The role of category- and exemplar-specific experience in ensemble processing of objects

Oakyoon Cha, Randolph Blake, Isabel Gauthier

Department of Psychology, Vanderbilt University, Nashville, TN, USA

Author Note

This work was supported by the Centennial research fund and the David K. Wilson Chair research fund, both at Vanderbilt University, and by a grant from the National Science Foundation (BCS-1840896).

The datasets generated during and/or analyzed during the current study are available in the Open Science Framework repository, https://osf.io/96gf7/

The authors declare no conflict of interest.

Correspondence concerning this article should be addressed to Oakyoon Cha, Department of Psychology, Vanderbilt University, 111 21st Avenue South, Nashville, TN 37240, USA. E-mail: oakyoon@gmail.com

Accepted version

In press, Attention, Perception and Psychophysics

Abstract

People can relatively easily report summary properties for ensembles of objects, suggesting that this information can enrich visual experience and increase the efficiency of perceptual processing. Here, we ask whether the ability to judge diversity within object arrays improves with experience. We surmised that ensemble judgments would be more accurate for commonly experienced objects, and perhaps even more for objects of expertise like faces. We also expected improvements in ensemble processing with practice with a novel category, and perhaps even more with repeated experience with specific exemplars. We compared the effect of experience on diversity judgments for arrays of objects, with participants being tested with either a small number of repeated exemplars or with a large number of exemplars from the same object category. To explore the role of more prolonged experience, we tested participants with completely novel objects (random-blobs), with objects familiar at the category level (cars), and with objects with which observers are experts at subordinate-level recognition (faces). For objects that are novel, participants showed evidence of improved ability to distribute attention. In contrast, for object categories with long-term experience, i.e., faces and cars, performance improved during the experiment but not necessarily due to improved ensemble processing. Practice with specific exemplars did not result in better diversity judgments for all object categories. Considered together, these results suggest that ensemble processing improves with experience. However, the role of experience is rapid, does not rely on exemplar-level knowledge and may not benefit from subordinate-level expertise.

Keywords: Ensemble perception, Object recognition, Diversity judgment, Experience, Expertise

Introduction

We often encounter visual scenes containing multiple objects of the same category, such as faces in a classroom or a cluster of apples in a produce section. Some decisions require impressions of overall properties for such ensembles. For instance, a speaker may try to discern whether or not her audience is generally bored based on their facial expressions and a wheat farmer has to decide whether it is time to harvest his field based on the size and color of the seed heads. What determines someone's ability to make these kinds of judgments about a congregation, i.e., ensemble, of objects? Based on research in object recognition (Gauthier, 2018), we imagine that both domain-general and domain-specific visual abilities would contribute to performance on ensemble judgments. Recent work reveals the existence of a domain-general object recognition ability, *o*, independent from general intelligence (Richler et al., 2018), accounting for a substantial amount of the variability in performance across tasks and object categories. This recognition ability also correlates with performance on ensemble judgments (Gauthier, Sunday, Tomarken, & Cho, in press). In addition, it is natural to wonder whether expertise with a specific category of objects could influence performance. Compared to individuals without specialized training, seasoned lecturers should excel at inferring boredom in an audience and experienced wheat farmers should be more likely to know whether their grains are sufficiently ripe for harvest.

There is indeed evidence that performance in ensemble judgments for one category of objects (e.g., faces) is particularly well predicted by performance on individual recognition judgments with the same category (Haberman, Brady, & Alvarez, 2015), but other work suggests that most of this shared variability may not be specific to faces but rather generalizes across complex objects (Chang & Gauthier, 2020). Critically however, past research has not considered the importance of domain- *and* task-specific experience: the requisite experience for a specific kind of ensemble judgment might entail more than just extensive experience with exemplars of the relevant category. An immigration officer, despite having lots of exposure to individual faces, might perform no better than most other people in judging the *average* age or *average* facial expression within an array of faces because this kind of ensemble judgment is not part of the officer's training. The same could be said for judging the variety among an array of cars, a judgment one is rarely asked to make and one that may not be facilitated by the ability to identify individual cars.

To be sure, people can reliably judge ensemble properties of a variety of visual features, including simple ones such as average size (Chong & Treisman, 2003), average orientation (Dakin & Watt, 1997; Parkes, Lund, Angelucci, Solomon, & Morgan, 2001), average color (Maule, Witzel, & Franklin, 2014), size variance (Cha, Blake, & Gauthier, 2020) as well as complex ones such as average facial identity (de Fockert & Wolfenstein, 2009; Neumann, Schweinberger, & Burton, 2013), average emotional expression (Haberman & Whitney, 2007), average gender (Haberman & Whitney, 2009), and race/gender diversity (Phillips, Slepian, & Hughes, 2018). However, little work has examined the role of experience on ensemble processing. This is partly due to the way ensemble properties have been defined, especially when it comes to ensemble judgments with more complicated features. Earlier studies investigated ensemble judgments for simple features that can be scaled quantitatively (e.g., circle size, Gabor orientation), and the use of simple features allowed those studies to define ensemble properties with arithmetically specifiable properties, such as *mean* and *variance* (for review, see Whitney & Yamanashi Leib, 2018). It seems unlikely that experience with simple features would

vary greatly among individuals (perhaps excepting that unique population of researchers who explicitly deploy those features in their laboratory settings). Perhaps, then, studies of ensemble processing with simple features such as circle size or Gabor orientation may have unwittingly ignored the impact of experience in ensemble processing.

When it comes to complex object features such as facial expression or the identity of cars, however, arithmetic formulation of properties like average and variance can be challenging. One popular workaround deployed in scaling complex features entails morphing, i.e., producing a series of transformed images comprising varying degrees of blends between two anchor-point images (e.g., a male and a female). Morphing can be implemented to create images varying along several dimensions (e.g., gender and age), and morphing among all pairs of three images has been used to create a circular emotion/identity morph space (e.g., Haberman & Whitney, 2010). Morphing, however, creates problems that could complicate the study of ensemble judgments (cf. ZeeAbrahamsen & Haberman, 2018) and the role of experience in particular. First, morphing two exemplars of objects other than faces can result in unrealistic images. A morph combining 25% of a neutral face image and 75% of a happy face image will look like a moderately happy face, but a morph combining 25% of a sedan image and 75% of an SUV image can produce a vehicle unlike anything you've ever seen. This limitation of morphing may have discouraged explorations of ensemble judgments with non-face complex objects (but see Chang & Gauthier, 2020; Gauthier et al., in press), because few categories are as homogeneous in their configuration as faces. And faces are also different from most categories because people have ample experience with both their individual recognition and the processing of their ensemble properties. Second, morphing requires labor-intensive manual landmark matching between pairs of anchor-point images. For this reason, most studies use morphs created from a few faces, and repeat them over the course of the experiment (Bai, Yamanashi Leib, Puri, Whitney, & Peng, 2015; Elias, Dyer, & Sweeny, 2017; Haberman et al., 2015; Haberman & Whitney, 2007, 2009; Im et al., 2017; Li et al., 2016; Wolfe, Kosovicheva, Yamanashi Leib, Wood, & Whitney, 2015; Yamanashi Leib et al., 2014). Repeatedly sampling stimuli from a morph continuum based on a few faces appears equivalent to sampling from a circular space of orientation, but in fact it allows only very minimal coverage of the actual multidimensional space of face identities or expressions, which is much larger. These two problems make it difficult to manipulate experience with object ensembles at both the category level (faces vs. novel objects) and the exemplar level (familiar faces vs. novel faces). Only a few studies have used non-facial object ensembles (Chang & Gauthier, 2020; Gauthier et al., in press; Sweeny, Haroz, & Whitney, 2013; Yamanashi Leib, Kosovicheva, & Whitney, 2016) and a handful of studies have used large number of face identities (de Fockert & Wolfenstein, 2009; Neumann et al., 2013; Phillips et al., 2018; Yang, Yoon, Chong, & Oh, 2013). Note, however, those previous studies do not address whether learning impacts the ability to extract statistical properties within an ensemble of objects, and it is that question which motivated our experiment.

