Ensemble perception of faces with naturalistic occlusions

Hayden Schill Hendley Department of Psychology, University of California, Riverside, CA, USA

Natalia K. Pallis Hassani Department of Psychology, University of California, Los Angeles, CA, USA

Timothy F. Brady Department of Psychology, University of California, San Diego, CA, USA

The visual system takes advantage of redundancy in the world by extracting summary statistics, a phenomenon known as ensemble perception. Ensemble representations are formed for low-level features like orientation and size and high-level features such as facial identity and expression. Whereas recent research has shown that the visual system forms intact ensemble representations even when faces are partially occluded via solid bars, how ensemble perception is impacted with the addition of naturalistic objects such as face masks or sunglasses is largely unknown. To investigate this, we conducted a series of experiments using continuous report tasks in which faces (either varying in identity or expression) were partially occluded with a surgical mask or sunglasses and participants had to report the average face using a face wheel. We found evidence that participants could still accurately extract the average even when a significant portion of it was occluded with either face masks or sunglasses. In a second experiment, however, we found performance was worse when the face wheel was variable trial to trial. Thus part of the preservation of performance in occlusion arises from the visual system learning the features of the particular face wheel being used. Overall, our results suggest that the visual system is able to establish robust ensemble representations for faces with naturalistic occlusions, but that robustness appears to be supported at least partially by learning information about the particular features that are informative for a given set of faces.

Introduction

As we move about the world, we are bombarded with far more information than our visual system can process at one time. Yet we can successfully navigate and process key information, in part because the visual system has developed tools and heuristics to counter our limited capacity systems. Research suggests ensemble perception may be one way we bypass the severe capacity limitations of visual working memory and attentional systems (Alvarez, 2011; Haberman & Whitney, 2012; Whitney, Haberman, & Sweeny, 2014; Wolfe & Horowitz, 2004). Ensemble perception takes advantage of redundancy in the world (e.g., blades of grass, faces in a crowd) by computing the statistical average of visual information quickly and with minimal processing power (Alvarez & Oliva, 2009; Haberman, Brady & Alvarez, 2015). Increasing the number of items on the screen, for instance, has a detrimental impact on the number of individual items we can reliably process, but does not strongly affect our ability to extract the mean (Chong & Treisman, 2003; Haberman et al., 2015). The resulting average is often called the ensemble percept or ensemble representation and is a useful heuristic to guide decisions and behavior without needing to process every individual instance of redundant visual information.

Ensemble representations support our visual percept at many levels of visual processing, operating on simple features such as size, orientation, and motion, but also more complex features like facial identity, expression, gender, and ethnicity (Haberman & Whitney, 2007; Haberman, et al., 2015; Haberman, Lee, & Whitney, 2015; Leib et al., 2014; Smith, Cottrell, Gosselin, & Schyns, 2005; Tanaka & Farah, 1993). Even for high-level, complex information such as faces, the ensemble percept can be extracted without the need to fully process—or even be aware of—individual items (Haberman & Whitney, 2009; Wolfe, Kosovicheva, Leib, Wood, & Whitney, 2015). For instance, participants can still recognize changes in the average face even when failing to recognize which individual changes are driving this change (Haberman & Whitney, 2011) and can extract the average when unable to recognize individual faces because of crowding or noise (Fischer & Whitney, 2011). Further, participants can extract average walking directions from crowds of people despite individual

Citation: Hendley, H. S., Pallis Hassani, N. K., & Brady, T. F. (2025). Ensemble perception of faces with naturalistic occlusions. *Journal of Vision*, *25*(10):5, 1–15, https://doi.org/10.1167/jov.25.10.5.

https://doi.org/10.1167/jov.25.10.5 Received June 27, 2024; published August 8, 2025 ISSN 1534-7362

walking directions being highly variable (Sweeny, Haroz, & Whitney, 2012), demonstrating how ensemble perception can overcome variations in viewpoint and direction in real-world contexts. Given the ubiquity of ensemble perception across many types of information and its ability to relay useful information without needing to process all of the information present, ensemble extraction may be a fundamental mechanism supporting our visual experience.

Incoming visual information is noisy and, in the real world, often only partially intact, such as when objects are partially occluded by another object (e.g., your car partially obstructed by another parked vehicle) or when the viewpoint masks relevant features (e.g., a face turned to the side). Given these real-world complexities, there is a growing body of work investigating how perception of individual items, as well as ensemble perception, is impacted by the presence of incomplete or limited information. Several studies using stimuli that support amodal completion, which prompt the visual system to perceive the "whole" despite part of the stimulus being absent or occluded, have found that ensemble representations of both low- and high-level information maintain precision and accuracy despite the incomplete information (Haberman & Ulrich, 2019). In particular, Haberman and Ulrich (2019) found evidence that the ensemble representation of both facial identity and expression can still be derived when portions of faces are covered with artificial black bars, but that this occurred more accurately when the bars appeared to be "in front" of the faces (allowing amodal completion) but not when the bars appeared to be "behind" the faces, suggesting a role of surfaces. Interestingly, the researchers did not find evidence that individuals create a "completed" face to use in their representation of the missing information; in other words, forming the ensemble percept was not contingent on representing the missing information (Haberman & Ulrich, 2019). The current study expands on this work by investigating how ensemble perception is impacted by types of occlusions that occur more naturally in the world—sunglasses and face masks—where there is no explicit manipulation of surface-based representations.

Since the COVID-19 pandemic, there has been a significant increase in research on how more ecologically valid or naturalistic occlusions impact the perception of individual faces (Lander & Saunders, 2023). This work has found that face masks generally decrease our ability to recall or recognize the identity of a face (Carragher & Hancock, 2020; Freud, Stajduhar, Rosenbaum, Avidan, & Ganel, 2020), although some information such as a sense of familiarity with famous faces remains intact (Carlaw, Huebert, McNeely-White, Rhodes, & Cleary, 2022). Carlaw et al. (2022) also found that recognition impairment was greater when the faces were wearing sunglasses compared to masks, which aligns with a large body of research showing the importance of eyes in recognizing faces (Davies, Ellis, & Shepherd, 1977; McKelvie, 1976).

The challenge that such occlusions pose to individual face recognition occurs at least in part because there is less evidence observers can amass in order to make a decision (Carragher, Towler, Mileva, White, & Hancock, 2022; McKelvie, 1976), leading to masked face recognition relying more heavily on individual features opposed to holistic processing (Freud et al. 2020). It remains unclear whether pooling information across multiple faces helps reduce the impact of masks and other naturalistic occlusions, potentially leading to less impairment to ensemble perception compared to individual face perception.

Using an artificial occluder, like a black bar, may in many cases objectively remove the same information as a more naturalistic occluder, like a mask. However, it may not always be solely the objective removal of information that limits performance in face recognition tasks (e.g., the same information is removed in both the "in front" and "behind" conditions in Haberman and Ulrich, 2019, but they result in very different performance). For example, Carragher et al. (2022) assessed whether a training strategy designed to help observers maximize their awareness and use of available features (a strategy known to be used by professional face examiners) might mitigate the impairment face masks cause on identification and communication. They found that even brief diagnostic feature training (i.e., telling participants to focus on the available information—the other facial features) improved recognition performance significantly (Carragher et al., 2022). This research suggests that for individual faces with naturalistic occlusions, observers learn to take advantage of available facial feature information to help overcome the decrease in available usable information from the occluder. Occlusions that have their own meaning, like masks or sunglasses, may therefore give viewers insight into other information about the face, such as affective states, that impact performance above and beyond the objective removal of information. Nojavanasghari, Hughes, Baltrušaitis, and Morency (2017), for instance, tested individual facial perception with hand occlusion on the basis that hand occlusion can provide information about the state and situation of the face they are viewing. This study revealed that images with hand occlusions over regions of the face with discriminative power, such as the eyes and mouth, were easier to recognize than images with hand occlusions over the outer edge and nose of the face. Likewise, we suggest that naturalistic occlusion using a mask or sunglasses might impact perception differently than artificial bars because of the situational information they hold. This effect is equally important in ensemble perception, as situation and state can impact our ability to average identity and, particularly, the emotion of faces.

