



Contribution of a common ability in average and variability judgments

Oakyoon Cha^{1,2} · Randolph Blake¹ · Isabel Gauthier¹

Accepted: 3 July 2021 / Published online: 19 July 2021
© The Psychonomic Society, Inc. 2021

Abstract

People can make judgments about statistical properties of visual features within groups of objects, such as the average size, size variability, or numerosity of circles. Emerging from recent work is the view that these kinds of visual estimations, collectively dubbed ensemble perception, rely on independent abilities that are specific to the statistical property being estimated (mean, variance, range, numerosity). Here we revisit evidence for the claim that different statistical judgments (i.e., average and variability) for the same object feature are based on independent abilities. We tested a large sample of people, using a pre-registered open-ended sequential design to avoid ending up with weak evidence. We estimated the shared variance in ability across different ensemble judgments, with common constraints for the different tasks. We found that the abilities to judge the average size and the size variability for an array of circles are positively correlated, even after controlling for the ability to discriminate the size of single circles. Our results refute the idea that judgments of average and variability for the same object feature rely on completely independent abilities.

Keywords Ensemble perception · Average · Variability · Individual differences

People can derive statistical properties of different features when they look at groups of objects. Imagine looking at a group photograph from a wedding reception. Among other things, you can derive an overall sense of the group's emotional expression, while also estimating variability in the age of the individuals. Laboratory studies of these kinds of judgments confirm that people can successfully judge properties of ensembles of items (e.g., average, variability), whether those judgments focus on simple features (e.g., average size – Chong & Treisman, 2003; average orientation – Parkes et al., 2001; orientation variance – Morgan et al., 2008) or on multidimensional features (e.g., average emotion – Haberman & Whitney, 2007; race/gender diversity – Phillips et al., 2018). See Whitney and Yamanashi Leib (2018) for a review of those studies. Not surprisingly, people vary in the

accuracy of their ensemble judgments (e.g., Haberman et al., 2015), and studying the relations between individual differences across a variety of ensemble judgments offers an opportunity to understand the relation among the abilities underlying these decisions (Gauthier, 2018; Vogel & Awh, 2008).

Haberman et al. (2015) published a seminal study of the factors influencing performance on ensemble judgments. Testing about 100 college students with different stimuli spanning a range of complexity, they found shared variance in performance among tasks where participants judged “averages” of simple features (orientation and color). They also found shared variance for tasks involving complex features (e.g., face identity/expression). But, importantly, correlations were small or absent when relating performance with simple features to performance involving facial features. Haberman and colleagues thus concluded that ensemble perception involves at least two different abilities, supporting statistical judgments for ensembles comprising simple versus complex objects.

This work on individual differences for central tendency judgments inspired studies of the specificity of ensemble judgment abilities for different types of statistical properties (i.e., mean, variance, range, numerosity). Those studies employed multiple tasks that required judgments about different

✉ Oakyoon Cha
oakyoon.cha@sungshin.ac.kr

¹ Department of Psychology, Vanderbilt University, Nashville, TN, USA

² Department of Psychology, Sungshin Women's University, 2 Bomun-ro 34da-gil, Seongbuk-gu, Seoul 02844, Republic of Korea

ensemble properties, measuring the correlation between performance on these judgments (mean vs. numerosity/range – Khvostov & Utochkin, 2019; mean vs. numerosity – Utochkin & Vostrikov, 2017; mean vs. variance – Yang et al., 2018). In the three studies, weak correlations across participants between judgments of different ensemble properties were taken to imply independence of mechanisms for estimating different ensemble properties. One study (Khvostov & Utochkin, 2019) considered weak trial-by-trial correlations for judgments of different ensemble properties as evidence of independent mechanisms.¹ From these works emerged the view that ensemble processing is supported by distinct perceptual-cognitive processes, the engagement of which depends not only on the complexity of relevant object features, but also on the nature of the statistical judgment being made.

Here we revisit the claim that judgments about different ensemble properties of a given object set (e.g., average and variability) are not correlated, and thus rely on entirely distinct abilities, which implies the absence of common mechanisms. For two reasons, we question this claim. The first reason has to do with sample size: the six studies (from three articles) that explicitly examined correlations in peoples' ability to estimate different summary statistics (Khvostov & Utochkin, 2019; Utochkin & Vostrikov, 2017; Yang et al., 2018) each reported data for 30 or fewer people. All three articles deployed frequentist statistical tests, which require sufficient power and cannot provide support for the null hypothesis. Two of the studies (Khvostov & Utochkin, 2019; Utochkin & Vostrikov, 2017) added Bayes Factors (BFs), which can estimate support for a null hypothesis. However, most of the reported BFs were between .33 and 3, values often considered too low to support unequivocal conclusions (Jeffreys, 1961). Even those BFs smaller than .33 only provided moderate evidence in favor of the null (ranging from .265 to .330). The sensitivity of frequentist statistics, the precision of estimates, and the ability of Bayesian statistics to provide conclusive evidence all depend on sample size, albeit in different ways (Schönbrodt & Wagenmakers, 2018). As a group, these studies had poor precision, and in several cases, claims were made based on weak evidence.

