The Neural Representation of Ensemble Mean.

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Psychology

by

Kyle W. Killebrew

Dr. Gideon P. Caplovitz/Dissertation Advisor

May, 2019
THE GRADUATE SCHOOL

We recommend that the dissertation prepared under our supervision by

Kyle W. Killebrew

Entitled

The Neural Representation of Ensemble Mean

be accepted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Gideon P. Caplovitz, Ph. D , Advisor

Marian E. Berryhill, Ph. D , Committee Member

Lars Strother, Ph. D , Committee Member

Timothy J. Vickery, Ph. D. , Committee Member

Holly Hazlett-Stevens, Ph.D , Committee Member

Benjamin D. Young, Ph. D., Graduate School Representative

David W. Zeh, Ph.D., Dean, Graduate School

May-2019
Abstract

Our perceptual systems are capacity limited by the bottleneck of attention and, as a result, we can only process a limited amount of information at any given time. In order to help overcome this limitation, our perceptual systems can quickly summarize and extract information over a large area of visual space. In other words, we have the remarkable ability to extract the ‘gist’ of a scene or group of objects. Ensemble encoding, a proxy of gist perception, is the ability to rapidly extract the average feature of a group of items. For example, people can extract the average orientation, size, direction of motion, hue, and even facial expression among a group of similar objects. This ability has been demonstrated behaviorally many times using many different experimental paradigms. However, little is known about how these ensemble averages are extracted and how they are neurally encoded. We predict that if there is a representation of the ensemble, we can measure it in response to systematically varying the average feature of a group objects using high-density electroencephalography (hdEEG) and functional magnetic resonance imaging (fMRI). Specifically, the current series of experiments attempts to identify the neural correlates and temporal dynamics of ensemble encoding of orientation and size as well as measuring changes to that representation by manipulating spatial attention and the type of averaging task performed by the participant. In experiment 1, we measured neural adaptation to repeated presentations of adapting ensembles with a reference average orientation and size and test ensembles of progressively larger or more tilted averages using fMRI repetition suppression. In
experiment 2, we used hdEEG to measure evoked potentials in response to ensembles of framed ellipses with different mean sizes and orientations. We then performed univariate and multivariate analysis in an attempt to find differences over time between the signals of these ensembles. In experiment 3, we attempted to tease out the effects of attention and relevant averaging task on the representation of these ensemble averages. We used a multiplexed frequency tagging oddball paradigm in which we ‘tagged’ ensembles by flickering them at specific frequencies. We then transform the EEG waveforms from the time to frequency domain using a fast Fourier transform and measure the resulting amplitude to the specific presentation frequency. Although we do see some results consistent with the view of ensemble encoding as a rapid parallel process, our results largely show no consistent differentiable response in the neural signal between ensembles of different levels. Our data are most consistent with a theory of ensemble encoding as an encoding strategy as opposed to a pre-attentive, automatic, and parallel process. More work will need to be done in order to make a firmer conclusion about the neural representation of the ensemble average.

Keywords: Ensemble encoding, summary statistics, attention, hdEEG, fMRI, MVPA.
Acknowledgments

Science is a collaborative process. As such, I would never have been able to make it this far without the seemingly unlimited amount of support and guidance I’ve received from so many people along the way. First, I would like to thank my mentor and role model Dr. Gideon Caplovitz. His endless support, guidance, and patience has been invaluable. I would also like to thank Dr. Marian Berryhill who has been like a second advisor to me throughout my graduate career.

No lab would produce much without a cast of great individuals. I want to thank every member of the C-Lab, both past and present. Specifically, Dr. Christopher Blair, Dr. Gennadiy Gurariy, and Taissa Lytchenko have been a constant source of support and have helped me through countless problems, both in and out of lab.

Finally, I would not have made it very far without the steadfast support of my family. I would like to thank my mother for pushing me to excel throughout my life. My father, for providing valuable advice to help guide me in times of difficulty. Lastly, my grandparents for always being there to talk whenever I needed them.
**Table of Contents**

Introduction.................................................................................................................................................. 1

General Introduction and Specific Aims ..................................................................................................... 1

Aim 1: Identify neural correlates of ensemble averaging using fMRI adaptation................................................. 2

Aim 2: Explore the pattern of activity in processing feature averages over time using EEG and MVPA................................................................. 3

Aim 3: Identify the effects of attention and task on the representation of the average using EEG and a frequency tagging oddball paradigm............................................. 3

A history of perceiving the average ........................................................................................................... 4

Features of ensemble encoding ................................................................................................................ 6