In addition to occlusions, the precision of ensemble representations is systematically affected by the variability present in the display. When items in a set are more similar to each other, ensemble estimates become more precise - whether averaging colors (Maule & Franklin, 2015), sizes (Solomon, Morgan, & Chubb, 2011), or faces (Haberman et al., 2015). This relationship between set variability and ensemble precision appears to be a fundamental property of ensemble coding, reflecting both the statistical uncertainty inherent in estimating a mean from noisy samples and the neural mechanisms that implement ensemble processing (Alvarez, 2011; Whitney et al., 2014). In ecological contexts, this sensitivity to variability is typically adaptive—when items in the environment are highly similar (like leaves on a single tree), their mean provides a meaningful summary, but when items are highly variable (like different species of trees in a forest), the mean becomes less representative of any individual item.

In the current series of experiments, we examined our ability to extract the ensemble with naturalistic occlusions. Using a blocked design in Experiment 1, we explored the impact of two types of occlusions (face masks and sunglasses) on our ability to represent the ensemble of two types of facial information (identity and expression). The variability in the set of faces was also varied, creating some conditions where extracting ensemble information was more difficult and less useful because the items varied too much for the ensemble to be a useful summary. In Experiment 2, we interleave conditions and stimulus sets together to assess whether access to the same available facial features over time impacts ensemble perception. As described earlier, this kind of facial feature learning has been shown to contribute to face processing of individual faces with masks or sunglasses (Carragher et al., 2022). Finally, we include a single-face condition, which serves as a baseline to determine whether adding these occluders disrupt ensemble perception differently than in the perception of a single face. The results of the current experiments add to a growing literature that looks at how face masks impact face perception (Lander & Saunders, 2023), while also contributing to our understanding of ensemble perception broadly. To preview our results, we find that participants can derive the ensemble of faces despite added naturalistic occlusions but that there is some cost and evidence of a significant role of feature learning.

Experiment 1

In Experiment 1, we explored the impact of face masks and sunglasses on the ensemble representation of facial identity and expression. Although there is growing research on the impact of occlusions such as face masks on individual face perception (Lander & Saunders, 2023), the current study is one of the first to investigate the influence of naturalistic occlusions on ensemble perception.

Experiment 1 consisted of four across-participant conditions, with different stimuli, which we call Experiments 1A, 1B, 1C, and 1D: a face wheel that varies in identity, with masks; a face wheel that varies in identity, with sunglasses; a face wheel that varies in expression, with masks; and a face wheel that varies in expression, with sunglasses. In each of these, there were individual item and ensemble within-participant conditions, both with and without these occlusions.

Experiment 1 methods

Participants

One hundred fifteen total participants completed Experiment 1 (39 in 1A, 18 in 1B, 29 in 1C, and 14 in 1D). The mask conditions (1A and 1C) were run first, with a goal of approximately 40 participants, and the sunglasses conditions were run second, with a goal of approximately 20 participants. However, participants were recruited for each for a fixed period of time, rather than a fixed number of participants, and not all participants who registered submitted data, resulting in small differences in participant number. No participants were excluded. All participants were University of California San Diego undergraduate students who gave informed consent and were compensated with course credit. Informed consent procedures were approved by the Institutional Review Board of the University of California, San Diego.

Stimuli and design

Participants were presented with a single face or groups of four faces varying in either identity (Experiments 1A and 1B) or expression (Experiments 1C and 1D). Our stimulus sets were the same as those used in Haberman and Ulrich (2019) from the Harvard Face Database, which consists of 360 linearly interpolated face morphs based on three male faces. Face morphs were separated into either identity or expression "units," corresponding to approximately equal steps in a 360° circular stimulus space. For Experiments 1A and 1C (mask occluders), we made a copy of each face wheel (identity and expression) and added standard white surgical masks onto the faces, which covered the lower half of the face. For Experiments 1B and 1D, we added opaque black sunglasses onto a copy of both wheels. Taken together, Experiment 1 had six face wheels—identity wheel with no occlusion, identity wheel with masks, identity wheel with sunglasses, expression wheel with no occlusion,

expression wheel with masks, and expression wheel with sunglasses.

In each sub-experiment, we tested performance on a single face versus the ensemble and occlusion versus no occlusion. This led to each sub-experiment consisting of four counterbalanced blocks: single face (occluded or unoccluded) and ensemble (occluded or unoccluded). The unoccluded conditions serve as a baseline to examine the extent to which adding naturalistic occlusions impacts our ability to accurately represent the identity or expression of a single or average group of faces. Each face was presented in grayscale and approximately 125 × 125 pixels in size. The mean of each set was chosen randomly on each trial.

Procedure: The same procedure was followed for experiments 1A, 1B, 1C, and 1D. On each trial in an ensemble block, participants were told to report the average identity or expression of a set of faces. Four faces appeared on the screen for 1 second, followed by a 250-ms interstimulus interval, and then a single test face appeared on the center of the screen (Figures 1 and 2). The test face always matched the condition of the previous set (e.g., if four faces with masks were shown, the test face had a mask; if four faces with sunglasses were shown, the test face

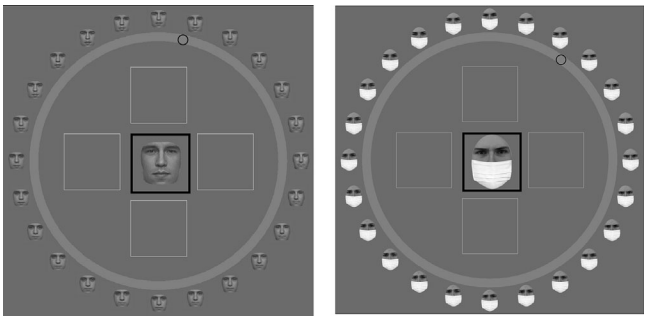

Figure 2. The response screen consisting of a "face wheel." Participants were told to move their mouse around the wheel until the middle test face matched what they felt represented the average of the faces presented on this trial (for ensemble trials) or the exact face shown (for single-item trials). The test faces always had the same occlusion as the faces shown in the trial (i.e., if the face(s) shown in the trial were masked, then the test face was masked).

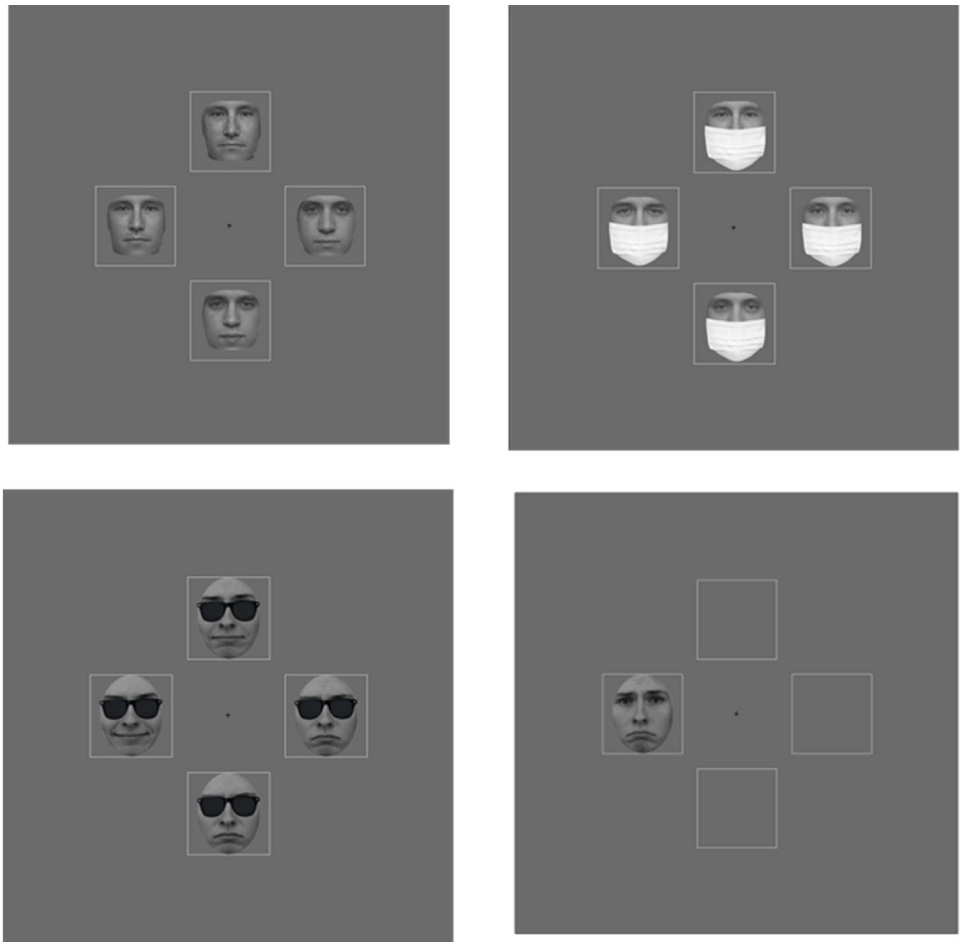

Figure 1. A selection of some of the key trial types in Experiment 1. The top left shows an unoccluded ensemble identity trial. The top right shows an ensemble identity trial with masks. The bottom left shows an ensemble expression trial with sunglasses. The bottom right shows a single face unoccluded expression trial.