The second reason we believe this question should be revisited concerns a design issue. In each of those six studies, the estimation tasks differed depending on the ensemble property being assessed. For instance, Yang et al. (2018) required participants to estimate and compare the variance of two ensemble displays but asked them to estimate the average of a single ensemble display by asking them to compare it to a

single stimulus. Performance on variance judgments could have reflected participants' ability to maintain an ensemble array in visual working memory as well as their ability to estimate the ensemble property of interest (i.e., variance). In contrast, performance on average judgments would have only reflected participants' ability to estimate the ensemble property of interest (i.e., average), and thus the ability to maintain an ensemble array in visual working memory would not be captured as shared variance in the correlation. These task differences, together with small sample sizes, may limit the likelihood of detecting a correlation between abilities to estimate different kinds of summary statistics.

To remedy these shortcomings, we tested a large sample of people in tasks involving estimation of two statistical properties of the same visual feature, setting the same constraints for both tasks. We opted for an open-ended sequential design using Bayesian statistics (Schönbrodt et al., 2017) to ensure that we would obtain conclusive evidence. Moreover, we raised the bar for interpreting correlations. Unlike prior studies mentioned above, we measured perceptual ability with a single stimulus and controlled for this source of variability. This partial correlation approach also allowed better control for other sources of variability that are not associated with ensemble judgments, compared to other approaches such as using a correlation of presumably uncorrelated tasks as a baseline (e.g., Haberman et al., 2015). Controlling for variability in performance in single stimulus size judgments can control for the most plausible source of spurious shared variance while at the same time controlling for other potential sources of correlation that could arise between *any* two tasks, such as differences in motivation.

In a sample of 75 participants, we measured perceptual ability to judge: (i) the size of a single circle, (ii) the variability in size within an array of circles, and (iii) the average size within an array of circles. We then computed the partial correlations between performance on the average and variability estimation tasks, controlling for participants' perceptual ability in judgments with single circles. We found that variability and average estimation abilities for the same feature were positively correlated even when individual differences in the ability to judge the size of individual objects were accounted for. This result suggests the presence of common mechanisms contributing to ensemble processing of different statistical summaries, over and above those that support judgments about single items.

Methods

Participants

We used the Sequential Bayes Factor (SBF) design to determine when to stop data collection (Schönbrodt et al., 2017). In

¹ We do not consider this method optimal for claims of independence. Trial-by-trial correlations in errors for two different kinds of ensemble judgments could be either small or large depending on how the two ensemble properties are manipulated in a set of trials. In other words, difficulties in judging different ensemble properties are not necessarily the same for the same ensemble array and could be either correlated or uncorrelated across trials depending on the manipulation.

an SBF design, researchers collect data and compute BFs after the pre-determined minimum sample size was met until a pre-determined level of evidence for/against hypotheses is reached. We used a BF of 3 as a threshold of evidence in favor of a correlation, and BF of .33 as a threshold against a correlation (Jeffreys, 1961). Following our pre-registered protocol (<https://osf.io/78w2t>)², we computed BFs when we reached the sample size of 75 and stopped data collection as the BF of interest was larger than 3. Specifically, we collected data from 122 paid participants via Amazon Mechanical Turk (MTurk), among which 47 performed no better than chance (defined by the one-sided 95% confidence interval of a binomial distribution centered at 50% chance) for at least one task – they were excluded because their results could inflate correlations.³ This left 75 participants for analyses (38 females, 36 males, including one who preferred not to report; age: $M = 42.28$ years, $SD = 12.18$). We only recruited MTurk workers with US IP addresses who had been approved for more than 100 tasks and for more than 95% of the tasks they completed. All procedures were approved by the Vanderbilt University Institutional Review Board in charge of overseeing protection of human subjects. Informed consent was obtained prior to the experiment.