The following paragraphs explain the line of reasoning that guided the design of this experiment. We begin by explaining the task used to assess people's ability to distinguish arrays of objects differing in the extent of diversity among those object arrays. We then introduce the three object categories (between-subjects factor) we used in this diversity task, explaining our rationale for their selection. And we introduce the second between-subjects factor that could influence participants' exemplar-specific experience, i.e., exemplar repetition during the experiment. Finally, we discuss one within-subjects

factor, i.e., spatial arrangement of items within object arrays, designed to assess whether improvement in diversity judgments was related to improved ability to distribute attention (i.e., ensemble processing; see Chong & Treisman, 2005).

Diversity Judgment Task.

We devised an objective, 2-interval forced choice (2-IFC) task that required participants to judge which one of two successively presented object arrays portrayed more diversity among the exemplars of those objects (Cha et al., 2020). For instance, imagine an array that includes six images of cars, each being different from one another. Without making inferences about the similarity of individual pairs of cars, such a six-item array would constitute a maximally diverse array. Now imagine an array of six cars in which three of the six are duplicates of one another, with the remaining three being unique. Such an array is less diverse than the array with six unique exemplars. From trial to trial, the participants' task is to view two successively presented arrays of objects and then pick which one of the two arrays contains more discriminably different object exemplars (i.e., which array exhibits greater diversity).

Why does this task involve ensemble perception? According to Whitney and Yamanashi Leib (2018), ensemble processing entails deriving a specified statistical property that characterizes a distribution of values of a given stimulus attribute within a stimulus array. Many ensemble perception studies have focused on central tendency as the property of interest, but as Whitney and Yamanashi Leib underscore, variance among items in a set can also reflect an important, behaviorally relevant property characterizing ensembles of objects. Variability (e.g., variance, range, diversity), like central tendency, is not specifiable based on the value of a given, single item but rather is a property that emerges from sampling over multiple items within a set. In that sense, the diversity task used in this study provides a valid, meaningful index of ensemble perception.

Object Categories.

In a between-subjects design, we tested diversity discrimination performance using three object categories that plausibly differ in terms of the extent of prior visual experience people have with those objects: faces, cars, and random blobs. We assume that participants would have considerable everyday experience making decisions about individual and groups of faces, whereas they would have ample experience attending to individual cars but less experience judging properties of car ensembles. We also assumed that participants would have virtually no experience viewing random blobs individually or in arrays. These putative category differences, to the extent that they are valid, allow us to ascertain how diversity discrimination varies with the familiarity with a given category (faces and cars relative to random blobs), and with greater expertise at the subordinate level (faces relative to cars)[1].

[1] We note that face recognition likely differs from that of cars in at least two ways that are relevant: greater expertise at the subordinate level, and greater experience with ensemble judgments.

Exemplar Repetition.

The second between-subjects factor was the variety of exemplars experienced during a sequence of trials. One group of participants performed hundreds of trials making diversity judgments about ensembles created from the same small set of images (6 exemplars; high-repetition group), whereas another group performed the same number of trials with ensembles of exemplars drawn from an image set sufficiently large to ensure that no two arrays ever had the same combination of 6 images (low-repetition group). Through the experiment, category-specific experience was comparable between the high- and low-repetition groups, but exemplar-specific experience grew much higher in the high-repetition group.

Spatial Arrangement of Items.

We sought to assess improvement in diversity judgments as participants' exemplar-specific experience grew and, if so, whether the improvement was related to improved ensemble processing. To assess improvement, we compared accuracy of judging diversity during the first and second halves of an extended testing session. To test whether the improvement was related to improved ensemble processing, we manipulated the spatial arrangement of arrays in which some of the items were duplicates of the same exemplar (i.e., less diverse array). In clustered-duplicate trials, duplicates of the same exemplar were placed in spatially adjacent locations whereas in scattered-duplicate trials the duplicates were distributed in a more haphazard fashion within the array. Based on the Gestalt principle of proximity, we reasoned that the clustered duplicate items in an array were more likely to be grouped and the array could be easily identified as a less diverse array. Thus clustered-duplicate trials would be easier than scattered-duplicate trials, but would not benefit much with improvement in the ability to distribute attention to all array items evenly. If participants get better at distributing attention to array items, they would show more improvement in the scattered- than in the clustered-duplicate trials.

Methods

Participants

We recruited 126 participants via the Vanderbilt University SONA Systems who were tested in the laboratory (VU participants), and 340 paid participants via the Amazon Mechanical Turk who were tested online (MTurk participants). VU participants received either a course credit or a payment. We only recruited MTurk participants with US IP addresses who had been approved for more than 100 tasks and for more than 95% of the tasks they completed on MTurk. Six out of 126 VU participants (4.76%) and 92 out of 340 MTurk participants (27.06%) performed no better than chance in the second half of the experiment (defined by 95% departure from 50% chance in a binomial distribution), and they were excluded from further analyses.[2] We chose this criterion because it allowed for

[2] Visual inspection of the recorded click responses (described in the Procedure section below) suggested that a substantial number of the excluded MTurk participants may not have sustained sufficient attention during the task. For instance, throughout the experiment

participants who needed some practice to improve at the task, but it also required participants to keep trying for the entire experiment. Table 1 shows the number of participants included in analyses and their respective demographic information depending on the experimental group to which they were assigned and the pool from which they were recruited (see Design). In total, we analyzed data from 125 participants who completed the face diversity judgment task, 127 participants who performed the car diversity task, and 116 participants who performed the diversity task using random blobs. Informed consent was obtained prior to the experiment and all procedures were approved by the Vanderbilt University Human Research Protection Program.

Table 1. Number of participants included in analyses and their respective demographic information.

Object	Repetition	VU participants	MTurk participants
Faces	High	30 (24 females and 6 males; age: M = 19.67, SD = 1.51)	33 (12 females and 21 males; age: M = 39.57, SD = 9.90)
	Low	30 (22 females and 8 males; age: M = 20.13, SD = 1.64)	32 (17 females and 15 males; age: M = 33.67, SD = 8.72)
Cars	High	30 (19 females and 11 males; age: M = 20.82, SD = 2.23)	33 (20 females and 13 males; age: M = 35.31, SD = 10.92)
	Low	30 (23 females and 7 males; age: M = 24.29, SD = 8.04)	34 (17 females and 17 males; age: M = 32.88, SD = 8.88)
Random blobs	High	-	57 (27 females and 30 males; age: M = 36.23, SD = 9.48)
	Low	-	59 (34 females and 25 males; age: M = 41.48, SD = 13.42)

Going into this project, we had little basis for deriving an expected effect size for the interactions among combinations of object category (face/car/random blob), exemplar repetition (high/low repetition), block (first/second half), and duplicate-item arrangement (clustered/scattered). Therefore, we used an adaptive Bayesian procedure, starting with 30 participants per experimental group and assessing with Bayesian analyses whether we had enough data to claim support for or against the interactions of interest ($BF_{inclusion} < .33$ or $BF_{inclusion} > 3$; Jeffreys, 1961). We first recruited 30 participants each for the 4 experimental groups (face/car × high/low repetition) via Vanderbilt University SONA Systems, and then recruited additional participants via MTurk to hasten data collection. As we achieved comparable reliability online as in the lab for faces and car, all participants in the random blob diversity judgment task were recruited via MTurk.

