AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

ISSN: 0096-3445

2024, Vol. 153, No. 6, 1517–1536
https://doi.org/10.1037/xge0001582

A General Ability for Judging Simple and Complex Ensembles

Ting-Yun Chang[1], Oakyoon Cha[2], and Isabel Gauthier[1]
[1] Department of Psychology, Vanderbilt University
[2] Department of Psychology, Sungshin Women's University

People can report summary statistics for various features about a group of objects. One theory is that different abilities support ensemble judgments about low-level features like color versus high-level features like identity. Existing research mostly evaluates such claims based on evidence of correlations within and between feature domains. However, correlations between two identical tasks that only differ in the type of feature that is used can be inflated by method variance. Another concern is that conclusions about high-level features are mostly based on faces. We used latent variable methods on data from 237 participants to investigate the abilities supporting low-level and high-level feature ensemble judgments. Ensemble judgment was measured with six distinct tests, each requiring judgments for a distinct low-level or high-level feature, using different task requirements. We also controlled for other general visual abilities when examining how low-level and high-level ensemble abilities relate to each other. Confirmatory factor analyses showed a perfect correlation between the two factors, suggesting a single ability. There was a unique relationship between these two factors beyond the influence of object recognition and perceptual speed. Additional results from 117 of the same participants also ruled out the role of working memory. This study provides strong evidence of a general ensemble judgment ability across a wide range of features at the latent level and characterizes its relationship to other visual abilities.

Public Significance Statement
People can summarize what a similar group of objects looks like to quickly gain general information about what they are seeing. This work suggests that some people are better than others at making such judgments, and that the same ability predicts performance for judgments about simple visual features or those about complex object shapes. This ability is also related to performance recognizing individual objects. This suggests visual abilities apply across a variety of situations and can therefore predict performance in many real-world settings.

Keywords: ensemble perception, individual differences, object recognition

Driving on a highway, we constantly gauge the average speed of the surrounding cars to drive safely. At an ice cream shop, we estimate the flavors of ice cream by the diversity of colors in the freezer. People can make such judgments about ensembles of objects for a host of visual features, ranging from lower level features like size (Ariely, 2001), aspect ratio (Elias & Sweeny, 2020; Sweeny et al., 2021), lightness (Bauer, 2017; Takano & Kimura, 2020), and orientation (Alvarez & Oliva, 2009), to higher level features such as car models (Cha et al., 2021; Chang & Gauthier, 2022) and facial expression and identity (De Fockert & Wolfenstein, 2009; Haberman & Whitney, 2007). People can also evaluate groups of objects to consider summaries of abstract features, such as the lifelikeness or economic value of objects (Yamanashi Leib et al., 2016, 2020) and the attractiveness of faces (Luo & Zhou, 2018). Moreover, these ensemble judgments can be made with a focus on different summary statistics, including average (Alvarez & Oliva, 2009; Ariely, 2001; Im & Chong, 2014), variance/diversity (Albers et al., 2014; Haberman, Lee, & Whitney, 2015; Solomon, 2010; Ward et al., 2016), and numerosity (Khvostov & Utochkin, 2019; Utochkin & Vostrikov, 2017). Given that ensemble judgments can apply to a wide range of features and summary statistics, we ask to what extent a single ability supports performance on these tasks.

Isabel Gauthier https://orcid.org/0000-0002-6249-4769

This work was supported by the David K. Wilson Chair Research Fund from Vanderbilt University awarded to Isabel Gauthier and the 2021–2022 Taiwanese Overseas Pioneers Grants (TOP Grants) for PhD candidates from the Ministry of Science and Technology, Taiwan, awarded to Ting-Yun Chang. The authors declare no conflicts of interest. All materials, data, and analysis code are publicly available online at https://Figshare.com through this link: https://doi.org/10.6084/m9.figshare.24236950.v2.

Ting-Yun Chang served as lead for formal analysis and writing–original draft and contributed equally to writing–review and editing. Oakyoon Cha served in a supporting role for writing–original draft and writing–review and editing. Isabel Gauthier served as lead for writing–review and editing and served in a supporting role for formal analysis and writing–original draft. Ting-Yun Chang, Oakyoon Cha, and Isabel Gauthier contributed equally to conceptualization.

Correspondence concerning this article should be addressed to Ting-Yun Chang, Department of Psychology, Vanderbilt University, 111 21st Avenue South, Nashville, TN 37240, United States. Email: ting-yun.chang@hotmail.com

Specifically, we focus on the contrast between low-level and high-level features and examine whether the ability supporting ensemble judgments at each level is shared, after abstracting away from specific task demands.

A landmark study by Haberman, Brady, and Alvarez (2015) first addressed the question of ensemble perception (EP) abilities from the perspective of individual differences. In a series of separate studies, they surveyed color and orientation as low-level features, as well as facial expression and face identity as high-level features. An ensemble task required participants to adjust an anchor matching the average feature of the four objects presented in a simultaneous array. A matched single-item task showed the same arrays but required a matching judgment about a single cued stimulus. Looking at any given feature, performance in the ensemble judgments and in the single-item judgments were correlated ($rs \geq .43$, $ps < .01$). Also, performance on the ensemble judgments was correlated ($rs \geq .42$, $ps < .01$) across pairs of low-level features and across pairs of high-level features. But critically, little or no correlation was found across ensemble judgments for low- and high-level features ($rs \leq .29$, nonsignificant). Haberman et al. concluded that EP is not general for all features. They proposed that the extent to which ensemble representations engage the same process depends on the complexity of the relevant features. At the very least, their data suggest that performance on ensemble judgments for facial features may not be strongly related to that for other visual features.

Although this early work only measured ensemble judgments for complex items that were faces, other studies have since provided evidence for a shared ability supporting ensemble judgments for a variety of other nonface complex features. Chang and Gauthier (2022) reported strong correlated performance ($rs \geq .41$, Bayes factor, $BF_{10}s > 10$) in judging mean identity for distinct complex object categories such as birds, planes, and cars, even after controlling for single object recognition (OR) in each category. These correlations across complex object categories suggested for the first time that EP for high-level features like identity is not specific to faces and is to a large extent domain-general for categories other than faces. In a later study (Sunday et al., 2022), participants completed three visual tasks, including ensemble mean identity judgment, with six distinct complex object categories spanning from novel objects like greebles (Gauthier & Tarr, 1997) to familiar objects like birds. A single factor accounted for the performance on these tasks with all complex object categories. These results suggest that the high/low distinction proposed by Haberman, Brady, and Alvarez (2015) could be put to the test in work that used a range of nonface objects.

Although Haberman, Brady, and Alvarez (2015) only used faces as complex objects, their proposed low-level ensemble ability was supported by pairwise correlations across judgments of the average color of dots and the average orientation of Gabors and triangles. Some work with small samples challenged these results, reporting only minimal correlations between judgments of average line orientation and average length (Yörük & Boduroglu, 2020). However, other studies with larger samples found correlations among ensemble judgments for different low-level features. For example, Cha et al. (2022) asked participants to estimate different summary statistics for circle sizes, the mean size, and the variance of size. They found that these different summary judgments were positively correlated, after controlling for the ability to judge the size of single circles ($r = .45$, $BF_{10} > 10$). Similarly, Kacin et al. (2021, 2022) reported a strong positive correlation ($rs \geq .62$, $BF_{10s} > 10$) between judging the average length and average orientation of line ensembles. In sum, although a dissociation for low- and high-level features has not been tested extensively since Haberman, Brady, and Alvarez (2015), there is support for a shared ability for different features of similar complexity, be it low-level or high-level.

Most of what we know about these questions is based on pairwise correlations (Cha et al., 2022; Chang & Gauthier, 2022; Haberman, Brady, & Alvarez, 2015; Kacin et al., 2021, 2022). Inferences based on pairwise correlations, even when they depend on a large set of such results, are intrinsically limited. Pairwise correlations cannot provide strong evidence of abilities that can generalize beyond the particular pair of tasks being compared. The correlation between a specific pair of tests can reflect shared task demands for the two specific tests in a way that may not generalize to other tests meant to tap a similar construct but that use a different task format. Whereas shared task demands may inflate the correlation, measurement errors can deflate it. Interpreting several pairwise correlations collectively is also challenging, because the common variance among several distinct pairs of tests does not necessarily have the same source, and thus transitive inferences cannot be made.

Most individual differences studies of EP have focused on comparing performance in two versions of the same task that differ only in the feature property to summarize, such as contrasting the mean length with the mean orientation of lines (Kacin et al., 2021, 2022), or contrasting the mean circle size with the variance of circle size (Cha et al., 2022). In Sunday et al. (2022), a latent variable was assessed in the same mean estimation task (ME) for six object categories, three novel and three familiar. This work suggested that a common factor accounts for between 34% and 69% of the variance in each ensemble test and approximately 42% overlap with a latent factor for OR. However, this study only used a single-task format across all six EP tests. This lack of variability in task demands may inflate the correlations between EP tests while at the same time reducing the generality of the latent factor extracted, limiting its correlation with other latent variables.

The goal of the current study is to characterize the relationship between EP for low-level and EP for high-level features, in a latent variable framework. In a latent variable approach (T. A. Brown, 2015; Miyake, 2001), unobservable constructs such as abilities are estimated based on multiple measures (i.e., indicators), which are assumed to be independent aside from the common influence of the construct. Ideally, indicators of a construct should be as diverse as possible to avoid a biased approximation of the factor and therefore better generalize to other conditions (Little et al., 1999; Whitely, 1983). We use confirmatory factor analysis (CFA; T. A. Brown, 2015; Tomarken & Waller, 2005) to examine the correlation between latent variables. Relative to zero-order correlations and regression, this approach limits the influence of construct-irrelevant variance and measurement error. Specifically, we asked: do different low-level EP tasks load onto one latent variable and different high-level EP tasks load onto another latent variable? If so, how strongly are these two latent variables correlated?

Based on the findings in a recent study (Chang et al., 2024) that two EP tasks not using the same stimuli, summary statistics, or task format could tap into a common EP ability, we created a series of tests to measure performance in different ensemble judgments for low- and high-level features of various kinds of simple and

complex objects. Because many studies have revealed correlations between EP and OR (Cha et al., 2022; Chang et al., 2024; Haberman, Brady, & Alvarez, 2015; Sweeny et al., 2015), we measured domain-general OR, the ability to make within-category discriminations for objects regardless of familiarity and reliance on short- or long-term memory (Richler et al., 2019; Smithson et al., 2024; Sunday et al., 2022). We also chose to control for perceptual speed (PS; Ekstrom et al., 1976). PS is a facet of processing speed (Carroll, 1993) and represents the efficiency of processing visual information and executing elementary visual operations such as rapidly scanning and matching symbols, pictures, or words (Ackerman, 1988, 1990; Ackerman & Cianciolo, 2000; Ackerman et al., 2002). PS is a construct related to general intelligence (Ackerman et al., 2002; Redick et al., 2013; Salthouse & Babcock, 1991) and is well-positioned as a potential source of shared variance between EP and OR. In prior studies (Richler et al., 2017, 2019), measures of general intelligence like Raven's progressive matrices (Raven, 2000) have either not correlated strongly with general OR or were not able to account for all of the stable variance in its measurement (Smithson et al., 2024).

Method

Participants

This study consisted of ten different tests over the course of two sessions. The study was completed online, and all participants were recruited for monetary reward through the online participant recruitment platform Prolific.co. The study was approved by Vanderbilt University Institutional Review Board. Based on the observed pairwise correlation of .4 between very different EP tasks (Chang & Gauthier, 2022), we needed at least 200 participants to reach the desired .8 power for estimating two latent EP factors with six indicators using structural equation modeling (Cohen, 1988; Soper, 2023; Westland, 2010). In addition, correlations stabilize around 250 participants under most conditions (Schönbrodt & Perugini, 2013). We recruited 300 participants in Session 1 to allow potential dropout and exclusion. Participants were selected based on reporting English proficiency and a previous submission approval rate of at least 95%. Session 2 was made available to all participants who completed Session 1 after 72 hr of their submission for Session 1. There was a window of 7 days to complete Session 2. Two-hundred and thirty-seven participants returned for Session 2. One attention check trial was embedded in each test to estimate data quality, except for the two speeded PS tests. We prespecified a criterion to exclude data from participants that failed more than two out of four attention check trials in a session. None of the participants was excluded from data analyses, since they all passed at least half of the attention check trials in the study. The results presented here were derived from the data for all 237 participants, 71 self-reported as men, 161 women, and five others as gender-variants; $M_{age} = 35.90$; $SD = 9.83$; range = (18, 55), and 117 among the 237 participants who voluntarily completed the additional two working memory (WM) tests, 39 self-reported as men, 78 as women; $M_{age} = 37.99$; $SD = 9.47$; range = (18, 55).

Procedure

Participants completed a total of 10 tests in two sessions: six EP tests, two PS tests, and two OR tests (see Table 1 for a summary for task information about the tests). In Session 1, participants completed five tests in the following order: the identical pictures test (Ekstrom et al., 1976), the mean matching task with birds (mean matching—bird), the ME with orientation (mean estimation—Orientation), the diversity comparison task with lightness (diversity comparison—lightness), and the object matching tests with asymmetric greebles. In Session 2, the following five tests were administered in the following order: the hidden pattern test (Ekstrom et al., 1976), the mean matching task with aspect ratio (mean matching—aspect ratio), the ME with transformers (mean estimation—transformer), the diversity comparison task with ziggerins (diversity comparison—ziggerin), and the novel object memory test with symmetric greebles (OR—greeble). To sum up, each session consisted of one PS test, three EP tests, and one OR test.

To avoid confounding presentation order with individual differences, the trials within each test were in the same order for all participants, roughly ordered from easiest to the most difficult based on pilot results. At the beginning of each session, participants were told that there would be five different tests in the session and that they should pay close attention to the instructions before each test. Each session took approximately 30 min to complete. Participants were allowed to take breaks between tests.