wore sunglasses, etc.). As participants moved their mouse along the face wheel surrounding the test face (see Figure 2), the identity or expression of the test face morphed to match the corresponding location on the wheel. Each degree of change on the circle corresponded to a degree of change in either the identity or expression of the face. Once participants clicked on the location on the wheel that best matched their perceived average of the four preceding faces, they were shown a blank screen and told to click a small black cross in the middle of the screen to initiate the next trial. Performance was measured as the distance in degrees on the circle from where the participant clicked (the perceived average) to the actual mean of the set.

Single-face trials were largely the same as the ensemble trials, except only one face appeared on the screen on each trial. The location of the presented face was randomly chosen to appear in one of the four locations the ensemble set appeared in. Participants were told to replicate their perception of the single face on the test face.

Each of the four blocks contained 80 trials. The total time for each sub-experiment was approximately 40 minutes. Participants were given instructions before each block, noting whether they would be viewing a single face or ensemble block and whether the faces would have masks, sunglasses, or no occlusions.

Analysis and data clustering: In each sub-experiment, error was used as an index of performance, calculated as the circular standard deviation of the distance (in degrees) between the face the participant chose as the average with the true average. Smaller error (i.e., less distance between the true average and the perceived average) indicates a stronger or more precise representation of the ensemble percept or single face. This measure has been shown to correlate with model-based measures of accuracy in working memory tasks (Schurgin, Wixted, & Brady, 2020). Chance performance in the circular standard deviation measure is 81°. We calculated this by sampling completely uniform angles 10,000 times and calculating the circular standard deviation for each sample.

To test whether error changed with the addition of naturalistic occluders, both in the single face and ensemble conditions, we performed a series of null hypothesis significant *t*-tests and Bayesian *t*-tests (Rouder, Speckman, Sun, Morey, & Iverson, 2009). To calculate a Bayes factor, we applied the BayesFactor 0.9.12-2 R package which calculates a BF corresponding to a one-sample *t*-statistic. The use of Bayes factors is grounded in Jeffreys (1961) foundational framework for quantifying evidence in hypothesis testing. We used the default effect size for the alternative hypothesis (Cauchy with scale $r = \sqrt{2}/2$).

In a typical design of ensemble perception experiments, items are always clustered in feature space on every trial, which results in every trial having a well-defined ensemble (i.e., the average is no longer representative if the variability of the set is too high), but also creates a strong dependency between the faces (Maule & Franklin, 2015). In contrast, the four faces presented at encoding in ensemble blocks in Experiment 1 were randomly chosen, as they were for single face trials, meaning trials varied in how well-defined the "average" face was. Sometimes the four faces presented on a given trial might be clustered closer to the average (an easier trial), maximally separated along the wheel (a nearly impossible trial), or somewhere in between. The decision to generate the items randomly was chosen because it was the same way the single item was generated in the 1 item condition (e.g., nothing unique was done to control the items on the ensemble trials), but the effect this had on the availability of the ensemble information in ensemble trials was not intentional. However, it did offer an opportunity to ask an additional question not often offered by a typical ensemble perception design: we can also analyze the data by how clustered the faces were on the wheel (i.e., how difficult the ensemble trials were) and examine the impact of clustering distance on ensemble performance. While there has been some research looking at set similarity in ensemble perception for faces (Cabeza, Bruce, Kato, & Oda, 1999), none has assessed this with naturalistic occlusions.

We separated the data based on the mean resultant vector length of the four faces, which provided a measure of how clustered the faces were on the face wheel (and thus how easy an average would be to calculate). This allowed us to examine ensemble perception only on trials where the ensemble average was well-defined. If all four faces were exactly the same as each other, the additive vector would be 1. If the four faces were as far apart as they could be (e.g., spaced every 90°), the vector would be zero. When the four faces are clustered close together on the wheel, we expect ensemble report error to be reduced. When the four faces are located further apart from the mean, we anticipate the error to be greater as it is objectively difficult to choose a spot on the wheel that averages four almost equidistant locations. To visualize this, we parsed the data into thirds, which correspond to vector lengths of 0–0.25, 0.25–0.5, and 0.5–1, since 0.25 and 0.5 approximately correspond to the 33% and 66% percentile of mean resultant lengths in our experiment. See Figure 3 for a visualization of the error distributions based on the three clusters, using Experiment 1A's data as a case study. In each of the four sub-experiments below, when analyzing the ensemble data, we show all data but focus our analysis on trials where the mean resultant vector was greater than 0.5, which corresponds to a moderate cluster (where the ensemble is well-defined). This is because asking about the ensemble in trials where the items are

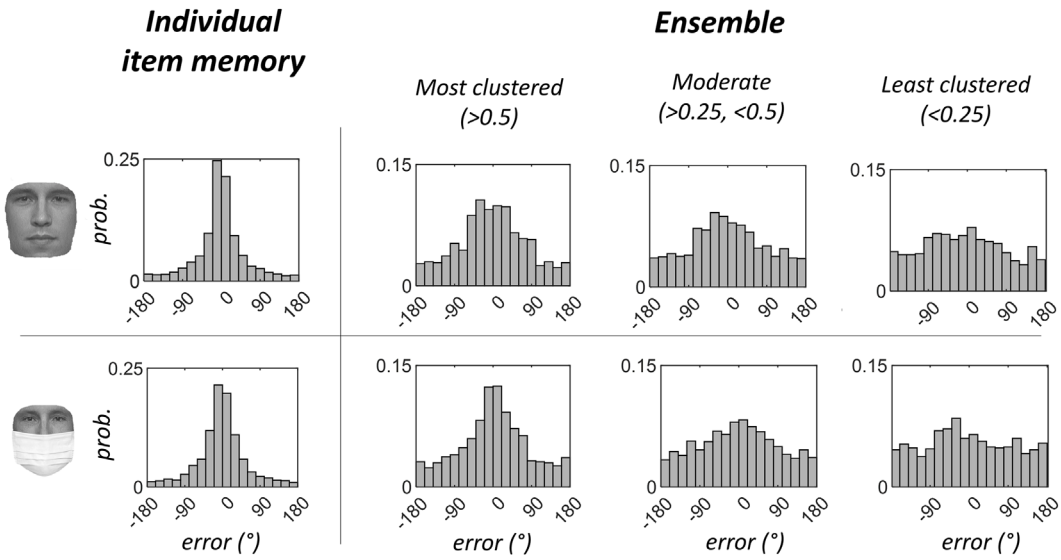

Figure 3. Histogram distribution of error from Experiment 1A as an illustration of performance in the individual item condition and the three levels of clustering in the ensemble task. Top row shows non-masked faces, and bottom shows masked faces. Errors near 0 are indicative of good performance.

not at all clustered is unlikely to be a sensitive measure of ensemble perception.

Experiment 1 results

Experiment 1A results—Identity masks

Experiment 1A examined the impact of masks on ensemble perception for facial identity. This sub-experiment had four blocks, counterbalanced across participants: single face (masked, no mask) and ensemble (masked, no mask).