Stimuli

Face images with neutral expression were sourced from the Chicago Face Database (Ma, Correll, & Wittenbrink, 2015), and modified to minimize differences in non-facial features (hair, beard) and in low-level image properties (skin tone, contrast). For each image, facial landmark points were detected (Zhu & Ramanan, 2012), a facial contour was determined using a smoothed polygon comprising the landmark points, and an image area outside the facial contour was removed. We manually discarded images where this procedure failed to remove non-facial features, leaving 474 face images of different identities. We then

---

some of the excluded MTurk participants clicked on random locations rather than the locations of response buttons (see Procedures).

converted color images to grayscale and matched the median luminance of all images. Since all non-facial features such as hair and beard were removed, median luminance corresponds to respective skin tone of faces shown in the images. Finally, we matched luminance histograms of all images using SHINE toolbox (Willenbockel et al., 2010). Six images were randomly selected and used in practice trials, and the remaining 468 images were used in experimental trials.

Car images were downloaded from the Motor Trend© magazine website, with the stipulation that all images portray a front driver-side view of a car seen against a transparent background and that the portrayed models were manufactured within the last 5 years. We converted color images to grayscale and composed two sets of car images based on the grayscale images' luminance profiles. One set comprised 147 images of white cars and the other set comprised 149 images of black cars. Within each set, car images were divided into 6 subsets of ~25 images, grouping cars of the same brand within a subset. On each trial, one image was selected from each set, preventing the selection of more than one image from the same brand. This manipulation helped to keep car images in an array reasonably discriminable from one another. We used 6 images of gray cars in practice trials.

Random blob images were generated with MATLAB (MathWorks, Natick, MA) using the following procedure. First, we made 900 frequency-filtered noise images whose pass band was centered at 3.6 cycle/image. These noise images were saturated to have the Michelson contrast of 40%, and windowed using a circular aperture with blurred edges. Then, we created 6 subsets using k-means clustering on a pixel-by-pixel basis. This clustering procedure assigned 900 images into 6 subsets while maximizing pixel-by-pixel similarity among images within the same subset and minimizing the similarity among images across different subsets. For each subset, we selected 25 representative images, i.e., 25 images that were most different from images in the other subsets. The remaining 750 images were discarded since they bear similarity with images in more than one subset, compared to any of the selected images. This procedure ensured that images in different subsets were easily discriminable from one another. The subsets were used in the same way as the car image subsets. Technical details on each procedure, along with runnable MATLAB codes, are available in the OSF repository (https://osf.io/96gf7/).

Procedure

In the diversity judgment task, participants viewed two successively presented arrays of objects (faces/cars/blobs), and then judged in which one of the two arrays of objects was more diverse (Fig. 1A). The two arrays contained 6 items portraying objects of the same category, each presented for 700 ms with a 600 ms blank interval separating the two presentations. This exposure duration was chosen based on pilot testing aimed at finding a duration sufficient to promote reasonably high but non-asymptotic percent-correct performance on this difficult task; the blank interval duration was selected to preclude masking interactions between the two successive exposures. Following each trial participants received visual feedback in the form of a brief message ("correct" in green text or "incorrect" in red text). For both intervals of each trial, 6 items were arrayed around a virtual circle centered in a fixation mark, with the exact positions of the items jittered to prevent participants from selectively attending to given locations. On each trial, one of the two arrays had 6 different exemplars (more diverse array; Fig. 1B, left) and the other array had 6 items among which 3 or 4 items were identical (less diverse array; Fig. 1B, middle

and right). To make less diverse arrays, we selected one item from the more diverse array and replaced 2 or 3 items with the selected item (duplicate item), and then item locations in the less diverse array were shuffled while duplicate-item locations were constrained by the spatial arrangement condition. Thus in a single trial, exemplars in the less diverse array were always a subset of six exemplars in the more diverse array, and the locations of the same exemplars varied between the two arrays. Participants were not informed about either the number of different exemplars or the number of duplicate items in each array, but they were given examples of more diverse and less diverse arrays during practice trials, to ensure that they understand what was meant by the more diverse array.

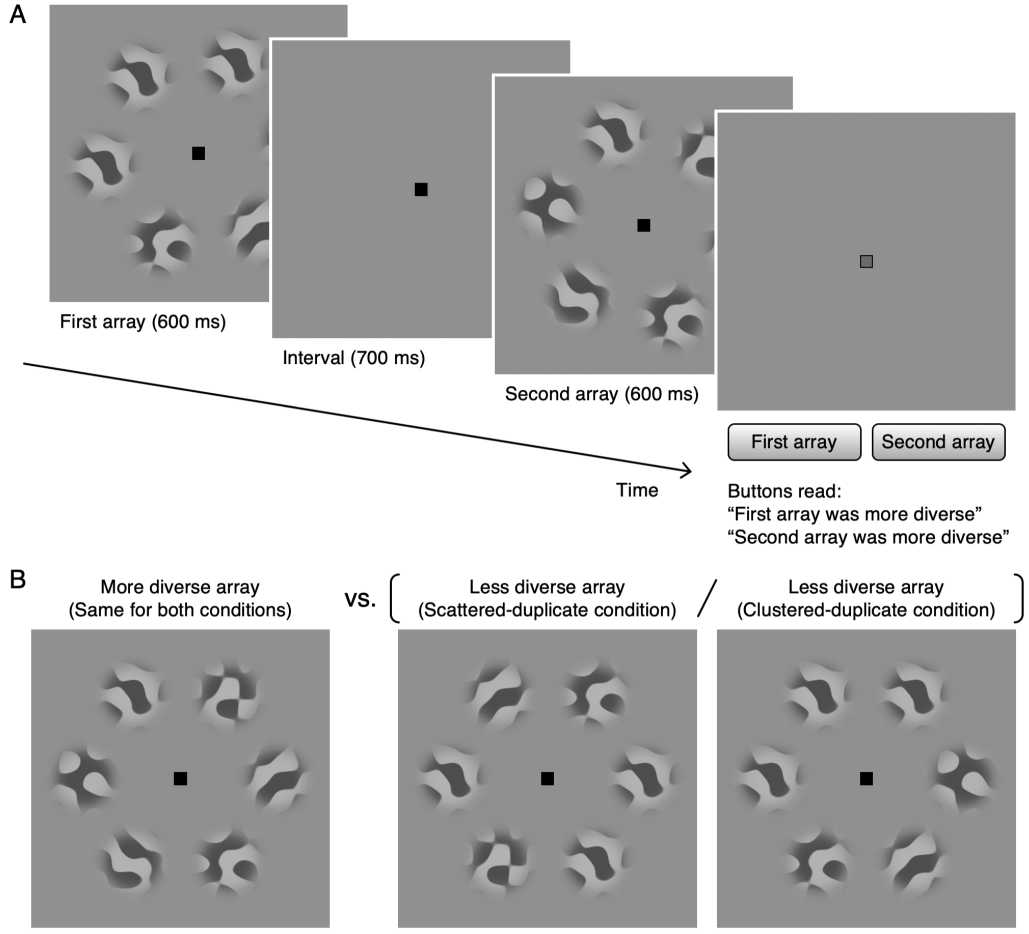

Fig. 1. Examples of displays and stimuli. Random blobs are shown in examples; other stimulus conditions were the same in terms of configuration except that six faces or six cars were shown rather than six random blobs. **(A)** Participants viewed two arrays of six items, and then responded which array had more diverse items. **(B)** On a given trial, one array was the more diverse array, and the other array was the less diverse array. The less diverse array could have two different arrangements depending on whether or not duplicate items were clustered together.