Stimuli

In the single-circle-size task (Fig. 1a), participants saw two circles one after the other on each trial, each with a diameter of any of nine values (35.90, 36.99, 38.06, 39.11, 40.15, 41.17, 42.19, 43.18, 44.17 pixels) equally distanced on a psychological scale (Teghtsoonian, 1965). In the size-variance task and the average-size task, participants saw arrays of six circles, with the size of each circle randomly sampled from a uniform probability distribution and then normalized to yield the desired average and standard deviation (SD). In the two ensemble tasks, the average diameter of six circles in an array was one of five possible values (35.90, 38.06, 40.15, 42.19, 44.17 pixels) and the SD of psychological sizes of the six circles was one of five values (18.75, 37.5, 56.25, 75, 93.75 pixels)^{1,52}; note that the psychological size is proportional to the diameter to the 1.52nd power and thus the unit for psychological size is pixels^{1.52}; Teghtsoonian, 1965). Circle sizes in the single-circle-size task were determined in a pilot experiment so that the average task performance would approximate that used in the ensemble tasks. In an individual

differences study like ours, we are not interested in interpreting differences between conditions/tasks; instead, we are focused on the correlation across performance for different tasks, in the same set of participants. Thus, while average performance in each task is not important, the analyses require that each task provide a reliable measure of its target construct. The use of small step sizes in the single-circle-size task help ensure this task would be sufficiently hard to avoid ceiling and provide sensitive measurement of the differences among participants. Each open circle was drawn as a D4 contour (whose cross-sectional luminance profile was defined with fourth derivative of Gaussian function; Wilkinson et al., 1998) of 100% contrast on a uniform gray rectangle.

In the single-circle-size task, each stimulus appeared in a random location on an imaginary circular contour, 164 pixels in radius, centered around a fixation mark, with the contour's exact position jittered radially within ± 12 pixels. In the other two tasks, the location of one circle was determined in the same way as in the single-circle-size task, and the other five circles were placed 60° apart from one another on the same imaginary circular contour. On each trial, the exact positions of individual circles varied slightly (within ± 12 pixels radially).

All three tasks were administered online in a single session, meaning that screen size and viewing distance, while not precisely controlled, were presumably the same across tasks for any given participant. The image size of the uniform gray rectangle was 600×500 pixels and fit within a single page of web browser without scrolling on a relatively low-resolution monitor ($1,024 \times 768$ pixels). All stimuli were presented within an HTML5 webpage generated by and hosted on a custom-built web application server, and the stimulus presentation was controlled by JavaScript aided by jQuery 3.4.0. Before starting the experiment, participants were advised to choose a screen on which they could see the entire gray rectangle and instruction text without having to scroll.

Design and procedures

To avoid order effects contributing to individual differences (Goodhew & Edwards, 2019; Mollon et al., 2017), we tested all participants with the same task order (single-circle-size task, size-variance task, and average-size task) and the same order of trials within each task.

For all three tasks, each trial included two successive displays presented for 700 ms each, with a 600-ms interval between presentations (Fig. 1). When the second display disappeared, two response buttons appeared below the gray rectangle. Feedback in the form of a green (“correct”) or a red (“incorrect”) banner was given on every trial. We selected an exposure duration longer than those used in early ensemble studies (e.g., 50–200 ms in Chong & Treisman, 2003), based on evidence that longer presentation durations can benefit

² This experiment is the second of two with a very similar design and similar results. The first experiment was deemed *a posteriori* to have a potential flaw. Procedures and results from that experiment are fully available online for the sake of transparency; see Open Practices Statement at the end of this paper.

³ Visual inspection of the recorded click locations suggested that a substantial number of the excluded participants may not have paid sufficient attention to the online experiment. For instance, some of the excluded participants clicked on random locations rather than on one of the response buttons throughout the experiment.