We first analyzed the impact of adding face masks on our perception of a single face. We found that our ability to accurately represent a single face was not significantly impacted by the addition of a mask ($t(38) = 0.76$, $p = 0.45$, $dz = 0.12$). The data are approximately 4.42 times more likely under the null hypothesis (BF01 = 4.42). Next, we asked whether the addition of face masks hurt our ability to extract the average facial identity of the group of four faces varying in identity. Focusing only on trials with well-defined ensembles (resultant vector > 0.5), we found that our ability to extract the statistical average of facial identity was also not significantly impacted by the additional mask ($t(38) = 0.78$, $p = 0.44$, $dz = 0.13$), which suggests that ensemble perception is also robust to the added occlusion. The data are 4.31 times more likely under the null (BF01 = 4.31).

Experiment 1B results—Identity sunglasses

Experiment 1B examined the impact of sunglasses on identity ensemble perception. This sub-experiment again had four blocks, counterbalanced across participants: single face (sunglasses, no sunglasses,) and ensemble (sunglasses, no sunglasses).

Similar to Experiment 1A, we first analyzed the impact of adding sunglasses on our perception of a single face. We found that, similar to masks, our ability to accurately represent a single face was not significantly impacted by the addition of sunglasses ($t(38) = 0.99$, $p = 0.33$, $dz = 0.16$) and the data are approximately 2.68 times more likely under the null (BF01 = 2.68). Next, we asked whether the addition of sunglasses hurt our ability to extract the average identity of a set of four faces varying in identity. Similar to Experiment 1A, our ability to extract the statistical average of facial identity (on trials with resultant vectors > 0.5) was not significantly impacted by the additional pair of sunglasses ($t(38) = 1.66$, $p = 0.11$, $dz = 0.27$), and the data are approximately 1.64 times more likely under the null (BF01 = 1.64), which provides additional evidence that ensemble perception is robust to the added occlusion.

Experiment 1C results—Expression masks

Experiment 1C examined the impact of face masks on expression ensemble perception. This sub-experiment again had four blocks, counterbalanced across participants: single face (masked, no mask) and ensemble (masked, no mask).

Our findings were consistent with Experiments 1A and 1B. In analyzing the impact of face masks on our perception of a single face, we found no significant change in performance with the addition of a mask ($t(38) = 1.45$, $p = 0.15$, $dz = 0.23$). The data are

approximately 1.98 times more likely under the null (BF01 = 1.98). Next, we asked whether the addition of face masks hurt our ability to extract the average facial expression of the set. Similar to previous experiments, we found that our ability to extract the statistical average of facial expression on trials with resultant vectors > 0.5 was not significantly impacted by the additional mask ($t(38) = 1.02$, $p = 0.31$, $dz = 0.16$). The data were 3.53 times more likely under the null (BF01 = 3.53).

Experiment 1D results—expression sunglasses

Experiment 1D examined the impact of sunglasses on expression ensemble perception. This sub-experiment had four blocks, counterbalanced across participants: single face (sunglasses, no sunglasses) and ensemble (sunglasses, no sunglasses).

The findings of Experiment 1D were consistent with 1A–1C. We first analyzed the impact of adding sunglasses on our perception of a single face. We found that our ability to accurately represent a single face was not impacted by the addition of a pair of sunglasses ($t(38) = 0.18$, $p = 0.86$, $dz = 0.03$) and the data are approximately 4.15 times more likely under the null (BF01 = 4.15). We next asked whether the addition of sunglasses hurt our ability to extract the average facial expression of the set on trials with resultant vectors >0.5. Similar to the experiments above, we again found that our ability to extract the statistical average of facial expression was not significantly impacted by the additional pair of sunglasses ($t(38) = 0.26$, $p = 0.80$, $dz = 0.04$), and the data were 5.55 times more likely under the null (B01 = 5.55). See Figure 4 for a summary of all sub-experiment results. Taken together, these results suggest that ensemble perception is not impacted significantly by the addition of face masks or sunglasses. Performance remained quite accurate even in the presence of occlusions.

Across experiments

Experiments 1a–1d examined how face masks and sunglasses affect our ability to perceive facial identity and emotional expressions, both for single faces and groups of faces (ensemble perception). Although each individual experiment focused on specific aspects of face perception (identity or emotion) under different viewing conditions (unoccluded vs. masked or with sunglasses), we can also consider them together. Thus we conducted an overall analysis to understand the broader patterns across all of these experiments. We analyzed the combined data using a linear mixed-effects model with mask type (unoccluded, mask, or sunglasses), face wheel type (identity or emotion), and task type (single-face or ensemble judgment) as fixed effects and subjects as a random effect. Results revealed a significant main effect of task type ($F(1,395) = 143.05$, $p < 0.001$), with ensemble judgments showing higher error rates than single-face judgments ($\beta = 16.09°$). Neither the main effect of mask type ($F(2,395) = 0.55$, $p = 0.58$) nor face wheel type ($F(1,395) = 0.71$, $p = 0.40$) reached significance, indicating similar performance across occlusion conditions and across identity and emotion judgments, Bayesian analyses using the method of Wagenmakers (2007) provided strong evidence for these null effects, with the data being 230 times more likely under the null hypothesis than under an effect of mask type (BF01 = 230.19), and 14 times more likely under the null hypothesis than under an effect of face wheel type (BF01 = 14.0). Together, these results suggest that in these circumstances, people are very robust to partial occlusion of faces, maintaining similar levels of performance whether faces are unoccluded, wearing masks, or wearing sunglasses. This resilience holds true both for perceiving individual faces and for extracting ensemble information from groups of faces, though ensemble judgments were consistently more challenging than remembering a single individual face.

Although our primary analyses focused on highly clustered trials, we also examined how clustering affected performance across all conditions in Experiment 1 using a mixed-effects model. This analysis revealed a strong main effect of clustering ($F(2,591) = 37.21$, $p < 6.0 \times 10^{-16}$), with Bayesian analysis providing overwhelming evidence for this effect (BF10 > 7.5×10^{20}). In contrast, we found no evidence for an effect of mask type ($F(2,591) = 0.84$, $p = 0.43$), with the data being over 5.7×10^{7} times more likely under the null hypothesis (BF01 = 56,710,567). Importantly, there was also no interaction between clustering and mask type ($F(4,591) = 0.35$, $p = 0.84$, BF01 = 178,972), suggesting that the impact of clustering on performance was similar regardless of whether faces were occluded or not. These results validate our decision to focus on well-clustered trials in our primary analyses, as clustering has a substantial impact on performance, while also demonstrating that the null effect of occlusion we observed in well-clustered trials generalizes across all clustering levels.

Experiment 2: Feature learning

Experiment 1 revealed a surprising degree of robustness in ensemble perception: despite occlusions removing substantial face information, participants showed little to no impairment in their ability to extract averages of both identity and expression. This robustness occurred despite the fact that masks and sunglasses covered major diagnostic regions of

