SD = .17$; 5 s: $M = .51$, $SD = .14$), $F(1, 46) = 1.10$, $p = .30$, $\eta_G^2 = .02$. There was no significant interaction between duration and face type, $F(1, 46) < 1$, $\eta_G^2 = .0001$.

Confidence-accuracy correlations

As in Experiment 1, we first plotted accuracy against each confidence level for member and average test faces in two durations in Experiment 2. When the confidence rating increased, accuracy of identifying both the member and average faces also increased (Fig. 7A).

The repeated-measures ANOVA on the Fisher Z-transformed gamma correlations (the transformed value was infinite in one condition for one participant) showed a

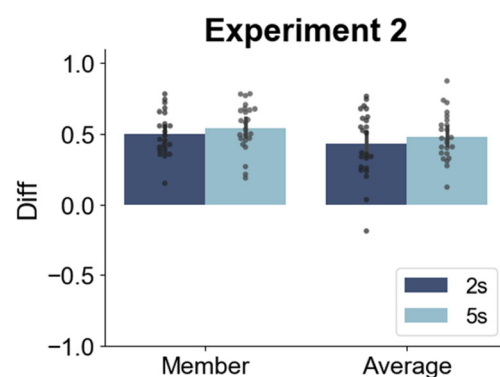


Fig. 6 Difference scores (the difference of the probability of judging “yes” to the test face between the matching and non-matching conditions) when duration was 2 s (dark blue) and 5 s (light blue), for two face types separately in Experiment 2. The error bar represents the 95% confidence interval of the mean, and each dot represents data for each participant

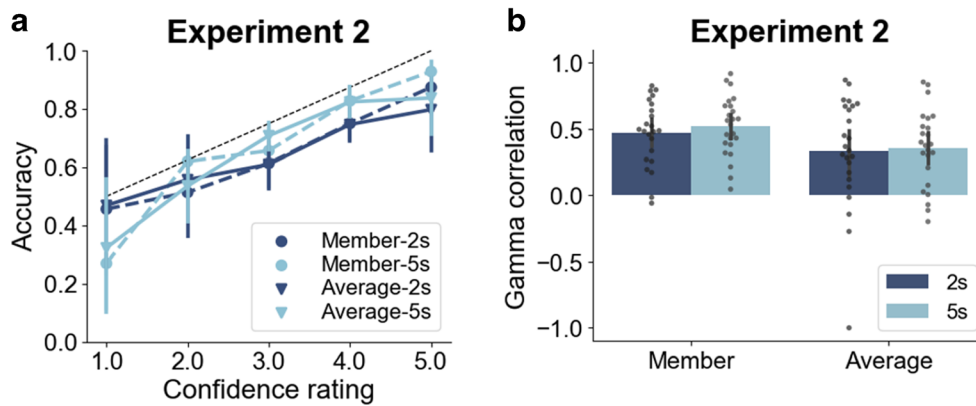


Fig. 7 (a) Accuracy against each confidence level for member (dashed line) and average test face (solid line) when duration was 2 s (dark blue) and 5 s (light blue) in Experiment 1. The error bar represents the 95% confidence interval of the mean. In each figure, a dotted black line has been drawn for reference purposes. The line is drawn in such a way that

the lowest confidence rating corresponds to accuracy at chance level and the highest confidence rating corresponds to perfect accuracy (accuracy equals one). (b) Gamma correlations when duration was 2 s (dark blue) and 5 s (light blue) in each face type condition in Experiment 2

marginally significant main effect of face type, $F(1, 44) = 3.83, p = .06, \eta_G^2 = .03$. The gamma correlations were larger for discriminating member faces compared to average faces. The main effect of duration and the interaction between face type and duration was not significant, $F(1, 44) < 1, \eta_G^2 = .0005; F(1, 44) = 1.27, p = .27, \eta_G^2 = .01$. One-sample t-tests showed that gamma correlations were significantly greater than zero for member faces in both 2-s ($M = .47, SD = .25$) and 5-s durations ($M = .52, SD = .22$), $t(22) = 9.09, p < .0001, d = 1.90$; $t(23) = 11.40, p < .0001, d = 2.33$. They were also significantly above zero for average faces when duration was 2 s ($M = .33, SD = .42$) and 5 s ($M = .36, SD = .29$), $t(23) = 3.92, p = .0007, d = .80$; $t(23) = 11.40, p < .0001, d = 2.33$ (Fig. 7B).