EP Tasks: Mean Estimation, Diversity Comparison, and Mean Matching

We measure EP with three distinct task formats. We use a ME, a diversity comparison task, and a mean matching task. Each task is applied to both simple and complex features. Mean estimation requires participants to estimate the mean feature of a group of objects, gauging performance by the proximity of their response to the actual average. Diversity comparison assesses participants' ability to discern the diversity of the target feature between two sets of objects, with performance indexed by accuracy. Mean matching, a novel approach, has participants determining whether two ensembles share the same mean feature, also measured by accuracy. Unlike mean estimation, which focuses on the average within a single group, mean matching requires a comparison of averages across groups. Both diversity comparison and mean matching tasks offer the advantage of not necessitating morphed stimuli, allowing for a broader range of exemplars and reducing repetition. This minimizes learning or memory effects during the task. This range of tasks enhances the generalizability of our EP construct within the realm of explicit EP tasks.

ME—Orientation (Kacin et al., 2021)

Each trial began with a white square to fixate at the center of the screen. After 1,000 ms, the fixation square turned black for 500 ms. A line ensemble (12 lines varying in their clockwise orientation, see Figure 1) was then shown for 200 ms, followed by a response display with five oriented lines (18°, 31.5°, 45°, 58.5°, and 72° clockwise from the vertical orientation in respective sequence). Participants chose the mean orientation by clicking on it. The lines (all of them 80 pixels long) could appear oriented clockwise from 4.5° to 85.5° depending on the designated mean orientation (18°, 31.5°, 45°, 58.5°, or 72°) and the *SD* (9° or 13°). There were 10 practice trials with feedback (correct/incorrect) before a block of 75 experimental trials with no feedback. One attention check was included (it said "click here"). Performance was indexed with mean absolute

Table 1
Information About the Tests Used

Test	Ability	Task requirement	Response
Mean estimation—orientation	EP_{simple}	Estimate mean orientation of 12 lines presented for 200 ms	5AFC
Diversity comparison—lightness	EP_{simple}	Choose the more diverse in lightness in pairs of 16-square ensembles presented for 500 ms	2AFC
Mean matching—aspect ratio	EP_{simple}	Decide if two 6-ellipse ensembles, each presented for 300 ms, have the same mean aspect ratio	2AFC
Mean estimation—transformer	$EP_{complex}$	Estimate mean identity of four Transformers presented for 1 s	6AFC
Diversity comparison—ziggerin	$EP_{complex}$	Choose the more diverse in identity in pairs of six ziggerin ensembles presented for 500 ms	2AFC
Mean matching—bird	$EP_{complex}$	Decide if two 6-bird ensembles, each presented for 1 s, have the same mean identity	2AFC
Identical pictures	PS	Given a target picture and select the identical picture as correctly and quickly as possible in 1.5 min	5AFC
Hidden patterns	PS	Given a target pattern and repeatedly respond if it is embedded in all subsequent images as correctly and quickly as possible in 3 min	2AFC
Object matching	OR	Given a target greeble and choose from two distractors regardless of viewpoint	3AFC
Novel object memory	OR	Studied six greebles and recognized them against two distractors regardless of viewpoint	3AFC
Binding	WM	Memorize two to six arbitrary word–number pairs and recall given a cue	AFC
Operation span	WM	Evaluated four to eight equations and memorize letters alternately presented then recall letters in order	Free recall

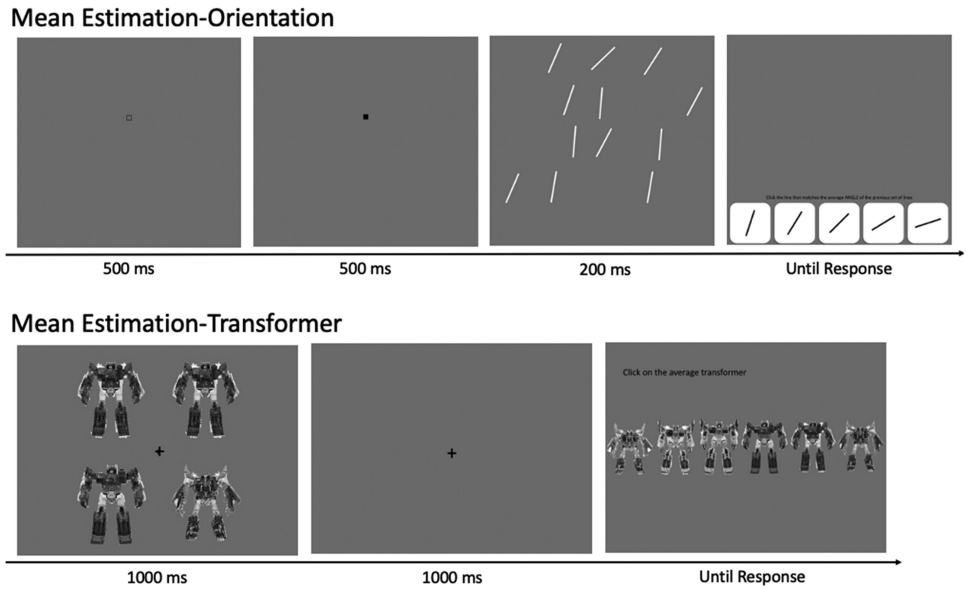
Note. AFC = alternative forced choice; EP = ensemble perception; PS = perceptual speed; OR = object recognition; WM = working memory (used in the supplemental study).

error in degree. For example, if the correct mean on a trial was Option 3 (45°) and the participants responded Option 4 (58.5°), then the trial would be scored as a 13.5°. Chance performance was 21.42°. That is because depending on which of the five orientations is the correct answer, starting with 18°, the average error when guessing is 27°, 18.9°, 16.2°, 18.9°, and 27°, and the number of trials with these correct answers, respectively, was 14, 15, 14, 18, and 14. Accordingly, to compute chance we multiply the number for each kind of trial by the random error in each case and obtain a guessing of 21.42° (expected if someone randomly chose any of the five answers on all trials).

Mean Estimation—Transformer (Chang & Gauthier, 2022)

This test showed an array of four morphed images of Transformer robots for 1 s (see Chang & Gauthier, 2022, for details of the morphing procedure that used three original transformer images). A trial began with a 1,000 ms study array of four Transformers. Following

Figure 1
Schematics for the ME With Orientation (Top) and Transformer (Bottom)



Note. The instruction for response reads "Click the line that matches the average ANGLE of the previous set of lines" in orientation test. The instruction for response reads "Click on the average transformer" in transformer test. Note that the presentation of these displays had been adjusted for demonstration purposes and was not the same as used in the actual study. ME = mean estimation task.

a 1,000-ms interstimulus interval, a response array consisting of six transformer morphs was shown and participants chose the morph that best matched the average. The six morphs were presented in a fixed order in the response arrays, although starting from a different morphs across trials. There were four practice trials with feedback, followed by 50 trials with no feedback, and one attention check. Performance was indexed by the mean absolute distance from the correct answer in degree. For instance, if a participant responded Option 2 on a trial in which the correct answer was 3, scored the performance on this trial as $|2 - 3| \times 60 = 60$. The smaller the absolute distance indicates the better performance. Chance performance was 90°. This is because the six answers are on a circle, and so the possible distances from the correct answer are always 0, 1, 2, 3, 2, and 1, which is 1.5 on average (and $1.5 \times 60 = 90°$).

Diversity Comparison—Lightness. Each trial began with a 400-ms fixation cross presented at the center of the screen (Figure 2). Two ensembles consisting of 16 gray squares that vary in lightness appeared on the screen sequentially, for 500 ms each with an interstimulus internal (ISI) of 400 ms between them. Next, participants were given a two-alternative forced choice between "1st Array was more diverse" or "2nd Array was more diverse." Performance on this test was indexed by the accuracy across all trials, with a chance performance of .50.

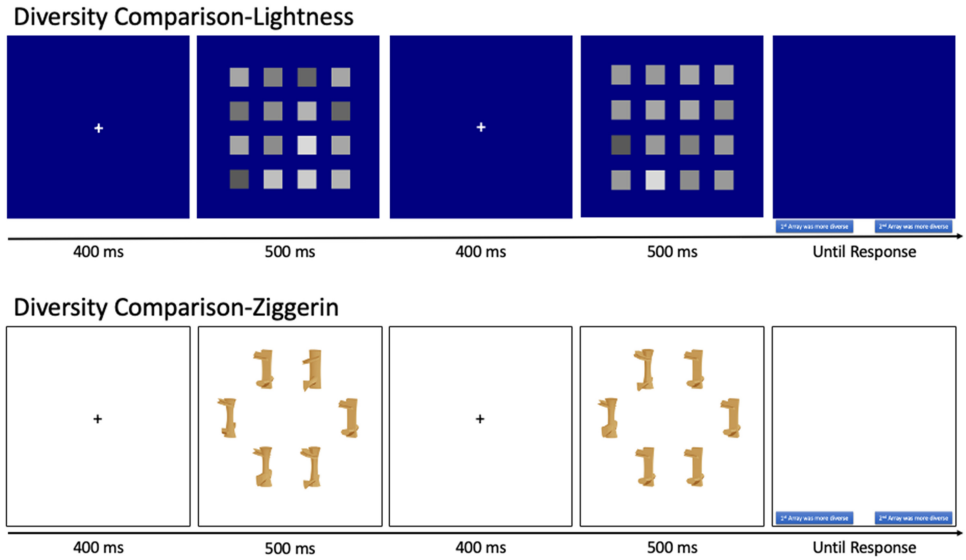
The background color of all displays was set to be a dark blue that excludes red and green, RGB (0, 0, 128), so that the luminance was low enough that even the darkest square could be easily distinguished from the background. The size of each square was 50 × 50 pixels and they were separated by 40 pixels. Consider the darkest gray (black) as 0%, RGB (0, 0, 0), and the brightest gray (white) is 100%, RGB (255, 255, 255), the lightness of the gray squares ranged from 30% to 95%, in 5% increments, for a total of 14 levels. The lightness of a given ensemble may have a mean of level 5, 6, 7, 8, 9, and 10 with an SD of 1.5, 2, 2.5, 3, 3.5, and 4 levels. The lightness of the squares ranged from eight to 12 levels within an ensemble, with the lightness level for each square in the ensemble randomly sampled from a uniform distribution within this range. We constrained the two ensembles in the same trial to have the same mean and range of lightness. Based on pilot results, the difference in *SD* between the two ensembles in a trial was set to either 0.5 or 1 on different trials, to prevent ceiling and floor performance while minimizing the number of repeated lightness level in an ensemble. Participants were given 10 practice trials with feedback on accuracy before the 71 test trials without feedback. One attention check trial was included.

Diversity Comparison—Ziggerin. At the beginning of a trial, a fixation cross appeared at the center of the screen for 400 ms, followed by the presentation of the first ensemble of six ziggerin objects for 500 ms (Figure 2). After another 400-ms fixation cross, the second ensemble of six ziggerins was shown for 500 ms. Participants then responded either "1st array was more diverse" or "2nd array was more diverse" by mouse click. A total of 84 distinct ziggerin objects were used to create the ensembles. Both ensembles in a trial had the same repeated ziggerins. Specifically, one ensemble included a ziggerin repeated two times (a more diverse ensemble) while the same ziggerin was repeated four times in the other ensemble (a less diverse ensemble). Object location was shuffled across ensembles, ensuring that detecting a change between ensembles or a single repetition in one ensemble was not sufficient to perform the task. There were 10 practice trials (with a longer 800 ms presentation time) with feedback (correct/incorrect) before 64 trials with no feedback. One trial used a "Click Here" button as an attention check. Performance on this test was indexed by the accuracy across the 64 formal trials. Chance performance is 0.50.

Mean Matching—Aspect Ratio. Each trial began with a 400-ms fixation square, followed by a first ensemble of six ellipses presented for 300 ms (Figure 3). Following a 500-ms ISI, a second

Figure 2

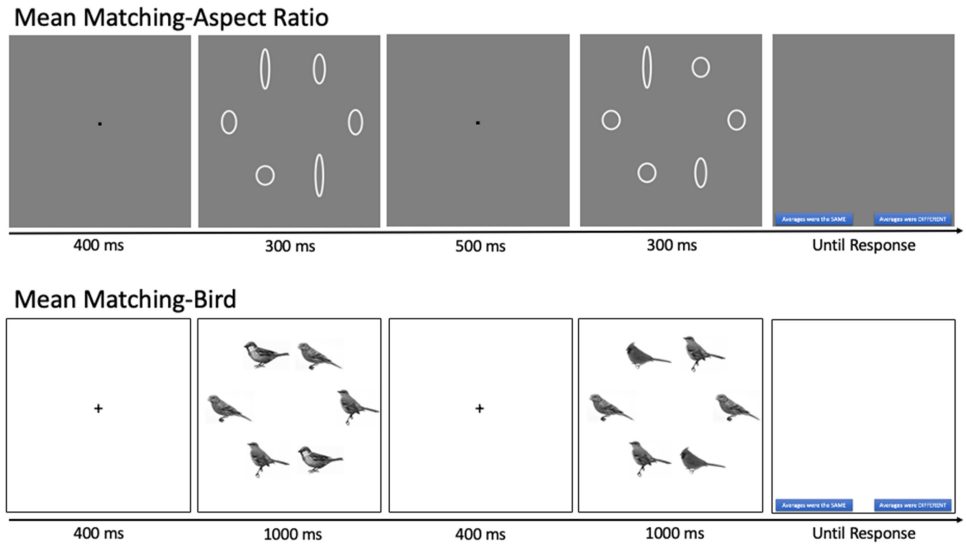
Schematics for the Diversity Comparison Task in Lightness (Top) and Ziggerin (Bottom) Tests



Note. The two buttons shown at the bottom of the response display read "1st array was more diverse" and "2nd array was more diverse" from left to right, respectively. Note that these ensembles are enlarged for the purpose of illustration. See the online article for the color version of this figure.

