9-2023

# Examining Metacognitive Access to Low-Level Ensemble Representations

Vladimir Mudragel
*The Graduate Center, City University of New York*

# Examining Metacognitive Access to Low-Level Ensemble Representations

by Vladimir Mudragel

A master's thesis submitted to the Graduate Faculty in Cognitive Neuroscience in partial fulfillment of the requirements for the degree of Master of Science, The City University of New York.

2023

APPROVAL

Examining Metacognitive Access to Low-Level Ensemble Representations

by

Vladimir Mudragel

This manuscript has been read and accepted for the Graduate Faculty in Cognitive Neuroscience in satisfaction of the thesis requirement for the degree of Master of Science.

Approved: June 2023

Tatiana Emmanouil, Advisor

Robert Duncan, Second Reader

THE CITY UNIVERSITY OF NEW YORK

## Abstract

Examining Metacognitive Access to Low-Level Ensemble Representations

by

Vladimir Mudragel

Advisor: Dr. Tatiana Emmanouil

Ensemble perception is a process that allows our sensory systems to rapidly extract summary information about the stimuli in the environment. For example, we are able to get a sense of the average number of items in a group of similar items (Burr & Ross, 2008; Halberda, Sires, & Feigenson, 2006) or the average size of a group of similar shapes of different sizes (Ariely, 2001). It is theorized that the qualitative result of ensemble perception is that it provides a gist impression of the current environment, which then enables attentional processes in the brain to determine which parts of the environment should next be attended to (Alvarez, 2011). Metacognition is a cognitive process in which thoughts, mental performance, or sensory representations are consciously accessed and analyzed. Because ensemble representations form so rapidly, it is not known whether they can be consciously accessed by metacognitive processes. Here, we explored whether or not ensemble representations can be metacognitively accessed by having participants report the average angle of an array of lines with semi-random angles (a low-level visual ensemble) and then asking them to rate their levels of confidence in their answers. We then calculated the degree of correlation between their performance on the ensemble task and their reported confidence levels, along with some other measures, both at the individual and group levels. Although there were some weak correlations between some of the measures, we did not find that there was any correlation between task performance and confidence.

## Table of Contents

## List of Figures

## Part 1: Introduction

The human brain excels at converting raw electromagnetic radiation within a certain range of wavelengths (roughly 400 nm to 800 nm (Gegenfurtner & Kiper, 2003)) into distinct and temporally contiguous visual scenes. With over 50% of the brain's cortical mass being dedicated to visual processing (Felleman & Van Essen, 1991), there is constantly a massive amount of activity occurring in the occipital, or visual, cortex. Information travels in a hierarchical fashion through, and within, a series of distinct cortical regions in the occipital cortex that represent that information with increasing complexity and scope. For example, cells in the primary visual cortex (V1) deal primarily with detecting particular wavelengths in the electromagnetic spectrum (Zeki, 1983); detecting the presence or absence, the length, and the orientation of a line or edge of a shape; the direction in which that line or edge is moving, if it's moving across the receptive field; as well as collectively representing one half of the fields of view from both eyes in each hemisphere, arranged in V1 into ocular dominance columns (Hubel & Wiesel, 1977). Each cell in V1 that represents the fovea has a very small receptive field (~1˚, or 0.33 mm on the retina) (Hubel & Wiesel, 1977); while cells higher in the visual hierarchy, in regions like V3 and V4, encode information about whole shapes in the visual field and context-independent color values of those shapes, respectively, over progressively larger receptive fields (1-20˚ (Rousselet, Thorpe, & Fabre-Thorpe, 2004)). Simultaneously, each processing region in this functional and neuroanatomical hierarchy sends information back down to the regions that feed into it, as well as to other regions of the brain, modulating them in turn. This feedback signaling maintains continuity between regions involved with processing fine details at the early stages of the visual processing pathways and regions involved with processing more abstract conceptual information higher up in the visual processing hierarchy. Concurrently, along with

the signaling pathways between the functional regions of the occipital lobe, there are also signals being sent back and forth between the occipital lobe and other brain regions devoted to other sensory processing modalities, memory, emotion, and executive functions. In this dizzying series of computational interactions between basic perceptual information, attention, and memory, our brains are able to convert that raw information into what we visually experience in every waking moment of every day of our lives.

All that raw sensory information is funneled through the mechanism of attention, which determines what part of the visual scene "stands out" to us at a given moment. Attention consists of a central focus of attention and a peripheral diffuse region (Alvarez, 2011). The focus of attention is where stimuli are most closely examined by neural sensory processing mechanisms, and therefore those stimuli are most accurately represented in the brain. The diffuse region surrounding the focus of attention also conveys information about the items within it to processing mechanisms, but with less and less accuracy the further they are located from the focus of attention. While it is typical for there to be one focus of attention at a time, it is possible for multiple items in a visual scene to be attended to in parallel (Howe et al., 2010; Cavanagh & Alvarez, 2005; Halberda, Sires, & Feigenson 2006; Friedman-Hill & Wolfe 1995; Chong & Treisman, 2005; Pylyshyn & Storm, 1988), producing multiple foci of attention, though the accuracy with which they are represented decreases as a function of how many are attended to at one time (Alvarez, 2011). On average, an individual can attend to 3 to 5 items at once (Pylyshyn & Storm, 1988) in a given visual scene, but that can vary based on several factors, such as the complexity of the stimulus (Horowitz et al., 2007), the depth of plane in which the stimulus can be found (Viswanathan & Mingolla, 2002), the speed with which a moving stimulus is moving (Alvarez & Franconeri, 2007), the speed the observer is moving in relation to the stimulus

(Thomas & Seiffert, 2010), and individual differences in the observers' physiologies (Oksama & Hyona, 2004). The simpler, closer, or slower a given stimulus is, the easier it is to attend to and the more of it is represented, allowing for more items to be attended to at once with precision. Conversely, the more complex, far away, or mobile a stimulus is, the more difficult it is to attend to and the fewer items can be attended to at once under those conditions. For example, Alvarez and Franconeri (2007) found that subjects could attend to up to eight objects if they were travelling at low speeds, but only one of the objects if they travelled at high speeds. This relatively low limit on the number of items that can be simultaneously attended-to by attentional systems in the brain is surprising, but well-established. Surprising because it fails to explain how it is that we are able to see, not 3 to 5 disembodied semi-stationary items at a time, but a full visual scene. This suggests that there is some other process at work, besides attention, that creates that background scene, in effect populating that diffuse region surrounding the focus of attention. A possible candidate for this process is a more recently described mechanism, called "ensemble perception".

**Ensemble Perception**

Ensemble perception takes advantage of grouped, redundant stimuli in the visual field, or ensembles, and extracts summary information about them automatically and preattentively (Oriet & Brand, 2013; see also Treisman, 2006). Because the items are redundant, but rarely truly identical, this summarized ensemble representation represents all of the items in the ensemble more accurately than any one of them can (Alvarez 2011). The more items in an ensemble, the more accurately they are represented by that summary information. Everyday visual examples of these kinds of stimuli can be the leaves on a tree, a flock of birds, stones in a stone wall, berries on a bush, or a crowd of people. Ensemble perception was first described in 2001, by Dan