As in Experiment 1, we also computed the gamma correlations for “yes” (positive) and “no” (negative) responses separately (Fig. 8). Five participants had no incorrect responses in some of the conditions, and one participant gave the same level of confidence ratings (level 4) for all “yes” responses

in one condition, and thus gamma correlations could not be computed for these conditions. For the rest, the gamma correlations had extreme values (-1 or 1) among 11 participants under some conditions. The repeated-measures ANOVA on the Fisher Z-transformed gamma correlations revealed a marginally significant main effect of face type, $F(1, 32) = 3.91, p = .06, \eta_G^2 = .02$. There was an interaction between face type and response type, $F(1, 32) = 10.23, p = .003, \eta_G^2 = .06$. The main effects of duration and response type and the other interactions were not significant, $ps > .17$. Gamma correlations for the positive decisions of member faces ($M = .62, SD = .29$) were significantly larger than those of the average faces ($M = .42, SD = .33$), $p = .0005$, while gamma correlations for the negative decisions did not differ significantly between the member ($M = .43, SD = .52$) and average faces ($M = .41, SD = .38$), $p = .35$. For the member faces, gamma correlations were larger for the positive than the negative responses, $p = .002$; while for the average face, gamma correlations did not differ significantly between the positive and negative

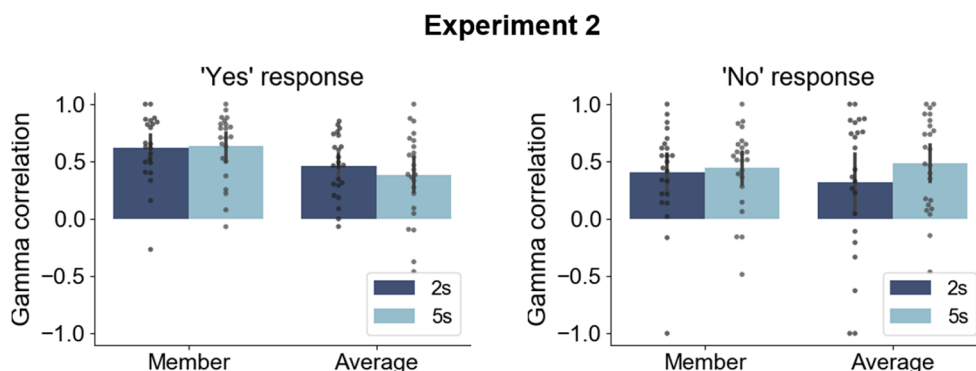


Fig. 8 Gamma correlations for member and average faces, and for “yes” and “no” responses separately, when duration was 2 s (dark blue) and 5 s (light blue) in Experiment 2. The error bar represents 95% confidence interval of the mean, and each dot represents data for each participant

responses, $p = .25$. Gamma correlations for the positive and negative decisions were above zero under both duration and face-type conditions, $t_s > 2.51$, $p_s < .02$.

Discussion

In Experiment 2, we used separate individual and average face identification tasks where participants were required to identify whether the test face was the member or the average face, respectively. The results showed that participants identified member faces better than average faces. The correlations between confidence and accuracy were positive in explicitly identifying the average and the member faces, suggesting that people have insight into both average face and individual face identification. The overall gamma correlations showed that metacognition of individual face identification was better than that of average face perception. Interestingly, when we look at the response-specific gamma correlations, CA relations were larger for member than average faces when participants gave “yes” responses; when they gave “no” responses, CA relations did not differ between member and average faces. This discrepancy suggests that participants might have different criteria when making positive and negative responses, and the corresponding confidence ratings (Maniscalco & Lau, 2014). In the present study, we found that metacognitive processing differed for positive and negative responses, and the comparisons of insight to individual and average face processing depended on the response type. We did not observe substantial effects of duration on overall or response-specific gamma correlations in Experiment 2, however. The slope of accuracy against confidence level seems to be steeper (Fig. 7A) when duration was longer, but since the percentages of trials having the lowest confidence ratings were extremely low (Supplementary Table S2, OSM), the impact of these points would be very limited. When we directly compared the CA relations of individual face identification between the two experiments, we did observe a duration effect, that is, gamma correlations for individual face identification increased with longer duration (OSM). Similar to Experiment 1, the results of duration on metacognition were mixed, and we need to be cautious about the conclusion we draw from this result.