Figure 3

Schematics for the Mean Matching Task in Aspect Ratio Test (Top) and Bird Test (Bottom)



Note. The two buttons shown at the bottom of the response display read "Averages were the SAME" and "Averages were DIFFERENT" from left to right, respectively. Note that the ensembles in this figure were enlarged for demonstration purpose. See the online article for the color version of this figure.

ensemble of six ellipses was shown for 300 ms. Next, participants clicked on "Averages were the SAME" or "Averages were DIFFERENT." There were 10 practice trials in which ensembles were shown for 500 ms with feedback (correct/incorrect), prior to a formal block of 55 trials. There was one attention check trial in which participants saw a "Click Here" button. The average accuracy over 55 trials was used to index performance.

People are less sensitive to the mean aspect ratio of ellipse sets when the sets contain both tall and flat ellipses compared to sets that are comprised of all tall or all flat ellipses (Elias & Sweeny, 2020). To maximize individual differences in this task, trials only included tall ellipses. Note that we assumed that the arithmetic mean is computed for the aspect ratio on these ellipse ensembles, which would not be optimal for cross-boundary ensembles that include both tall and flat ellipses (and for which the geometric mean would be more accurate).

We used a total of 16 levels of aspect ratio to create the ellipse ensembles, including the following aspect ratios (in log scale): 0.095, 0.182, 0.337, 0.470, 0.642, 0.742, 0.876, 0.993, 1.099, 1.194, 1.308, 1.411, 1.504, 1.609, 1.705, and 1.792. With the area of all ellipses fixed to be 314.159 pixels^2 ($\pi \times 10^2$), the major (long) radius and minor (short) radius of the 16 ellipses ranged from 10 to 25 pixels and from four to 10 pixels, respectively. Arrays consisted of six white ellipses placed over a 600×650 gray background, RGB (128, 128, 128). The overall size of the ensembles roughly spanned from 200 to 300 pixels in width and length. The range for each ensemble was constrained to be larger than 10 levels. For "same" trials, the two ensembles consisted of the same ellipses with their locations shuffled. For "different" trials, the range and standard deviation of the two ensembles were the same so participants had to rely on the mean. The difference ranged from 2.4 (*the hardest*) to 3 (*the easiest*) levels between the mean aspect ratio of ensemble pairs within a trial.

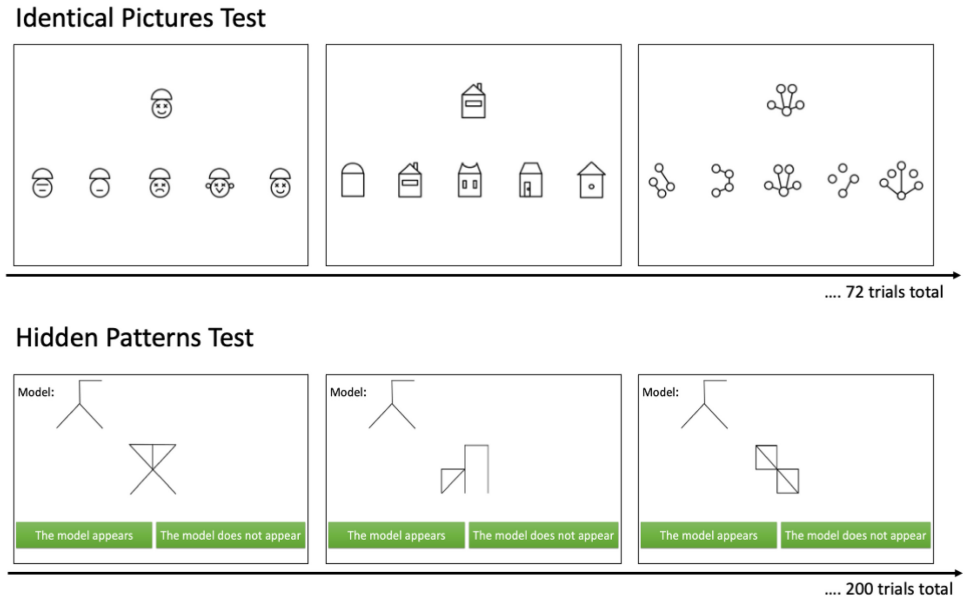
Mean Matching Task—Bird. On each trial, a 400-ms fixation cross was shown, followed by two ensembles of six birds presented sequentially, each for 1 s, with a 500-ms ISI (Figure 3). Participants then clicked on either "Averages were the SAME" or "Averages were DIFFERENT." In the instructions, participants saw an example of an average identity of two birds with their 50%–50% morph, but no morphs were used in the subsequent 10 practice trials and the 50 test trials. Performance was indexed by average accuracy across all 50 test trials. There was one attention check trial. We used a total of 14 perching birds to create the ensembles, including finches, sparrows, cardinals, crows, and phoebes. To make each array, three birds were selected and each was repeated twice. In the "same" trials, the two bird ensembles consisted of the same images shuffled in new positions. In the "different" trials, one bird pair was replaced by a different bird pair, and images were also shuffled in new positions.

PS Tests

We adapted two classic tests that were originally paper-and-pencil tests to measure PS. The first is the identical pictures test (Ekstrom et al., 1976). On each trial, a target object is shown at the top of the screen. At the bottom are five test objects, one of which matches the target object. Participants must select the target object correctly as fast as they can. There were four practice trials with feedback on accuracy, and 72 test trials in which no feedback was given. The target picture for each trial was all unique and nonrepeated (see Figure 4 top panel). We modified one item, which resembled a head with a conical hat and with narrow eyes as depicted by a backslash and a forward slash. The triangles and slashes were replaced with half circles and Xs to avoid racial associations (see the first trial in Figure 4 top panel). Performance was indexed with (the number of correct response) minus (the number of incorrect response), as achieved in a 1.5-min limit. Participants were told to be as correct and as quick as possible.

Figure 4

Sample Trials of the Identical Pictures Test (Top) and the Hidden Patterns Test (Bottom)



Note. The two buttons at the bottom of each display in the hidden patterns test read "The model appears" and "The model does not appear," from left to right, respectively. See the online article for the color version of this figure.

The second test was adapted from the hidden patterns test (Ekstrom et al., 1976). Participants were given one "model" pattern to recognize and then, on each trial, reported whether or not this pattern was embedded within a picture (see Figure 4, bottom panel). There was only one pattern to recognize throughout the test, which remained in the top-left corner of the screen as a reminder. There were 10 untimed trials with feedback, followed by a block of 200 trials. Performance on this test is indexed with (the number of correct response) minus (the number of incorrect response) within 3 min. Both accuracy and speed were emphasized.

OR Tests

Two tests were used to measure OR. Following prior work (Chang & Gauthier, 2022; Gauthier & Fiestan, 2023), the tests use novel objects and different task demands, and their aggregate provides an estimate of the latent factor *o*, as measured with CFA with a larger set of measures (Smithson et al., 2024). One measure was the novel object memory test—greeble (Richler et al., 2017; Figure 5A). In this test, participants completed three study-test blocks where they first studied six target symmetric greebles for 20 s in a study phase and then on each trial, chose which of three greebles was one of the six targets. In the first test phase, there were six trials in which the target greebles always appeared exactly the same as studied. In the second test phase, there were six more such trials, along with 12 trials where visual noise was added to all greebles. Upon completion of the second block, participants were explicitly informed that the target greebles would be presented in a different viewpoint in the next phase, which included 12 trials without noise and 12 trials with noise. There was one attention check trial in the second test phase where a target greeble was placed among two objects from other very different novel objects (not greebles). Percent correct over 48 trials was used to index performance on this test. Chance level was 0.33.

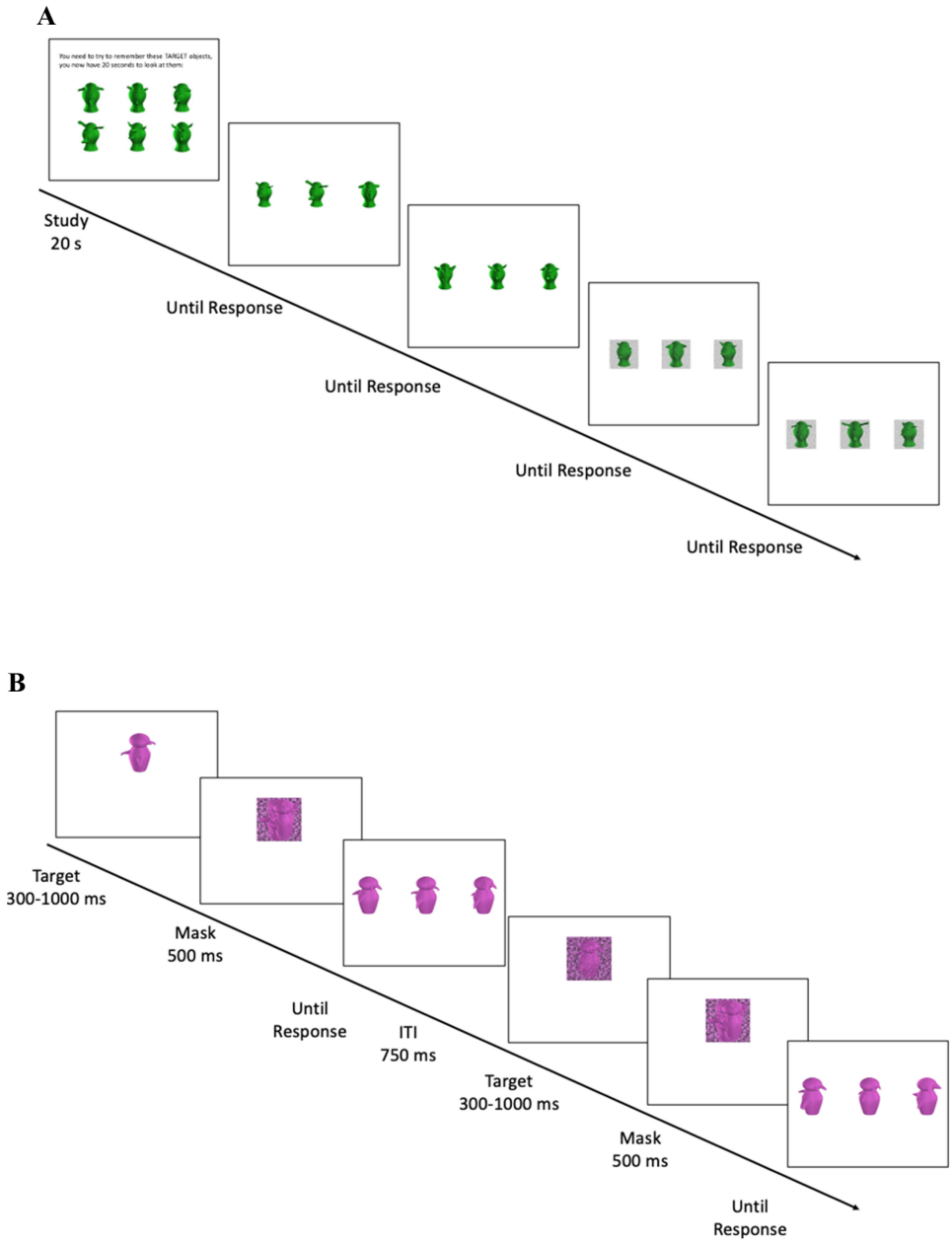
The second measure was the object matching test—greebles (Figure 5B). This uses asymmetric greebles, which have a different part configuration of parts and are considered a different basic-level category compared to symmetric greebles (Richler et al., 2019). On each trial, participants saw a target greeble for a short duration (300–1,000 ms depending on the trial). After a 500-ms mask made of scrambled greeble parts and noise, the participants selected the target greeble among three greebles, regardless of viewpoint differences. Options remained on the screen until the response. Targets and distractors did not repeat during the trials. The intertrial interval was 750 ms. Before the 48 formal trials, three practice trials with feedback (correct/incorrect) were included. The first eight trials showed the three options in the same viewpoint as the target (which appeared for 300 ms). The next 16 trials showed the three options in a different viewpoint as the target (which appeared for 500 ms). The next 12 trials showed the three options in the same viewpoint as the target (shown for 750 ms and with Perlin noise added). The final 12 trials showed the three options in a different viewpoint as the target (1,000 ms, with noise). One attention check trial included only one asymmetric greeble (the target) and the other two options were very different objects. Performance was indexed by mean accuracy across all 48 trials, with chance level being 0.33.

Transparency and Openness

This study was not preregistered. We reported how we screened participants on the online recruitment platform. We did not exclude any participants who completed all tests in this study

Figure 5

Sample Trials for the Novel Object Memory Test (Top) and the Object Matching Test (Bottom)



Note. ITI = intertrial interval. See the online article for the color version of this figure.

and provided the criteria for determining and excluding outliers. All measures for ensemble judgment, OR, and PS we administered are reported in the article. These tests are all available online at https://Figshare.com and we welcome requests for relevant materials. We conducted all correlational analyses and CFAs using JASP, Version 0.18.0 (JASP Team, 2023), R, Version 4.2.1 (R Core Team, 2023), and the lavaan package, Version 0.6-12 (Rosseel, 2012). Data and analysis code written in R are publicly available (Gauthier et al., 2023).