Ariely, who showed that people can accurately report the mean size of a set of circles of various sizes, presented simultaneously, even if the stimulus array was shown for less than 100 ms, and even when they could not report on the size of individual circles. Since then, many other attributes of sets of items have been shown to be encoded as ensemble representations, allowing for gist impressions of their variability to be calculated. Along with average size (Ariely, 2001; Chong & Treisman, 2003; Chong & Treisman, 2005) these include representations of elementary attributes, such as brightness (Bauer, 2009), orientation (Ariely, 2001; Haberman, Brady, & Alvarez, 2015; Parkes et al., 2001; Dakin & Watt, 1997), color (Haberman, Brady, & Alvarez, 2015), the amount of items in an array (Burr & Ross, 2008; Halberda, Sires, & Feigenson, 2006), and the average location of objects in a scene (Alvarez & Olivia, 2008). Ensemble representations also have been shown to be encoded for intermediate attributes, like speed and direction of motion of items in an array (Sweeney, Haroz, & Whitney, 2012; Alvarez & Olivia, 2008; Watamaniuk, Sekuler, & Williams, 1989; Watamaniuk & Duchon, 1992; Watamaniuk, 1993), and estimation of the mean of a set of numbers (Smith & Price, 2010); and more complex attributes, such as higher-order spatial statistics (Alvarez & Olivia, 2008; Olivia & Torralba, 2007; Im, Tiurina, & Utochkin, 2021; Utochkin, 2015), average face identity (de Fockert & Wolfenstein, 2009; Neumann, Schweinberger, & Burton, 2013; Haberman, Brady, & Alvarez, 2015), average gender in a crowd of faces (Haberman & Whitney, 2007), average emotion in a crowd of faces (Fischer & Whitney, 2011; Haberman & Whitney, 2007; Haberman & Whitney, 2009), and emotional variance in a crowd of faces (Haberman, Lee, & Whitney, 2015). Though much of the research into ensemble perception focuses on visual ensembles, these representations can also be found in other sensory modalities. Common auditory examples, also referred to as “sound textures”, include running water, a swarm of insects, a crackling fire, a

chorus of birds, an outdoor crowd, a rattlesnake's rattle (McDermott, Schemitch, & Simoncelli, 2013; Zhai et al., 2020), or simply a series of tones (Albrecht, Scholl, & Chun, 2012). In both auditory and visual sensory modalities, stimuli can go through ensemble processing both if they are encountered simultaneously or sequentially (Albrecht & Scholl, 2010; Smith & Price, 2010; Haberman, Harp, & Whitney, 2009; Piazza et al., 2013; McDermott, Schemitch, & Simoncelli, 2013; Zhai et al., 2020) (For a discussion of these different kinds of visual and auditory representations, see Whitney & Yamanashi Leib, 2018).

In each of these instances the amounts of stimuli can, and usually do, far exceed the established limitations on attention. This suggests that there may be two types of sensory processing systems involved in processing incoming information: an attention-based system and a non-attention-based system. Surely, they must interact to produce our continuous and coherent experience of the world around us. How would that work? Could it be that all information first goes through some ensemble perception filter before moving on to higher-order processing (similar to Broadbent's filter theory (Broadbent, 1958) and Treisman's attenuation theory (Treisman, 1964))? Is each ensemble calculated as a separate entity, or are all ensembles present in the whole scene calculated together? How is this information incorporated in the brain's calculations over the course of sensory processing? How does it influence other brain functions, such as cognition or memory? These are questions that were explored by several studies, including the current one.

**Ensemble Perception and Attention**

The question of how attention is deployed when calculating ensemble statistics was explored by Oriet and Brand, in 2013. They conducted two experiments involving tasks that relied on attending to a target ensemble, interspersed with distractor ensembles. In the first

experiment, participants were randomly assigned to one of two groups: one that was told to focus on vertical lines of varying lengths or one that was told to focus on horizontal lines of varying lengths (for the vertical lines group, the distractor sets consisted of horizontal lines; while for the horizontal lines group, the distractor sets consisted of vertical lines). For each group, the task was to determine which side of the screen, left or right, contained a target group of lines with a greater average length, after seeing the array for 200 ms. The trials were separated into three different conditions: two control conditions and one experimental condition. In the "single set" control condition, only the target set was presented on both sides of the screen, without any distractors. This was done in order to account for the possibility of selection difficulty, arising from the participants having to navigate two sets of stimuli and two sets of distractors. In the "matched" control condition, the target set was presented, interspersed with the distractor set on both sides of the screen. In both of these control conditions, the sets were presented in random positions within fixed grids on both sides of the screen. However, both target sets had an identical average length. This was also the case for the distractor sets. Therefore, in the matched condition, the lines in all four sets had the same average length. The difference appeared in the experimental, or "mismatched", condition. There, as in the matched condition, the average length of the left and right target groups was identical, however, the average length of the distractor group on the left of the screen was either -40%, -20%, 20% or 40% larger than the distractor group on the right side of the screen. The distractor group on the right side still had the same average length as the two target groups. The researchers predicted that the judgement of which side of the screen contained the larger target group would be biased by the degree that the distractor group was larger or smaller than the target group on the left side. This is indeed what happened.

While differences between the two control groups were not significant, the experimental groups showed that when the distractor set on the left was smaller than the relevant set, participants were less likely to indicate that the target set on that side was larger, while the opposite was true when the distractor set was larger. The likelihood also varied according to the degree of difference the distractor set had with the other sets. Oriet and Brand reasoned that it was possible that the presentation time of 200 ms was too short to allow participants to differentiate between target and distractor sets. Therefore, they then repeated this procedure in a second experiment, but altered various parts of it to allow participants to spend an unlimited amount of time looking at each set. The second experiment also revealed that there was a significant effect based on an inflated or deflated average length in the left distractor ensemble, relative to the target ensemble, which influenced the participants' judgement of the target group. This study showed that all items in the visual field are automatically averaged together, regardless of whether any item or ensemble is meant to be attended to or not, indicating that the whole process of ensemble perception is, itself, pretattentive (for a wider discussion of this, see also Treisman, 2006).

Oriet and Brand were also careful to distinguish this work from other similar studies that reported segregation of ensembles before averages were calculated, stating that those studies compared ensembles where multiple attributes, such as color and orientation, were tested at once. They countered these studies by citing Moutoussis and Zeki (1997), which said that awareness of color may occur roughly 60 ms sooner than awareness of orientation (Oriet & Brand, 2013). It follows, then, that this temporal difference can produce separate ensemble representations about the same set of stimuli, leading to that segregation.

**Ensemble Perception and Domain Specificity**

That potential segregation of ensembles was explored in another study, by Haberman, Brady, and Alvarez, in 2015. In this study, they wanted to explore the question of whether ensembles at different levels of complexity were built out of some general ensemble mechanism, or if ensemble mechanisms were domain-specific. They hypothesized that if there was a correlation in individual subjects' performances between different kinds of ensembles, representing different processing domains, then that would suggest a global mechanism. However, if there was correlation in performance between different ensembles within the same domain, but not across different domains, then it may be the case that ensemble mechanisms are domain-specific. To test this, they measured the degree of correlation in performance using different kinds of visual stimuli over the course of eight experiments. In the first experiment, they compared a high-level ensemble (average face identity) to a low-level ensemble (average Gabor patch orientation), while also comparing individual low-level stimuli and individual high-level stimuli (a single Gabor patch or a single face), in order to determine how individual representations relate to ensemble representations. In a trial where only one face or only one Gabor patch was cued, participants had to adjust a test face or a test Gabor, using a computer mouse, to match the cued stimulus. In a trial where the whole set of faces or Gabor patches was cued, participants had to adjust the test face or test Gabor patch to the average identity or orientation of the whole set. For the remaining seven experiments, they repeated Experiment 1, but using different kinds of ensembles and attributes in order to determine how performance varied across many different domains.

Because Experiment 1 compared both single items and ensembles, they needed to ensure that participants were actually looking at the center of the screen. For that reason, Experiment 1 was conducted in the lab, where they could ensure fixation on the center of the screen using eye

tracking software. Since the rest of the experiments involved comparing two different kinds of ensembles, instead of an ensemble and a single item of the same kind, eye fixation wasn't necessary, so they were performed online. Due to this, Experiment 2 was simply a replication of Experiment 1, but in an online environment, using only the ensemble tasks. Experiment 3 also repeated Experiment 1, but with different high-level and low-level ensembles as stimuli: average facial expression in a set consisting of the same face with varying facial expressions compared to the average color of a set of isosceles triangles. Experiment 4 compared two low-level statistics using ensembles consisting of the same kinds of objects: the average orientation of a set of isosceles triangles and the average color of a set of isosceles triangles. Experiment 5 compared two different kinds of low-level ensembles and statistics: the average orientation of a set of Gabor patches and the average color of a set of dots. Experiment 6 compared two different kinds of low-level ensembles, using the same statistic: the average orientation of a set of Gabor patches and the average orientation of a set of isosceles triangles. Experiment 7 compared two high-level ensembles: the average identity of a set of faces and the average emotional expression in a set of the same face with varying expressions. Experiment 8 was meant to establish a theoretical ceiling of correlation and a theoretical floor of correlation and was split into two parts. Experiment 8a compared the average orientation between two sets of Gabor patches, where one set of patches hat a high frequency of lines, while the other had a low frequency. Since these tasks shared both the type of stimulus and the type of statistic, they were expected to have the highest degree of correlation of any of the tasks, establishing a theoretical "ceiling" of correlation. Experiment 8b compared performance on the average orientation of a set of Gabor patches and an adaptive letter span memory task. The letter span task was chosen because it is a working memory task that requires minimal visual processing, and especially does not require

ensemble processing. As such, the degree of correlation between these two tasks was expected to be around that of chance, establishing a theoretical "floor" of correlation.