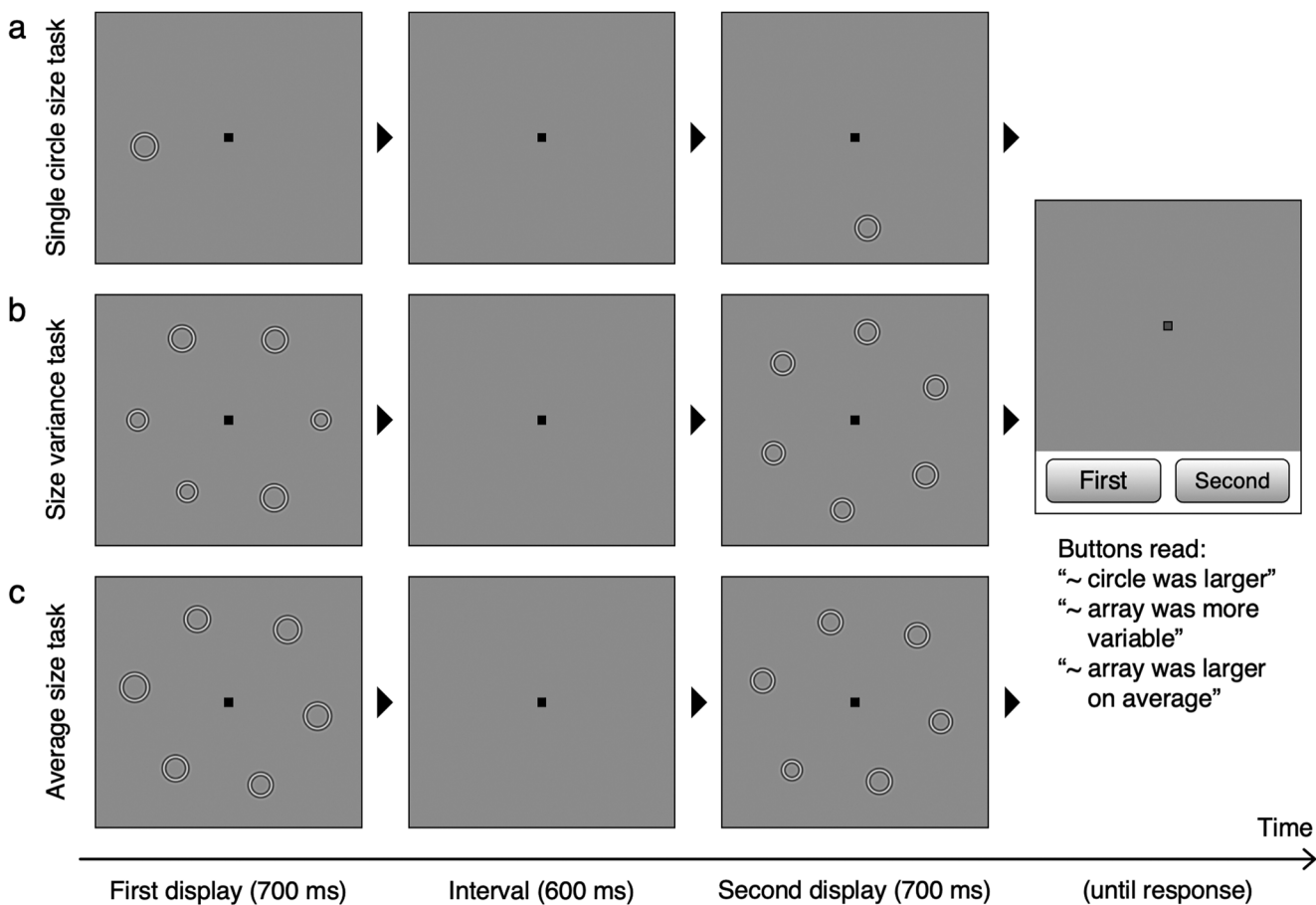


Fig. 1 Example displays of a single trial in (a) a single-circle-size task, (b) a size-variance task, and (c) an average-size task. Response displays were the same for all three tasks with the exception of texts on the two response buttons, which read “First/second circle was larger” in the single-circle-size task, “First/second array was more variable” in the size-variance task, and “First/second array was larger on average” in the

average-size task. In three examples shown in this figure, the first circle/array is larger/more variable/larger on average. Sizes and configurations of circles in this figure are proportional to those in the actual experiment. The gray background rectangle expanded larger and the texts on the response buttons were smaller in the actual experiment

judgments of average circle size (Gorea et al., 2014; Whiting & Oriet, 2011) and that ensemble judgments are not particularly influenced by foveating individual items (Wolfe et al., 2015). We also wanted to allow sufficient time for processing variability as well as average. Display-by-display instructions for a single trial and practice trials were given before each task; for instance, for the size-variance task, participants viewed an array of circles with small variance in their sizes, with a written text that the circle sizes were similar to one another and that an array of circles with very different sizes would be more variable than this array.

There were 80 experimental trials in each of the three tasks. In the single-circle-size task, one circle had the size of level 5 (40.15 pixels) out of nine sizes, and the other circle could have the size of level 1, 2, 3, 4, 6, 7, 8, or 9 (35.90, 36.99, 38.06, 39.11, 41.17, 42.19, 43.18, or 44.17 pixels, respectively). In the size-variance task, one array had the psychological size SD of level 3 (56.25 pixels^{1.52}) out of five SDs, and the other array could have the SD of level 1, 2, 4, or 5 (18.75, 37.5, 75, or 93.75 pixels^{1.52}, respectively); and the average sizes of the

two arrays were the same in the same trial. In the average-size task, one array had the average size of level 3 (40.15 pixels) out of five average sizes, and the other array could have the average size of level 1, 2, 4, or 5 (35.90, 38.06, 42.19, or 44.17 pixels, respectively); and the SDs of the two arrays were the same in the same trial.