Results

Descriptive Statistics, Reliability, and Normality

Table 2 summarizes the descriptive statistics, normality, and reliability estimates for all tests (note that for simplicity, we now refer to all tests using only the feature used—e.g., bird or orientation—as those are unique identifiers). For the EP and OR tests, Guttman's λ_2 (Guttman, 1945) was computed to index the reliability of the

Table 2
Descriptive Statistics and Reliability for Each Test

Factor	Test	Mean (*SD*)	Min	Max	Skewness	Kurtosis	Reliability
EP_{simple}	Orientation	6.10° (2.30°)	1.8°	18°	1.74***	8.31***	$\lambda_2 = .88$
	Lightness	0.69 (0.10)	0.44	0.97	−0.13	2.87	$\lambda_2 = .75$
	Aspect ratio	0.73 (0.11)	0.42	0.90	−0.72***	2.92	$\lambda_2 = .71$
$EP_{complex}$	Transformer	57.61° (15.68°)	26.09°	101.74°	−0.39*	2.50	$\lambda_2 = .80$
	Ziggerin	0.64 (0.11)	0.32	0.94	−0.03	2.56	$\lambda_2 = .75$
	Bird	0.71 (0.10)	0.43	0.93	−0.39*	2.88	$\lambda_2 = .59$
PS	Identical pictures	50.88 (8.22)	12	70	−0.78***	4.60***	$r = .80$
	Hidden patterns	103.72 (36.56)	−3	175	−0.63***	3.04	$r = .95$
OR	Object matching	0.69 (0.13)	0.33	0.94	−0.64***	2.99	$\lambda_2 = .78$
	Novel object memory	0.51 (0.14)	0.21	0.96	0.32*	2.64	$\lambda_2 = .78$

Note. Positive value for skewness indicates positively skewed distribution; a value larger than 3 for kurtosis indicates leptokurtic distribution, a value smaller than 3 indicates platykurtic distribution, compared to normal distribution (*skewness* = 0, *kurtosis* = 3). Results from D'Agostino skewness test and Anscombe–Glynn kurtosis test are reported. Reliabilities reported here were estimated after item analyses. Min = minimum; Max = maximum; EP = ensemble perception; PS = perceptual speed; OR = object recognition.
* $p < .05$. *** $p < .001$.

measurement (Callender & Osburn, 1979; Oosterwijk et al., 2016). To assess the reliability for the two PS tests, the score was computed once for the first half of the time limit and once for the second half (i.e., first 45 s and last 45 s for identical pictures test; first 90 s and last 90 s for hidden patterns test). The Spearman–Brown formula was applied to compute corrected split-half reliability (W. Brown, 1910; Spearman, 1910).

Prior to multivariate analyses, item analyses were conducted to remove trials that did not correlate with the total test score, to improve reliability of each measure (see Table 2). This was helpful because many of the tests were new or relatively new in this study. Removing problematic trials in this manner, without regard to correlations with other tests, ensures better indicators before proceeding to latent variable modeling. The number of trials removed for lightness, aspect ratio, transformer, ziggerin, bird, and object matching was 7 (9.9%), 3 (5.5%), 4 (8.0%), 3 (4. 7%), 8 (16%), and 2 (4.2%), respectively. We excluded from analyses a total of five scores across all tests that were below chance and more than 3 *SD*s below the standardized average score for each test (only 0.21% of the data). We chose this criterion to preserve the variability of the data as much as possible since all participants had passed more than half of the attention checks and none performed below chance on more than half of the tests. This suggested that poor performance was more likely because of lower abilities rather than to inattention.

Zero-Order Correlations

Prior to computing correlations, errors in the orientation and transformer tasks were multiplied by −1 before correlations, so that all expected correlations would be positive. In Table 3, we report the zero-order correlations between each individual measure. JASP statistical software (JASP Team, 2023) was used to conduct Bayesian analyses. We report BF_{+0} in favor of the presence of a positive correlation relative to a null hypothesis (H_0) of no correlation between tests. We interpret $logBF_{+0}$ using the (Kass & Raftery, 1995) guidelines, with values of 0.5–1 considered substantial evidence for H_0, values of 1–2 considered strong evidence for H_0, and values above 2 are considered decisive evidence for H_0.

The largest correlation was found between the two PS measures. The two OR tests were also correlated as expected from previous studies (Richler et al., 2019; Smithson et al., 2024; Sunday et al., 2018). All EP tests, except for lightness, correlated with all PS and OR tests. The three low-level EP tests (orientation, lightness, and aspect ratio) were positively correlated with one another, with strong support relative to no correlation ($logBF_{+0}s > 1$). Similarly, pairs of high-level EP tests (transformer, ziggerin, and bird) were positively correlated, with strong evidence relative to no correlation ($logBF_{+0}s > 1$). All pairwise correlations between a low-level test and a high-level test (inside the black square in Table 3) were positive, with substantial to decisive support ($logBF_{+0}s > 0.5$).

To compare the strength of within-domain (high or low) correlations and between-levels correlations, we examined the average across the six within-domain correlation coefficients ($r_{within} = .27$) and the average across nine between-domain correlation coefficients ($r_{between} = .29$) (all Fisher z-transformed before averaging). There was no significant difference between the two types of correlations ($z = .32$, $p = .37$), suggesting that performance intercorrelated to a similar extent for tests that required ensemble judgment for within-domain and between-domain objects.

Gender and Age Effects

Given our diverse sample, we conducted exploratory analyses on the effects of age and gender for each test. Specifically, PS often declines with age while there is mixed evidence for gender differences (Salthouse, 1992, 1996, 2000). OR with familiar categories like cars and planes can correlate with age (McGugin et al., 2012; Sunday et al., 2019). Age does not predict OR ability measured with multiple tests and novel objects (Smithson et al., 2024). There are gender differences for OR with certain categories, with an advantage for women for natural categories (e.g., butterflies and leaves) and an advantage for men with manmade categories (e.g., cars and planes, McGugin et al., 2012). However, there is no gender difference in OR with novel objects or tests with varied categories (Sunday et al., 2022). So far, we know of no evidence of a relation between EP and age or gender. We were curious whether EP for Transformer toys would show an advantage in men, which was found in two different OR tasks with this category (Ryan & Gauthier, 2016; Sunday et al., 2022).

Table 4 summarizes the mean performance on all measures by gender, the correlation of each test performance with age, and corresponding BF_{10}. We report BF_{10} in favor of the presence of a correlation

Table 3
Pairwise Correlations of All Tests, as Indexed With Pearson's Correlation r

Factor	Test	1	2	3	4	5	6	7	8	9
EP_{simple}	1. Orientation	—								
	2. Lightness	.22[b]	—							
	3. Aspect ratio	.33[c]	.23[b]	—						
$EP_{complex}$	4. Transformer	.23[c]	.14[a]	.30[c]	—					
	5. Ziggerin	.29[c]	.28[c]	.43[c]	.37[c]	—				
	6. Bird	.29[c]	.25[c]	.32[c]	.21[b]	.24[c]	—			
PS	7. Identical pictures	.22[b]	.11	.21[b]	.18[a]	.36[c]	.17[a]	—		
	8. Hidden patterns	.33[c]	.07	.35[c]	.21[b]	.38[c]	.20[b]	.55[c]	—	
OR	9. Object matching	.34[c]	.07	.30[c]	.37[c]	.28[c]	.21[b]	.42[c]	.44[c]	—
	10. Novel object memory	.29[c]	.11	.33[c]	.34[c]	.30[c]	.20[b]	.26[c]	.30[c]	.35[c]

Note. The alternative hypothesis specifies that the two tests are positively correlated. EP = ensemble perception; PS = perceptual speed; OR = object recognition; BF = Bayes factor.
[a] $LogBF_{+0} > 0.5$ (substantial evidence). [b] $LogBF_{+0} > 1$ (strong evidence). [c] $LogBF_{+0} > 2$ (decisive evidence).

relative to a H_0 of no correlation between tests. Following Jeffreys (1939), we considered a BF_{10} between 1 and 3 as inconclusive evidence, a BF_{10} between 3 and 10 as moderate evidence, and a BF_{10} greater than 10, strong evidence. Finally, a BF_{10} below .33 was considered evidence in favor of the null, which indicated no correlation. Due to small sample size for individuals who self-identified as gender variant, we ran Bayesian independent samples *t* tests comparing data from men and women. mean estimation–transformer was the only case with strong evidence for a gender difference, unexpectedly with women outperforming men.[1]

Results from correlational analyses showed no evidence of an effect of age on EP tests. In line with prior work, we found strong evidence for negative correlations for both PS measures. Both OR measures also correlated negatively with age, with moderate Bayesian support.

Uncovering the Factor(s) Supporting Ensemble Judgment Performance

We found support for a general ensemble ability from zero-order correlations and regression analysis. We then went on to use CFA to more accurately describe these relationships at the latent level. We compared two models: one with separate latent variables for $EP_{complex}$ and for EP_{simple}, and another with a single latent variable EP supporting ensemble judgments for all tasks.

In Model 1 (a model of separate abilities), orientation, lightness, and aspect ratio loaded onto a simple ensemble factor EP_{simple}, whereas transformer, ziggerin, and bird loaded onto a complex ensemble factor $EP_{complex}$. The two latent variables for simple and complex objects were then related. In Model 2 (a model of single ability), each of the six ensemble judgment tests served as an indicator loading onto the same latent variable EP. For both models, we also included correlated error terms for different ensemble judgment tasks, to control for effects because of shared task specifics (i.e., between the two mean estimation tests, between the two diversity comparison tests, and between the two mean matching tests). We denoted these models Model 1.ce and Model 2.ce. Therefore, the only difference between Models 1 and 1.ce and between Models 2 and 2.ce was the added correlated errors.

To conduct CFAs, we used the R library lavaan (Rosseel, 2012). Since the performance index scores for all tests were continuous, we used the maximum likelihood method (specified as the "MLR" estimator in lavaan) for model parameter estimation. MLR estimator also provides estimates robust to missing data and nonnormality of observed variables. We used several indices to evaluate the goodness of model fit, including the chi-square test statistic (and its *p* value), the root-mean-square error of approximation (RMSEA; Steiger, 2016; Steiger & Lind, 1980), and the comparative fit index (CFI, Bentler, 1990) that lavaan generates by default. A nonsignificant chi-squared test is preferred as it indicates that the actual data match the proposed model. We used .01, .05, and .08 to suggest excellent, good, and mediocre fit with RMSEA, respectively (MacCallum et al., 1996), and a CFI larger than .90 (Bentler, 1990) to infer good fit.

To examine how well each test measured the relevant latent variables, we examined factor loadings, or path coefficients (ranging from 0 to 1), for each relationship between a latent variable and an indicator. All factor loadings were estimated freely in our CFA analyses. The higher the factor loading, the more of that indicator's variance is explained by that factor. Significance tests on the factor loadings should ideally be significant, indicating the effective measurement of the latent variable. For model comparison, we consulted Akaike Information Criterion (Akaike, 1973) and Bayesian Information Criterion (Raftery, 1995; Schwarz, 1978), two indices that take into consideration the degree of model parsimony and complexity. A better model is indicated by a smaller Akaike information criteria or Bayesian information criteria. Scaled chi-square difference tests were conducted using a likelihood ratio test in *lavaan* to compare nested models (Satorra & Bentler, 2001). A nonsignificant result suggests the parsimonious model.

Table 5 summarizes the goodness of fit indices for all models tested in this study. The models testing for separate abilities (see Model 1 in Figure 6 for model structure), showed evidence of good fit, both with (Model 1.ce, CFI = .981 and RMSEA = .055) or without correlated errors (Model 1, CFI = .994 and RMSEA = .02), suggesting the possibility of a factor for EP_{simple} tests and another factor for $EP_{complex}$ tests. Based on a nonsignificant chi-squared difference test, $\chi^2_{difference}(3) = 0.69$, $p = .88$, between Model 1 and Model 1.ce, and

[1] There was no gender difference on ME-Transformer in Chang et al. (2024) ($BF_{10} = .21$) and in Sunday et al. (2021) ($BF_{10} = .18$). Since the stimuli used in ME-Transformer are derived from a small number of three exemplars, this could restrict the domain coverage for this category and thus less likely to tap expertise. However, it is not clear why the reverse relationship was observed here.

Table 4
Mean Performance by Gender and Correlations With Age

Test	Gender analysis: Men $N = 70/71$	Gender analysis: Women $N = 160/161$	Gender analysis: Gender-variant $N = 5$	Age analysis: Pearson's correlation r with age
Orientation	5.81° (1.70°)	6.24° (2.53°)	5.472° (1.72°)	−.002
Lightness	0.68 (0.10)	0.70 (0.10)	0.63 (0.07)	.08
Aspect ratio	0.73 (0.11)	0.73 (0.11)	0.77 (0.06)	−.01
Transformer	62.04° (16.84°)	55.36° (14.80°)[b]	67.04° (13.67°)	−.12
Ziggerin	0.63 (0.11)	0.65 (0.12)	0.62 (0.09)	−.16
Bird	0.69 (0.10)	0.73 (0.10)	0.69 (0.02)	.08
Identical pictures	51.16 (8.04)	50.75 (8.44)	51.40 (4.39)	−.40[b]
Hidden patterns	110.25 (36.08)	100.42 (37.06)	117.20 (7.98)	−.35[b]
Object matching	0.70 (0.13)	0.69 (0.13)	0.64 (0.15)	−.21[a]
Novel object memory	0.50 (0.14)	0.51 (0.15)	0.48 (0.15)	−.19[a]

Note. The alternative hypothesis for the Bayesian independent t tests specifies that the two samples (men and women) are not the same in their performance on each test. We marked the significance for BF_{10} from Bayesian independent samples t test in the Women column. The alternative hypothesis for the age effects is that there is no positive correlation between test performance and age. Sample sizes for men and women changed across tests because five data points were excluded because of poor performance. BF = Bayes factor.
[a] $3 < BF_{10} < 10$ (moderate evidence). [b] $BF_{10} > 10$ (strong evidence).

nonsignificant correlated error estimates, we focus on the more parsimonious Model 1, without correlated errors. All factor loadings were significant. We fixed the factor variances in Model 1 to examine the correlation between the two factors. The correlation between EP_{simple} and $EP_{complex}$ was significant and equaled 1.05, an estimation of a perfect correlation with some error. This suggests that the two factors were essentially the same and should be combined.