The results of all these experiments painted an interesting picture. In Experiment 1, they found a significant correlation between subjects' performances on individual face identification tasks and average face identification tasks, as well as a significant correlation between individual Gabor patch orientation tasks and average Gabor orientation tasks. However, there was no correlation between the face identification tasks and the Gabor orientation tasks, individual or ensemble. This showed that performance in identifying individuals in a set predicted performance in identifying the average of a set made up of either high- or low-level stimuli, but performance in identifying either individual or the average of the whole set of low-level stimuli did not predict performance with high-level stimuli, and vice versa. This was verified in Experiment 2, which again showed a low level of correlation between the average face identity task and the average Gabor orientation task. There was also a low degree of correlation between the high- and low-level tasks in Experiment 3. There was, however, a high degree of correlation in the following four experiments, where the stimulus domains were either both high-level or both low-level. The specific statistic being calculated did not appear to make any difference. These results suggest that there is a functional separation between performance in calculating statistics about high-level visual ensembles and low-level visual ensembles. Rather than there being one overarching mechanism for calculating ensemble statistics, it appears that each type of ensemble is calculated locally, at the level where that kind of stimulus is encoded. Put another way, as neural representations from lower processing domains are combined as features in more complex neural representations in higher processing domains, ensemble representations are also calculated in parallel from that same information.

**Functional Benefits of Ensemble Perception**

As stated previously, ensemble processing is useful because it can create an impression of a set of relatively redundant stimuli that is more accurately represents the whole set than any individual member can. This is because the process of averaging automatically cancels out random deviations from the mean any individual in the group may have. One benefit of this is that it can allow the brain to rapidly categorize groups of items in a visual scene. The distribution of the individual items averaged in the ensemble along each feature dimension, or "feature distribution", can serve to inform the visual system of the general layout of the scene. As described by Duncan and Humphreys (1989), in a visual search task, where a target stimulus needs to be found among distractor stimuli, it become more difficult to find the target stimulus if there is a high amount of heterogeneity among the distractors. However, Yurevich and Utochkin (2014) found that if there is a smooth feature distribution of distractors, rather than a sporadic one, it instead makes it easier to identify the target stimulus. Based on this observation, they proposed that there is some threshold of dissimilarity they referred to as "segmentability". If the feature distribution of a set of features is closer together, they are more likely to be counted as one heterogeneous group by the visual system. However, if the feature distributions of groups of stimuli are far enough apart, they are then "segmentable", and are considered by the visual system to belong to different groups. An example of this can be seen in berries growing on a bush: the ripe ones tend to be red, while the unripe ones tend to be green. As a feature distribution, the colors of the berries would be presented as clustering around two peaks instead of one, thus their segmentability allows for the recognition that they belong to two groups instead of one. Other feature distributions can be calculated along all the other dimensions mentioned in the introduction, across low, intermediate, and high processing levels. The closer a given item in

an ensemble is to the ensemble's mean, in a given dimension, the more likely it is to be considered part of a group (Yurevich & Utochkin, 2014; Im, Tiurina, & Utochkin, 2021). Conversely, the further from the mean the value from an item is, the less likely it is to be considered part of the group (Khayat & Hochstein, 2018; 2019).

Besides implicit categorization, ensemble percepts can also facilitate change detection and outlier detection over the whole visual scene. Because they are not as well-represented as the items in the focus of attention, changes in individual items in an unattended ensemble tend not to be noticed by the visual system. This leads to the phenomena of change blindness (Simons & Levin, 1997) and inattentional blindness (Rock et al., 1992). However, if the change is such that it alters the summary statistic of the ensemble in one or more of its relevant dimensions, or its ensemble structure, then the change is noticed (Alvarez & Olivia, 2009), even though it occurs outside the focus of attention. This same property is what enables ensemble representations to facilitate outlier detection (Cavanaugh, 2001; Cohen, Dennett, & Kanwisher 2016). If an item is sufficiently different from the rest of the items around it, it is more likely to stand out in the scene.

By allowing the visual system to roughly group stimuli according to their feature distributions in a scene, as well as detect outliers and statistical changes in the features of a scene, all outside the focus of attention, ensemble perception allows for a representation of the general patterns that can be found in any given scene. These patterns can then guide the focus of attention in its visual search of the scene in an efficient and intuitive way. They can also provide the visual system with clues as to what kind of environment is being observed by combining low-level features and then using them to categorize the whole scene (Olivia & Schyns, 1997; Schyns & Olivia, 2004; Triesman, 2006; Cohen, Dennett, & Kanwisher, 2016). Without this

ready-made contextualization of the scene, attention would be "flying blind", unable to orient the items it focuses on within the larger context of the scene. There would be no ability to detect outliers or changes in the scene (Noe & O'Reagan, 2000; O'Reagan, 1992; Alvarez, 2011), unless the observer happened to be focusing on them when they occurred.

**Ensemble Perception, Anatomy, Reverse Hierarchies, and Metacognition**

There are several neuroanatomical sites potentially related to visual ensemble processing. Since neurons in regions lower in the visual processing hierarchy, such as V1 and V2, have the smallest receptive fields (Hubel & Wiesel, 1977, Reynolds & Chelazzi, 2004), it has been proposed that ensembles themselves are calculated by the pooling of incoming signals from those lower regions (Haberman & Whitney, 2012). That same pooling is also what allows for each of the cells in higher processing areas in the hierarchy to have larger receptive fields, representing a larger area of the visual scene (Rousselet, Thorpe, & Fabre-Thorpe, 2004). Some higher regions that have been implicated in ensemble perception include V3, which is involved in processing the orientations of groups of lines; the parahippocampal place area, which is responsible for identifying a given scene; the retrosplenial cortex, which recognizes how a scene is laid out; the occipital place area, which recognizes the navigability of a given scene; as well as the fusiform cortex, which is specialized to recognize faces and emotional information they convey (Tark et al., 2021; Cohen, Dennett, & Kanwisher, 2016).

Interestingly, Tark and colleagues (2021) were able to identify, in one of a series of fMRI studies, that while signals in the visual cortex travelled from V1, to V2 and V3, their activity would vary depending on task requirements. In this study, participants were briefly shown a set of 36 Gabor patches and were asked to either report the average of the whole set or to report the orientation of a single pre-cued Gabor patch among the other Gabor patches. These two different

kinds of trials were presented either in a task paradigm that was meant to prioritize mean orientation identification or in a paradigm that was meant to prioritize single orientation identification. In reality, both task paradigms had both tasks interspersed within them. The data were analyzed using Representational Fidelity, a measure that compared the angles being represented in a given brain region to a previously calculated encoding model, based on each voxel's orientation selectivity. After analyzing the data, Tark and her colleagues found that during the mean orientation task, when the mean orientation was reported, there was a low amount of fidelity in V1, a moderate amount of fidelity in V2, and a high amount of fidelity in V3. However, when a single Gabor patch was pre-cued during the mean orientation task, there was very little fidelity in all three regions. Meanwhile, during the single orientation task, when the single Gabor patch was pre-cued, there was that same steady increase in fidelity from V1 to V3. The angles were being progressively more accurately represented as they travelled up the hierarchy. However, when the single Gabor patch was not pre-cued in the single orientation task, and the mean orientation had to be reported, there was still a moderate amount of fidelity in V1 and V3, though not as much in V2. They reasoned that this was happening because the entire ensemble of lines was being represented in V3, regardless of its task relevance. This is similar to the conclusion reached by Oriet and Brand (2013) discussed above. The difference is that now this was demonstrated anatomically, using brain scanning technology.