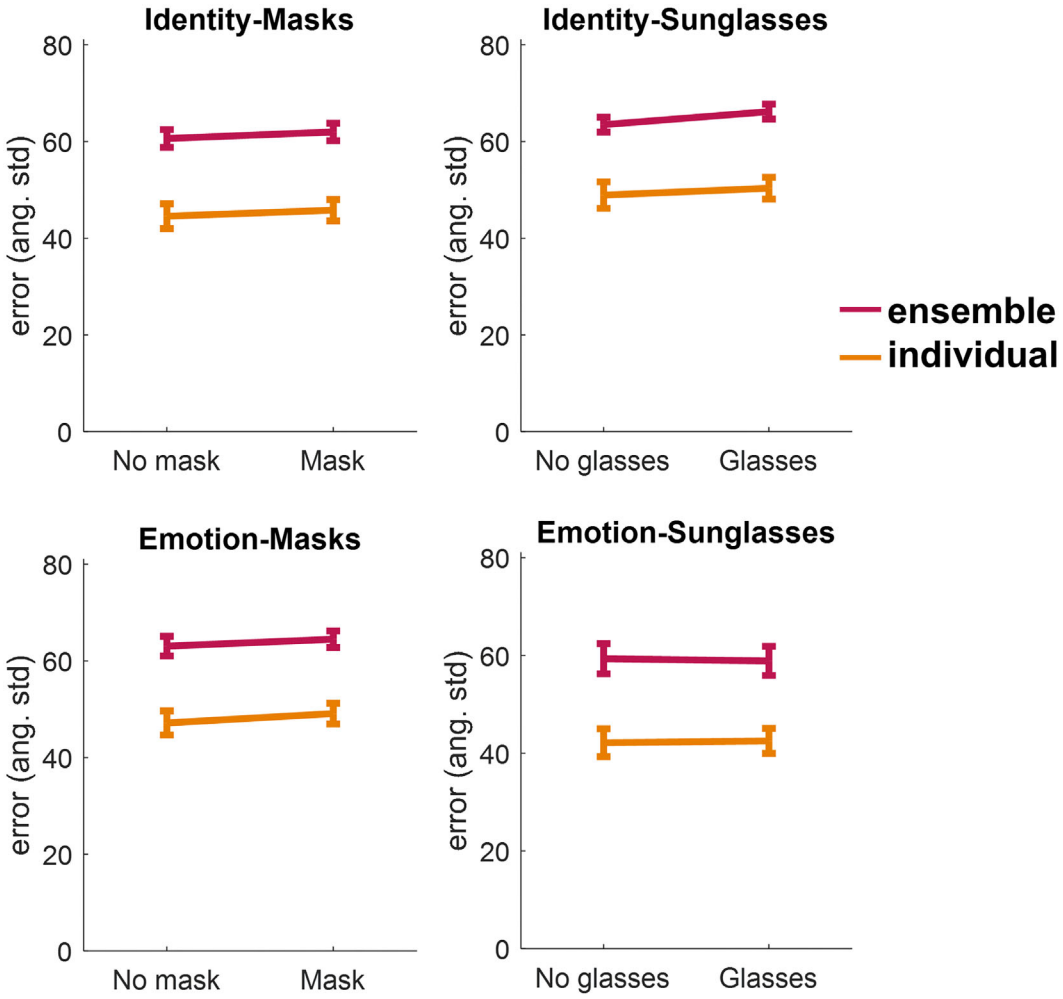

Figure 4. Top left: The results of Experiment 1A showing the impact of adding a mask on extracting the average identity of a group of faces (red) versus the identity of a single individual (orange). Top right: Results of Experiment 1B showing the impact of adding sunglasses on extracting the average identity of a group of faces (red) versus the identity of a single individual (orange). Bottom left: Results of Experiment 1C showing the impact of adding a mask on extracting the average expression of a group of faces (red) versus the identity of a single individual (orange). Bottom right: Results of Experiment 1D showing the impact of adding sunglasses on extracting the average expression of a group of faces (red) versus the identity of a single individual (orange).

the face—regions known to be critical for individual face perception. What might explain this preserved performance?

One possibility is that in Experiment 1's blocked design, where participants viewed the same type of stimuli (e.g., masked faces or faces with sunglasses) for 80 consecutive trials, they were able to develop and optimize their processing strategy. When viewing a series of masked faces chosen from the same stimulus wheel, viewers may have learned which features remained informative (e.g., the eyes for masked faces, the mouth for faces with sunglasses) and weighted those features more heavily when computing the ensemble. This kind of strategic adaptation to available information has been documented in individual face recognition, where observers learn to maximize use of visible features to overcome mask-related impairments (Carragher et al., 2022).

Experiment 2 was designed to test this feature-learning hypothesis in two ways. First, we replicated Experiment 1's blocked conditions but with consistently spaced ensembles (rather than variable spacing) to provide a more sensitive test of occlusion effects. Second, we added an interleaved condition where different types of trials (masked vs. sunglasses, identity vs. expression) were randomly intermixed, making it impossible to maintain a consistent feature-based strategy. If the robustness we observed in Experiment 1 relied on learning which features to attend to, performance should suffer more from occlusions when such learning is disrupted by trial-by-trial variation in the relevant features.

Methods

Participants

One hundred seventy-one total participants completed the experiment 2 (33 in 2A, 38 in 2B, 36 in 2C, 37 in 2D, and 41 in 2E). No participants were excluded. We aimed for approximately 30 participants per condition but again ran the experiment for a fixed amount of time rather than to a fixed number of participants. Participants were University of California San Diego undergraduate students. All participants gave informed consent and were compensated with course credit. Informed consent procedures were approved by the Institutional Review Board of the University of California, San Diego.

Stimuli, design, and procedure

Similar to Experiment 1, participants were presented with sets of faces varying in either emotion or identity and either wearing a mask, sunglasses, or no occlusion. Because we primarily wanted to compare blocked versus interleaved designs to assess the impact of feature learning, we did not include a single-face condition in Experiment 2. Similar to Experiment 1, participants completed two conditions out of eight possible conditions in the blocked experiment: ensemble identity (masked and unmasked, or sunglasses and no sunglasses) and ensemble expression (masked and unmasked, or sunglasses and no sunglasses). Thus, these sub-experiments were similar to Experiments 1A–1D but with two differences: every ensemble trial had the same clustering of faces (rather than varied similarity, and thus varied difficulty), and there were no individual face trials.

In the interleaved sub-experiment, all of the trials from the blocked experiment were interleaved together, such that participants did not know which type of trial type came next. Each participant in the interleaved sub-experiment completed all six trial types. A participant might have a trial asking them to average four faces with masks varying in identity, followed by four faces wearing sunglasses varying in expression, followed by four faces varying in expression with no occlusion, and more. This design discouraged participants from learning the features of any particular stimulus wheel as they did not know whether that information would be useful on any given trial. We predict that if participants are maximizing their use of knowledge about the available features to create the ensemble percept, then disrupting their ability to as easily learn the stimulus features will lead to a drop in performance.

The procedure for each trial was largely the same as in Experiment 1, where participants were shown an ensemble of four faces for one second followed by a 250 ms interstimulus interval and test face, where participants indicated the best representation of the perceived average of the four previous faces. Unlike Experiment 1 where spacing varied randomly, Experiment 2 used fixed spacing between the faces in each ensemble display. Faces appeared at −45°, −15°, 15°, and 45° relative to the mean, creating sets with a well-defined average (centered at 0°) and consistent variability across trials. This fixed spacing ensured that every trial had the same objective difficulty level in terms of computing the ensemble average, which is in line with previous experiments and corresponds to a circular resultant vector length of 0.84. We again used the circular standard deviation as our main measure of performance.

Experiment 2 consisted of five sub-experiments (2A–2E). The blocked design sub-experiments (2A–2D) each tested one type of occlusion (masks or sunglasses) with one type of face wheel (identity or expression), with participants completing 80 trials per condition. The interleaved design (2E) combined all conditions, with participants completing trials from all possible combinations of occlusion type and face wheel type. Although there were eight possible combinations of conditions (2 wheel types × 2 occlusion types × 2 occlusion states), in the interleaved condition these were presented as six trial types because each occlusion type was only compared to its own unoccluded baseline. Each participant in the interleaved condition completed 390 total trials.

Results

The results of Experiment 2 are displayed in Figure 5. We first analyzed the data from the blocked experiment, asking whether face masks or sunglasses hurt ensemble performance (2A–2D). Note the difference in the blocked version of Experiment 2 compared to Experiment 1 was that the clustering of faces presented was fixed in all conditions at a strong level of clustering (and no individual face trials were present). We subsequently analyzed data from the interleaved experiment (2E) and then compared blocked versus interleaved designs. Overall, we find costs to occlusion in many conditions, but find that interleaving results in major costs to performance compared to blocked—consistent with the idea that learning the features of a particular wheel helps support performance in the blocked conditions.

Experiment 2A blocked results—Identity masks

We first analyzed whether the added occlusions impacted our ability to extract the average identity.