In Model 2 and Model 2.ce, we specified one factor for all EP tests (see Figure 6). Fit indices also showed decent fit whether the correlated errors were estimated (CFI = .985 and RMSEA = .04) or not (CFI = .987 and RMSEA = .04), with all factor loadings significant, suggesting a single factor EP. Again, we selected Model 2 because none of the estimated correlated errors were significant in Model 2.ce. Chi-square difference test also supported the use of Model 2, $\chi^2_{difference}(3) = 0.55$, $p = .91$. Adding correlated errors did not improve model fit but would increase parameter estimation. A nonsignificant chi-squared difference test, $\chi^2_{difference}(1) = 0.29$, $p = .59$, on nested models Model 1 and Model 2 further suggested that using one factor rather than two factors did not impair model fit. These findings collectively showed that EP_{simple} and $EP_{complex}$ should be combined as one-factor EP.

The Role of PS and OR in the Relationship Between Low-Level EP and High-Level EP

We tested Model 3 in which EP_{simple}, PS, and OR predicted $EP_{complex}$ (Model 3, Table 5). This model allowed us to examine how much of the variance in $EP_{complex}$ could be uniquely accounted for by each of these predictor factors after controlling for their intercorrelations. Model 3 showed good fit (CFI = .952 and RMSEA = .05, see Figure 7). The correlations between predictor factors were all significant. All factor loadings were significant as

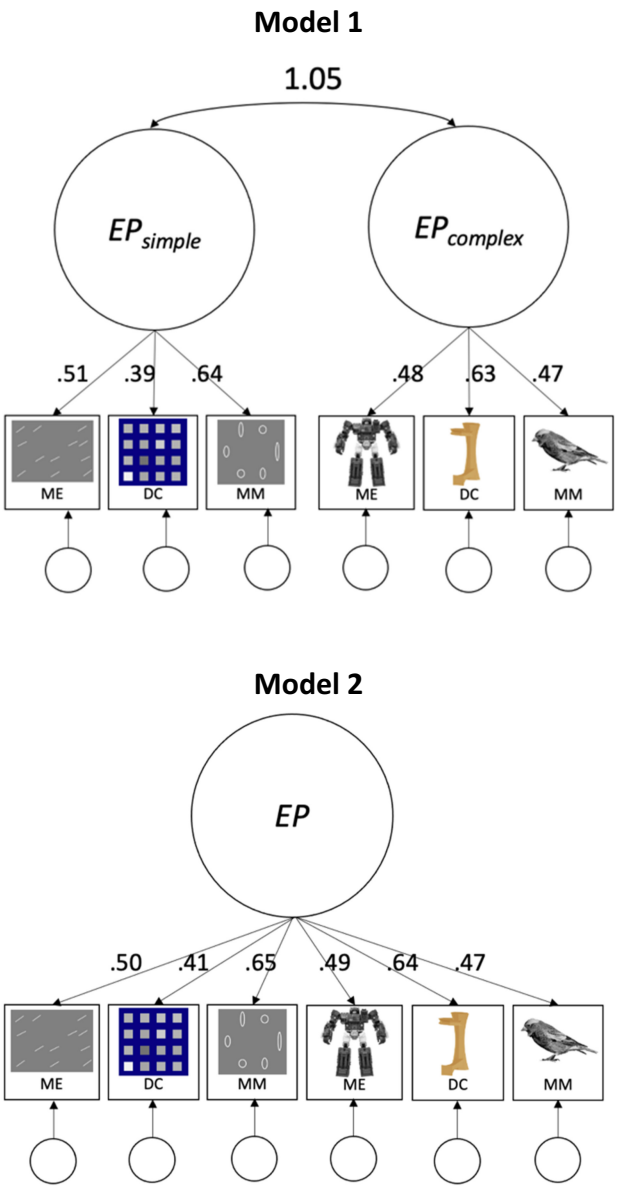
Table 5
Model Fit Indices for Each Tested Model

Model	Description	*df*	Robust χ^2	*p*	Robust CFI	Robust RMSEA 90% CI	SRMR	Robust AIC	Robust BIC
1	Freely correlated model with factors EP_{simple}, $EP_{complex}$	8	9.17	.33	.994	.02 [0.00, 0.08]	.027	−1,861.51	−1,795.62
2	One EP factor model	9	8.56	.2	.987	.04 [0.00, 0.09]	.027	−2,050.51	−1,977.68
1.ce	Freely correlated model with factors EP_{simple}, $EP_{complex}$ + correlated errors	5	8.81	.12	.981	.06 [0.00, 0.11]	.026	−1,856.06	−1,779.76
2.ce	One EP factor model + correlated errors	6	8.82	.18	.985	.04 [0.00, 0.01]	.027	−1,857.89	−1,785.06
3	Structural model predicting $EP_{complex}$ with factors EP_{simple}, PS, and OR	29	50.55	<.01	.952	.05 [0.03, 0.08]	.043	1,400.93	1,525.78
4	Structural model predicting $EP_{complex}$ with factors EP_{simple}, PS, OR, and WM	49	51.67	.3	.982	.03 [0.00, 0.07]	.056	3,758.2	3,876.97
5	Freely correlated three-factor model testing task effect	6	8.82	.18	.985	.04 [0.00, 0.10]	.027	−1,857.89	−1,785.06

Note. CFI = comparative fit index (good fit indicated as >.90); RMSEA = root mean square error of approximation (the cutoffs for excellent, good, and mediocre fit are .01, .05, and .08); SRMR = standardized root mean square residual; AIC = Akaike information criteria (better fit indicated by smaller AIC); BIC = Bayesian information criteria (better fit indicated by smaller BIC); EP = ensemble perception; PS = perceptual speed; OR = object recognition; WM = working memory.

Figure 6

The Structural Equation Model for Model 1 and Model 2 (Standardized Solution)



Note. All tests are indicated by rectangles are observed variables. EPs stand for the latent variables that cause the observed correlations among relevant tests and are indicated by large circles. Measurement errors associated with the tests are shown in small circles. In Model 1, the tests assessing performance with simple and complex objects load onto two distinct latent variables EP_{simple} and $EP_{complex}$, whereas in Model 2, all tests load onto one single latent variable EP. All factor loadings were significant ($p < .001$). EP = ensemble perception; ME = mean estimation task; DC = diversity comparison task; MM = mean matching task. See the online article for the color version of this figure.

well. We found a strong unique relationship of EP_{simple} to $EP_{complex}$, with EP_{simple} explaining 81% of the variance in $EP_{complex}$ (indicated by the standardized factor loading 0.90 squared). However, neither PS nor OR significantly predicted unique variance in $EP_{complex}$.

Interim Discussion

From the analyses so far, we learned that the performance on all six EP tests can be accounted for by a single EP factor, whether the decision is based on low-level or high-level feature. There is strong evidence for a positive correlation between low- and high-level EP after controlling for age-adjusted OR and PS. CFA further shows a unique relationship between low-level EP and high-level EP when general visual abilities such as OR and PS are controlled.

We consistently found very high correlations between all (except for diversity comparison task—lightness) measures but could not explain entirely the relationship between low-level and high-level EP using OR and PS. To understand more about the variance common to low- and high-level EP, we decided to measure WM as another general ability that could account for shared variance. WM represents the ability to maintain and retrieve information and to control attention under conditions of interference (Engle, 2002)—it encompasses the constructs of short-term memory and executive attention (Cowan, 2008; Engle et al., 1999). In prior work on domain-general OR, we controlled for general cognitive skills by estimating fluid intelligence (Chow et al., 2021; Richler et al., 2019). Here we use an alternative which is to combine PS (a visual component of processing speed) with WM, which each has found some support as an explanation for general intelligence (Mashburn et al., 2024). PS and WM were only modestly correlated but each contributed unique variance to general intelligence (Ackerman et al., 2002; Redick et al., 2013; Salthouse & Babcock, 1991). Even when measured in nonvisual tasks, WM can account for individual differences in other complex visual tasks (e.g., categorical learning, Lewandowsky, 2011). We expected to explain more of the variance in EP by considering the contribution from WM. We therefore invited the initial 237 participants back to complete a binding task and an operation span task to measure WM.

Additional Study and Results

Participants

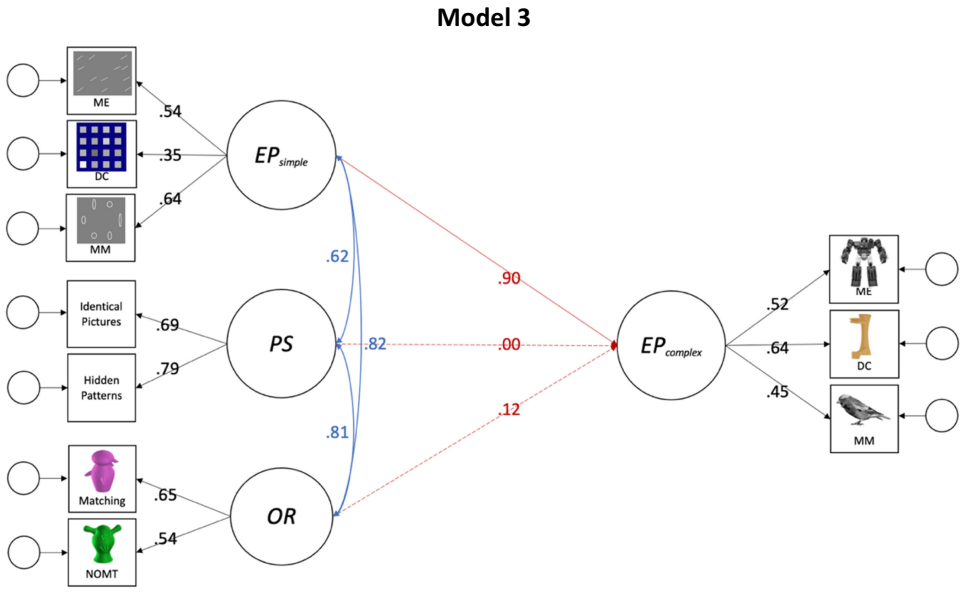
A total of 117 from the initial 237 participants volunteered upon invitation to complete the additional two WM tests, 39 self-reported as men, 78 women; $M_{age} = 37.99$; $SD = 9.47$; range = (18, 55). Participants completed the two WM tests in a fixed order, first the binding test then the operation span test.

Two WM Tests

We used two tests modeled by Wilhelm et al. (2013) to measure WM (see Figure 8, top panel). The binding test measures how well participants remember random word–number pairs. At the beginning of each trial, there was a learning phase in which participants were presented with a sequence of two to six word–number pairs (e.g., actor—16), each lasting on screen for 3 s. Afterwards, either a word or a number showed up on screen and participants had to match the corresponding number or word that was paired with it in the sequence by mouse click. There was a total of 17 trials, progressing from the easiest to the most demanding trials. The number of trials at load levels of two to six was 2, 4, 5, 4, and 2, respectively. While only one pair in the sequence was tested for recall in the first 15 trials, participants were tested on each of the six pairs in the sequence in the last two trials. The response was not speeded and no

Figure 7

Graphical Representation of Model 3 (Standardized Solution)



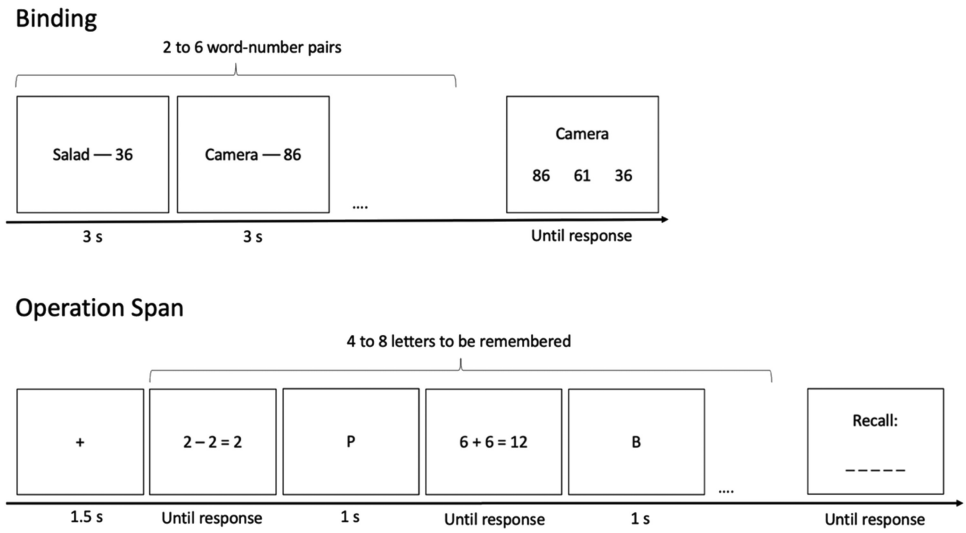
Note. This model tests the unique relationships of EP_{simple}, PS, and OR, to $EP_{complex}$. Blue (curved) paths are factor correlations. Red paths (those pointing to the $EP_{complex}$ factor) show significant (solid line) and nonsignificant (dotted line) predictors. All factor correlations and factor loadings were significant ($p < .001$). EP = ensemble perception; PS = perceptual speed; OR = object recognition; ME = mean estimation task; DC = diversity comparison task; MM = mean matching task; NOMT = novel object memory test. See the online article for the color version of this figure.

feedback on accuracy was provided. The words were all nonabstract nouns, for example, drum, game, salad, etc., and the numbers were all two digits. To encourage temporary binding in WM, a word/number sometimes appeared multiple times and was paired with a different number/word across trials. Performance was indexed by the mean accuracy.