With this 2-IFC task, participants must decide in which of two successive presentations a more diverse set of exemplars was portrayed. To make that judgment requires mental comparison of the impression acquired from the first and the second presentations. The task, in other words, involves short-term memory. We preferred this task rather than a 2-alternative spatial forced choice version wherein two pairs of six-item

arrays are presented simultaneously on either side of a central fixation point. That procedure would have introduced other challenges including variable spatial resolution at different retinal eccentricities and the necessity for participants to treat simultaneously presented items from the same object category as members of two separate ensembles. The 2-IFC procedure and display layout we designed avoid those problems. Moreover, learning to maintain information from multiple items in working memory may be part of learning to perform ensemble judgments, as suggested by Dubé and Sekuler (2015).

VU participants performed the task individually on a Mac mini (Apple, Cupertino, CA) computer housed inside a quiet room. Stimuli were presented using MATLAB with Psychophysics Toolbox Version 3 extension (Brainard, 1997; Pelli, 1997), and participants responded using a computer keyboard. VU participants completed 4 blocks of 96 trials each, with mandatory short breaks between blocks. Prior to the experiment, participants were given 2 practice trials with display-by-display instructions followed by 4 more practice trials.

MTurk participants performed the task using a web browser on the computer of their choice, and they responded using a computer mouse by clicking on one of the response buttons (Fig. 1A). During the task, locations of clicks anywhere in the web browser were recorded. MTurk participants completed 4 blocks of 64 trials each, with mandatory short breaks between blocks. We decided to reduce the number of trials to keep the duration of the online experiment under 30 minutes. Since reducing the number of trials might influence statistical power by reducing the correlation between the different conditions of a repeated factor (Faul, Erdfelder, Lang, & Buchner, 2007), we based our decision on the correlations between percent-correct performance of scattered- and clustered-duplicate trials performed by VU participants (Fig. 2). At the start of the experiment, participants were given 2 practice trials with display-by-display instructions followed by 4 more practice trials.

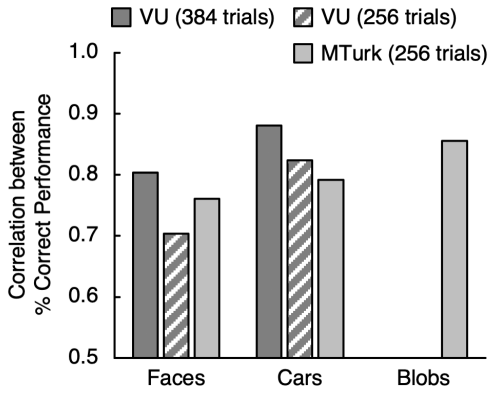

Fig. 2. Correlations between percent-correct performance of scattered- and clustered-duplicate trials. For VU participants who completed 384 trials, we estimated correlations twice, first with all 384 trials, and then with the first 256 trials. MTurk participants completed 256 trials, and all 256 trials were used to estimate correlations.

Design

In each diversity judgment task (face/car/random blob), participants in the high-repetition group viewed object images sampled from a pool of 6 images, and participants in the low-

repetition group viewed object images sampled from a pool of 351 face images (VU participants), 468 face images (MTurk participants), 147 white car images, 149 black car images, or 150 blob images. To analyze all the data together, we used the first 256 trials from VU participants and all 256 trials from MTurk participants. We analyzed the results after splitting 256 trials into two blocks of 128 trials each (first and second halves).

There were two types of duplicate-item arrangements. In the scattered-duplicate trials, one or two items were placed in-between the duplicate items (Fig. 1B, middle), and in the clustered-duplicate trials, all duplicate items were spatially clustered together (Fig. 1B, right). If performance improved in the second half and the improvement was larger for the scattered- than for the clustered-duplicate trials, we could interpret that as evidence of improved ability to distribute attention to array items (i.e., improved ensemble processing). On the other hand, if the improvement was similar for the scattered- and clustered-duplicate trials, the improvement would not be necessarily related to improved ensemble processing.

Thus, for the main analysis, we had two between-subjects factors, object category (face/car/random blob) and exemplar repetition (high/low), and two within-subjects factors, block (first/second half) and duplicate-item arrangement (clustered/scattered). Between-subjects factors were used to manipulate participants' prior experiences and experience over the time course of the experiment, and within-subjects factors were used to assess improvement in diversity judgments and whether the improvement could be related to improved ensemble processing.

In addition to the design described above, we manipulated one more aspect of the face diversity judgment task, for VU participants. Participants viewed face images sampled from a pool of either 6 images (high-repetition group) or 351 images (low-repetition group) in the first 3 blocks (totaling 288 trials), and then in the fourth block, participants in both high- and low-repetition groups viewed face images sampled from a new pool of 117 images. Accordingly, participants in the high-repetition group learned the diversity judgment task with only 6 exemplars for 3 blocks, and then performed the task with completely different exemplars in the fourth block. We compared results of the high- and the low-repetition groups between the third and fourth blocks (i.e., before/after switching the stimulus pool) to determine whether repetition influenced performance. It should be noted that this manipulation did not influence our main analysis, since the first 256 trials were completed before switching the stimulus pool.

Analysis

We used JASP statistics software (JASP Team, 2020) to conduct Bayesian analyses. To address whether performance improved over the time course of the experiment and whether the amount of improvement differed between the clustered- and scattered-duplicate conditions, we split the 256 trials into two equal blocks of trials, and focused on the interaction terms that include block (first/second half) and duplicate-item arrangement (clustered/scattered). To assess the strength of support for or against statistical terms of interest, we report $BF_{inclusion}$ calculated across matched models, and we evaluated whether the statistical terms of interest were included in the best model from Bayesian repeated-measures ANOVA. In addition, we report the top-3 models and their BF_{10} for comparison purpose.

Results and Discussion

Correlations among Repeated Measures

We assessed the quality of the online and the laboratory-based data by computing correlations between percent-correct performance of the clustered-duplicate and the scattered-duplicate conditions, i.e., the within-subjects factor in our experiment. Those percent-correct values were derived from 256 trials and from 384 trials per participant for the lab-tested participants and from 256 trials for the MTurk participants. Our decision to reduce the number of trials in the MTurk implementation was based on the correlations measured in the in-lab part of the study which was completed before we started the MTurk data collection. This decision was guided by knowledge of the reliance of statistical power of Bayesian inferences with repeated-measures ANOVA design (Nathoo & Masson, 2016). The correlations calculated with the first 256 trials from the VU participants (.70 for faces, .82 for cars) were comparable to the correlations calculated with the whole 384 trials (.80 for faces, .88 for cars), and the data from MTurk participants achieved similar levels of correlations (.76 for faces, .79 for cars, .86 for blobs; Fig. 2).

Results Overview

The graphs in Figure 3 summarize percent-correct performance for the conditions of interest in this experiment, and several conspicuous patterns of results emerge. First, as seen in all four panels, performance improved significantly ($BF_{10} = 4.01 \times 10^{30}$) in the second half ($M = 80\%$, $SD = 9\%$) relative to performance in the first half ($M = 75\%$, $SD = 9\%$) for all object categories and in both high- and low-repetition groups. Experience mattered, which does not surprise us given the similarity of items within each object category and the challenging nature of the diversity task. Second, performance in the clustered-duplicate condition ($M = 80\%$, $SD = 10\%$) was significantly better ($BF_{10} = 4.94 \times 10^{51}$) than was performance in the scattered-duplicate condition ($M = 75\%$, $SD = 8\%$). Again, not a surprising result because clustered duplicate arrays are likely to be more salient owing to the potency of proximity as a force in object grouping (e.g., Palmer, 2002) which is known to promote ensemble integration[3].

[3] Grouping cues including color (Brady & Alvarez, 2011), similarity, proximity, and common region (Corbett, 2017), surface properties (Cha, Blake & Chong, 2018), and category membership (Elias & Sweeny, 2020) all serve to improve the precision of ensemble formation.