Analysis

MATLAB (MathWorks, Natick, MA) was used to calculate accuracy scores (proportion correct), residual accuracy scores (the difference between the observed value on a task and the value predicted in the model that controlled for single-circle-size performance), and their split-half reliabilities. To provide a stronger test for the correlation between the two ensemble tasks, we controlled for accuracy in the single-circle-size task for each of the two ensemble judgments and calculated a partial correlation. Specifically, we regressed size-variance task accuracy on single-circle-size task accuracy, average-size task accuracy on single-circle-size task accuracy, and then

analyzed the correlation between residuals from the two regressions (i.e., correlated residual accuracy scores). In addition, we computed disattenuated correlations obtained by dividing an observed correlation by the square root of the product of the two measures' reliabilities. A disattenuated correlation provides an estimate of what a correlation would be without measurement error (Nunnally, 1970; Osborne, 2002). For the split-half reliabilities, we calculated the correlations between split-halves using 20 random sets of split-halves (with the same numbers of stimulus/array pairs in each half), and then corrected them using the Spearman-Brown correction. We reported the average of the Spearman-Brown corrected correlations from 20 split-halves. We used JASP statistics software (JASP Team, 2019) for Bayesian analyses, and reported BF_{10} (i.e., likelihood of the observed data when two measurements are correlated compared to when the two measurements are not correlated) for each correlation, computed using a stretched beta prior width of 1. As a rule of thumb, one may consider $BF > 30$ as “very strong,” $BF > 10$ as “strong,” and $BF > 3$ as “moderate” evidence for a correlation (Lee & Wagenmakers, 2013; adjusted from the heuristic proposed originally by Jeffreys, 1961; as discussed in Stefan et al., 2019). Likewise, one may consider $BF < .03$ as “very strong,” $BF < .01$ as “strong,” and $BF < .33$ as “moderate” evidence against a correlation.

Results

The mean and SD of accuracy scores from the 75 participants who satisfied prescreening for inclusion in the analyses are shown in Table 1, along with reliabilities of accuracy and residual accuracy scores. Each ensemble task was correlated with single-circle judgments (size-variance task: $r = .56$, 95% C.I. = [.38; .70], $BF_{10} = 77161.98$; average-size task: $r = .38$, 95% C.I. = [.17; .56], $BF_{10} = 41.66$). We found very strong evidence for a correlation between the size-variance and the average-size tasks (Fig. 2a; $r = .45$, 95% C.I. = [.25; .61], $BF_{10} = 419.67$), indicating that the shared variance (R^2) between the two tasks was 20%. Given that the reliability of the two tasks sets the upper bound of correlation,⁴ the disattenuated correlation is .64, 95% C.I. = [.48; .76], indicating shared variance of 40% without measurement error. Crucially, performance in the two ensemble tasks was positively correlated even after controlling for performance with single circles (Fig. 2b; $r = .31$, 95% C.I. = [.09; .50], $BF_{10} = 4.74$; disattenuated $r = .54$, 95% C.I. = [.36; .68]). This means that there is approximately 29% shared variance after disattenuation (approximately 9%

⁴ Task reliabilities set the upper bound of correlation (square root of the product of the measurements' reliabilities) we could expect due to measurement error (Nunnally, 1970). For instance, the maximum correlation between the two ensemble tasks could be .69 for accuracy scores and .63 for residual accuracy scores.

Table 1 Mean, SD, reliability of accuracy scores, and the reliability of residual accuracy scores

	Accuracy score			Residual accuracy score
	<i>M</i>	<i>SD</i>	Reliability	Reliability
Single-circle-size task	.85	.06	.63	-
Size-variance task	.76	.09	.72	.53
Average-size task	.79	.08	.69	.62

shared variance before disattenuation). In other words, correlated variance between these two aspects of ensemble perception – perceived central tendency and perceived variability – cannot be explained solely on the basis of performance variability among judgments of individual size.