General discussion

In this study, we explored whether human observers can metacognitively introspect face-averaging processes and asked whether that process would be impacted by the presentation time of faces. Metacognition was measured as the confidence-accuracy (CA) relation. Using the implicit average task, we found that there were large proportions of “present” judgments for the average face that was morphed across the previously presented faces, replicating the previous results (de

Fockert & Wolfenstein, 2009; Neumann et al., 2013). Our novel findings showed a negative CA relation for average faces when duration was 2 s. When the presentation time of faces increased to 5 s, the CA relations for average faces increased from negative to zero. On the other hand, when averaging was explicitly required, participants could identify whether the test face was the average of member or non-member faces above chance level, and there was a positive CA relation for identifying the average faces.

Individuals were more confident about their incorrect identification responses than the correct ones for average faces, which resulted in a negative CA relation in Experiment 1. Previous studies have already found that there exists a strong tendency to judge the never-shown average item to be present among the multiple items in the set, for both low-level features (e.g., sizes, Ariely, 2001) and high-level objects (e.g., faces, de Fockert & Wolfenstein, 2009; Haberman & Whitney, 2007). In addition, perception and memories for the individual items are also biased to the average information of them (Brady & Alvarez, 2011; Griffiths, Rhodes, Jeffery, Palermo, & Neumann, 2017; Sama, Nestor, & Cant, 2019). The current study took one step further. We found that the confidence-accuracy correlations were significantly below zero for the average faces under the 2-s duration condition, and participants were confident that the unseen average face was present. A high confidence rating indicates a high degree of match between a specific stimulus and an image in memory (Bernbach, 1971; also see Sauer, Brewer, & Weber, 2008). Thus, our new results suggest that individuals had the (wrong) knowledge that the average identity matched one of the member faces. When duration increased to 5 s, the CA relations for the average faces increased, but they were not significantly different from zero. There are two factors that would contribute to the enlarged CA relations. One is that participants became less confident for the incorrect endorsement for the matching average faces, and another is that participants became more confident for the correct rejection of the non-matching average faces. Our results supported the latter factor (Fig. 4; also see OSM). Interestingly, with longer time to process the face sets, participants were still quite confident about their incorrect selection of the average face that was morphed across four study faces. They displayed awareness for some elements of the studied faces, but they were not aware that this specific face had not been studied. The near-zero CA relation for average faces confirmed the lack of insight into the implicit average perception. On one hand, participants had more endorsement for the member than the average faces, and even had generally higher confidence ratings for the former than the latter (OSM), implying the ability to perceptually discriminate these two kinds of stimuli; on the other

hand, our results suggest that they might not be aware of the differences between the member and average faces at a conscious level, and they seemed to lack the knowledge about the potential averaging process when averaging is not explicitly required.

When participants needed to explicitly judge whether the test face was the average of the previously presented faces, as in Experiment 2, performance was above chance level but worse than identification of member faces. In previous studies using different averaging tasks (Haberman & Whitney, 2009; Ji, Rossi, & Pourtois, 2018; Yamanashi Leib et al., 2014), a similar level of individual face coding and ensemble face coding has been found. The better performance of identifying the member than the average faces in the current study might be due to the longer processing time we provided. It has been shown that individual representations of face identities could be advantageous over ensemble representations when the former become sufficiently detailed (Li et al., 2016; Neumann et al., 2017). More importantly, our new results showed that people were more confident about their correct identification responses than the incorrect ones for the average faces, implying that they did have insight into average face perception, when averaging is explicitly required. This finding contrasts with the lack of insight observed in Experiment 1 when averaging was judged implicitly. There are large numbers of studies demonstrating that participants can extract summary statistics implicitly and explicitly (Whitney & Leib, 2018). A recent study showed similar dissociations between implicit and explicit ensemble representations (Hansmann-Roth, Kristjansson, Whitney, Chetverikov, 2020). To the best of our knowledge, the current study is the first to move beyond ensemble perception and to explore metacognition of the averaging processes using retrospective confidence judgments. It appears that ensemble perception can be an unconscious process and can be open to metacognitive introspection at the same time. These two aspects are not necessarily contradictory. The average face matched the mean impression from the four target faces in the set, and thus participants responded with relatively high confidence. The factor of averageness was an interference factor that would lead to incorrect responses in Experiment 1 but it was an explicit requirement in Experiment 2. In addition, the average face was needed to be discriminated from member faces in Experiment 1, but only from non-matching averages in Experiment 2. Ensemble perception is quite flexible (Haberman & Whitney, 2009; Whitney & Leib, 2018), and metacognition about it might be dependent on the nature of the judgment required.