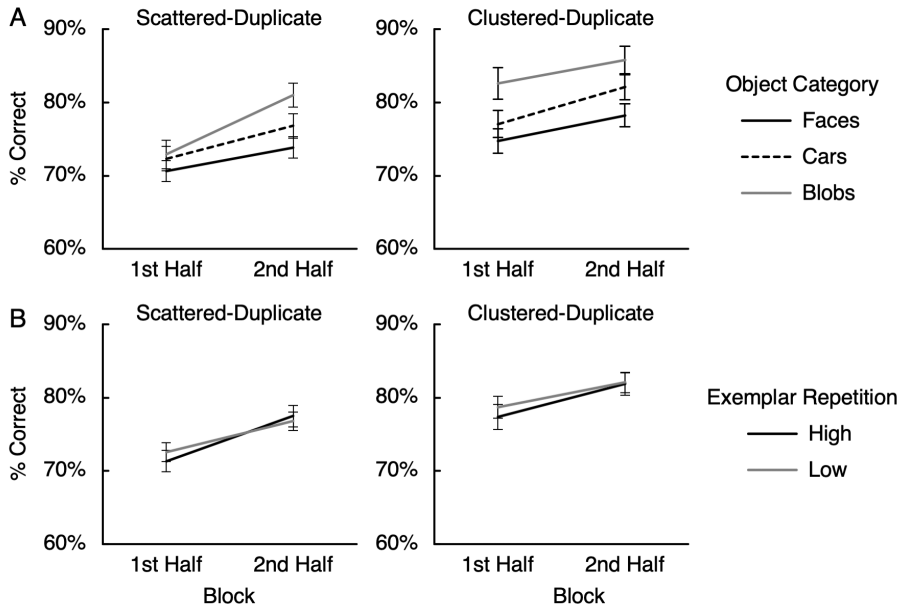

Fig. 3. Participants' performance from the first 256 trials of VU participants and all 256 trials of MTurk participants. **(A)** Participants' accuracies are plotted against the block. Separate plots show accuracies for scattered- and clustered-duplicate conditions, and separate lines in each plot show accuracies for different object categories (between-subjects factor). **(B)** Participants' accuracies are plotted against the block. Separate plots show accuracies for scattered- and clustered-duplicate conditions, and separate lines in each plot show accuracies for high- and low-repetition groups (between-subjects factor). Note that the whole set of data is used to plot (A) and the same set of data is used to plot (B). In all plots, error bars indicate 95% confidence intervals.

Turning next to the results for different object categories, we found that participants' overall performance (i.e., percent-correct for all conditions in all blocks) was better with random blobs (M = 81%, SD = 9%) than with cars (M = 77%, SD = 9%; BF_{10} = 13.18), and better with cars than with faces (M = 74, SD = 7%; BF_{10} = 4.32). However, we refrain from interpreting those differences because we made no attempt match the three categories in terms of overall difficulty or discriminability. Instead, we used pilot testing of the stimulus conditions for each condition separately to ensure that items within each separate category yielded performance that avoided ceiling or floor effects in both first and second halves. Thus, any comparisons across object categories must be contingent on interaction effects with other manipulated variables (i.e., interaction among *Factor(s) of interest* × *Block* × *Duplicate-item arrangement*). In terms of exemplar repetition, overall performance (i.e., percent-correct for all conditions in all blocks) achieved by participants who repeatedly viewed a small number of exemplars on every trial (M = 77%, SD = 9%) was indistinguishable (BF_{10} =.13) from the performance of participants who viewed numerous different exemplars throughout testing (M = 78%, SD = 8%).

Improvements Depending on the Object Category and Exemplar Repetition

We designed the conditions of this experiment to learn whether *improvement in diversity judgments* varied with the nature of the objects comprising arrays (i.e., cars vs. faces vs. random blobs). The answer to this question is evident in left panel of Figure 3A, where we see that the solid gray line (blobs) shows steeper slope than the other two lines (faces and cars) in the scattered-duplicate condition. At the same time slopes of all three lines in the

right panel are comparable to slopes of the solid and dashed black lines (faces and cars) in the left panel, suggesting that improvement in diversity judgments for blobs, not for faces and cars, could be better explained by improved ensemble processing. To assess this trend statistically, we submitted data from all three tasks to Bayesian repeated-measures ANOVA, with and two between-subjects factors, *Object category* (face/car/blob) and *Exemplar repetition* (high vs. low), and two within-subjects factors, *Block* (first vs. second half) and *Duplicate-item arrangement* (clustered vs. scattered). All top-3 models include *Object category* × *Block* × *Duplicate-item arrangement* interaction term (Table 2, shown in italic), and Bayes factors of the models including this interaction term were 72.59 times larger on average across matched models, implying that a larger improvement in the scattered- than in the clustered-duplicate condition was found with blobs, but not with faces and cars. This interaction was driven by different improvement depending on duplicate-item arrangements (i.e., interaction between *Block* and *Duplicate-item arrangement*) in the random blob diversity judgment task. Performance improvement was similar in the scattered- and clustered-duplicate conditions for the face diversity judgment task ($BF_{inclusion}$ = .14) and for the car diversity judgment task ($BF_{inclusion}$ = .15). For the random blob diversity judgment task, however, participants showed larger improvement in the scattered- than in the clustered-duplicate condition ($BF_{inclusion}$ = 963.87). In other words, this relatively unfamiliar stimulus set enjoyed larger benefit in the scattered-duplicate condition only, suggesting that this benefit was related to the improved ability to distribute attention. For ensemble judgments of simple features, it has been proposed that pooling across early-stage representations of individual items could have the virtue of dampening the impact of uncorrelated noise within those representations and, thus, improving ensemble perception (Alvarez, 2011; Sweeny et al., 2013). Perhaps, novel objects have noisier representations than familiar objects, with more room for improvement in ensemble perception.

Table 2. Top-3 models from Bayesian repeated-measures ANOVA (all participants).

Model terms	**BF_{10}**
Object category + Block + Duplicate-item arrangement + [Object category × Block] + [Object category × Duplicate-item arrangement] + [Block × Duplicate-item arrangement] + [*Object category* × *Block* × *Duplicate-item arrangement*]	3.10×10^{101}
Object category + Exemplar repetition + Block + Duplicate-item arrangement + [Object category × Block] + [Object category × Duplicate-item arrangement] + [Exemplar repetition × Block] + [Block × Duplicate-item arrangement] + [*Object category* × *Block* × *Duplicate-item arrangement*]	1.37×10^{101}
Object category + Exemplar repetition + Block + Duplicate-item arrangement + [Object category × Block] + [Object category × Duplicate-item arrangement] + [Block × Duplicate-item arrangement] + [*Object category* × *Block* × *Duplicate-item arrangement*]	6.66×10^{100}

In contrast, exemplar repetition did not influence diversity discrimination, as illustrated by the black and gray lines almost overlapping in both scattered- and clustered-duplicate plots in Figure 3B. Performance improvement was similar in the scattered- and clustered-duplicate conditions regardless of the exemplar repetition ($BF_{inclusion}$ for *Exemplar repetition* × *Block* × *Duplicate-item arrangement* = .10). In addition, we found evidence against a four-way interaction among *Object category*, *Exemplar repetition*, *Block*, and *Duplicate-item arrangement* ($BF_{inclusion}$ = .08). One might surmise that the effects of exemplar repetition would be more pronounced in the random blob diversity

judgments where we found evidence of improved ensemble processing. We repeated the same analyses with data from the random blob diversity judgment task only, and found that participants who learned the diversity judgment task with a limited number of blobs showed very similar patterns of results to participants who learned the task with a large number of blobs ($BF_{inclusion}$ for *Exemplar repetition* × *Block* × *Duplicate-item arrangement* = .19).