From low- to high- level feature averages .................................................................................................. 7

Other summary statistics .......................................................................................................................... 9

Stimulus factors influencing ensemble encoding ....................................................................................... 12

The neural representation of the average.................................................................................................. 17

Proposed mechanisms in ensemble encoding.......................................................................................... 17

Attentional requirements of ensemble processing.................................................................................... 21

The neural correlates of the average......................................................................................................... 23

What cognitive resources are needed to extract the average? ................................................................. 27

Strong ensemble encoding ..................................................................................................................... 28

Weak ensemble encoding 1 ....................................................................................................................... 31

Weak ensemble encoding 2 ....................................................................................................................... 31
Ensemble encoding as a strategy ................................................................. 32
Outstanding questions and proposed experiments ........................................ 33
General methodology and stimulus parameters ........................................... 34
Introduction ................................................................................................. 34
General stimulus parameters ........................................................................ 35
Controlling confounds for overall ensemble configuration ............................ 38
Psychophysical experiment: Identifying individual’s thresholds for ensemble averaging of two stimulus features ................................................................. 41
Introduction and experimental paradigm ....................................................... 41
Data analysis .................................................................................................. 42
Experiment 1: Identify neural correlates of ensemble averaging using fMRI
repetition suppression .................................................................................. 43
Introduction .................................................................................................... 43
Methods ........................................................................................................ 44
Participants ..................................................................................................... 44
Apparatus ....................................................................................................... 44
MRI apparatus and scanning procedure ......................................................... 45
Retinotopy and anatomical scans .................................................................. 45
Experimental scans ....................................................................................... 46
Stimuli ............................................................................................................. 47
Experimental procedure and task design ....................................................... 47
Behavioral data analysis ................................................................................ 49
Experiment 2: Exploring the pattern of activity in processing feature averages across time using EEG and MVPA

Introduction

Methods

Participants

Apparatus

EEG Data Acquisition

Stimuli

Experimental procedure and task design

Behavioral data analysis

EEG data preprocessing

Univariate analysis

Multivariate analysis
## Experiment 3: Identify the effects of attention and task on the representation of the average using EEG and a frequency tagging oddball paradigm.

### Introduction

Introduction

### Methods

Methods

#### Participants

Participants

#### Apparatus

Apparatus

#### EEG Data Acquisition

EEG Data Acquisition

#### Stimuli

Stimuli

#### Experimental procedure and task design

Experimental procedure and task design

#### The multiplexed frequency tagging oddball paradigm

The multiplexed frequency tagging oddball paradigm

#### Behavioral data analysis

Behavioral data analysis

#### EEG data preprocessing

EEG data preprocessing

#### Carrier frequency analysis

Carrier frequency analysis

#### Sanity check: Are participants awake?

Sanity check: Are participants awake?
Sanity check: Are participants paying attention? ........................................107
Oddball frequency analysis........................................................................108
Feature levels vs Δ levels ..............................................................................108
Frequency tag vs. index value analysis..........................................................109
Four types of ensemble encoding................................................................110
Strong ensemble encoding..........................................................................110
Weak ensemble encoding 1 .........................................................................111
Weak ensemble encoding 2 .........................................................................113

Results ..............................................................................................................115
Behavioral results............................................................................................115
Carrier frequency analysis............................................................................117
\( T_{\text{circ}}^2 \) analysis......................................................................................118
Are participants paying attention? .................................................................120
Oddball analysis...............................................................................................123
Strong ensemble encoding............................................................................123
Feature level analysis......................................................................................123
Δ level analysis ...............................................................................................126
Weak ensemble encoding 1 .........................................................................129

Is the allocation of spatial attention needed to extract the ensemble, regardless
of attend feature? ............................................................................................131

Feature level analysis ......................................................................................131
Δ level analysis ...............................................................................................136
Is the presence of an explicit averaging task necessary to extract the ensemble, regardless of spatial attention? ................................. 141

Feature level analysis ................................................................................................................. 142

Δ level analysis ............................................................................................................................ 146

Weak ensemble encoding 2 ........................................................................................................... 151

No task attended vs. no task unattended ...................................................................................... 152

Feature level analysis ................................................................................................................. 152

Δ level analysis ............................................................................................................................ 158

Task attended vs. Task unattended .............................................................................................. 162

Feature level analysis ................................................................................................................. 163

Δ level analysis ............................................................................................................................ 168

Task attended vs No task unattended ......................................................................................... 172

Feature level analysis ................................................................................................................. 173

Δ level analysis ............................................................................................................................ 178

Experiment 3 summary ................................................................................................................. 183