Discussion

Estimation abilities for judging average and variability of circle sizes within an array of different-sized circles are positively correlated, even after controlling for perceptual discrimination abilities of the size of individual circles. This result suggests the involvement of common mechanisms when judging different statistical properties of ensembles of items differing along a simple dimension (i.e., size). This conclusion differs from that of prior studies with smaller samples and that in many cases obtained inconclusive evidence for the null hypothesis (Khvostov & Utochkin, 2019; Utochkin & Vostrikov, 2017; Yang et al., 2018). Some of the prior work observed correlations in the same range as ours (Khvostov & Utochkin, 2019; mean size and range: $r = .35$), but were not significant and were associated with a BF traditionally considered inconclusive. While it can be difficult to know what resulted in prior low correlations, we speculate that differences in procedures between tasks, such as asking participants to compare the variance of two arrays versus the mean size of an array to a single item, may have complicated the measurements. Performance on a given task nearly always reflects contributions from additional factors than just the construct that performance aims to operationalize (e.g., ensemble-processing ability in the current study), and this affects the interpretation of any study trying to interpret correlations between sets of individual tasks. Future work could deploy a latent variable approach (Bollen, 2002), using the common variance for tasks with different procedures, to estimate the ability relevant to each construct (e.g., average and variability estimation) before testing whether they relate. The present results suggest that such a future study certainly could be fruitful.

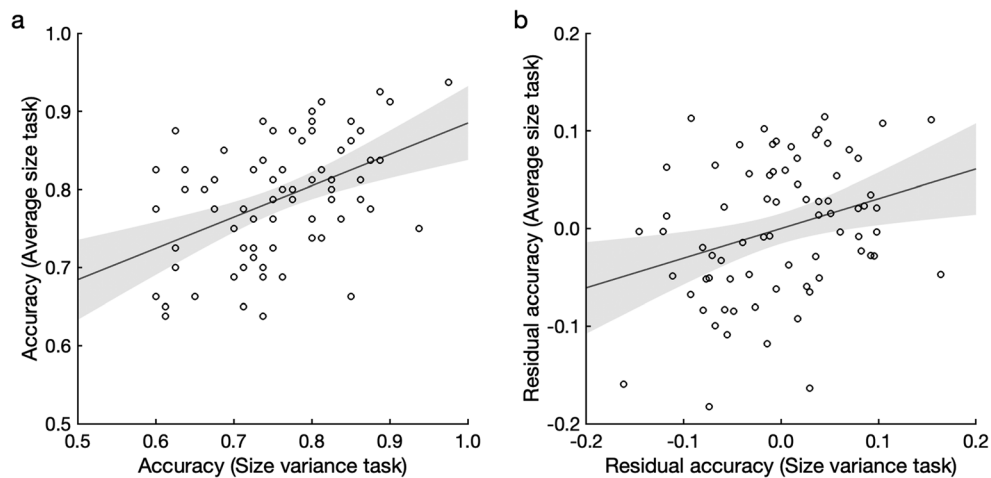


Fig. 2 Correlation between accuracy on the size-variance task and on the average-size task (a) before and (b) after controlling for the single-circle-size task accuracy. Dark gray lines show linear regression lines, and shaded areas indicate 95% confidence intervals for the regression lines

To be clear, we are not claiming that judgments of average and variability depend *only* on a common ability. Even after correction for measurement error, the shared variance between the two tasks (R^2) was 40%. It is always difficult to interpret a moderate correlation, and so our strongest possible conclusion is based on evidence against one extreme claim, which has been made in prior work (Khvostov & Utochkin, 2019; Utochkin & Vostrikov, 2017; Yang et al., 2018), that these judgments rely on independent abilities. With regards to how to think about the effect size we obtain, we offer the following considerations. First, we conservatively excluded participants so that the individuals who did not make a reasonable effort in both tasks do not inflate the correlations. Because it can be hard to distinguish such individuals from participants who tried but were poor at both tasks, this could only lead us to underestimate the effect size. Second, the size a correlation needs to be interpreted as a function of the maximum correlation that *could* be obtained given measurement error. When we disattenuate the correlations based on the reliability of the residual accuracy scores, the estimates are relatively high ($r = .54$). Third, the size of a correlation is sometimes expected to exceed what has been called the “crud factor,” i.e., a correlation between one of the tasks and some other task that it is not expected to relate to (e.g., ensemble judgment and verbal memory task; Haberman et al., 2015). However, the “crud factor” argument can be difficult to defend (Orben & Lakens, 2020), especially for moderate correlations between disparate tasks that can be the basis of strong domain-general factors in intelligence (Deary et al., 2010) or visual ability (Richler et al., 2019). Instead of comparing our effect size to the size of an effect in a different set of subjects, we measured performance on another task (single-circle-size judgment) in the same individuals so that we can control for the specific variance that could contribute to our effect of interest.⁵ Our partial correlation therefore controls for general factors like motivation, intelligence, and visual acuity, in a way that

comparison to a crud factor cannot. How the remainder of variance in each task should be apportioned will require further work.