Evidently, the ability to discern diversity among arrays of objects does not benefit from seeing the same sets of specific exemplars during testing. This finding may strike some readers as remarkable (e.g., "how can seeing the same faces over and over again have essentially no impact on perception?"). We surmise that this seemingly counterintuitive finding is attributable to the basis of the diversity judgment task. The majority of ensemble perception studies assess some aspect of central tendency, i.e., a property indexed by a particular stimulus quality (e.g., average size or most predominant color). Diversity, the ensemble property we have focused on, is a property based on the incidence of differences among items. It is not essential, we conjecture, for one to discern the particulars of array items to gain a global impression of the diversity of those items. This conjecture seems in line with the finding that people viewing two successive arrays of faces, some of which change in facial expression from the first to the second presentation, can derive a statistical sense of how those arrays differ in expression without being able to specify which particular faces in those arrays have changed (Haberman & Whitney, 2011).

Effect of Switching the Stimulus Pool During Testing

In addition, we asked whether learning the diversity judgment task with a very small number of repeated exemplars (6 faces) would impact participants' diversity judgments upon being tested with an entirely new set of exemplars. Recall that 60 VU participants tested with faces experienced a switch in block 4, after the first 3 blocks, to a new, large pool of faces. First, we looked at the overall performance in blocks 3 and 4, and found the same patterns as in the main analysis: better performance (BF_{10} = 1553.25) in the clustered-duplicate condition (*M* = 79%, *SD* = 9%) than in the scattered-duplicate condition (*M* = 75%; *SD* = 8%) and no difference (BF_{10} = .27) between the high-repetition group (*M* = 77%; *SD* = 8%) and the low-repetition group (*M* = 77%; *SD* = 8%). We then submitted the data from these VU participants to Bayesian repeated-measures ANOVA, with one between-subjects factor, *Exemplar repetition* (high vs. low), and two within-subjects factors, *Block* (block 3 vs. block 4) and *Duplicate-item arrangement* (clustered vs. scattered). Participants in the high-repetition group performed as well after switching the stimulus pool (i.e., in block 4 compared to block 3) as participants in the low-repetition group ($BF_{inclusion}$ for *Exemplar repetition* × *Block* = .21), they showed similar magnitudes of improvement for the clustered- and scattered-duplicate conditions regardless of the exemplar repetition ($BF_{inclusion}$ for *Exemplar repetition* × *Block* × *Duplicate-item arrangement* × = .23), and none of the top-3 models include any interaction term (Table 3). In other words, participants exposed to a very limited set of face images during the initial trials of testing nonetheless maintained their level of performance when switched to a new, much larger set of face images, and that level of performance was equivalent to that achieved by the group of participants exposed to that large set of face images from the outset. This finding, too, may seem surprising, but it is consistent with our tentative conclusion concerning the absence of an exemplar repetition effect: people are basing their

diversity judgments on the perceived variety of items without recourse to the identities of the individual items.

Table 3. Top-3 models from Bayesian repeated-measures ANOVA (VU participants tested with faces).

Model terms	**BF_{10}**
Duplicate-item arrangement	2728.15
Exemplar repetition + Duplicate-item arrangement	981.15
Block + Duplicate-item arrangement	390.08

Because we were surprised that switching from a small to a large set had so little effect, we collected some data from 18 participants[4] so that we could ask them about their impressions after the experiment. They performed one block of 96 trials using 6 faces, followed by 96 trials where we switched without warning to a different set of 117 faces. Participants were asked at the end to estimate how many faces were used before and after the break. One outlier estimated 100 followed by 92 faces. The estimates of the other participants were higher for the second block than the first (block 1: $M = 5.24$, $SD = 2.05$; block 2: $M = 13.18$, $SD = 7.42$; $BF_{10} = 397.20$). These estimates suggest that participants may be able to notice a difference between few vs. many, at least with faces, but that they greatly underestimated the number of different faces when a large number of faces were used – that is, as we increased the number of faces almost 20-fold, participants estimated the increase to be less than 3-fold. In other words, participants appear to get a gist for the diversity of faces in this task despite poor memory of, and no benefit from, repetition of specific faces.

Comparison between VU and MTurk Participants

We compared performance achieved by VU and MTurk participants (only relevant for faces and cars, as performance with random blobs was only tested on MTurk participants). For the face diversity judgment task, VU and MTurk participants showed very similar overall performance levels (VU participants: $M = 75\%$, $SD = 7\%$; MTurk participants: $M = 74\%$, $SD = 7\%$; $BF_{10} = .29$). We submitted the face diversity judgment task data to Bayesian repeated-measures ANOVA with two between-subjects factors, *Pool* (VU vs. MTurk) and *Exemplar repetition* (high vs. low), and two within-subjects factors, *Block* (first vs. second half) and *Duplicate-item arrangement* (clustered vs. scattered), and found evidence against any interaction with *Pool* (all $BF_{inclusion} < .33$). For the car diversity judgment task, VU participants ($M = 80\%$, $SD = 8\%$) performed better than MTurk participants ($M = 75\%$, $SD = 8\%$; $BF_{10} = 55.02$), and the pattern of results is more complicated (Fig. 4A). We submitted the car diversity judgment data to Bayesian repeated-measures ANOVA and found that Bayes factors favor the inclusion of two interaction terms, *Pool* × *Exemplar repetition* × *Block* ($BF_{inclusion} = 2.58$, see Fig. 4B) and *Pool* ×

[4] We ran this experiment with the same apparatus used for the face diversity judgment task ran in the lab, with three exceptions. All participants were assigned to the high-repetition group. The experimental program ran block 1, and then after the break, jumped to block 4 (i.e., block after the switch). After that, participants were given two text input boxes and required to type in the total number of different faces they saw before and after the break. Informed consent was obtained prior to the experiment and all procedures were approved by the Vanderbilt University Human Research Protection Program.

Exemplar repetition × *Block* × *Duplicate-item arrangement* ($BF_{inclusion}$ = 2.00). Since support for both interaction terms was anecdotal ($BF_{inclusion}$ < 3), we will not dwell on these interactions. However, it is interesting to consider a potential reason why testing with cars might be more sensitive to the source of participants.

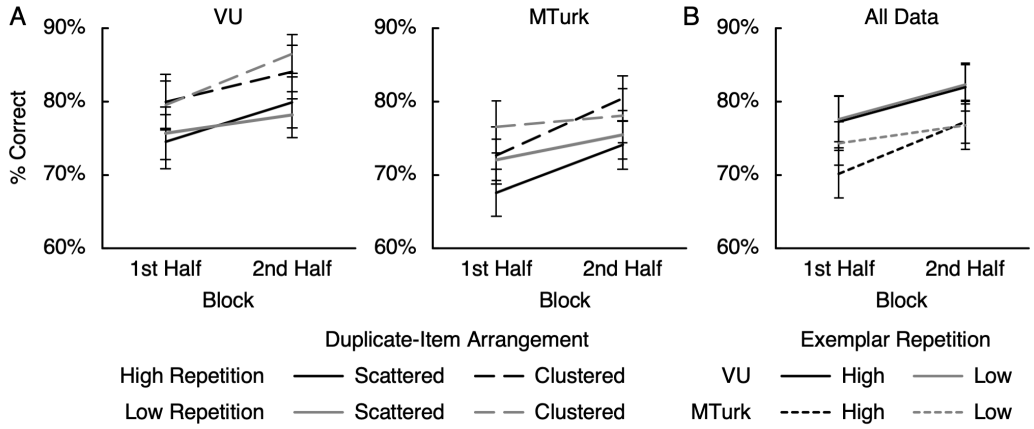

Fig. 4. Car diversity judgment task performances for VU and MTurk participants. **(A)** Separate plots show performance for participants recruited from different pools, and separate lines in each plot show performance for different experimental conditions. **(B)** VU and MTurk participants' performance are plotted together, without considering different arrangement conditions. In all plots, error bars indicate 95% confidence intervals.