On the other hand, when the test face was the true member that was present previously in the set, positive CA correlations were observed under both duration conditions in two experiments, suggesting that a certain amount of insight to face identification did exist, which was consistent with previous findings from eyewitness studies (Brewer & Wells, 2011;

Sporer et al., 1995). The insight into *fine-grained* face identification, as when the average face had to be discriminated from member faces in Experiment 1, is poorer, however, relative to what is found in Experiment 2 where member and average face task was blocked (OSM). Previously, some researchers have found that the CA relations were smaller when the similarity between the target and the foils was high compared to the low-similarity condition (Brewer & Wells, 2006; Brigham, 1990). The average face, which is averaged across all the member faces in the target set, is necessarily similar to any of the member faces in the current study. When the foil face was perceptually similar to the target faces (e.g., the average of the member faces), the CA correlations dropped significantly. Admittedly, we did not control or directly measure perceptual similarities between multiple targets and between targets and test faces in the current study. Considering the potential modulatory role of face similarity, future studies could systematically manipulate this factor to further explore metacognition of face averaging and face identification.

Previously, researchers proposed an optimality hypothesis for the CA relationship, in that it becomes stronger when the processing condition is better, for example under longer exposures (Bothwell et al., 1987; Deffenbacher, 1980), but our findings regarding the duration effect on CA relations were not unambiguous. In Experiment 1 using the member identification task, CA relations increased with increasing duration when we analyzed the member and average faces separately; however, there was no significant effect of duration shown by an overall gamma correlation that treated the average faces as one kind of non-member faces. In Experiment 2 using separate member and average face identification tasks, we did not observe duration effects on CA relations. The response type (positive or negative decisions) did not modulate the duration effects in either of the experiments. Extending the presentation time of 2 s to 5 s might not be long enough to enable participants to make more appropriate confidence estimates. Some previous eyewitness studies found that the CA relation was stronger when exposure was short relative to long (12 s vs. 45 s, Memon et al., 2003; 5 s vs. 90 s, Palmer et al., 2013). It is also likely that there are stimulus-specific CA relations (Maniscalco & Lau, 2014), as we found a different impact of duration on matching and non-matching average faces in Experiment 1. In addition, the gamma correlation as a measure of CA relations has its own limitations (Masson & Rotello, 2009). For instance, it may be impacted by response biases. If the response biases or the processing strategies (Ying, Burns, Choo, & Xu, 2020) were different between conditions or across participants, they would bias the results of CA relations. Moreover, some researchers have distinguished metacognition sensitivity and metacognition efficiency (Fleming & Lau, 2014). The gamma correlations we used in the current study represent metacognition sensitivity, which differs with task difficulty, for example. When task performance is better,

metacognition sensitivity is usually better as well. Metacognition efficiency controls the effect of task difficulty and is ideally constant across different levels of task performance. In this case, metacognition efficiency would be hypothesized not to be impacted by duration of face sets. It would be interesting in future studies to use shorter (e.g., 500 ms) presentation time as a baseline, and some other response bias-free measures of metacognition sensitivity (e.g., the receiver operating characteristics analysis, Benjamin & Diaz, 2008; Fleming et al., 2010) and of metacognition efficiency (e.g., relative meta-d' measures, Maniscalco & Lau, 2012, 2014) to further examine metacognition of (average) face identification and its potential modulations.