The operation span test (Figure 8, bottom panel) measures the ability to control and execute attention when there is interference or distraction on a task. On each trial, participants had to remember a series of letters, presented one by one, while being distracted by solving task-irrelevant math problems in between. Each trial began with a 1.5 s fixation cross at the center of the screen, followed by

Figure 8

Schematic Representations of the Binding Task (Top) and the Operation Span Test (Bottom)



a simple arithmetic equation (e.g., 8 + 9 = 19). Participants first evaluated whether the equation was correct and then memorized a following number that was presented for 1 s. After a series of four to eight such operations, they recalled all of the letters in the order in which they were presented. A letter would only count as correct if it was recalled correctly in the correct position in the sequence. For example, "PBXVC" was scored as 60% accurate for a letter sequence "PBXCV." Incorrect responses to the equation would not influence the scoring. There were three practice trials with one for each three-, four-, and five-letter loading, respectively, where feedback on accuracy was given. Afterward, participants completed 15 trials without feedback, sorted from four-letter to eight-letter loading, three for each load level. Performance was indexed by the mean proportion of letters recalled in their correct position across all trials.

Summary Results for All Tasks, for Participants Who Did the WM Tests

Table 6 presents the descriptive statistics, normality, and reliability for all the 12 tests performed by the 117 participants. Both WM tests showed a good range of performance and acceptable reliabilities. The distributions and reliabilities on the tests excluding WM tests were comparable to the initial data (Table 2). To improve the reliability, we removed three trials from the binding test because they correlated negatively to the rest of the test. No trial was discarded from the operation span test. The same trials were excluded from the other tests as reported earlier. We applied the same data exclusion criterion and excluded one score from the analyses across all 12 tests that were below chance and 3 *SD*s lower than the standardized average score on the test.

Gender Differences and Age Effects for the Additional Study

Since we found strong evidence for an age effect on the PS and OR measures, we also examined the correlations between our measures and age (Table 7). There was strong evidence for a negative correlation between age and the two PS measures. For both OR and WM measures, there was only inconclusive evidence for an age effect. There is mixed evidence for age effects on OR. The absence of an age effect on WM measures may be because the age range of our sample (18–55 years) was not sensitive enough to detect the effect. Previous reports on age-related WM declines were found in older samples up to 99 years old (Oberauer, 2005; Salthouse, 1992; Salthouse & Babcock, 1991). No gender difference was found for any test.

Zero-Order Correlations Between Abilities in the Additional Study

There is decisive evidence that the binding test and operation span test are positively correlated ($r = .28$, $\log BF_{+0} = 3.05$). Based on this positive correlation and the positive correlations between tests that measured the same general ability in our large sample of 237 participants (Table 3), we used aggregate scores to indicate each ability and examined their zero-order correlations (Table 8). Results showed strong and decisive evidence of a positive correlation for every pair of abilities ($\log BF_{+0} > 1$), supporting the importance to control these general visual abilities in the investigation of the relationship between low- and high-level EP.

Commonality Between Low-Level and High-Level EP After Controlling General Abilities

We tested Model 4 in which EP_{simple}, PS, OR, and WM uniquely predicted $EP_{complex}$ (Figure 9). Model 4 showed good fit (CFI = .982 and RMSEA = .03, Table 5). There were strong correlations between predictor factors. However, EP_{simple} still uniquely predicted $EP_{complex}$ above and beyond the other general abilities and accounted for approximately 69% of the variance in $EP_{complex}$ (indicated by the standardized factor loading 0.83 squared). Although the supplemental study only had 117 participants, results from both the multiple regression and CFA suggested a unique contribution from EP_{simple} to $EP_{complex}$ over and above what was accounted for by PS, OR, and WM.

Discussion

Haberman, Brady, and Alvarez (2015) used individual differences to uncover the structure of mechanisms underlying EP. They found

Table 6
Summary for the 12 Tests From the Subset of 117 Participants

Factor	Test	Mean (*SD*)	Min	Max	Skewness	Kurtosis	Reliability
EP_{simple}	Orientation	6.13° (2.20°)	1.80°	14.94°	−1.37***	5.88***	$\lambda_2 = .85$
	Lightness	0.68 (0.12)	0.33	0.97	−0.41	3.39	$\lambda_2 = .79$
	Aspect ratio	0.74 (0.12)	0.42	0.90	−0.74**	2.76	$\lambda_2 = .76$
$EP_{complex}$	Transformer	56.82° (16.56°)	26.09°	97.80°	−0.42*	2.41	$\lambda_2 = .83$
	Ziggerin	0.64 (0.12)	0.37	0.94	−0.10	2.29	$\lambda_2 = .76$
	Bird	0.71 (0.11)	0.45	0.93	−0.20	2.54	$\lambda_2 = .64$
PS	Identical pictures	50.88 (8.22)	12	70	−0.89***	4.73**	$r = .82$
	Hidden patterns	101.08 (37.98)	−3	175	−0.56*	3.00	$r = .95$
OR	Object matching	0.70 (0.13)	0.33	0.94	−0.57*	2.97	$\lambda_2 = .77$
	Novel object memory	0.50 (0.13)	0.21	0.85	0.18	2.35	$\lambda_2 = .75$
WM	Binding	0.62 (0.16)	0.21	1	0.19	2.73	$\lambda_2 = .71$
	Operation span	0.73 (0.15)	0.31	1	−0.43*	3.02	$\lambda_2 = .86$

Note. Positive value for skewness indicates positively skewed distribution; a value larger than 3 for kurtosis indicates leptokurtic distribution, a value smaller than 3 indicates platykurtic distribution, compared to normal distribution (*skewness* = 0, *kurtosis* = 3). Results from D'Agostino skewness test and Anscombe–Glynn kurtosis test are reported. Reliabilities reported here were estimated after item exclusions. Min = minimum; Max = maximum; EP = ensemble perception; PS = perceptual speed; OR = object recognition; WM = working memory.
* $p < .05$. ** $p < .01$. *** $p < .001$.

Table 7
Mean Performance by Gender and Correlations With Age From the Subset of 117 Participants

Test	Gender analysis: Men $N = 39$	Gender analysis: Women $N = 78$	Age analysis: Pearson's correlation r with age
Orientation	5.87° (1.63°)	6.25° (2.43°)	.03
Lightness	0.66 (0.12)	0.70 (0.11)	.05
Aspect ratio	0.73 (0.11)	0.74 (0.12)	−.10
Transformer	61.98° (17.16°)	54° (15.78°)	−.15
Ziggerin	0.63 (0.12)	0.64 (0.12)	−.18
Bird	0.69 (0.10)	0.72 (0.11)	.09
Identical pictures	49.85 (8.39)	50.06 (9.29)	−.42***
Hidden patterns	104.51 (36.46)	99.36 (38.84)	−.44***
Matching	0.69 (0.13)	0.70 (0.13)	−.24
Novel object memory	0.48 (0.12)	0.52 (0.14)	−.16
Binding	0.62 (0.16)	0.61 (0.17)	.06
Operation span	0.72 (0.14)	0.74 (0.15)	−.23

Note. The alternative hypothesis for the Bayesian independent t tests specifies that the two groups (men and women) are not the same in their performance on each test. No test was significant as suggested by BF_{10}. The alternative hypothesis for the age effects is that there is no correlation between test performance and age. BF = Bayes factor.
*** $BF_{10} > 10$ (strong evidence).

that performance correlated strongly between tasks with both low-level features (color and orientation) and with both high-level features (facial expression and identity). By contrast, little to no correlation was found between one task that used faces and tasks that used low-level features. These findings suggested a dissociation between EP for low-level and high-level feature domains. Here, we tested this proposed dissociation at the latent level. We adopted a latent variable approach to remedy possible measurement errors and shared task effects that might confound interpretation from pairwise correlations, and define abilities by controlling for construct-irrelevant variance. To improve the generality of latent factors, we used six EP tests paired with six distinct features from three simple and three nonface complex objects. CFA suggested a single factor supporting the performance on all six EP tests. We also found unique variance between low-level and high-level EP that was not entirely explained by PS, OR, and WM.

We aimed to characterize the relationship between EP for low-level and high-level features. Our results offer a completely different picture of the structure of these abilities, in contrast to previous work that suggested distinct abilities. Several design choices may have influenced our results. First, while Haberman, Brady, and Alvarez (2015) used only faces as a complex object, we used nonface objects to measure high-level EP. Although the relationship between face EP and face recognition is still unclear, there is extensive evidence that EP and OR are related and that individual recognition of faces is minimally correlated with the individual recognition of other objects (Ćepulić et al., 2018; McGugin et al., 2012). Therefore, we chose to stay away from using faces to represent all complex objects. There is an increasing number of reports that people can make ensemble judgments on various complex objects, including familiar categories like cars (Cha et al., 2021; Chang & Gauthier, 2022; Yamanashi Leib et al., 2016, 2020) and novel categories like greebles (Sunday et al., 2022). The three complex objects we used here (transformer, ziggerins, and birds) have all been tested in EP tasks and yielded measurements with good reliability in prior work (Chang & Gauthier, 2022; Sunday et al., 2022). They are visually distinct both within domain and across domains, and they collectively span different dimensions such as familiarity and animateness. This allowed us to approximate a less biased high-level EP ability that may better generalize to other complex objects (Little et al., 1999; Whitely, 1983).

Table 8
Pairwise Correlations Between Factors, as Indexed With Pearson's Correlation r

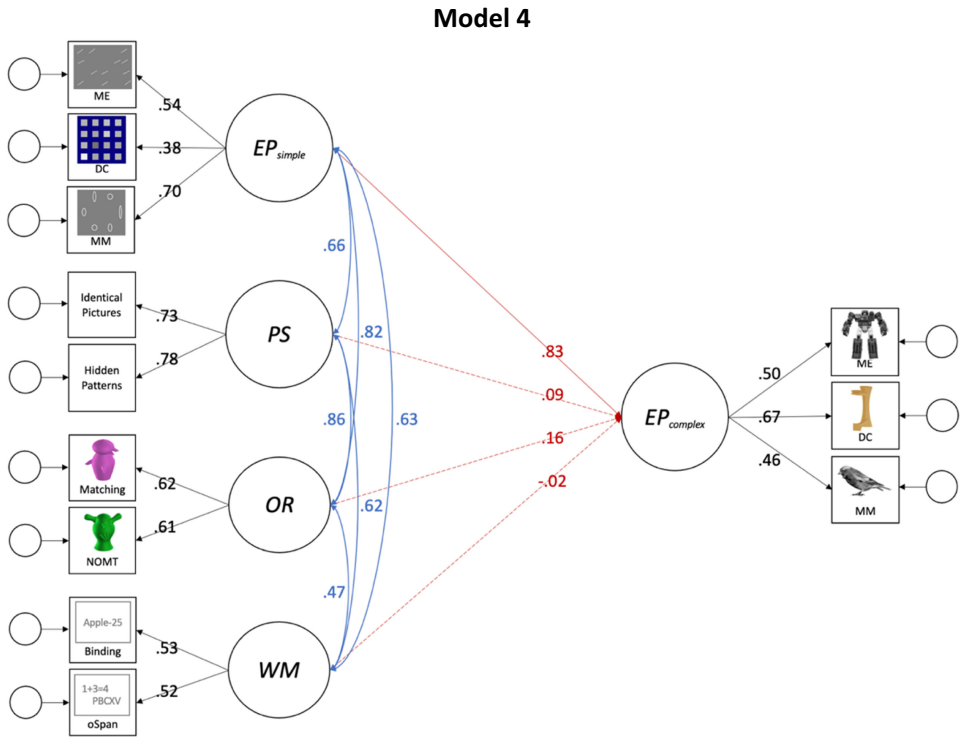
Factor	1	2	3	4	5	6
1. EP	—					
2. EP_{simple}	.89[b]	—				
3. $EP_{complex}$	.87[b]	.58[b]	—			
4. PS	.51[b]	.41[b]	.49[b]	—		
5. OR	.50[b]	.40[b]	.48[b]	.53[b]	—	
6. WM	.36[b]	.34[b]	.30[b]	.35[b]	.23[a]	—

Note. The alternative hypothesis specifies that the two tests are positively correlated. EP = ensemble perception; PS = perceptual speed; OR = object recognition; WM = working memory; BF = Bayes factor.
[a] $LogBF_{+0} > 1$ (strong evidence). [b] $LogBF_{+0} > 2$ (decisive evidence).

Second, the hypothesis that low-level EP and high-level EP are two distinct abilities was originally supported by positive correlations between pairs of tasks at the same level, and by a lack of correlation between pairs of tasks at different levels of complexity. This is the first time that this hypothesis is tested in a latent factor model. We chose indicators for each EP factor that were highly heterogeneous. The set of three indicators did not overlap in either the feature, the object that carried the feature, or the specific task demands. This was a choice made to test the hypothesis that there is a general EP ability for each of the low- and high-level features. By using three heterogeneous indicators, we defined two EP factors that were already general in nature. Imagine if we had measured low-level EP with three MEs for judging the orientation of lines, Gabors, and triangles. These three indicators would be more highly intercorrelated than our present set, but the resulting low-level EP factor would likely be less general and correlate less strongly with a high-level EP factor, especially if this factor was also less general. This

Figure 9

Graphical Representation of Model 4 (Standardized Solution)



Note. This model tests the unique relationships of EP_{simple}, PS, OR, and WM, to $EP_{complex}$. Blue (curved) paths are factor correlations. Red paths (those pointing to the $EP_{complex}$ factor) show significant (solid line) and nonsignificant (dotted line) predictors. All factor correlations and factor loadings were significant ($p < .001$). EP = ensemble perception; PS = perceptual speed; OR = object recognition; WM = working memory; ME = mean estimation task; DC = diversity comparison task; MM = mean matching task; NOMT = novel object memory test. See the online article for the color version of this figure.

reminds us that the choice of indicators for constructs is not simply an optimization process (i.e., the set of most highly intercorrelated indicators is not necessarily the best, Little et al., 1999) and reflects a theoretical decision (Gustafsson, 2002).