It is well established that expertise impacts object recognition of single items. For example, relative to car novices, car experts recognize specific cars more rapidly than do novices (Curby & Gauthier, 2009), they process cars holistically (Bukach, Phillips, & Gauthier, 2010) and they exhibit higher visual short-term memory capacity for arrays of cars (Curby, Glazek, & Gauthier, 2009). While we made no effort to manipulate or to measure car expertise in our participants, there is some evidence that the younger pool of participants (age of the VU participants: *M* = 20.80, *SD* = 3.44; MTurk participants: *M* = 38.72, *SD* = 11.26; see Table 1), were better overall in the car task. Our comparison of performance with faces for similar pools of participants from VU and MTurk point to no difference between groups, suggesting that the testing platform (lab vs. online) and the response mode (pressing a key vs. clicking) cannot account for the differences found with cars. Prior work using a different task with cars suggested that an older sample of participants tested online was likewise disadvantaged relative to younger participants, presumably because familiarity with different models of cars may be highly dependent on age (Sunday, Lee, & Gauthier, 2018). The car images that we used in the present experiment were all drawn from vehicles manufactured during the last 5 years. Perhaps the older MTurk participants would have performed better with car models on the road when they started driving or bought their first car. In any case, regardless of the reason for the difference in performance, it had limited (inconclusive) effects on the patterns of learning. There was a trend for older participants to improve more over the course of the experiment in the high-repetition than in the low-repetition group, whereas repetition had no influence on learning for younger participants. While this was neither predicted nor strongly supported by inferential statistics, it suggests that an explicit manipulation of expertise in the future could be informative.

Implications and Limitations

One implication of the present study is that our version of the diversity judgment task could be a valuable tool for studying object ensembles. Participants were able to perform this task with categories that vary greatly in terms of prior experience, and they improved over the course of the experiment with all categories. Importantly, the diversity judgment task allows the use of objects that vary over multiple dimensions, instead of being artificially constrained by morphing. The diversity task, in other words, allows a more realistic object space within which to work, one that better taps into natural domains of faces or cars. With morphed images, features in every dimension change monotonically and altogether as the morphs progress from one exemplar to the other. Thus, these morphed images only occupy a thin vector within a multi-dimensional object space. This is likely to allow observers to attend to a single feature dimension even with complex objects. In other work on category learning, critical results depend on the kind of morph space used in the experiments (Folstein, Gauthier, & Palmeri, 2012).

Finally, we want to mention two limitations of the present study. First, we relied on larger improvement in the scattered- than in the clustered-duplicate condition to test whether performance improvement was attributable to the increased ability to distribute attention to all items evenly. There are, multiple possible concomitants to ensemble processing any of which might be amenable to improvement with experience. For instance, participants might have become better at manipulating incomplete representations of individual items. This ability would be important in ensemble judgments, especially with a task limited by short presentation durations such as ours, but our manipulation of spatial arrangement could not capture improvement in this ability. It is possible that participants' improvement in diversity judgments of faces and cars could be related to changes in ensemble processing that could not be captured with our spatial arrangement manipulation. Second, as we used the diversity judgment task, where incidences of differences among exemplars are important, we cannot tell whether or not our findings will apply to other types of ensemble tasks, such as those requiring average judgments. To explore this question will require a task that allows deriving visual representation of central tendency from object exemplars residing in multi-dimensional space. As we used the "diversity" judgment task to test people's ability to estimate variability, perhaps one may devise a task where participants report the "mode" of objects, which will serve as one type of central tendency.

Conclusion

We asked participants to judge the diversity of exemplars within ensembles of objects. This is a rather different judgment from the one employed in the majority of ensemble perception studies, which typically require judgments of the central tendency for arrays of objects. One might imagine that estimating central tendency involves the extra steps of pooling and compressing information not required in estimates of variability. But this difference does not invalidate 'variability' as an ensemble process, and one that could be quite useful in real-world settings (see Whitney & Yamanashi Leib, 2018 for examples). We recently found evidence that in the case of object size, judgments of diversity and of central tendency rely partially on a common ability (Cha et al., 2020). With this in mind, we sought to learn whether diversity judgments improve with practice, whether improvements vary depending on category- and/or exemplar-specific experiences and, if

so, whether the improvements could be related to improved ensemble processing. Thus we compared improvements in performance with three categories of objects for which people have varying degrees of familiarity: faces, cars, and random blobs. We also manipulated the redundancy of exemplars for diversity judgments throughout a series of trials: participants in one group viewed visual images sampled from a very small pool (6 images), and participants in another group viewed visual images sampled from approximately 150 images or more. What did we find?

Larger performance improvement in the scattered-duplicate condition compared to the clustered-duplicate condition occurred only for the random blobs and not for cars or faces, suggesting that improvement in diversity judgments for cars and faces were not necessarily associated with improved ensemble processing (i.e., improved ability to distribute attention). We surmise that experience during the experiment did not facilitate ensemble processing of faces or cars because participants were already quite familiar with these two categories of stimuli before being exposed to them in the experiment. Furthermore, the fact that participants' diversity judgments of random blobs improved steeply in the scattered-duplicate condition during the course of a one-half hour experiment suggests that category knowledge useful for diversity judgments can be acquired fairly rapidly.

At the same time, we found that repeated exposure to the same six exemplars did not afford participants a special advantage for diversity judgments of these exemplars. Among the three different object categories we tested, performance on diversity judgments was comparable between the participants who experienced high- vs. low-repetition exposures. Importantly, we found evidence supporting a null effect of repetition, even for novel objects. Moreover, participants who repeated diversity judgments with the same six faces showed no sign of degraded performance when they were subsequently tested with completely new sets of faces.

Altogether, our results suggest that participants' diversity judgments are performed without reliance on detailed knowledge of specific exemplars experienced over the course of the task. Instead, this ability may be based on abstract properties of an object category derived from exemplars that are not necessarily familiar. Ensemble processing is believed to enrich and facilitate our visual experience, promoting representations of redundant information from groups of objects (Alvarez, 2011; Oliva & Torralba, 2006). Still, our results, supported by conclusive Bayesian support for a null effect of repetition, suggest that for complex real-world objects, it may not make a difference whether a few exemplars or many form the basis for learning to judge diversity within object ensembles.

Acknowledgements This work was supported by the Centennial research fund and the David K. Wilson Chair research fund, both at Vanderbilt University, and by a grant from the National Science Foundation (BCS-1840896). We are grateful for the comments by the Editor and two referees of our paper.

Open Practices Statement The datasets generated during and/or analyzed during the current study are available in the Open Science Framework repository, https://osf.io/96gf7/

References

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences*, *15*(3), 122-131. https://doi.org/10.1016/j.tics.2011.01.003

Bai, Y., Yamanashi Leib, A., Puri, A., Whitney, D., & Peng, K. (2015). Gender differences in crowd perception. *Frontiers in Psychology*, *6*:1300, 1-12. https://doi.org/10.3389/fpsyg.2015.01300

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

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*(4), 433-436. https://doi.org/10.1163/156856897X00357

Bukach, C. M., Phillips, W. S., & Gauthier, I. (2010). Limits of generalization between categories and implications for theories of category specificity. *Attention, Perception, & Psychophysics*, *72*(7), 1865-1874. https://doi.org/10.3758/APP.72.7.1865

Cha, O., Blake, R., & Chong, S. C. (2018). Composite binocular perception from dichoptic stimulus arrays with similar ensemble information. *Scientific Reports*, *8*(1), 1-13. https://doi.org/10.1038/s41598-018-26679-9

Cha, O., Blake, R., & Gauthier, I. (2020). *Judgments of average and variability within object ensembles rely on a common ability*. Manuscript submitted for publication.