In sum, we provide evidence for the first time that people have insight into face averaging when it is required explicitly. In contrast, however, participants did not demonstrate insight into the identity of face averages when they needed to be discriminated from previously displayed face identities. We found that people do have metacognition of face identification to some extent, but the insight to fine-grained face identification is poor. Metacognition is considered a hallmark of intelligent behavior (Deroy, Spence, & Noppeney, 2016), and is closely related to awareness and consciousness (Clifford, Arabzadeh, & Harris, 2008; Persaud, McLeod, & Cowey, 2007). The current study represents an important starting point to explore the relationship between awareness and ensemble perception.

Acknowledgements We thank Robson Chow and Natasha Banger for stimuli preparation and data collection. This work was supported by a grant from the Hong Kong Research Grants Council (HKU17608519).

Open Practice Statement The datasets generated and analyzed during the current study are available in Open Science Framework at https://osf.io/9c2vm/?view_only=6c40617ce5d546e0aa5a7a2dd2bfe25a. The materials of the experiment are available upon contacting the corresponding author.

References

- Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, 12(2), 157–162. <https://doi.org/10.1111/1467-9280.00327>
- Arizpe, J. M., Saad, E., Douglas, A. O., Germine, L., Wilmer, J. B., & DeGutis, J. M. (2019). Self-reported face recognition is highly valid, but alone is not highly discriminative of prosopagnosia-level performance on objective assessments. *Behavior Research Methods*, 51(3), 1102–1116. <https://doi.org/10.3758/s13428-018-01195-w>
- Benjamin, A. S., & Diaz, M. (2008). Measurement of relative metamnemonic accuracy. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of Metamemory and Memory* (73–94). New York, NY: Psychology Press.
- Bernbach, H. A. (1971). Strength theory and confidence ratings in recall. *Psychological Review*, 78(4), 338–340. <https://doi.org/10.1037/h0031034>
- Bex, P. J., & Makous, W. (2002). Spatial frequency, phase, and the contrast of natural images. *Journal of the Optical Society of America A*, 19(6), 1096–1106. <https://doi.org/10.1364/JOSAA.19.001096>
- Bobak, A. K., Mileva, V. R., & Hancock, P. J. (2018). Facing the facts: Naive participants have only moderate insight into their face recognition and face perception abilities. *Quarterly Journal of Experimental Psychology*, 174702181877614. <https://doi.org/10.1177/1747021818776145>
- Bornstein, B. H., Deffebacher, K. A., Penrod, S. D., & McGorty, E. K. (2012). Effects of exposure time and cognitive operations on facial identification accuracy: A meta-analysis of two variables associated with initial memory strength. *Psychology, Crime and Law*, 18(5), 473–490. <https://doi.org/10.1080/1068316X.2010.508458>
- Bothwell, R. K., Deffenbacher, K. A., & Brigham, J. C. (1987). Correlation of eyewitness accuracy and confidence: Optimality hypothesis revisited. *Journal of Applied Psychology*, 72(4), 691–695. <https://doi.org/10.1037/0021-9010.72.4.691>
- Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science*, 22(3), 384–392. <https://doi.org/10.1177/0956797610397956>
- Brewer, N., & Wells, G. L. (2011). Eyewitness identification. *Current Directions in Psychological Science*, 20(1), 24–27. <https://doi.org/10.1177/0963721410389169>
- Brewer, N., & Wells, G. L. (2006). The confidence-accuracy relationship in eyewitness identification: Effects of lineup instructions, foil similarity, and target-absent base rates. *Journal of Experimental Psychology: Applied*, 12(1), 11–30. <https://doi.org/10.1037/1076-898X.12.1.11>
- Brigham, J. C. (1990). Target person distinctiveness and attractiveness as moderator variables in the confidence-accuracy relationship in eyewitness identifications. *Basic and Applied Social Psychology*, 11, 101–115. https://doi.org/10.1207/s15324834basps1101_7
- Clifford, C. W. G., Arabzadeh, E., & Harris, J. A. (2008). Getting technical about awareness. *Trends in Cognitive Sciences*, 12(2), 54–58. <https://doi.org/10.1016/j.tics.2007.11.009>
- de Fockert, J., & Wolfenstein, C. (2009). Rapid extraction of mean identity from sets of faces. *The Quarterly Journal of Experimental Psychology*, 62(9), 1716–1722. <https://doi.org/10.1016/j.cub.2007.06.039>
- Deffenbacher, K. A. (1980). Eyewitness accuracy and confidence: Can we infer anything about their relationship? *Law and Human Behavior*, 4, 243–260. <https://doi.org/10.1007/BF01040617>
- Deroy, O., Spence, C., & Noppeney, U. (2016). Metacognition in multi-sensory perception. *Trends in Cognitive Sciences*, 20(10), 736–747. <https://doi.org/10.1016/j.tics.2016.08.006>
- Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, 8, 1–9. <https://doi.org/10.3389/fnhum.2014.00443>
- Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science*, 329(5998), 1541–1543. <https://doi.org/10.1126/science.1191883>
- Goodman, L., & Kruskal, W. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, 49(268), 732–764. <https://doi.org/10.2307/2281536>
- Griffiths, S., Rhodes, G., Jeffery, L., Palermo, R., & Neumann, M. F. (2017). The average facial expression of a crowd influences impressions of individual expressions. *Journal of Experimental Psychology: Human Perception and Performance*. <https://doi.org/10.1037/xhp0000446>
- Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology*, 17(17), 751–753. <https://doi.org/10.1016/j.cub.2007.06.039>
- Haberman, J., & Whitney, D. (2009). Seeing the mean: ensemble coding for sets of faces. *Journal of Experimental Psychology: Human*