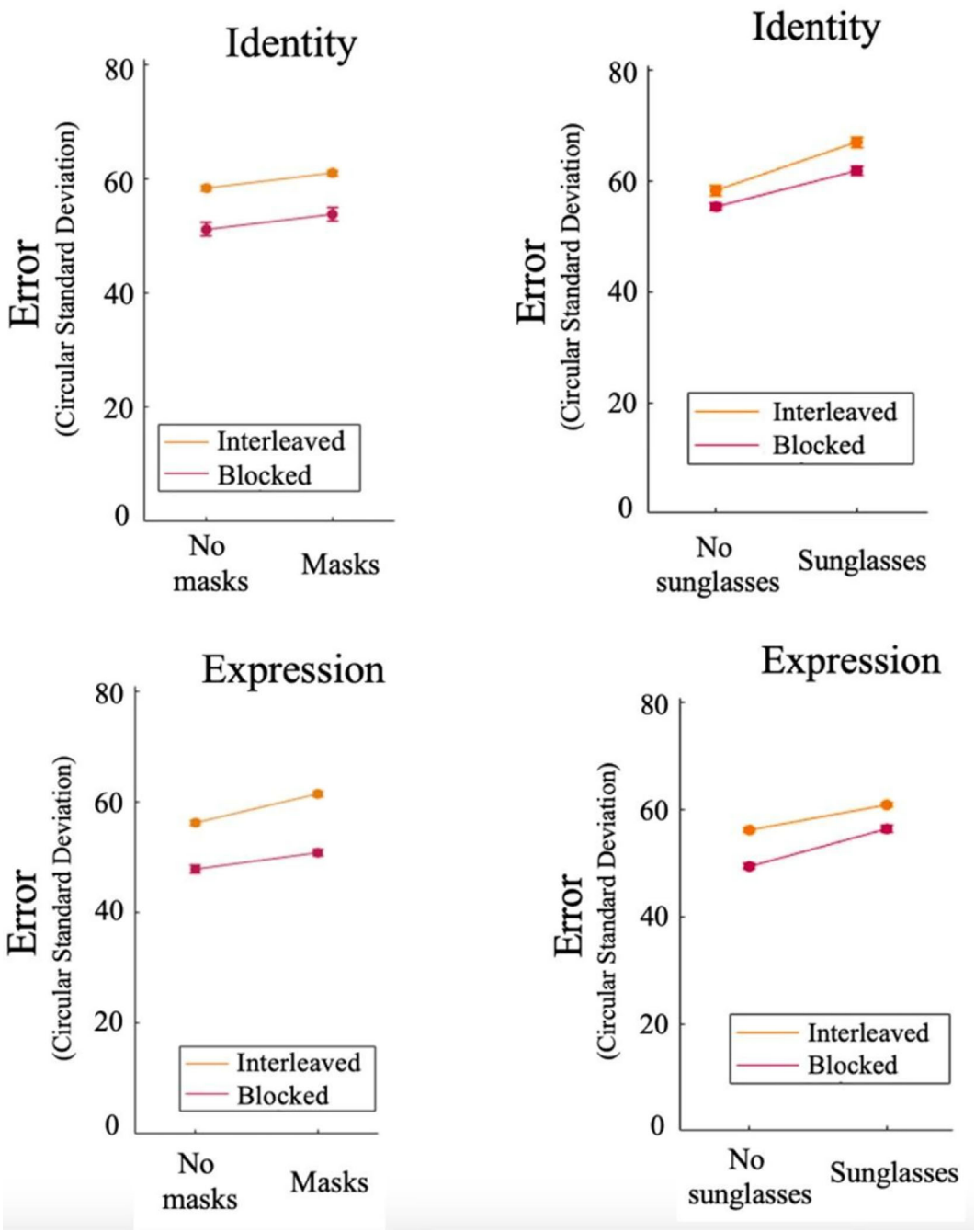

Figure 5. Above: Results within the identity wheels. Below: Results from expression face wheels. The yellow bars are the results from the interleaved experiment and the red bars are the results of the blocked experiments.

Similar to Experiment 1, we found that adding masks did not significantly harm our ability to extract the average identity of a group of faces ($t(31) = 1.08$, $p = 0.286$, $dz = 0.02$). Performance was still very strong in conditions with masked occlusions. For instance, blocks with face masks were still far better than chance performance (Chance is 81°; masks vs. chance: $t(31) = 10.99$, $p < 0.001$, $dz = 1.94$).

Experiment 2B blocked results—Identity sunglasses

Unlike in Experiment 1, adding sunglasses did lead to a small but reliable impairment in our ability to extract the average identity ($t(37) = 4.61$, $p < 0.001$, $dz = 0.97$), differing from the masked identity results. However, similar to the masked identity results, blocks with sunglasses were far better than chance performance

(sunglasses vs. chance: $t(37) = 8.32$, $p < 0.001$, $dz = 1.35$).

Experiment 2C blocked results—Expression masks

Next, we analyzed how the addition of face masks impacted our ability to extract the average expression. Unlike in Experiment 1, here we found a small but significant cost associated with masked occlusions ($t(35) = 2.28$, $p = 0.029$, $dz = 0.28$). Ensemble expression performance with masks is much better than chance performance (masks vs. chance: $t(35) = 11.70$, $p < 0.001$, $dz = 1.95$).

Experiment 2D blocked results—Expression sunglasses

Finally, we analyzed the addition of sunglasses on our ability to extract the average expression. We found sunglasses to have a large negative impact on performance ($t(36) = 6.10$, $p < 0.001$, $dz = 1.11$). Ensemble expression performance with sunglasses was much better than chance performance (sunglasses vs. chance: $t(36) = 13.78$, $p < 0.001$, $dz = 2.26$).

Blocked results summary

Comparing across Experiments 2A–2D again using a mixed-effects linear model, we find reliable effects of both occlusion condition ($F(2,425) = 11.9$, $p < 0.0001$) and wheel type ($F(1,425) = 12.7$, $p < 0.001$), unlike in Experiment 1. Post-hoc comparisons revealed that performance was significantly worse in the sunglasses condition compared to the no-occlusion condition ($\beta = 8.16$, $SE = 1.67$, $t(425) = 4.88$, $p < 0.0001$), whereas the face mask condition did not differ significantly from the no-mask condition ($\beta = 1.60$, $SE = 1.76$, $t(425) = 0.91$, $p = 0.36$). Overall performance was lower for the emotion wheel compared to the identity wheel ($\beta = -4.58$, $SE = 1.28$, $t(425) = -3.56$, $p < 0.001$), in contrast to Experiment 1 where no such difference was observed.

Thus, apart from the impact of masks on extracting the average identity, adding occlusions in all other conditions reliably hurt performance, unlike in Experiment 1. This difference between experiments, addressed more in the discussion, is likely due to a feature learning strategy in Experiment 1 that is, by design, not available in Experiment 2. However, other notable differences between Experiment 1 and the blocked condition of Experiment 2 were the presence of very difficult ensemble trials in Experiment 1 (with no clustering), and the presence of single-face trials in Experiment 1. These single-face trials may have provided additional opportunities for participants to learn which facial features remained informative under different occlusion conditions, potentially contributing to the even smaller effect of occlusions observed in Experiment 1 compared to Experiment 2's blocked condition.

Notably, however, even though we observe a cost of occlusion in the Blocked conditions, this cost is far from chance performance in all conditions, suggesting that ensemble perception is still robust despite being influenced by the disruption in information. Overall, the blocked conditions found a small but largely robust impairment in ensemble perception with occlusions.

Experiment 2E: Interleaved results

Next, we analyzed the results from the interleaved experiment. In this experiment, all participants completed blocks that had all 8 trial types intermixed together to discourage the ability to learn the particular features of any stimulus wheel. There was an impact of occlusions on the ability to extract the average identity within the interleaved condition. Pairwise *t*-tests revealed that both masks and sunglasses had a significantly detrimental effect on ensemble performance for facial identity (adding masks: $t(40) = 2.91$, $p = 0.006$, $dz = 0.45$; adding sunglasses: $t(40) = 5.10$, $p < 0.001$, $dz = 0.80$). However, similar to the blocked experiments, performance in the occluded conditions was still reliably greater than chance (masks vs. chance: $t(40) = 9.79$, $p < 0.001$, $dz = 1.53$; sunglasses vs. chance: $t(40) = 10.10$, $p < 0.001$, $dz = 1.58$), suggesting that even in interleaved conditions, the occlusions did not fully impair ensemble perception.

Next, we looked at the impact of occlusions on the ability to extract the average expression within the interleaved condition. Similar to identity, pairwise *t*-tests revealed that both masks and sunglasses had a significantly detrimental effect on ensemble performance for facial expression (adding masks: $t(40) = 4.85$, $p < 0.001$, $dz = 0.76$; adding sunglasses: $t(40) = 4.81$, $p < 0.001$, $dz = 0.75$). However, similar to the blocked experiments, performance in the occluded conditions was still reliably better than chance (masks vs. chance: $t(40) = 10.60$, $p < 0.001$, $dz = 1.66$; sunglasses vs. chance: $t(40) = 11.60$, $p < 0.001$, $dz = 1.81$).