A framework for how ensemble perception may fit into larger perceptual processes can be found in Reverse Hierarchy Theory (RHT). First proposed by Ahissar and Hochstein (Hochstein & Ahissar, 2002; Ahissar & Hochstein, 2004), just as in the classical view, RHT states that, while signals travel up the sensory processing hierarchy, all of their component parts are integrated, the representations of the stimuli becoming more complex and more integrated as

they travel up the hierarchy, with more and more of the visual field being represented by individual cells as the signals travel up the hierarchy. Conscious awareness of the visual scene does not occur until the signals arrive at the highest levels of the processing hierarchy. That is where the overall gist impression of the scene is first detected. This is where RHT diverges from other theories of consciousness, which have suggested that further processing of the scene that would allow for the perception of details would occur in higher regions or in the parietal cortex. Instead, Ahissar and Hochstein argue that in order to perceive the finer details of a scene, attention needs to be deployed to specific areas of the scene, directed, perhaps, by regions in the parietal cortex or the pulvinar (Hochstein & Ahissar, 2002), by selecting cells in the higher regions of the hierarchy and causing them fire back down into lower levels in the hierarchy. This, in turn, causes the lower-level cells to fire again, causing the impression of the area of the scene those cells represent to become more robust, making them stand out more from the background, in more detail. With enough iterations of this, the cells lower in the hierarchy become more likely to fire, through long-term potentiation, which adds even more weight to the signal coming from them. As attention shifts from one focus to the other, different parts of the visual field become attenuated in this way. With training, fast detection of details in a scene becomes easier and easier, but in different ways, depending on how far the stimulus can travel up the hierarchy and how much those higher regions can then influence the cells lower in the hierarchy that feed their projections into them.

In a series of experiments, Ahissar and Hochstein (2004) demonstrated, first, that when participants were presented with an array of lines at identical angles, they were able to detect a single line that was at a different angle from the rest of the lines. When they modulated the amount of time the stimulus array was presented before a mask was presented (stimulus-to-mask

onset asynchrony, or SOA), they found that the amount of time affected the flexibility with which participants could detect the line. When the SOA was long (more than 250 ms), participants could detect the line regardless of its location, orientation, or distance from the fixation point at the center of the array. On trials where the SOA was shorter than 250 ms, participants tended to only be able to detect it if it was in the same location from trial to trial, had the same orientation, and was closer to the fixation point. Ahissar and Hochstein reasoned that this was because the signal had not yet had time to travel all the way up the cortical hierarchy, so it did not get to have an opportunity to integrate with higher level objects and concepts. As the longer SOA's allowed for that, the detection of the outlying stimulus could be generalized across the whole visual field. This was accomplished because higher-level neurons had enough time to bias lower processing levels in the hierarchy through their feedback projections and increase their sensitivity to the outlier stimulus.

This was supported in subsequent experiments where Ahissar and Hochstein trained participants further by presenting the array with long and short SOAs in an interleaved manner. Initially, detection was low in both conditions, but over time, a telling pattern emerged. At first, participants began to improve only on the longer, easier, SOA detection tasks. They only began to improve on the trials with shorter SOAs once their performance had significantly improved on the trials with long SOAs. What this showed is that information about the anomalous stimulus was first integrated at higher processing regions, which adapted to it in multiple dimensions, allowing them to deploy attention to that stimulus more flexibly. Once that was achieved, the attentional deployment to more and more regions of the visual field strengthened the signals coming from those regions and trained the neurons further and further down in the hierarchy to fire when they detected the outlier. Because the higher-level regions were increasing the weight

of incoming signals through their feedback projections, participants could soon detect the outlier stimulus after the short SOAs, as well as they could after the long SOAs, even though the signal still did not have the opportunity to travel all the way up the hierarchy. In this way, RHT explains how ensemble perception is calculated – as a bottom-up pooling of signals through consecutive processing regions, leading to a conscious overall undetailed impression of the scene – and how it serves to then direct attention to particular parts of the scene, scrutinizing them, allowing us to experience those parts in detail. That initial generalized entry into consciousness is characterized also by our subjective recognition that we are conscious of it. That recognition that we are conscious of something is a prime example of what is called "metacognition".

Metacognition is our ability to consciously evaluate and make judgements about our own cognitive or perceptual performance (Nelson, 1996). Simply put, it is our ability to "think about thinking". Whenever someone wonders how well they can remember some piece of information, assesses their strategy for computing a math problem in their heads, or realizes that their thoughts have drifted from what they meant to be thinking about, they are utilizing metacognition. If a person can make metacognitive judgements about a neurological operation, such as how well they remember something, or how well they can see something, that means that that information is accessible to consciousness and can be attended to. If they cannot, then the information is not accessible. For example, a person cannot assess how well the representation of line orientations used in constructing the overall representation of a three-dimensional object performed in its integration, independent of the other processing domains that constructed that final impression. Even though those elements had to be identified and integrated by the brain to create the final impression, our inward-facing metacognitive eye is blind to them.

This process of self-referential conscious access can be measured in several ways, depending on what is being studied. Typically, the degree of metacognitive ability is determined by comparing objective performance to the subject's evaluation of their performance. Metacognition in purely cognitive tasks, such as memory tasks, is probed by assessing the participant's subjective Judgement of Learning (JOL) measure; while on perceptual tasks, it is the participant's degree of confidence in how well they perceived a stimulus that speaks to how much they were able to consciously access the representation in question (Bonder & Gopher, 2019). If they are more confident, that certainty is due to how well they were able to perceive the stimuli. This is the metric we employed in our current study.

**The Current Study: Metacognitive Access to Ensemble Representations**

Because ensemble processing happens outside of our conscious awareness, it remains unknown whether ensemble representations themselves are accessible to metacognition. For instance, it could be the case that they are completely inaccessible, and only become indirectly apparent to cognitive processes because of how they bias the overall sensory input and subsequent cognitive operations based on it. On the other hand, if the ensemble representations can be accessed by metacognition, then discreet judgements could be made specifically about the summary information they provide.

Some work has been done to explore this type of relationship. In one study, it was shown that there can be metacognitive access to ensemble representations in higher levels of the visual processing hierarchy, like our ability to detect average faces (Ji & Hayward, 2021). However, they also found that this metacognitive access was conditional on whether or not participants were asked to identify the average face in a group of faces. When they were asked to identify the average face, confidence was high and the participants were able to report it. In conditions where

they were asked to recall if a particular face was present in a set of faces they were previously presented, their answers tended to be biased towards the average face, but their confidence was low. Ensemble perception persisted, preattentively, but without metacognitive access.

As for visual processing in regions lower in the processing hierarchy, which tend to encode simpler features, rather than entire objects, other researchers have probed other kinds of cognitive access into ensemble perception, like value associations (Dodgson & Raymond, 2020), but no one has yet investigated whether ensemble representations calculated lower in the visual hierarchy can be accessed through metacognition, when the task requires it, in the way ensemble representations higher in the hierarchy can. This is the focus of the present study. In order to determine whether visual ensemble perception can be accessed by metacognition, we measured the accuracy of participants' estimates of the average angle of a rapidly presented set of lines with randomly assigned angles. We then probed their metacognitive access to the ensemble representation by asking them to rate their confidence in their answer after each array. Here, we define metacognition as the ability to access the output of the ensemble representation, rather than the steps involved in creating it. If, as we suspected, the participants' confidence scores were correlated with the accuracy of their answers (that is, inversely correlated with their degree of absolute error), then low-level visual ensemble representations would be accessible to metacognition. If there was no correlation at all, then that would mean there is no metacognitive access to that lower-level ensemble. Our results demonstrated, that there was no correlation between participants' accuracy and how confident they felt about their performance.

## Part 2. Methods

### Participants

This experiment was made freely available to users on Prolific.co, an online platform for psychological experiments. There were 75 participants (37 men, 38 women, ages ranging from 19 to 65), a number determined by a sample size power analysis. 7 participants were excluded because the confidence values they reported were less than one standard deviation away from the mean of all of the standard deviations for confidence values reported by each participant in the study. This meant that they were selecting the same or very similar confidence values for each trial, suggesting they were only selecting a value to move the experiment along, rather than reporting their actual confidence levels. Each participant was paid $4.50 once they completed the experiment.

We used Psychopy Builder v2021.2.0 to design the experiment, which we then uploaded to Pavlovia.org. Participants accessed and completed the experiment via a link to Pavlovia, posted on Prolific. In order to maintain anonymity, they were assigned numbers other than their Prolific IDs upon starting the experiment in Pavlovia. These were the only numbers we used to identify participants in our analysis. By design, Prolific prevents participants from participating in a study more than once.