General Discussion ....................................................................................................................... 185

The representation of the ensemble .............................................................................................. 186

The neural correlates of the average .......................................................................................... 187

Temporal dynamics of ensemble encoding .................................................................................. 189

Is ensemble encoding automatic and pre-attentive? ................................................................. 190

Ensemble encoding as a strategy ................................................................................................. 193

What is the neural representation of the average and why do we extract it? ........................... 195
Limitations ................................................................. 197

Conclusions ........................................................................ 199

References ............................................................................. 200
List of Figures

Figure 1. Hypothetical predictions in support of ensemble encoding for each proposed hypothesis. .......................................................... 28

Figure 2. Hypothetical data for experiment 1 for the strong ensemble encoding hypothesis. .......................................................... 30

Figure 3. Example ensemble stimuli for minimum and maximum size and orientation levels. .......................................................... 37

Figure 4. Example participant psychometric functions. ................................................. 43

Figure 5. Experiment 1 trial sequence for one block. ................................................. 49

Figure 6. Early visual dorsal and ventral regions of interest drawn for one participant. ............................................................................ 53

Figure 7. Experiment 1 behavioral results. ................................................................. 58

Figure 8. ROI analysis results for orientation .......................................................... 61

Figure 9. ROI analysis results for size .................................................................... 62

Figure 10. ROI analysis results z-score normalized for orientation ....................... 63

Figure 11. ROI analysis results z-score normalized for size. ................................. 64

Figure 12. ROI analysis for orientation thresholded by orientation > size. ............ 66

Figure 13. ROI analysis for size thresholded by size > orientation ....................... 67

Figure 14. Whole brain contrast for orientation level 1 vs. levels 2, 3, and 4 ............ 69

Figure 15. Whole brain contrast for size level 1 vs. level 2, 3, and 4 ....................... 70

Figure 16. Experiment 2 example trial sequences. .................................................. 74

Figure 17. Experiment 2 behavioral results. ............................................................ 82
Figure 18. Experiment 2 test statistics plotted over time for all electrodes for orientation. .............................................................. 84

Figure 19. Experiment 2 test statistics plotted over time for all electrodes for size. 85

Figure 20. Orientation split-half correlations. .............................................................. 87

Figure 21. Size split half correlations.............................................................. 88

Figure 22. Differences between within and between conditions split half correlations for orientation.............................................................. 89

Figure 23. Differences between within and between conditions split half correlations for size.............................................................. 90

Figure 24. Split half correlations of the z-scored VEPs for the orientation task........ 91

Figure 25. Split half correlations of the z-scored VEPs for the size task.............. 92

Figure 26. Orientation LDA classification accuracy.................................................. 94

Figure 27. Size LDA classification accuracy.................................................. 95

Figure 28. Orientation Naïve Bayes classification accuracy.................................. 96

Figure 29. Size Naïve Bayes classification accuracy.................................. 97

Figure 30. Experiment 3 frequency tagging trial sequence.................................. 102

Figure 31. Multiplexed frequency tagging oddball paradigm.............................. 104

Figure 32. Experiment 3 behavioral results.............................................................. 117

Figure 33. Example $T_{\text{circ}}^2$ for one participant for orientation.......................... 119

Figure 34. Example $T_{\text{circ}}^2$ for one participant for size.............................. 120

Figure 35. Carrier frequencies: Attended-Unattended.................................. 122

Figure 36. Feature level linear contrast – Orientation.................................................. 125
Figure 37. Feature level linear contrast – Size. .......................................................... 126

Figure 38. Δ level linear contrast – Orientation. .......................................................... 128

Figure 39. Δ level linear contrast – Size. ................................................................. 129

Figure 40. Attended vs. unattended – Frequency tags – Feature levels – Orientation. .......................................................................................................................... 133

Figure 41. Attended vs. unattended – Frequency tags - Feature levels – Size. ....... 134

Figure 42. Attended vs. unattended – Index values - Feature levels – Orientation. 135

Figure 43. Attended vs. unattended – Index values – Feature levels – Size........... 136

44. Attended vs. unattended – Frequency tags - Δ levels – Orientation................. 138

45. Attended vs. unattended – Frequency tags - Δ levels – Size.............................. 139

46. Attended vs. unattended – Index values - Δ levels – Orientation...................... 140

47. Attended vs. unattended – Index values - Δ levels – Size................................. 141

Figure 48. Task vs. no task – Frequency tags – Feature levels – Orientation......... 143

Figure 49. Task vs. no task – Frequency tags - Feature levels – Size....................... 144