- Perception and Performance*, 35(3), 718–734. <https://doi.org/10.1037/a0013899>
- Hansmann-Roth, S., Kristjánsson, A., Whitney, D., & Chetverikov, A. (2020, March 5). Limits of perception and richness of behaviour: Dissociating implicit and explicit ensemble representations. <https://doi.org/10.31234/osf.io/3y4pz>
- Ji, L., Chen, W., Loeys, T., & Pourtois, G. (2018). Ensemble representation for multiple facial expressions: Evidence for a capacity limited perceptual process. *Journal of Vision*, 18(3), 1–19. <https://doi.org/10.1167/18.3.17>
- Ji, L., Rossi, V., & Pourtois, G. (2018). Mean emotion from multiple facial expressions can be extracted with limited attention: Evidence from visual ERPs. *Neuropsychologia*, 111, 92–102. <https://doi.org/10.1016/j.neuropsychologia.2018.01.022>
- Kelly, K. J., & Metcalfe, J. (2011). Metacognition of emotional face recognition. *Emotion*, 11(4), 896–906. <https://doi.org/10.1037/a0023746>
- Li, H., Ji, L., Tong, K., Ren, N., Chen, W., Liu, C. H., & Fu, X. (2016). Processing of individual items during ensemble coding of facial expressions. *Frontiers in Psychology*, 7(SEP), 1–11. <https://doi.org/10.3389/fpsyg.2016.01332>
- Lindsay, D., Read, J., & Sharma, K. (1998). Accuracy and confidence in person identification: The relationship is strong when witnessing conditions vary widely. *Psychological Science*, 9(3), 215–218. Retrieved from www.jstor.org/stable/40063282
- Livingston, L. A., & Shah, P. (2018). People with and without prosopagnosia have insight into their face recognition ability. *Quarterly Journal of Experimental Psychology*, 71(5), 1260–1262. <https://doi.org/10.1080/17470218.2017.1310911>
- Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, 21(1), 422–430. <https://doi.org/10.1016/j.concog.2011.09.021>
- Maniscalco, B., & Lau, H. (2014). Signal detection theory analysis of type 1 and type 2 data: Meta-d', response-specific meta-d', and the unequal variance SDT model. In *The Cognitive Neuroscience of Metacognition* (Vol. 9783642451, pp. 25–66). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-45190-4_3
- Masson, M. E. J., & Rotello, C. M. (2009). Sources of bias in the Goodman-Kruskal gamma coefficient measure of association: Implications for studies of metacognitive processes. *Journal of Experimental Psychology: Learning Memory and Cognition*, 35(2), 509–527. <https://doi.org/10.1037/a0014876>
- Memon, A., Hope, L., & Bull, R. (2003). Exposure duration: Effects on eyewitness accuracy and confidence. *British Journal of Psychology*, 94(3), 339–354. Retrieved from <http://search.proquest.com/psychology/docview/199650697/abstract/BCBAA07A53BD4255PQ/50?accountid=14777>
- Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, 95(1), 109–133. <https://doi.org/10.1037//0033-2909.95.1.109>
- Neumann, M. F., Ng, R., Rhodes, G., & Palermo, R. (2017). Ensemble coding of face identity is not independent of the coding of individual identity. *The Quarterly Journal of Experimental Psychology*, 1–27. <https://doi.org/10.1080/17470218.2017.1318409>