Our findings suggest at least three sources of variance in each ensemble task. The first source of variance is associated with single-item judgments, evidenced by the correlation between each ensemble task and the single-circle-size task. This variance is likely to reflect individual differences in perceptual abilities not necessarily related to ensemble perception, and thus controlled in the partial correlation analysis. The second source is associated with common abilities involved in both ensemble judgments. We may think of two possible mechanisms that might support these common abilities. One possibility is mechanisms involved in the computation required for the estimation of both average and variance. For instance, variance estimation may require average estimation, in a similar way as variance is calculated mathematically. The other possibility is mechanisms associated with common visual processes for similar visual input (i.e., multiple circles presented in an array), such as mechanisms for distributing attention to array items efficiently. Regardless of which of these mechanisms support common abilities, the partial correlation in our study suggests common abilities specific to ensemble perception for simple objects of the same category. Finally, the third source of variance comes from mechanisms that are not shared among different ensemble judgments. This pertains to any portion of the variance on a task that proves reliable (i.e., is not measurement error), but is not shared with any of the other tasks. Processes specific to the nature of each type of statistical property could contribute to this uncorrelated variance.

⁵ In other words, if Haberman et al. (2015) had measured verbal memory in the same subjects as those for two different ensemble tasks, they could have asked how much of the correlation between these tasks remained, after controlling for verbal memory. Because performance on each task combines several sources of variation, one cannot assume that all of the variance shared between tasks A and B overlaps with the variance shared between tasks B and C.

Outside of this line of research, there is some evidence that ensemble perception does not emerge from a “monolithic process,” to borrow the phrase used by Haberman et al. (2015). Despite considerable power, these authors found that average judgments might share essentially no variance for features that differ greatly in their complexity. In addition, while they found a correlation between ensemble judgments and single-item judgments, they did not measure and, therefore, did not control for, single-item performance in their subsequent experiments where they reported correlations among simple feature ensemble judgments. Even for stimuli of different categories, contributions from object recognition ability could inflate correlations, given strong evidence for domain-general object recognition ability (Richler et al., 2019). Using a latent variable framework and complex objects, recent work estimated about 42% shared variance between a domain-general factor for individual object recognition and a domain-general factor for ensemble perception of average identity (Sunday et al., 2020). This supports the importance of controlling for individual differences in object recognition when aiming to interpret the correlations between different ensemble judgment tasks.

We readily acknowledge the limitations of the present work. First, we only studied judgments of average size and variability in size, so it is imprudent to generalize to other types of ensemble judgments. Our goal was simply to address limitations of recent work on this topic and to raise the bar for future claims of full independence between mechanisms. More empirical exploration with other judgments (e.g., numerosity) and features (e.g., orientation) will be needed to obtain a more complete picture of the structure of ensemble-processing abilities. Second, our conclusions may be constrained by some of the specific aspects of our tasks, most notably the 700-ms exposure duration we used. That value was selected to ensure that participants had sufficient time to grasp the ensemble properties we were interested in (Gorea et al., 2014; Whiting & Oriet, 2011), but it could have unwittingly allowed participants to base their judgments on individual-item processing. Importantly, this would be expected to inflate the correlation between ensemble judgments and the single-circle-size judgments (i.e., variance controlled in the partial correlation). However, we found that this portion of the variance did not account for the main part of the correlation between ensemble tasks. It remains to be seen whether ensemble judgments performed with very brief exposure durations tap into the same ability as that deployed with longer duration exposures. It is also possible that a 700-ms presentation duration was sufficient for processing one property (e.g., average) but limited processing of another property (e.g., variability), leading to the underestimation of common variance between ensemble judgment abilities.

Studying individual differences in ensemble perception can enrich our understanding of the extent to which this ability

represents an important precursor to object recognition (Oliva & Torralba, 2006). Progress in this direction can benefit from methods that have long been applied to the study of intelligence or personality and are starting to deepen our understanding of visual cognition (Bollen, 2002; Richler et al., 2019).

Author note This work was supported by the Centennial Research Fund (Vanderbilt University) and by the David K. Wilson Chair Research Fund (Vanderbilt University).

Declarations

Conflicts of interest We have no conflicts of interest to disclose.