**Stimuli**

The stimuli consisted of a 4 x 4 array of 1-cm-long lines, with constituent angles having random orientations, ranging from 10° to 80°, in 5° increments, with 0° being a horizontal line. This range of angles was chosen in order to limit the difficulty of the task. The lines also had a randomized positional jitter in x and y directions of either 0, 25, or -25 pixels. Together, this ensured that no two arrays were identical, in terms of both the angle of each line and their positions in space. At a distance of roughly 57 cm from the screen, the visual angle of the arrays was 4° ± 1.3°.

## Design and Procedure

Participants were first shown a consent form to read, which detailed the content of the experiment. If they consented, they were taken to the next screen. If they did not consent, they were instead taken to the end of the experiment. They were also asked whether their data can be used for future research by the current researchers and/or by future researchers.

Next, participants were instructed to adjust a picture of a credit card on their screen using the arrow keys on their keyboards at home, so that it would match the size of their own credit cards. This size adjustment was then applied to the size of the presented stimuli and to their spacing, so that they would remain constant across different monitor sizes. After that, the participants were shown a screen explaining to them the contents of the experiment and the two tasks they would have to complete in each trial. Then they were taken through 5 practice trials of the task to get accustomed to it.

Following the practice trials, participants started the actual experimental trials. Each trial consisted of three steps. First, participants were shown the stimulus array for 200 ms, preceded and followed by a blank screen lasting 1 second. This was followed by a probe, where each participant used the left and right arrow keys on their keyboards to rotate a single line so that it matched what they believed was the average orientation of the set of lines they were just shown. Next, they were asked to rate their confidence in their assessment on a scale of 1 to 100 (as seen in Pallier et al., 2002 and Bonder & Gopher, 2019). This was repeated 60 times per participant (see Figure 1).

Once the trials were finished, participants were asked demographic questions about their age, ethnicity, and gender. Then they were thanked for their participation and given a completion

code so that they could get paid for their time. Measures of confidence, estimated average angle, actual average angle, and response times for each stage of each trial were recorded by Prolific. The answers to the demographic questions, as well as the angles and jitter of each line in each presented array were also recorded.

Once the data were collected, the participants' estimates of the average angle of each array were subtracted from the actual average angle of that array. The error of each participant was the absolute value of that difference; the higher the error, the further their answer was from the actual average angle. The key measure of this study was a correlation analysis between confidence and error. However, if there was not a correlation between those two particular measures, there could have been correlation between these measures and other aspects of the experiment. Therefore, along with this main measure, the correlation coefficients of confidence and probe adjustment reaction time; confidence and distance the probe was rotated; error and probe adjustment reaction time; and error and distance rotated were also calculated. Finally, in order to determine if the range of the angles presented in a given trial correlated with any of these measures, the correlation analyses were repeated for range and confidence; range and error; range and probe adjustment reaction time; and range and distance rotated. These analyses were performed at the individual level, and then averaged and performed at the group level.

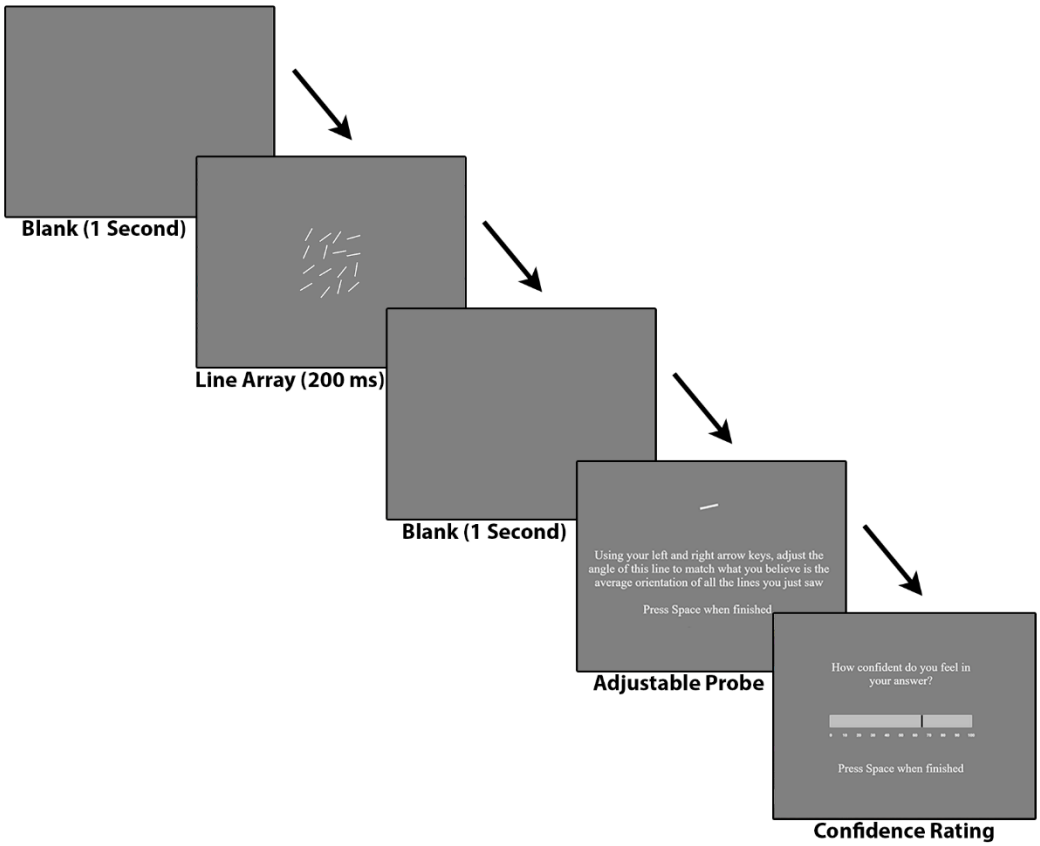


*Figure 1.* A representation of the order in which stimuli, probe adjustment task, and confidence rating task were shown to participants in each trial (text enlarged for legibility).

## Part 3. Results

The question this study was designed to investigate was whether the output of early ensemble perception was accessible to metacognition. Therefore, the key measure was a Pearson correlation coefficient between reported confidence levels and the absolute error between estimated averages and actual averages of all lines in each array, for each trial. We computed correlation coefficients between absolute error and confidence for each participant and then transformed those to Fisher z values for subsequent statistical analyses. The average Fisher z value between absolute error and confidence was low (Fisher z = -.011) and a two-tailed t-test, comparing Fisher z scores to 0 was not statistically significant ($t(67) = -.66$, $p = .509$). This indicated that, overall, performance on this low-level ensemble task did not correlate with the

degree of metacognition exhibited by the participants, though there were individuals who showed a moderate level of correlation (see Figure 2), indicating a potential role for individual differences in either overall confidence or accuracy in reporting ensemble representations.

Because other aspects of participants' performance were also measured, the analysis was extended to investigate whether or not any of them correlated at the group level. These included probe adjustment reaction time (PArt), the distance the probe was rotated (DR), and the range of angles presented in each trial. Fisher z scores were very low between all these measures as well (Confidence vs. DR (Fisher z = .006); Range of Angles vs. PArt (Fisher z = .016); Range of Angles vs. DR (Fisher z = .005); Confidence vs. PArt (Fisher z = -.103); Confidence vs. Range of Angles (Fisher z = -.047); Absolute Error vs. PArt (Fisher z = .049); Absolute Error vs. DR (Fisher z = .075); and Absolute Error v. Range of Angles (Fisher z = .048)).

However, when the two-tailed t-test was repeated on these comparisons, using Bonferroni adjusted alpha levels of .0056 per test (.05/9), only two of them yielded t-values that showed that they were significantly different than 0 (Confidence vs. PArt ($t(67) = -4.18$, $p < .001$) and Absolute Error vs. Range of Angles ($t(67) = 3.03$, $p = .004$)). The rest of the comparisons yielded t-values which indicated that they were not significantly different than zero at the group level (Confidence vs. DR ($t(67) = .25$, $p = .804$); Range of Angles vs. PArt ($t(67) = 1.03$, $p = .309$); Confidence vs. Range of Angles ($t(67) = -2.64$, $p = .010$); Absolute Error vs. PArt ($t(67) = 2.06$, $p = .044$); Absolute Error vs. DR: $t(67) = 2.63$, $p = .011$; and DR vs. Range of Angles ($t(67) = .27$, $p = .788$)).