Taken together, in the interleaved condition, we find a reliable cost to adding the occlusion, but we are still effectively able to extract the ensemble. Analyzing the interleaved condition using a mixed-effects linear model, we found a significant effect of occlusion condition ($F(2,242) = 6.35$, $p < 0.01$), but unlike in the blocked trials, only a marginal effect of wheel type ($F(1,242) = 2.89$, $p = 0.091$). Post-hoc comparisons

revealed that performance was significantly worse in both the sunglasses condition ($\beta = 6.71$, $SE = 1.89$, $t(242) = 3.55$, $p < 0.001$) and the face mask condition ($\beta = 3.94$, $SE = 1.89$, $t(242) = 2.08$, $p = 0.038$) compared to the no-occlusion condition. Although there was a trend toward better performance for the identity wheel compared to the emotion wheel ($\beta = -2.62$, $SE = 1.54$, $t(242) = -1.70$, $p = 0.091$), this difference did not reach statistical significance in the interleaved trials.

Blocked versus interleaved results

Combining data across blocked and interleaved trials revealed a highly significant effect of blocking condition ($F(1,670) = 36.42$, $p < 0.001$), with substantially higher error in the interleaved condition compared to blocked trials ($\beta = -6.56$, $SE = 1.09$, $t(670) = -6.04$, $p < 0.001$). This suggests participants performed significantly better when they could prepare for a specific wheel and task in the blocked condition. The overall analysis also confirmed the robustness of our other effects that were present or partially present in the blocked or interleaved alone: both occlusion type ($F(2,670) = 16.57$, $p < 0.0001$) and wheel type ($F(1,670) = 15.65$, $p < 0.0001$) remained significant predictors of performance. Specifically, collapsing across all data, both face masks ($\beta = 2.64$, $SE = 1.29$, $t(670) = 2.04$, $p < 0.05$) and sunglasses ($\beta = 7.30$, $SE = 1.27$, $t(670) = 5.76$, $p < 0.0001$) impaired performance relative to the no-occlusion condition, and participants showed better performance on the identity wheel compared to the emotion wheel ($\beta = -3.92$, $SE = 0.99$, $t(670) = -3.96$, $p < 0.0001$). There were no significant interactions among the three factors (interleaved vs. blocked, identity vs. emotion, and sunglasses vs. masks vs. unoccluded).

As in Experiment 1, we conducted Bayesian analyses to quantify the evidence for each main effect in our model. The analysis revealed extremely strong evidence for the blocking effect (BF10 = 1.49×10^6), indicating that the experimental design (blocked vs. interleaved) strongly influenced performance. We also found strong evidence for the effect of mask type (BF10 = 15,765), and strong evidence for the effect of face type (identity vs. emotion; BF10 = 88.01).

Overall, the biggest effect in Experiment 2 is a major cost of the interleaved design: subjects had much greater error in all conditions compared to the blocked design. These results suggest that part of the participants' success on extracting the ensemble average in the blocked conditions is partly based on learning feature information about the set, because those expectations help them in blocked conditions but not in the interleaved condition.

General discussion

The present study examines whether the ability to extract the average facial identity or expression is impacted when faces have naturalistic occlusions such as face masks or sunglasses, and what strategies participants might use to combat the added occlusion. We found that although there is some cost to adding masks or sunglasses, participants are still reliably able to extract ensemble averages from such occluded faces at far better-than-chance levels. When the presented faces are more variable across trials, removing the ability to easily learn exactly which features to attend to on each trial as in the interleaved condition of Experiment 2, there is an additional cost to performance, suggesting part of the robustness of ensemble perception in this context is that participants learn what features remain informative in the presence of the occluders and focus on those features. Taken together, our results suggest that ensemble perception for faces can still operate despite the obstruction of information, particularly if the type of occluder is reliably the same. Thus ensemble perception may be robust to occluders in naturalistic settings, like a room full of masked individuals or a sunny day outside with most people wearing sunglasses.

In Experiment 1, the spacing of the individual items in the ensemble set varied on each trial, such that the items of the ensemble might be more or less clustered around the mean. To account for this, we parsed the data into thirds, corresponding to how similar the items were. It is established that ensembles with less variance lead to better performance. For example, ensembles with less color variation make it easier to extract the average color (Maule & Franklin, 2015), and extracting the mean face is easier when the faces of the set are more similar to the average (Cabeza et al., 1999). We found the same effect in our data (e.g., Figure 3 is a visualization of a similar pattern in our experiment): the more clustered the faces were towards the mean, the more accurate the ensemble performance was.

Experiment 2 was designed to (1) replicate Experiment 1 with fixed ensemble spacing, and (2) test whether participants rely on a feature learning strategy to form the ensemble percept in situations with limited available information. In the blocked condition, unlike in Experiment 1, we found a small but reliable effect of occlusion, such that adding sunglasses or face masks hurt participants' ability to extract the average expression or identity of the set. However, in all cases, the decrease in performance still led to performance that was well above chance. Thus, although there is some cost to occlusion, the blocked condition in Experiment 2 still provides strong evidence that ensemble perception remains powerful and flexible despite the added occlusion. This finding expands work by Haberman and Ulrich (2019), who showed that

observers can still extract the average facial expression and identity from a group of faces partially occluded with horizontal black bars.

Experiment 2's interleaved design reduced the usefulness of learning the features of each stimulus wheel, by making that information less predictive. This allowed us to assess whether participants are maximizing information coming from the available information, a strategy known to be employed by expert face examiners (Carragher et al., 2022), by learning the features of the stimulus set. As illustrated in Figure 5, we found that interleaving trials from different face wheels and different occluders hurt performance more than adding masks or glasses hurt performance in blocked conditions. In other words, whatever is making ensemble perception robust to the added occlusions is strongly impacted by the interleaved design. This suggests that participants are learning information about the stimulus wheels in the blocked conditions and using those expectations to help form the ensemble statistic. For example, they likely learn to focus solely on the eyes in mask blocks and solely on the mouth in sunglasses blocks. The blocked design allows participants to develop and maintain consistent strategies specific to each condition, while the interleaved design disrupts not only feature learning but also the ability to maintain any consistent processing strategy, as participants must rapidly switch between different task demands and stimulus types on each trial. This suggests that the performance advantage we observed in blocked conditions may reflect both low-level perceptual learning of facial features and higher-level strategic adaptations to the task demands. This has implications for how we interpret results from studies using similar face wheels, which have often used just one or two face wheels presented in blocked conditions (e.g., Haberman & Ulrich, 2019). Such conditions are applicable to integrating information from a single face over time (e.g., during a conversation) but may be less applicable in situations where groups of faces vary.

Although there has recently been a significant expansion of work looking at the impact of face masks on our ability to process individual faces, very little work has looked at how naturalistic occlusions impact ensemble perception, despite its known role in supporting our visual experience. Despite occluding both the lower and top half of the face in the present study, further experimentation is needed to determine whether certain features are more heavily weighted in facial ensemble perception. Future studies, for instance, could incorporate neuroimaging to assess whether and how holistic processing is disrupted by the addition of naturalistic occlusions during ensemble perception, as the fusiform face area is critical to holistic face processing while the occipital face area plays a greater role in processing facial features rather than whole faces (Liu, Harris, & Kanwisher, 2010; Tsantani et al., 2021).

One particularly notable finding across our experiments concerns the relative impact of masks versus sunglasses on face perception. Although face masks cover a larger portion of the face than sunglasses, we found no evidence that they were more disruptive to performance. In fact, across experiments, the pattern of results suggested that if anything, sunglasses may be more disruptive than masks, although this effect was not consistently significant. This finding has practical implications given ongoing recent debates about the impact of mask-wearing on face perception and social interaction (e.g., Carragher & Hancock, 2020; Freud et al., 2020). Although both types of occlusions can impact face perception, our results suggest that common assumptions about masks being particularly disruptive to face perception may need to be reconsidered. The fact that masks were never more disruptive than sunglasses—a form of occlusion that society has long accepted without concern about its impact on face perception—suggests that adaptation to mask-wearing could be more feasible than initially assumed during the COVID-19 pandemic.