References

- Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology*, 53(1), 605–634. <https://doi.org/10.1146/annurev.psych.53.100901.135239>
- 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](https://doi.org/10.1016/S0042-6989(02)00596-5)
- Deary, I. J., Penke, L., & Johnson, W. (2010). The neuroscience of human intelligence differences. *Nature Reviews Neuroscience*, 11(3), 201–211. <https://doi.org/10.1038/nrn2793>
- 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>
- Goodhew, S. C., & Edwards, M. (2019). Translating experimental paradigms into individual-differences research: Contributions, challenges, and practical recommendations. *Consciousness & Cognition*, 69, 14–25. <https://doi.org/10.1016/j.concog.2019.01.008>
- Gorea, A., Belkoura, S., & Solomon, J. A. (2014). Summary statistics for size over space and time. *Journal of Vision*, 14(9), 22. <https://doi.org/10.1167/14.9.22>
- 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>
- JASP Team. (2019). JASP (Version 0.11.1) [Computer Software]. Available at <https://jasp-stats.org/>
- Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford University Press.
- Khvostov, V. A., & Utochkin, I. S. (2019). Independent and parallel visual processing of ensemble statistics: Evidence from dual tasks. *Journal of Vision*, 19(9), 3. <https://doi.org/10.1167/19.9.3>
- Lee, M. D., & Wagenmakers, E. J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.
- Mollon, J. D., Bosten, J. M., Peterzell, D. H., & Webster, M. A. (2017). Individual differences in visual science: What can be learned and what is good experimental practice? *Vision Research*, 141, 4–15. <https://doi.org/10.1016/j.visres.2017.11.001>
- Morgan, M., Chubb, C., & Solomon, J. A. (2008). A ‘dipper’ function for texture discrimination based on orientation variance. *Journal of Vision*, 8(11), 9. <https://doi.org/10.1167/8.11.9>

- Nunnally, J. C. (1970). *Introduction to psychological measurement*. McGraw-Hill.
- 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](https://doi.org/10.1016/S0079-6123(06)55002-2)
- Orben, A., & Lakens, D. (2020). Crud (re) defined. *Advances in Methods and Practices in Psychological Science*, 3(2), 238–247. <https://doi.org/10.1177/2515245920917961>
- Osborne, J. W. (2002). Effect sizes and the disattenuation of correlation and regression coefficients: Lessons from educational psychology. *Practical Assessment, Research, and Evaluation*, 8(1), 11. <https://doi.org/10.7275/0k9h-tq64>
- 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>
- 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>
- Schönbrodt, F. D., & Wagenmakers, E. J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, 25(1), 128–142. <https://doi.org/10.3758/s13423-017-1230-y>
- Schönbrodt, F. D., Wagenmakers, E. J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, 22(2), 322–339. <https://doi.org/10.1037/met0000061>
- Stefan, A. M., Gronau, Q. F., Schönbrodt, F. D., & Wagenmakers, E. J. (2019). A tutorial on Bayes Factor Design Analysis using an informed prior. *Behavior Research Methods*, 51(3), 1042–1058. <https://doi.org/10.3758/s13428-018-01189-8>
- Sunday, M.A., Tomarken, A., Cho, S.-J. & Gauthier, I. (2020). *Novel and familiar object recognition rely on the same ability*. Manuscript submitted for publication.
- Teghtsoonian, M. (1965). The judgment of size. *The American Journal of Psychology*, 78, 392–402. <https://doi.org/10.2307/1420573>
- Utochkin, I. S., & Vostrikov, K. O. (2017). The numerosity and mean size of multiple objects are perceived independently and in parallel. *PLoS One*, 12(9), e0185452. <https://doi.org/10.1371/journal.pone.0185452>
- Vogel, E. K., & Awh, E. (2008). How to exploit diversity for scientific gain: Using individual differences to constrain cognitive theory. *Current Directions in Psychological Science*, 17(2), 171–176. <https://doi.org/10.1111/j.1467-8721.2008.00569.x>
- Whitney, D., & Yamanashi Leib, A. (2018). Ensemble perception. *Annual Review of Psychology*, 69, 105–129. <https://doi.org/10.1146/annurev-psych-010416-044232>
- Whiting, B. F., & Oriet, C. (2011). Rapid averaging? Not so fast! *Psychonomic Bulletin & Review*, 18(3), 484–489. <https://doi.org/10.3758/s13423-011-0071-3>
- Wilkinson, F., Wilson, H. R., & Habak, C. (1998). Detection and recognition of radial frequency patterns. *Vision Research*, 38(22), 3555–3568. [https://doi.org/10.1016/S0042-6989\(98\)00039-X](https://doi.org/10.1016/S0042-6989(98)00039-X)
- 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. <https://doi.org/10.1167/15.4.11>
- Yang, Y., Tokita, M., & Ishiguchi, A. (2018). Is there a common summary statistical process for representing the mean and variance? A study using illustrations of familiar items. *i-Perception*, 9(1), 2041669517747297. <https://doi.org/10.1177/2041669517747297>

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