Along with looking at the distribution of correlation coefficients from each participant across all the measurements mentioned above, we also looked at how the average values of those measurements correlated to one another across all participants, with a Bonferroni adjusted alpha

level of .0056 per test (.05/9). Of the nine different correlations, however, only absolute error vs. distance rotated ($r(58) = .54$, $p < .001$), showed significant correlation at the group level. The rest did not (Absolute error vs. PArt ($r(58) = .26$, $p = .030$); Confidence vs. Error ($r(58) = .13$, $p = .284$); Confidence vs. PART ($r(58) = -.003$, $p = .981$); Confidence vs. DR ($r(58) = .19$, $p = .119$); Confidence vs. Range ($r(58) = -.10$, $p = .433$); Range vs. Error ($r(58) = .14$, $p = .247$); Range vs. PART ($r(58) = .06$, $p = .625$); Range vs. DR ($r(58) = .15$, $p = .233$)). This likely means that the further the probe was moved, the more it negatively affected performance.

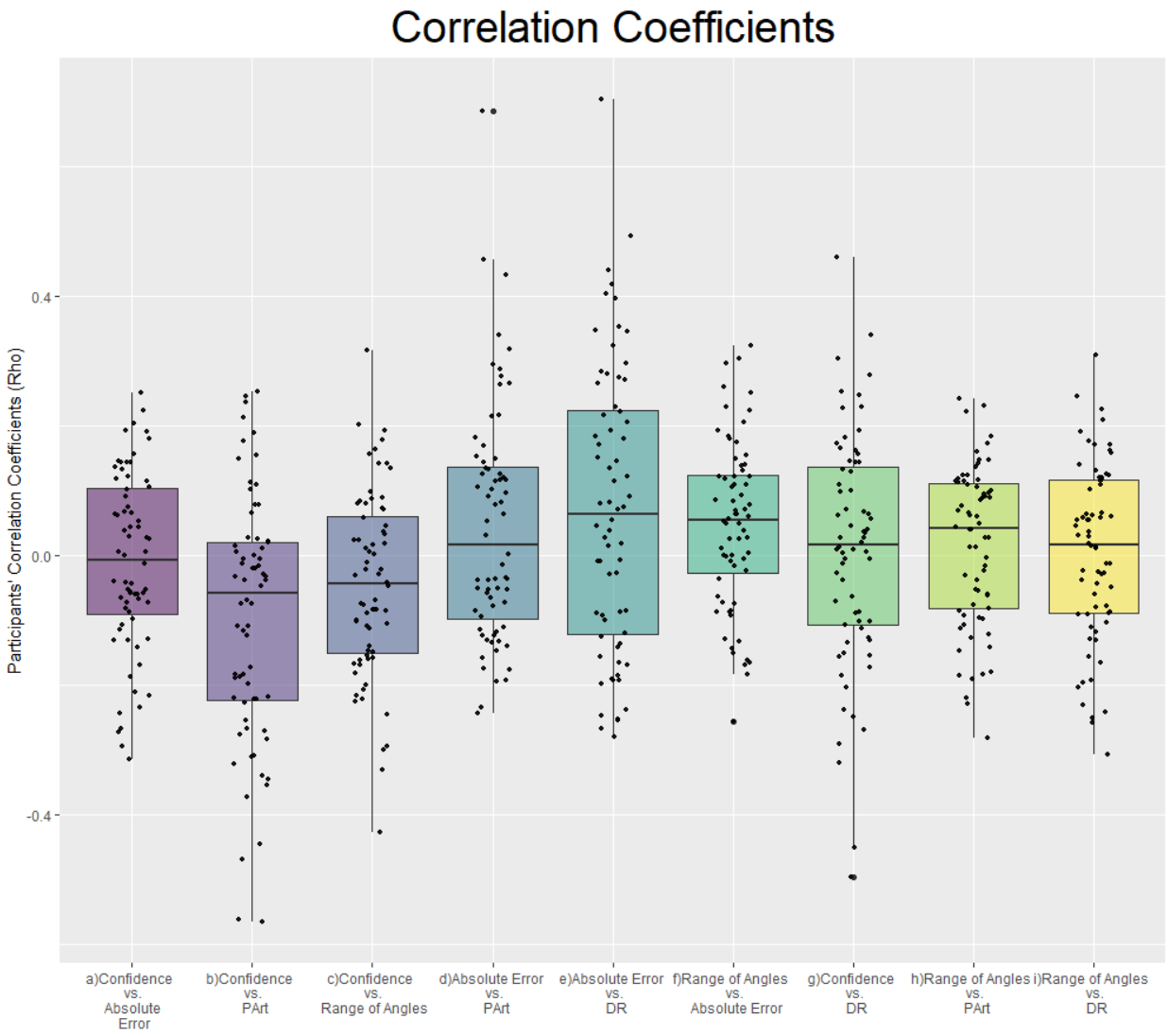


*Figure 2.* Box and Whisker plots depicting the distribution of correlation coefficients for each participant within the nine different correlation analyses that were conducted. While each of these plots reflect the results across all participants in the study, in each analysis, there were both participants who showed significant, but weak, correlations and participants who showed significant, but weak, anti-correlations.

## Part 4. Discussion

The data at the group level showed that there was no correlation between participants' level of confidence in their performance and the absolute error between the actual average angle of the lines presented and the estimates the participants gave. Though they should be interpreted with caution, and further studies are needed to fully explore a potential relationship between ensemble perception and metacognition, these results do not support that there is a relationship between ensemble processing and metacognition, at least not at this early stage of visual processing. We were therefore unable to reject the null hypothesis. This result is indirectly supported by studies that show that ensemble representations at early stages of visual processing are preattentive (Oriet & Brand, 2012), and therefore it is unsurprising that they were not accessed by metacognitive mechanisms. However, it is still interesting that correlations have been found at later stages of visual processing between ensemble perception and metacognition, depending on the type of task, as mentioned previously (Ji & Hayward, 2021).

Other sets of data collected in this study were also checked for any correlations between them. Besides confidence and absolute error, probe adjustment reaction time, the distance the probe was rotated, and the range of angles presented in each trial were also included. All of these measures were compared to one another. Though all of them yielded rho values and Fisher z values that were close to 0, a two-tailed t-analysis showed that only two comparisons were significantly different than 0. The first of these was the comparison between confidence and probe adjustment reaction time. Therefore, it can be said that there was a very weak anti-correlation at the individual level between those two measures. This means that, at a very small but significant rate, that either the more confident participants felt, the less time they took adjusting the probe; or the less time they spent adjusting the probe, the more confident they felt.

Second, the comparison between error and range of angles also had a t-value significantly different than 0, though the Fisher z value was still very low. This indicates that there was an extremely weak but significant correlation between those measures at the individual level, meaning that participants were very slightly more likely to make errors when the range of angles presented was wider. This result aligns well with previous literature suggesting decreased performance in ensemble perception tasks when the range of values along a feature dimension increases (Utochkin & Tiurina, 2014).

Normally, directionality cannot be established from a correlation coefficient. However, in the cases involving participants' performance and their adjustment of the probes, as there was no relationship found between metacognitive access at this level of processing, it can be inferred that participants would have had no awareness of how well they were doing at a given trial. Since participants were given no feedback on their performance, this meant that their amount of error could not influence how much they manipulated the probe, but their manipulation of the probe could influence their amount of error. The directionality can also be established in the case involving the range of the angles presented because they were randomized every trial, and so could not determine participants' levels of confidence.

While there was either no correlation, a very weak correlation, or a very weak anti-correlation across all measures, it is interesting that there were individuals in several of the comparisons who differed significantly from the mean either in a positive direction or a negative one. This could mean that individual differences may play a role in how these measures correlate in at least a subset of the general population. These outliers were most noticeable in the distribution of correlation coefficients between levels of confidence and how much they engaged with the probe (both in terms of time and the distance moved); as well as between absolute error

and how much they engaged with the probe (again, both in terms of time and distance moved). However, because they differed in both positive and negative directions, it is difficult to characterize the nature of these individual correlations.