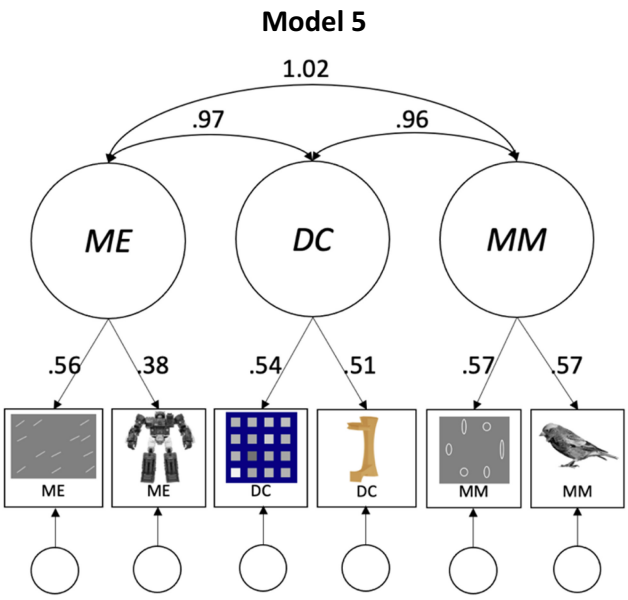
Third, the overlap between the indicators contributing to the low- and high-level EP factors was by necessity in task demands and not in the features. In a situation without time or money constraints, we could have paired all three low-level (or high-level) features with all three tasks, resulting in nine indicators for each factor. This "multitrait–multimethod" approach (Campbell & Fiske, 1959) would allow us to evaluate the strength of correlations when the same feature is estimated in different tasks relative to the strength of correlations when different features are estimated in the same task. When personality traits are measured using different methods (e.g., self-report vs. observation), we may not be interested in shared method variance and hope to see higher correlations driven by traits than methods. It is less clear what would be theoretically expected for EP abilities as there is debate both regarding common mechanisms for estimating different features (Haberman, Brady, & Alvarez, 2015; Sama et al., 2019; Yörük & Boduroglu, 2020) and different summary statistics (Cha et al., 2022; Khvostov & Utochkin, 2019; Utochkin & Vostrikov, 2017; Yang et al., 2018).

In the interpretation of our results, one could ask whether method variance is so important that it inflates the correlation between factors EP_{simple} and $EP_{complex}$. The zero-order correlations from our large sample showed no stronger correlation between a low-level and a high-level test using the same task (e.g., both mean estimations) than with different task (Fisher z-transformed $r_{same} = .28$ and $r_{different} = .29$, $z = -0.07$, $p = .47$). This suggests that performance intercorrelated to a similar extent for tests regardless of the type of task format. Furthermore, we examined how the three EP tasks related in a correlated three-factor model where the six EP tests each loaded onto their corresponding method factors (e.g., mean estimation—orientation and mean estimation transformer loaded onto a method factor mean estimation, see Figure 10, Model 5 in Table 5). We found significant factor loadings from each test for each method. Critically, there were very high correlations between these method factors ($rs > .96$, $ps < .001$). Thus, these results suggest that there is no clear hierarchical structure in the abilities measured by this set of six indicators driven by either feature complexity or task.

In summary, our results suggest that the ability common to a variety of ensemble judgments with low-level features is the same ability common to a variety of ensemble judgments with high-level features. This does not mean that this EP ability explains all the variance in performance on any type of ensemble judgment. For instance, we can estimate from the factor loadings in Model 2 (Figure 6) that the amount of the variance of our EP tests accounted for by the general factor EP ranged from 17% to 40% (indicated by the standardized

Figure 10

A Structural Equation Model Testing the Three Method Factors (Standardized Solution; CFI = .92; RMSEA = .08; Model 5 in Table 5)



Note. All factor correlations and factor loadings were significant ($p < .001$). Note that Model 5 would be equivalent to Model 2 if the correlations between method factors were constrained to be 1. CFI = comparative fit index; RMSEA = root-mean-square error of approximation; ME = mean estimation task; DC = diversity comparison task; MM = mean matching task. See the online article for the color version of this figure.

factor loadings squared). The remainder of the variance could be explained at the level of features or tasks as discussed above or by other dimensions not considered here.

What type of mechanism could explain this domain-general EP ability? Most authors introduce EP as a phenomenon that has been studied for a broad range of features, both low-level and more complex like the ones we have discussed here, up to some abstract dimensions like the economic value of an object (Yamanashi Leib et al., 2020). Authors who go on to propose and test a specific model of EP usually focus on a specific feature in their implementation, without explicitly returning to the question of how it may generalize. For instance, Baek and Chong (2020) introduce a distributed attention model for perceptual averaging of the size of circles, which posits that the precision of EP is influenced by both early and late noise in the visual system, modulated by the distribution of attention across the elements being averaged. It is easy to imagine the same model applied to other linear feature dimensions but perhaps not as clearly to stimulus sets that vary on multiple dimensions at once.

In this model, the concept of late noise is more likely than early noise to be a source of domain-general effects, as it is driven by the limits of memory, decision-making, and attentional processes. Interestingly, the influence of late noise should increase as the set size (and cognitive load) increases. Some authors (Corbett et al., 2023) have questioned our prior claims of domain-general influences in EP (Chang & Gauthier, 2022) on the basis that our set sizes may not have been large enough to tap "true" EP. However, at least according to Baek and Chong (2020), larger set sizes should increase the influence of late noise, and thus would only increase domain-general effects. Future work should investigate how the relative contribution of domain-general versus domain-specific effects varies as a function of set size.

It is also important to keep in mind that our tasks can be performed using a variety of strategies, although the participants who perform best should be those who integrate successfully across many items. This differs from tasks designed to support the very existence of EP as a process and used to demonstrate that EP can operate independently from the encoding of individual items. In a recent study (Hadar et al., 2022), people were more accurate at averaging numerical values or facial emotions after the induction of abstract thinking, compared to concrete thinking. Trait propensity to reason abstractly (e.g., Vallacher & Wegner, 1989) could be related to the domain-general component of EP.

We readily recognize several caveats of the present study. There could be a potential confound in the assessment of high-level features only on complex objects and low-level features only on simple objects. Assessing high-level features for simple objects might be tricky, but people are able to report both low-level and high-level features for complex objects. For instance, Yang et al. (2018) examined whether ensemble judgments focused on mean and variance are dissociable using strawberries and lollipops for size and orientation, respectively. However, given the very strong correlation between factors EP_{simple} and $EP_{complex}$, it seems unlikely that a stronger relationship could be found if the same objects were used for judgments at different levels.

Following Haberman, Brady, and Alvarez (2015), we began by investigating the structure of ensemble abilities by dividing visual features into low-level and high-level domains. However, there are also visual features considered to be midlevel about which people can make ensemble judgments, including size (Ariely, 2001; Chong & Treisman, 2003) and symmetry (Jerskey, 2020). Aspect ratio used as a low-level feature in our study can also be viewed as a midlevel feature because it remains unclear whether aspect ratio detectors (or size detectors, symmetry detectors) exist in early visual processing (Myczek & Simons, 2008; Whitney & Yamanashi Leib, 2018). Future studies may more systematically compare low-, mid-, and high-level ensemble judgments.

In all tested models, the six EP tests always loaded significantly onto the relevant latent factor. However, the lightness consistently loaded less on its corresponding factor compared with the other tests. One possibility is that, different from the other tests, the array of squares can be viewed as a grid-patterned object as opposed to an ensemble of 16 squares. Participants could then engage in texture perception (Balas et al., 2009; Morgan et al., 2001), and some individuals may do so more than others. When designing this test, we adjusted the squares to be large enough and sparse enough to try to avoid texture association. However, future studies could address this concern by including several tests that vary in their possible reliance on texture perception.

Another possible strategy that could be used in the diversity comparison task would be to focus on variance rather than diversity. Consider this hypothetical example: the variance of this set [1 1 1 6 6 6] is larger than the variance of this set [1 2 3 4 5 6], but the diversity of the first set is lower than that of the second set. In the diversity comparison—lightness task, diversity was defined as the number of distinct levels of lightness in a particular ensemble. Even though we provided examples to demonstrate what we meant by more and less diverse, different participants could have used different strategies.

Our tasks included only judgments for spatial ensembles with multiple objects presented on screen at the same time. Ensemble

judgments can be made on items presented at the same time and items presented sequentially (Albrecht & Scholl, 2010; Chong & Treisman, 2003; Haberman et al., 2009; Whiting & Oriet, 2011). There is conflicting evidence regarding whether averaging over space and time relies on the same mechanisms. A model-based psychophysical study showed that the thresholds for discriminating the mean size of eight circles were about the same regardless the circles were presented simultaneously or sequentially (Gorea et al., 2014). Another study with faces reported that participants sampled fewer faces to judge average gaze direction when the faces were presented simultaneously than sequentially, but that the same number of faces was sampled to judge head direction in the two presentation conditions (Florey et al., 2017). Sequentially presented items are weighted differently depending on factors that do not influence ensemble estimation across space, such as their positions in a sequence which may result in primacy and recency biases on the estimated ensemble statistics (Hubert-Wallander & Boynton, 2015; Yashiro et al., 2020). Therefore, averaging over time could also rely on different abilities from averaging across space, because of different constraints associated with each (e.g., divided attention in spatial ensembles, and visual WM capacity in temporal ensembles). The literature on EP covers such a large spectrum of tasks and features that it will likely require a large number of multivariate studies to adequately sample the space of the general ability we have proposed here.

Constraints on Generality

Our participants ranged from the age of 18 to 55, which may restrict our findings to cover only young to middle-aged adulthood. Our EP tasks measured judgments of mean and variance and may not apply to other properties (e.g., numerosity estimation). Our EP tasks were all explicit judgments and used arrays of simultaneously presented items, so the results may not apply to EP effects in more implicit tasks or tasks where items are presented sequentially.

References

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. *Journal of Experimental Psychology: General*, *117*(3), 288–318. https://doi.org/10.1037/0096-3445.117.3.288

Ackerman, P. L. (1990). A correlational analysis of skill specificity: Learning, abilities, and individual differences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(5), 883–901. https://doi.org/10.1037/0278-7393.16.5.883

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2002). Individual differences in working memory within a nomological network of cognitive and perceptual speed abilities. *Journal of Experimental Psychology: General*, *131*(4), 567–589. https://doi.org/10.1037/0096-3445.131.4.567

Ackerman, P. L., & Cianciolo, A. T. (2000). Cognitive, perceptual-speed, and psychomotor determinants of individual differences during skill acquisition. *Journal of Experimental Psychology: Applied*, *6*(4), 259–290. https://doi.org/10.1037/1076-898X.6.4.259

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In E. Parzen, K. Tanabe, & G. Kitagawa (Eds.), *Breakthroughs in statistics* (pp. 610–624). Springer New York. https://doi.org/10.1007/978-1-4612-0919-5_38

Albers, D., Correll, M., Gleicher, M., & Franconeri, S. (2014). Ensemble processing of color and shape: Beyond mean judgments. *Journal of Vision*, *14*(10), Article 1056. https://doi.org/10.1167/14.10.1056

Albrecht, A. R., & Scholl, B. J. (2010). Perceptually averaging in a continuous visual world: Extracting statistical summary representations over time. *Psychological Science*, *21*(4), 560–567. https://doi.org/10.1177/0956797610363543

Alvarez, G. A., & Oliva, A. (2009). Spatial ensemble statistics are efficient codes that can be represented with reduced attention. *Proceedings of the National Academy of Sciences*, *106*(18), 7345–7350. https://doi.org/10.1073/pnas.0808981106

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, *12*(2), 157–162. https://doi.org/10.1111/1467-9280.00327

Baek, J., & Chong, S. C. (2020). Distributed attention model of perceptual averaging. *Attention, Perception, & Psychophysics*, *82*(1), 63–79. https://doi.org/10.3758/s13414-019-01827-z

Balas, B., Nakano, L., & Rosenholtz, R. (2009). A summary-statistic representation in peripheral vision explains visual crowding [Article]. *Journal of Vision (Charlottesville, VA)*, *9*(12), Article 13. https://doi.org/10.1167/9.12.13

Bauer, B. (2017). Does Stevens's power law for brightness extend to perceptual brightness averaging? *The Psychological Record*, *59*(2), 171–185. https://doi.org/10.1007/BF03395657

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford Publications.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 1904–1920*, *3*(3), 296–322. https://doi.org/10.1111/j.2044-8295.1910.tb00207.x

Callender, J. C., & Osburn, H. G. (1979). An empirical comparison of coefficient alpha, Guttman's lambda-2, and MSPLIT maximized split-half reliability estimates. *Journal of Educational Measurement*, *16*(2), 89–99. https://doi.org/10.1111/j.1745-3984.1979.tb00090.x

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press.

Ćepulić, D.-B., Wilhelm, O., Sommer, W., & Hildebrandt, A. (2018). All categories are equal, but some categories are more equal than others: The psychometric structure of object and face cognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(8), 1254–1268. https://doi.org/10.1037/xlm0000511

Cha, O., Blake, R., & Gauthier, I. (2021). The role of category- and exemplar-specific experience in ensemble processing of objects. *Attention, Perception, and Psychophysics*, *83*(3), 1080–1093. https://doi.org/10.3758/s13414-020-02162-4

Cha, O., Blake, R., & Gauthier, I. (2022). Contribution of a common ability in average and variability judgments. *Psychonomic Bulletin & Review*, *29*(1), 108–115. https://doi.org/10.3758/s13423-021-01982-1

Chang, T.-Y., Cha, O., McGugin, R. W., Tomarken, A., & Gauthier, I. (2024). How general is ensemble perception? *Psychological Research*, *88*, 695–708. https://doi.org/10.1007/s00426-023-01883-z

Chang, T.-Y., & Gauthier, I. (2022). Domain-general ability underlies complex object ensemble processing. *Journal of Experimental Psychology: General*, *151*(4), 966–972. https://doi.org/10.1037/xge0001110

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

Chow, J. K., Palmeri, T. J., & Gauthier, I. (2021). Haptic object recognition based on shape relates to visual object recognition ability. *Psychological Research*, *86*(4), 1262–1273. https://doi.org/10.1007/s00426-021-01560-z

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). L. Erlbaum Associates.