Chang, T. -Y. & Gauthier, I. (2020). *Domain-general ability underlies complex object ensemble processing*. Manuscript submitted for publication.

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research*, *43*(4), 393-404. https://doi.org/10.1016/S0042-6989(02)00596-5

Chong, S. C., & Treisman, A. (2005). Attentional spread in the statistical processing of visual displays. *Perception & Psychophysics*, *67*(1), 1-13. https://doi.org/10.3758/BF03195009

Corbett, J. E. (2017). The whole warps the sum of its parts: Gestalt-defined-group mean size biases memory for individual objects. *Psychological Science*, *28*(1), 12-22. https://doi.org/10.1177/0956797616671524

Curby, K. M., & Gauthier, I. (2009). The temporal advantage for individuating objects of expertise: Perceptual expertise is an early riser. *Journal of Vision*, *9*(6):7, 1-13. https://doi.org/10.1167/9.6.7

Curby, K. M., Glazek, K., & Gauthier, I. (2009). A visual short-term memory advantage for objects of expertise. Journal of Experimental Psychology: *Human Perception & Performance*, *35*(1), 94-107. https://doi.org/10.1037/0096-1523.35.1.94

Dakin, S. C., & Watt, R. J. (1997). The computation of orientation statistics from visual texture. *Vision Research*, *37*(22), 3181-3192. https://doi.org/10.1016/S0042-6989(97)00133-8

de Fockert, J., & Wolfenstein, C. (2009). Rapid extraction of mean identity from sets of faces. *Quarterly Journal of Experimental Psychology*, *62*(9), 1716-1722. https://doi.org/10.1080/17470210902811249

Dubé, C., & Sekuler, R. (2015). Obligatory and adaptive averaging in visual short-term memory. *Journal of Vision*, *15*(4):13, 1-13. https://doi.org/10.1167/15.4.13

Elias, E., Dyer, M., & Sweeny, T. D. (2017). Ensemble perception of dynamic emotional groups. *Psychological Science*, *28*(2), 193-203. https://doi.org/10.1177/0956797616678188

Elias, E., & Sweeny, T. D. (2020). Integration and segmentation conflict during ensemble coding of shape. *Journal of Experimental Psychology: Human Perception & Performance*, *46*(6), 593-609. https://doi.org/10.1037/xhp0000733

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175-191. https://doi.org/10.3758/BF03193146

Folstein, J. R., Gauthier, I., & Palmeri, T. J. (2012). How category learning affects object representations: Not all morphspaces stretch alike. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *38*(4), 807-820. https://doi.org/10.1037/a0025836

Gauthier, I. (2018). Domain-specific and domain-general individual differences in visual object recognition. *Current Directions in Psychological Science*, *27*(2), 97-102. https://doi.org/10.1177/0963721417737151

Gauthier, I., Sunday, M. A., Tomarken, A. & Cho, S. J. (in press). o is the same for familiar and novel objects. *Journal of Vision*.

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*(2), 432-446. https://doi.org/10.1037/xge0000053

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology*, *17*(17), R751-R753. 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 & Performance*, *35*(3), 718-734. https://doi.org/10.1037/a0013899

Haberman, J., & Whitney, D. (2010). The visual system discounts emotional deviants when extracting average expression. *Attention, Perception, & Psychophysics*, *72*(7), 1825-1838. https://doi.org/10.3758/APP.72.7.1825

Haberman, J., & Whitney, D. (2011). Efficient summary statistical representation when change localization fails. *Psychonomic Bulletin & Review*, *18*(5), 855-859. https://doi.org/10.3758/s13423-011-0125-6

Im, H. Y., Chong, S. C., Sun, J., Steiner, T. G., Albohn, D. N., Adams, R. B., & Kveraga, K. (2017). Cross-cultural and hemispheric laterality effects on the ensemble coding of emotion in facial crowds. *Culture and Brain*, *5*(2), 125-152. https://doi.org/10.1007/s40167-017-0054-y

JASP Team. (2020). JASP (Version 0.12.1) [Computer Software]. Available at https://jasp-stats.org/

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

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*:1332, 1-11. https://doi.org/10.3389/fpsyg.2016.01332

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, *47*(4), 1122-1135. https://doi.org/10.3758/s13428-014-0532-5

Maule, J., Witzel, C., & Franklin, A. (2014). Getting the gist of multiple hues: Metric and categorical effects on ensemble perception of hue. *Journal of the Optical Society of America A*, *31*(4), A93-A102. https://doi.org/10.1364/JOSAA.31.000A93

Nathoo, F. S., & Masson, M. E. (2016). Bayesian alternatives to null-hypothesis significance testing for repeated-measures designs. *Journal of Mathematical Psychology*, *72*, 144-157. https://doi.org/10.1016/j.jmp.2015.03.003

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

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*, *155*, 23-36. https://doi.org/10.1016/S0079-6123(06)55002-2

Palmer, S. E. (2002). Perceptual grouping: It's later than you think. *Current Directions in Psychological Science*, *11*(3), 101-106. https://doi.org/10.1111/1467-8721.00178

Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience*, *4*(7), 739-744. https://doi.org/10.1038/89532

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437-442. https://doi.org/10.1163/156856897X00366

Phillips, L. T., Slepian, M. L., & Hughes, B. L. (2018). Perceiving groups: The people perception of diversity and hierarchy. *Journal of Personality and Social Psychology*, *114*(5), 766-785. https://doi.org/10.1037/pspi0000120

Richler, J. J., Tomarken, A. J., Sunday, M. A., Vickery, T. J., Ryan, K. F., Floyd, R. J., ... & Gauthier, I. (2019). Individual differences in object recognition. *Psychological Review*, *126*(2), 226-251. https://doi.org/10.1037/rev0000129

Sunday, M. A., Lee, W. Y., & Gauthier, I. (2018). Age-related differential item functioning in tests of face and car recognition ability. *Journal of Vision*, *18*(1):2, 1-17. https://doi.org/10.1167/18.1.2

Sweeny, T. D., Haroz, S., & Whitney, D. (2013). Perceiving group behavior: Sensitive ensemble coding mechanisms for biological motion of human crowds. *Journal of Experimental Psychology: Human Perception & Performance*, *39*(2), 329-337. https://doi.org/10.1037/a0028712

Whitney, D., & Yamanashi Leib, A. (2018). Ensemble perception. *Annual Review of Psychology*, *69*, 105-129. https://doi.org/10.1146/annurev-psych-010416-044232

Willenbockel, V., Sadr, J., Fiset, D., Horne, G. O., Gosselin, F., & Tanaka, J. W. (2010). Controlling low-level image properties: The SHINE toolbox. *Behavior Research Methods*, *42*(3), 671-684. https://doi.org/10.3758/BRM.42.3.671

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*(4):11, 1-13. https://doi.org/10.1167/15.4.11

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):26, 1-13. https://doi.org/10.1167/14.8.26

Yamanashi Leib, A., Kosovicheva, A., & Whitney, D. (2016). Fast ensemble representations for abstract visual impressions. *Nature Communications*, *7*(1):13186, 1-10. https://doi.org/10.1038/ncomms13186

Yang, J. W., Yoon, K. L., Chong, S. C., & Oh, K. J. (2013). Accurate but pathological: Social anxiety and ensemble coding of emotion. *Cognitive Therapy and Research*, *37*(3), 572-578. https://doi.org/10.1007/s10608-012-9500-5

ZeeAbrahamsen, E., & Haberman, J. (2018). Correcting "confusability regions" in face morphs. *Behavior Research Methods*, *50*(4), 1686-1693. https://doi.org/10.3758/s13428-018-1039-2

Zhu, X., & Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. (pp. 2879-2886). Retrieved from https://ieeexplore.ieee.org/document/6248014.