As mentioned in the Results section, we also calculated the correlation coefficients between the averages of each of the measurements taken, producing rho values that represented the entire group of participants. This differed from the initial measurements, which were based on correlation coefficients calculated from each value across all the runs for each participant. This produced a correlation coefficient for each participant, which we then converted to Fisher z scores in order to run further analyses on them. Conversely, the group correlation analysis was based on a dataset made up of a set of values that were the average values of each participant's performance on each of the different measurements. This allowed us to answer questions with a wider reach than just the participants in our study, such as "On average, do people with low absolute error tend to have higher confidence?" or conversely, "On average, do people with higher confidence tend to have lower absolute error?".

The correlation analysis at the group level showed, again, no correlation between confidence and performance at the group level. There was, however, a significant moderate correlation between absolute error and the distance the probe was rotated. This means that on average, performance tends to suffer if the probe is rotated further from its starting position during a given trial. Again, since there is no correlation between metacognitive judgements and performance on this task, it is likely that how much they interacted with the probe was not influenced by their performance on previous trials.

Because this experiment was carried out online, presumably on each participant's home computer, we had no way of ensuring that participants were actually focusing on the center of

the screen before each trial, as trials within the lab could, by using eye-tracking mechanisms (as in Haberman, Brady, & Alvarez, 2015). Therefore, the lack of correlation could be explained by inconsistent attention between trials or distraction by other objects in their fields of view. It is also possible that the trial time of 200 ms was too short, although this is unlikely, since other studies (Chong & Treisman, 2003) have shown that accurate ensemble estimates can be extracted from stimuli being presented for as little as 50 ms without any pattern masks. However, contrary to that, Whiting and Oriet (2011) have shown that the average can be extracted from arrays presented for 200 ms with pattern masks, but no shorter. Pattern masks (Exner, 1868) are useful for blocking higher-level visual processing of stimuli in memory tasks, allowing the study of earlier visual processing. We did not use them here because they are usually used in studies where the stimuli are presented for shorter durations than 200 ms, however a repetition of this experiment with longer and shorter exposure times, with or without pattern masks, could perhaps yield illuminating results.

Another possible confound of the study was that the possibility that the 4 x 4 array of lines was too small or was composed of too few elements. Other studies (Ariely, 2001; Robitaille & Harris, 2011; Dodgson & Raymond, 2020) have shown that increasing the number of items of the same kind in an array improves performance. However, other studies have achieved positive results using an array of 16 lines arranged in a 4 x 4 grid (Im, Tiurina, & Utochkin, 2021). In order to better assess whether or not lower-level visual ensemble processing can be accessed by metacognition, this study should be repeated in a controlled lab environment, with eye-tracking to ensure fixation before each trial, with minimal distractions. Other manipulations that could be attempted include assessing performance with ensembles consisting of more and fewer items, ensembles with wider and narrower ranges of angles, or arrays consisting of the same amount of

items, but taking up more of the visual field, with either a corresponding increase in size of each of the elements in the array, or not.

Another potential future manipulation could be in regard to the metacognitive aspect of the study. For instance, performance could be compared between one group that had a confidence rating task and another that did not, in a counterbalanced fashion across participants. Other studies have shown that there may be a reactivity effect from probing metacognition (Rhodes & Tauber, 2011), though whether it has a positive or negative effect on performance tends to vary from study to study (as discussed in Bonder & Gopher, 2019). Other studies have also shown improvement in performance, relative to metacognition, after participants were given feedback on their performance (Bonder & Gopher, 2019). This can be incorporated in future iterations of this study. Another future approach can involve separating participants into groups based on assessments of their overall levels of confidence in their cognitive performance. The complexity of the stimuli in the ensembles can also be manipulated between groups, so that a comparison can be made between performance and metacognitive access across lower- (lines and different-colored shapes) and higher-level (faces and complex objects) visual stimuli. Finally, in line with RHT and studies regarding perceptual learning carried out by Ahissar and Hochstein, metacognition can also be probed with respect to adaptation to low-level stimuli over time, with difficult (short SOAs) conditions being bootstrapped by easy (long SOAs) conditions.

## Part 5. Conclusion

This study showed that there was no correlation between early ensemble perception and metacognition, suggesting that participants were not able to subjectively access the average orientation of all lines presented in an array to make judgements about their own performance. There were also no correlations or very weak correlations between other aspects of the

experiment. However, slightly stronger correlations were found at the group level that suggested that spending too much time adjusting the probe or adjusting the probe over a greater distance can have detrimental effects on performance. Perhaps repetitions of this experiment in the future, using a variety of different parameters, may serve to elaborate how cognitive processes and early perception interact with one another. However, at present, the relationship between metacognition and ensemble perception remains poorly defined.

## References


Ahissar, M., & Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. Trends in cognitive sciences, 8(10), 457-464.

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.

Albrecht, A. R., Scholl, B. J., & Chun, M. M. (2012). Perceptual averaging by eye and ear: Computing summary statistics from multimodal stimuli. Attention, Perception, & Psychophysics, 74, 810-815.

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. Trends in cognitive sciences, 15(3), 122-131.

Alvarez, G. A., & Franconeri, S. L. (2007). How many objects can you track?: Evidence for a resource-limited attentive tracking mechanism. Journal of vision, 7(13), 14-14.

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.

Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. Psychological science, 19(4), 392-398. https://doi.org/10.1111/j.1467-9280.2008.02098.x

Ariely, D. (2001). Seeing sets: Representation by statistical properties. Psychological science, 12(2), 157-162.

Bauer, B. (2009). Does Stevens's power law for brightness extend to perceptual brightness averaging? The Psychological Record, 59, 171-185.

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. Psychological science, 22(3), 384-392.

Broadbent, D. E. (1958). Perception and Communication. London: Pergamon Press.

Bonder, T., & Gopher, D. (2019). The effect of confidence rating on a primary visual task. Frontiers in Psychology, 10, 2674. doi:10.3389/fpsyg.2019.02674

Burr, D., & Ross, J. (2008). A visual sense of number. Current biology, 18(6), 425-428. http://dx.doi.org/10.1016/j.cub.2008.02.052

Cavanagh, P. (2001). Seeing the forest but not the trees. nature neuroscience, 4(7), 673-674. https://doi.org/10.1038/89436

Cavanagh, P., & Alvarez, G. A. (2005). Tracking multiple targets with multifocal attention. Trends in cognitive sciences, 9(7), 349-354.

Chong, S. C., & Treisman, A. (2005). Statistical processing: Computing the average size in perceptual groups. Vision research, 45(7), 891-900. https://doi.org/10.1016/j.visres.2004.10.004

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

Cohen, M. A., Dennett, D. C., & Kanwisher, N. (2016). What is the bandwidth of perceptual experience? Trends in cognitive sciences, 20(5), 324-335. doi: 10.1016/j.tics.2016.03.006

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

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

Dodgson, D. B., & Raymond, J. E. (2020). Value associations bias ensemble perception. Attention, Perception, & Psychophysics, 82, 109-117.

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. Psychological Review, 96(3), 433–458. https://doi.org/10.1037/0033-295X.96.3.433

Exner, S. (1868). On the time necessary for a visual perception. Proceedings of meetings of the Imperial Academy of Sciences: Mathematical-Natural Scientific Classe, 58. (Ger).

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. Cerebral cortex (New York, NY: 1991), 1(1), 1-47.

Fischer, J., & Whitney, D. (2011). Object-level visual information gets through the bottleneck of crowding. Journal of Neurophysiology, 106(3), 1389-1398. http://dx.doi.org/10.1152/jn.00904.2010

Friedman-Hill, S., & Wolfe, J. M. (1995). Second-order parallel processing: visual search for the odd item in a subset. Journal of Experimental Psychology: Human Perception and Performance, 21(3), 531.

Gegenfurtner, K. R., & Kiper, D. C. (2003). Color vision. Annual review of neuroscience, 26(1), 181-206.

Haberman, J., Brady, T. F., & Alvarez, G. A. (2015). Individual differences in ensemble perception reveal multiple, independent levels of ensemble representation. Journal of Experimental Psychology: General, 144(2), 432. https://doi.org/10.1037/xge0000053

Haberman, J., Harp, T., & Whitney, D. (2009). Averaging facial expression over time. Journal of vision, 9(11), 1-1. http://dx.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), 16-16. https://doi.org/10.1167/15.4.16

Haberman, J., & Whitney, D. (2012). Ensemble perception: Summarizing the scene and broadening the limits of visual processing. From perception to consciousness: Searching with Anne Treisman, 339-349.