- Neumann, M. F., Schweinberger, S. R., & Burton, A. M. (2013). Viewers extract mean and individual identity from sets of famous faces. *Cognition*, 128(1), 56–63. <https://doi.org/10.1016/j.cognition.2013.03.006>
- Palermo, R., Rossion, B., Rhodes, G., Laguesse, R., Tez, T., Hall, B., ... McKone, E. (2017). Do people have insight into their face recognition abilities? *Quarterly Journal of Experimental Psychology*, 70(2), 218–233. <https://doi.org/10.1080/17470218.2016.1161058>
- Palmer, M. A., Brewer, N., Weber, N., & Nagesh, A. (2013). The confidence-accuracy relationship for eyewitness identification decisions: Effects of exposure duration, retention interval, and divided attention. *Journal of Experimental Psychology: Applied*, 19(1), 55–71. <https://doi.org/10.1037/a0031602>
- Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., Lindeløv, J. (2019). PsychoPy2: experiments in behavior made easy. *Behavior Research Methods*, 51, 195–203. <https://doi.org/10.3758/s13428-018-01193-y>
- Persaud, N., McLeod, P., & Cowey, A. (2007). Post-decision wagering objectively measures awareness. *Nature Neuroscience*, 10(2), 257–261. <https://doi.org/10.1038/nn1840>
- Sama, M. A., Nestor, A., & Cant, J. S. (2019). Independence of viewpoint and identity in face ensemble processing. *Journal of Vision*, 19(5), 2. <https://doi.org/10.1167/19.5.2>
- Sauer, J. D., Brewer, N., & Weber, N. (2008). Multiple confidence estimates as indices of eyewitness memory. *Journal of Experimental Psychology: General*, 137(3), 528–547. <https://doi.org/10.1037/a0012712>
- Shah, P., Gaule, A., Sowden, S., Bird, G., & Cook, R. (2015). The 20-item prosopagnosia index (PI20): A self-report instrument for identifying developmental prosopagnosia. *Royal Society Open Science*, 2(6), 140343. <https://doi.org/10.1098/rsos.140343>
- Sporer, S. L., Penrod, S., Read, D., & Cutler, B. (1995). Choosing, confidence, and accuracy: A meta-analysis of the confidence-accuracy relation in eyewitness identification studies. *Psychological Bulletin*, 118(3), 315–327. <https://doi.org/10.1037/0033-2909.118.3.315>
- Ventura, P., Livingston, L. A., & Shah, P. (2018). Adults have moderate-to-good insight into their face recognition ability: Further validation of the 20-item Prosopagnosia Index in a Portuguese sample. *Quarterly Journal of Experimental Psychology*, 71(12), 2677–2679. <https://doi.org/10.1177/1747021818765652>
- Whitney, D., & Yamanashi Leib, A. Y. (2018). Ensemble perception. *Annual Review of Psychology*, 69, 12.1–12.25. <https://doi.org/10.1093/acprof:osobl/9780199734337.003.0030>
- Yamanashi Leib, A., Fischer, J., Liu, Y., Qiu, S., Robertson, L., & Whitney, D. (2014). Ensemble crowd perception: A viewpoint-invariant mechanism to represent average crowd identity. *Journal of Vision*, 14(8), 1–13. <https://doi.org/10.1167/14.8.26>
- Ying, H., Burns J. E. J., Choo, A. M., & Xu, H. (2020). Temporal and spatial ensemble statistics are formed by distinct mechanisms. *Cognition*, 195, 104128. <https://doi.org/10.1016/j.cognition.2019.104128>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.