Corbett, J. E., Utochkin, I., & Hochstein, S. (2023). *The pervasiveness of ensemble perception: Not just your average review*. Cambridge University Press.

Cowan, N. (2008). What are the differences between long-term, short-term, and working memory? *Progress in Brain Research*, *169*, 323–338. https://doi.org/10.1016/S0079-6123(07)00020-9

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

Ekstrom, R. B., French, J. W., Harman, H. H., & Derman, D. (1976). *Kit of factor-referenced cognitive tests*. Educational Testing Service.

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

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science: A Journal of the American Psychological Society*, *11*(1), 19–23. https://doi.org/10.1111/1467-8721.00160

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General*, *128*(3), 309–331. https://doi.org/10.1037/0096-3445.128.3.309

Florey, J., Dakin, S. C., & Mareschal, I. (2017). Comparing averaging limits for social cues over space and time. *Journal of Vision*, *17*(9), Article 17. https://doi.org/10.1167/17.9.17

Gauthier, I., Chang, T.-Y., & Cha, O. (2023). *Data for "A General Ability for Judging Simple and Complex Ensembles", by Chang, Cha & Gauthier*. Figshare. Dataset. https://doi.org/10.6084/m9.figshare.24236950.v2

Gauthier, I., & Fiestan, G. (2023). Food neophobia predicts visual ability in the recognition of prepared food, beyond domain-general factors. *Food Quality and Preference*, *103*, Article 104702. https://doi.org/10.1016/j.foodqual.2022.104702

Gauthier, I., & Tarr, M. J. (1997). Becoming a "Greeble" expert: Exploring mechanisms for face recognition. *Vision Research*, *37*(12), 1673–1682. https://doi.org/10.1016/S0042-6989(96)00286-6

Gorea, A., Belkoura, S., & Solomon, J. A. (2014). Summary statistics for size over space and time. *Journal of Vision*, *14*(9), Article 22. https://doi.org/10.1167/14.9.22

Gustafsson, J.-E. (2002). Measurement from a hierarchical point of view. In H. Braun, D. Jackson, & D. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 87–111). Routledge.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*(4), 255–282. https://doi.org/10.1007/BF02288892

Haberman, J., Brady, T., & Alvarez, G. (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., Harp, T., & Whitney, D. (2009). Averaging facial expression over time. *Journal of Vision*, *9*(11), Article 1. https://doi.org/10.1167/9.11.1

Haberman, J., Lee, P., & Whitney, D. (2015). Mixed emotions: Sensitivity to facial variance in a crowd of faces. *Journal of Vision*, *15*(4), Article 16. https://doi.org/10.1167/15.4.16

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

Hadar, B., Glickman, M., Trope, Y., Liberman, N., & Usher, M. (2022). Abstract thinking facilitates aggregation of information. *Journal of Experimental Psychology: General*, *151*(7), 1733–1743. https://doi.org/10.1037/xge0001126

Hubert-Wallander, B., & Boynton, G. M. (2015). Not all summary statistics are made equal: Evidence from extracting summaries across time. *Journal of Vision*, *15*(4), Article 5. https://doi.org/10.1167/15.4.5

Im, H. Y., & Chong, S. C. (2014). Mean size as a unit of visual working memory. *Perception*, *43*(7), 663–676. https://doi.org/10.1068/p7719

JASP Team. (2023). *JASP (0.18.0)* [Computer software]. https://jasp-stat.org/

Jeffreys, H. (1939). *Theory of probability*. The Clarendon Press.

Jerskey, G. (2020). *A glance at the mirror: Ensemble perception of symmetry* [Master's thesis]. City University of New York.

Kacin, M., Cha, O., & Gauthier, I. (2022). The relation between ensemble coding of length and orientation does not depend on spatial attention. *Vision*, *7*(1), Article 3. https://doi.org/10.3390/vision7010003

Kacin, M., Gauthier, I., & Cha, O. (2021). Ensemble coding of average length and average orientation are correlated. *Vision Research*, *187*, 94–101. https://doi.org/10.1016/j.visres.2021.04.010

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Khvostov, V. A., & Utochkin, I. S. (2019). Independent and parallel visual processing of ensemble statistics: Evidence from dual tasks. *Journal of Vision*, *19*(9), Article 3. https://doi.org/10.1167/19.9.3

Lewandowsky, S. (2011). Working memory capacity and categorization: Individual differences and modeling. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(3), 720–738. https://doi.org/10.1037/a0022639

Little, T. D., Lindenberger, U., & Nesselroade, J. R. (1999). On selecting indicators for multivariate measurement and modeling with latent variables: When "good" indicators are bad and "bad" indicators are good. *Psychological Methods*, *4*(2), 192–211. https://doi.org/10.1037/1082-989X.4.2.192

Luo, A. X., & Zhou, G. (2018). Ensemble perception of facial attractiveness. *Journal of Vision*, *18*(8), Article 7. https://doi.org/10.1167/18.8.7

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*(2), 130–149. https://doi.org/10.1037/1082-989X.1.2.130

Mashburn, C. A., Barnett, M. K., & Engle, R. W. (2024). Processing speed and executive attention as causes of intelligence. *Psychological Review*, *131*(3), 664–694. https://doi.org/10.1037/rev0000439

McGugin, R. W., Richler, J. J., Herzmann, G., Speegle, M., & Gauthier, I. (2012). The Vanderbilt Expertise Test reveals domain-general and domain-specific sex effects in object recognition. *Vision Research*, *69*, 10–22. https://doi.org/10.1016/j.visres.2012.07.014

Miyake, A. (2001). Individual differences in working memory: Introduction to the special section. *Journal of Experimental Psychology: General*, *130*(2), 163–168. https://doi.org/10.1037/0096-3445.130.2.163

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

Myczek, K., & Simons, D. J. (2008). Better than average: Alternatives to statistical summary representations for rapid judgments of average size. *Perception & Psychophysics*, *70*(5), 772–788. https://doi.org/10.3758/PP.70.5.772

Oberauer, K. (2005). Binding and inhibition in working memory: Individual and age differences in short-term recognition. *Journal of Experimental Psychology: General*, *134*(3), 368–387. https://doi.org/10.1037/0096-3445.134.3.368

Oosterwijk, P. R., Ark, L., & Sijtsma, K. (2016). *Overestimation of reliability by Guttman's λ 4, λ 5, and λ 6 and the greatest lower bound*. In The Annual Meeting of the Psychometric Society (pp. 159–172).

Raftery, A. E. (1995). Bayesian Model selection in social research. *Sociological Methodology*, *25*, 111–163. https://doi.org/10.2307/271063

Raven, J. (2000). The Raven's progressive matrices: Change and stability over culture and time. *Cognitive Psychology*, *41*(1), 1–48. https://doi.org/10.1006/cogp.1999.0735

R Core Team. (2023). *R: A language and environment for statistial computing*. R Foundation for Statistical Computing.

Redick, T. S., Shipstead, Z., Harrison, T. L., Hicks, K. L., Fried, D. E., Hambrick, D. Z., Kane, M. J., & Engle, R. W. (2013). No evidence of intelligence improvement after working memory training: A randomized, placebo-controlled study. *Journal of Experimental Psychology: General*, *142*(2), 359–379. https://doi.org/10.1037/a0029082

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

Richler, J. J., Wilmer, J. B., & Gauthier, I. (2017). General object recognition is specific: Evidence from novel and familiar objects. *Cognition*, *166*, 42–55. https://doi.org/10.1016/j.cognition.2017.05.019

Rosseel, Y. (2012). *Lavaan*: An R Package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Ryan, K. F., & Gauthier, I. (2016). Gender differences in recognition of toy faces suggest a contribution of experience. *Vision Research*, *129*, 69–76. https://doi.org/10.1016/j.visres.2016.10.003

Salthouse, T. A. (1992). *Mechanisms of age-cognition relations in adulthood*. L. Erlbaum Associates.

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review*, *103*(3), 403–428. https://doi.org/10.1037/0033-295X.103.3.403

Salthouse, T. A. (2000). Aging and measures of processing speed. *Biological Psychology*, *54*(1–3), 35–54. https://doi.org/10.1016/S0301-0511(00)00052-1

Salthouse, T. A., & Babcock, R. L. (1991). Decomposing adult age differences in working memory. *Developmental Psychology*, *27*(5), 763–776. https://doi.org/10.1037/0012-1649.27.5.763

Sama, M. A., Nestor, A., & Cant, J. S. (2019). Independence of viewpoint and identity in face ensemble processing. *Journal of Vision*, *19*(5), Article 2. https://doi.org/10.1167/19.5.2

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*(4), 507–514. https://doi.org/10.1007/BF02296192

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*(5), 609–612. https://doi.org/10.1016/j.jrp.2013.05.009

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Smithson, C. J. R., Chow, J. K., Chang, T.-Y., & Gauthier, I. (2024). Measuring object recognition ability: Reliability, validity, and the aggregate z-score approach. *Behavior Research Methods*. Advance online publication. https://doi.org/10.3758/s13428-024-02372-w

Solomon, J. A. (2010). Visual discrimination of orientation statistics in crowded and uncrowded arrays. *Journal of Vision*, *10*(14), Article 19. https://doi.org/10.1167/10.14.19

Soper, D. S. (2023). *A-priori sample size calculator for structural equation models* [Computer software]. https://www.danielsoper.com/statcalc

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 1904–1920*, *3*(3), 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

Steiger, J. H. (2016). Notes on the Steiger–Lind (1980) handout. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(6), 777–781. https://doi.org/10.1080/10705511.2016.1217487

Steiger, J. H., & Lind, J. C. (1980). *Statistically-based tests for the number of common factors*. Paper Presented at the Annual Meeting of the Psychometric Society, Iowa City, IA.

Sunday, M. A., Dodd, M. D., Tomarken, A. J., & Gauthier, I. (2019). How faces (and cars) may become special. *Vision Research*, *157*, 202–212. https://doi.org/10.1016/j.visres.2017.12.007

Sunday, M. A., Donnelly, E., & Gauthier, I. (2018). Both fluid intelligence and visual object recognition ability relate to nodule detection in chest radiographs. *Applied Cognitive Psychology*, *32*(6), 755–762. https://doi.org/10.1002/acp.3460

Sunday, M. A., Tomarken, A., Cho, S.-J., & Gauthier, I. (2022). Novel and familiar object recognition rely on the same ability. *Journal of Experimental Psychology: General*, *151*(3), 676–694. https://doi.org/10.1037/xge0001100

Sweeny, T. D., Bates, A., & Elias, E. (2021). Ensemble perception includes information from multiple spatial scales. *Attention, Perception, & Psychophysics*, *83*(3), 982–997. https://doi.org/10.3758/s13414-020-02109-9

Sweeny, T. D., Wurnitsch, N., Gopnik, A., & Whitney, D. (2015). Ensemble perception of size in 4–5-year-old children. *Developmental Science*, *18*(4), 556–568. https://doi.org/10.1111/desc.12239

Takano, Y., & Kimura, E. (2020). Task-driven and flexible mean judgment for heterogeneous luminance ensembles. *Attention, Perception, and Psychophysics*, *82*(2), 877–890. https://doi.org/10.3758/s13414-019-01862-w

Tomarken, A. J., & Waller, N. G. (2005). Structural equation modeling: Strengths, limitations, and misconceptions. *Annual Review of Clinical Psychology*, *1*(1), 31–65. https://doi.org/10.1146/annurev.clinpsy.1.102803.144239

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), Article e0185452. https://doi.org/10.1371/journal.pone.0185452

Vallacher, R. R., & Wegner, D. M. (1989). Levels of personal agency: Individual variation in action identification. *Journal of Personality and Social Psychology*, *57*(4), 660–671. https://doi.org/10.1037/0022-3514.57.4.660

Ward, E. J., Bear, A., & Scholl, B. J. (2016). Can you perceive ensembles without perceiving individuals?: The role of statistical perception in determining whether awareness overflows access. *Cognition*, *152*, 78–86. https://doi.org/10.1016/j.cognition.2016.01.010

Westland, J. C. (2010). Lower bounds on sample size in structural equation modeling. *Electronic Commerce Research and Applications*, *9*(6), 476–487. https://doi.org/10.1016/j.elerap.2010.07.003

Whitely, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*(1), 179–197. https://doi.org/10.1037/0033-2909.93.1.179

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

Whitney, D., & Yamanashi Leib, A. (2018). Ensemble perception. *Annual Review of Psychology*, *69*(1), 105–129. https://doi.org/10.1146/annurev-psych-010416-044232

Wilhelm, O., Hildebrandt, A., & Oberauer, K. (2013). What is working memory capacity, and how can we measure it? *Frontiers in Psychology*, *4*, Article 433. https://doi.org/10.3389/fpsyg.2013.00433

Yamanashi Leib, A., Chang, K., Xia, Y., Peng, A., & Whitney, D. (2020). Fleeting impressions of economic value via summary statistical representations. *Journal of Experimental Psychology: General*, *149*(10), 1811–1822. https://doi.org/10.1037/xge0000745

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

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), Article 2041669517747297. https://doi.org/10.1177/2041669517747297

Yashiro, R., Sato, H., Oide, T., & Motoyoshi, I. (2020). Perception and decision mechanisms involved in average estimation of spatiotemporal ensembles. *Scientific Reports*, *10*(1), Article 1318. https://doi.org/10.1038/s41598-020-58112-5

Yörük, H., & Boduroglu, A. (2020). Feature-specificity in visual statistical summary processing. *Attention, Perception, & Psychophysics*, *82*(2), 852–864. https://doi.org/10.3758/s13414-019-01942-x

Received October 3, 2023
Revision received February 12, 2024
Accepted February 22, 2024 ■