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

Haberman, J., & Whitney, D. (2009). Seeing the mean: ensemble coding for sets of faces. Journal of Experimental Psychology: Human Perception and Performance, 35(3), 718.

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. Current biology, 17(17), R751-R753. http://dx.doi.org/10.1016/j.cub.2007.06.039

Halberda, J., Sires, S. F., & Feigenson, L. (2006). Multiple spatially overlapping sets can be enumerated in parallel. Psychological science, 17(7), 572-576. http://dx.doi.org/10.1111/j.1467-9280.2006.01746.x

Hochstein, S., & Ahissar, M. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. Neuron, 36(5), 791-804.

Horowitz, T. S., Klieger, S. B., Fencsik, D. E., Yang, K. K., Alvarez, G. A., & Wolfe, J. M. (2007). Tracking unique objects. Perception & psychophysics, 69, 172-184.

Howe, P. D., Cohen, M. A., Pinto, Y., & Horowitz, T. S. (2010). Distinguishing between parallel and serial accounts of multiple object tracking. Journal of Vision, 10(8), 11-11.

Hubel, D. H., & Wiesel, T. N. (1977). Ferrier lecture-Functional architecture of macaque monkey visual cortex. Proceedings of the Royal Society of London. Series B. Biological Sciences, 198(1130), 1-59.

Im, H. Y., Tiurina, N. A., & Utochkin, I. S. (2021). An explicit investigation of the roles that feature distributions play in rapid visual categorization. Attention, Perception, & Psychophysics, 83, 1050-1069.

Ji, L., & Hayward, W. G. (2021). Metacognition of average face perception. Attention, Perception, & Psychophysics, 83, 1036-1048.

Khayat, N., & Hochstein, S. (2019). Relating categorization to set summary statistics perception. Attention, Perception, & Psychophysics, 81, 2850-2872.

Khayat, N., & Hochstein, S. (2018). Perceiving set mean and range: Automaticity and precision. Journal of vision, 18(9), 23-23.

McDermott, J. H., Schemitsch, M., & Simoncelli, E. P. (2013). Summary statistics in auditory perception. Nature neuroscience, 16(4), 493-498.

Moutoussis, K., & Zeki, S. (1997). Functional segregation and temporal hierarchy of the visual perceptive systems. Proceedings of the Royal Society of London. Series B: Biological Sciences, 264(1387), 1407-1414.

Nelson, T. O. (1996). Consciousness and metacognition. American psychologist, 51(2), 102.

Neumann, M. F., Schweinberger, S. R., & Burton, A. M. (2013). Viewers extract mean and individual identity from sets of famous faces. Cognition, 128(1), 56-63. http://dx.doi.org/10.1016/j.cognition.2013.03.006

Noë, A., & O'Regan, J. K. (2000). Perception, attention, and the grand illusion. Psyche, 6(15), 6-15.

O'regan, J. K. (1992). Solving the" real" mysteries of visual perception: the world as an outside memory. Canadian Journal of Psychology/Revue canadienne de psychologie, 46(3), 461.

Oliva, A., & Schyns, P. G. (1997). Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. Cognitive psychology, 34(1), 72-107.

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. Trends in cognitive sciences, 11(12), 520-527.

Oksama, L., & Hyönä, J. (2004). Is multiple object tracking carried out automatically by an early vision mechanism independent of higher-order cognition? An individual difference approach. Visual cognition, 11(5), 631-671.

Oriet, C., & Brand, J. (2013). Size averaging of irrelevant stimuli cannot be prevented. Vision Research, 79, 8-16.

Pallier, G., Wilkinson, R., Danthiir, V., Kleitman, S., Knezevic, G., Stankov, L., & Roberts, R. D. (2002). The role of individual differences in the accuracy of confidence judgments. The Journal of general psychology, 129(3), 257-299.

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

Piazza, E. A., Sweeny, T. D., Wessel, D., Silver, M. A., & Whitney, D. (2013). Humans use summary statistics to perceive auditory sequences. Psychological science, 24(8), 1389-1397.

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. Spatial vision, 3(3), 179-197.

Reynolds, J. H., & Chelazzi, L. (2004). Attentional modulation of visual processing. Annu. Rev. Neurosci., 27, 611-647.

Rhodes, M. G., & Tauber, S. K. (2011). The influence of delaying judgments of learning on metacognitive accuracy: a meta-analytic review. Psychological bulletin, 137(1), 131.

Robitaille, N., & Harris, I. M. (2011). When more is less: Extraction of summary statistics benefits from larger sets. Journal of vision, 11(12), 18-18.

Rock, I., Linnett, C. M., Grant, P., & Mack, A. (1992). Perception without attention: Results of a new method. Cognitive psychology, 24(4), 502-534. https://doi.org/10.1016/0010-0285(92)90017-V

Rousselet, G. A., Thorpe, S. J., & Fabre-Thorpe, M. (2004). How parallel is visual processing in the ventral pathway? Trends in cognitive sciences, 8(8), 363-370.

Schyns, P. G., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time-and spatial-scale-dependent scene recognition. Psychological science, 5(4), 195-200.

Simons, D. J., & Levin, D. T. (1997). Change blindness. Trends in cognitive sciences, 1(7), 261-267. https://doi.org/10.1016/S1364-6613(97)01080-2

Smith, A. R., & Price, P. C. (2010). Sample size bias in the estimation of means. Psychonomic Bulletin & Review, 17, 499-503.

Sweeny, T. D., Haroz, S., & Whitney, D. (2012). Reference repulsion in the categorical perception of biological motion. Vision research, 64, 26-34. https://doi.org/10.1016/j.visres.2012.05.008

Tark, K. J., Kang, M. S., Chong, S. C., & Shim, W. M. (2021). Neural representations of ensemble coding in the occipital and parietal cortices. NeuroImage, 245, 118680.

Thomas, L. E., & Seiffert, A. E. (2010). Self-motion impairs multiple-object tracking. Cognition, 117(1), 80-86.

Treisman, A. M. (2006). How the deployment of attention determines what we see, Visual Cognition, 14:4-8, 411-443, doi: 10.1080/13506280500195250

Treisman, A. M. (1964). Selective attention in man. British Medical Bulletin, 20(1), 12–16.

Utochkin, I. S. (2015). Ensemble summary statistics as a basis for rapid visual categorization. Journal of Vision, 15(4), 8-8. doi: 10.1167/15.4.8

Utochkin, I. S., & Tiurina, N. A. (2014). Parallel averaging of size is possible but range-limited: A reply to Marchant, Simons, and De Fockert. *Acta psychologica*, *146*, 7-18.

Viswanathan, L., & Mingolla, E. (2002). Dynamics of attention in depth: Evidence from multi-element tracking. Perception, 31(12), 1415-1437.

Watamaniuk, S. N. (1993). Ideal observer for discrimination of the global direction of dynamic random-dot stimuli. JOSA A, 10(1), 16-28. http://dx.doi.org/10.1364/JOSAA.10.000016

Watamaniuk, S. N., & Duchon, A. (1992). The human visual system averages speed information. Vision research, 32(5), 931-941. http://dx.doi.org/10.1016/0042-6989(92)90036-I

Watamaniuk, S. N., Sekuler, R., & Williams, D. W. (1989). Direction perception in complex dynamic displays: the integration of direction information. Vision research, 29(1), 47-59.

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, 105–129. https://doi.org/10.1146/annurev-psych-010416-044232

Yurevich, M. A., & Utochkin, I. S. (2014). Distractor heterogeneity effects in visual search are mediated by "segmentability". Journal of Vision, 14(10): 921, doi:10.1167/14.10.921. [Abstract]

Zeki, S. (1983). Colour coding in the cerebral cortex: the reaction of cells in monkey visual cortex to wavelengths and colours. *Neuroscience*, *9*(4), 741-765. https://doi.org/10.1016/0306-4522(83)90265-8

Zhai, X., Khatami, F., Sadeghi, M., He, F., Read, H. L., Stevenson, I. H., & Escabí, M. A. (2020). Distinct neural ensemble response statistics are associated with recognition and discrimination of natural sound textures. Proceedings of the National Academy of Sciences, 117(49), 31482-31493.