The results of the current study provide evidence that our visual system can extract the ensemble average of faces regardless of whether an added occlusion covers the upper or lower half of the face. However, further research needs to be done to determine the relative importance of individual facial features in the successful creation of an ensemble percept for faces. Such research would expand the literature looking at which facial features are most important for the recognition of individual faces (Davies et al., 1977; McKelvie, 1976; Smith et al., 2005). It also remains unclear whether naturalistic occlusions—which often are objects and have associated meaning themselves—impact ensemble perception of faces in a different way than artificial occlusions like black bars. Future studies could examine the impact of adding naturalistic objects that do not obscure facial features such as hats, tattoos, earrings, and more on ensemble performance, or strategies observers might use if only half of the ensemble had an added occlusion.

Given its ubiquity across many categories of visual information, ensemble perception is widely thought to serve as an important mechanism underlying visual perception. It is considered a key process used to bypass limited processing bottlenecks. The current study adds to our understanding of ensemble perception by showing its robustness to naturalistic occlusions and highlighting feature learning as a potential mechanism behind extracting the ensemble of faces with limited information. Our findings expand research showing that observers can learn to maximize awareness and use of available features to overcome the impairment face masks cause on single face recognition

(Carragher et al., 2022), and contribute to a growing research field examining how ensemble perception for faces is impacted in incomplete or noisy environments (Haberman & Ulrich, 2019).

Keywords: ensemble perception, face masks, summary statistics

Acknowledgments

Commercial relationships: none.
Corresponding author: Timothy F. Brady.
Email: tfbrady@ucsd.edu.
Address: Department of Psychology, University of California San Diego, McGill Hall, La Jolla, CA 92093, USA.

References

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences, 15*, 122–131.

Alvarez, G. A., & Oliva, A. (2009). Spatial ensemble statistics are efficient codes that can be represented with reduced attention. *Proceedings of the National Academy of Sciences United States of America, 106*, 7345–7350.

Cabeza, R., Bruce, V., Kato, T., & Oda, M. (1999). The prototype effect in face recognition: Extension and limits. *Memory & Cognition, 27*(1), 139–151.

Carlaw, B. N., Huebert, A. M., McNeely-White, K. L., Rhodes, M. G., & Cleary, A. M. (2022). Detecting a familiar person behind the surgical mask: Recognition without identification among masked versus sunglasses-covered faces. *Cognitive Research, 7,* 90, https://doi/10.1186/s41235-022-00440-3.

Carragher, D. J., & Hancock, P. J. B. (2020). Surgical face masks impair human face matching performance for familiar and unfamiliar faces. *Cognitive Research: Principles and Implications, 5*(1), 59, https://doi.org/10.1186/s41235-020-00258-x.

Carragher, D. J., Towler, A., Mileva, V. R., White, D., & Hancock, P. J. B. (2022). Masked face identification is improved by diagnostic feature training. *Cognitive Research: Principles and Implications, 7*(1), 30, https://doi.org/10.1186/s41235-022-00381-x.

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research, 43*, 393–404.

Davies, G., Ellis, H., & Shepherd, J. (1977). Cue Saliency in Faces as Assessed by the 'Photofit' Technique. *Perception, 6*(3), 263–269, https://doi.org/10.1068/p060263.

Fischer, J., & Whitney, D. (2011). Object-level visual information gets through the bottleneck of crowding. *Journal of Neurophysiology, 106*, 1389–1398.

Freud, E., Stajduhar, A., Rosenbaum, R. S., Avidan, G., & Ganel, T. (2020). *The COVID-19 pandemic masks the way people perceive faces* [Preprint]. PsyArXiv, https://doi.org/10.31234/osf.io/zjmr8.

Haberman, J. M., & Ulrich, L. (2019). Precise Ensemble Face Representation Given Incomplete Visual Input. *i-Perception, 10*(1), 1–15, https://doi.org/10.1177/2041669518819014.

Haberman, J., Lee, P., & Whitney, D. (2015) Mixed emotions: Sensitivity to facial variance in a crowd of faces. *Journal of Vision, 15*, 16–16, https://doi.org/10.1167/15.4.16.

Haberman, J., Brady, T. F., & Alvarez, G. A. (2015). Individual differences in ensemble perception reveal multiple, independent levels of ensemble representation. *Journal of Experimental Psychology: General, 144*, 432.

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology, 17*, R751–R753.

Haberman, J., & Whitney, D. (2009). Seeing the mean: Ensemble coding for sets of faces. *Journal of Experimental Psychology: Human Perception and Performance, 35*, 718–734.

Haberman, J., & Whitney, D. (2011). Efficient summary statistical representation when change localization fails. *Psychonomic Bulletin & Review, 18*(5), 855–859.

Haberman, J., & Whitney, D. (2012). Ensemble perception: Summarizing the scene and broadening the limits of visual processing. *From perception to consciousness: Searching with Anne Treisman, 33*.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Oxford University Press.

Lander, K., & Saunders, G. H. (2023). Face coverings: Considering the implications for face perception and speech communication. *Cognitive Research: Principles and Implications, 8*(1), 24, s41235-023-00479-w, https://doi.org/10.1186/s41235-023-00479-w.

Leib, A. Y., 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,* 26.

Liu, J., Harris, A., & Kanwisher, N. (2010). Perception of face parts and face configurations: An FMRI study. *Journal of cognitive neuroscience, 22*(1), 203–211, https://doi.org/10.1162/jocn.2009.21203.

Maule, J., & Franklin, A. (2015). Effects of ensemble complexity and perceptual similarity on rapid

averaging of hue. *Journal of Vision, 15*(4):6, 1–18.

McKelvie, S. J. (1976). The Role of Eyes and Mouth in the Memory of a Face. *The American Journal of Psychology, 89*(2), 311, https://doi.org/10.2307/1421414.

Nojavanasghari, B., Hughes, C. E., Baltrušaitis, T., & Morency, L. P. (2017). Hand2face: Automatic synthesis and recognition of hand over face occlusions. In *Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 209–215). IEEE.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*, 225–237.

Schurgin, M. W., Wixted, J. T., & Brady, T. F. (2020). Psychophysical scaling reveals a unified theory of visual memory strength. *Nature Human Behaviour, 4*(11), 1156–1172, https://doi.org/10.1038/s41562-020-00938-0.

Solomon, J. A., Morgan, M., & Chubb, C. (2011). Efficiencies for the statistics of size discrimination. *Journal of Vision, 11*(12), 13–13.

Sweeny, T. D., Haroz, S., & Whitney, D. (2012). Perceiving group behavior: Sensitive ensemble coding mechanisms for biological motion of human crowds. *Journal of Experimental Psychology: Human Perception and Performance, 38*(3), 1–15, https://doi.org/10.1037/a0028712.

Smith, M. L., Cottrell, G. W., Gosselin, F., & Schyns, P. G. (2005). Transmitting and decoding facial expressions. *Psychological Science, 16*, 184–189.

Tanaka, J. W., & Farah, M. J. (1993) Parts and wholes in face recognition, *The Quarterly Journal of Experimental Psychology, 46*(2), 225–245, https://doi.org/10.1080/14640749308401045.

Tsantani, M., Kriegeskorte, N., Storrs, K., Williams, A. L., McGettigan, C., & Garrido, L. (2021). FFA and OFA Encode Distinct Types of Face Identity Information. *The Journal of Neuroscience, 41*(9), 1952–1969, https://doi.org/10.1523/JNEUROSCI.1449-20.2020.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14*(5), 779–804.

Whitney, D., Haberman, J., & Sweeny, T. D. (2014). From textures to crowds: Multiple levels of summary statistical perception. *The New Visual Neurosciences,* 695–710.

Wolfe, J. M., & Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature Review Neuroscience, 5*, 495–501, https://doi.org/10.1038/nrn1411.

Wolfe, B. A., Kosovicheva, A. A., Leib, A. Y., Wood, K., & Whitney, D. (2015). Foveal input is not required for perception of crowd facial expression. *Journal of Vision, 15*, 11–11.