Figure 50. Task vs. no task – Index values - Feature levels – Orientation............... 145

Figure 51. Task vs. no task – Index values – Feature levels – Size.......................... 146

Figure 52. Task vs. no task – Frequency tags - Δ levels – Orientation.................... 148

Figure 53. Task vs. no task – Frequency tags - Δ levels – Size.............................. 149

Figure 54. Task vs. no task – Index values - Δ levels – Orientation....................... 150

Figure 55. Task vs. no task – Index values - Δ levels – Size................................. 151

Figure 56. No task attended vs. no task unattended – Frequency tags – Feature levels – Orientation. .............................................................. 154
Figure 57. No task attended vs. no task unattended – Frequency tags - Feature levels - Size..................................................................................................................................................155

Figure 58. No task attended vs. no task unattended – Index values - Feature levels – Orientation........................................................................................................................................................................156

Figure 59. No task attended vs. no task unattended – Index values – Feature levels – Size. ........................................................................................................................................................................................................................................157

Figure 60. No task attended vs. no task unattended – Frequency tags - ∆ levels – Orientation........................................................................................................................................................................................................................................159

Figure 61. No task attended vs. no task unattended – Frequency tags - ∆ levels – Size. ........................................................................................................................................................................................................................................................................................................160

Figure 62. No task attended vs. no task unattended – Index values - ∆ levels – Orientation........................................................................................................................................................................................................................................................................................................161

Figure 63. No task attended vs. no task unattended – Index values - ∆ levels – Size. ........................................................................................................................................................................................................................................................................................................162

Figure 64. Task attended vs. task unattended – Frequency tags – Feature levels – Orientation........................................................................................................................................................................................................................................................................................................164

Figure 65. Task attended vs. task unattended – Frequency tags – Feature levels – Size. ........................................................................................................................................................................................................................................................................................................165

Figure 66. Task attended vs. task unattended – Index values – Feature levels – Orientation........................................................................................................................................................................................................................................................................................................166

Figure 67. Task attended vs. task unattended – Index values – Feature levels – Size. ........................................................................................................................................................................................................................................................................................................167
Figure 68. Task attended vs. task unattended – Frequency tags – Δ levels – Orientation..........................................................169

Figure 69. Task attended vs. task unattended – Frequency tags – Δ levels – Size....170

Figure 70. Task attended vs. task unattended – Index values – Δ levels – Orientation.
............................................................................................................................................................................171

Figure 71. Task attended vs. task unattended – Index values – Δ levels – Size.......172

Figure 72. Task attended vs. no task unattended – Frequency tags – Feature levels – Orientation............................................................................................................................................................................174

Figure 73. Task attended vs. no task unattended – Frequency tags – Feature levels – Size. .................................................................................................................................................................................................................175

Figure 74. Task attended vs. no task unattended – Index values – Feature levels – Orientation....................................................................................................................................................................................................................176

Figure 75. Task attended vs. no task unattended – Index values – Feature levels – Size. ....................................................................................................................................................................................................................177

Figure 76. Task attended vs. no task unattended – Frequency tags – Δ Levels – Orientation....................................................................................................................................................................................................................179

Figure 77. Task attended vs. no task unattended – Frequency tags – Δ Levels – Size.
....................................................................................................................................................................................................................................................................................................................................................................................................................180

Figure 78. Task attended vs. no task unattended – Index values – Δ Levels – Orientation....................................................................................................................................................................................................................181

Figure 79. Task attended vs no task unattended – Index values – Δ Levels – Size.182
Figure 80. Individual Δ levels for the first three harmonics of the frequency tags in the orientation no task unattended condition.
Introduction

General Introduction and Specific Aims

Human vision is capacity limited by the bottleneck of attention and yet some information, that doesn’t make it past this bottleneck, still effects conscious perception and cognition (Alvarez, 2011a; Whitney & Yamanishi Leib, 2018). This is in part due to our ability to extract summary statistics about features present in the environment, like average orientation, size, facial expression, and life-likeness (Whitney & Yamanashi Leib, 2018). For example, many people have had the experience of being in front of a crowded room and detecting the audience’s overall mood or identifying where they are looking, despite not looking at any individual in particular. These ensemble statistics have been proposed to serve a variety of different functions, such as rapid scene processing (Utochkin, 2015; Brady, Shafer-Skelton, & Alvarez, 2017; Alvarez & Oliva, 2008), popout in visual search (Rosenholtz, Huang, Raj, Balas, & Ilie, 2012), and the richness of experience (Cohen, Dennett, & Kanwisher, 2016). Much less is known about the neural representations of ensemble averages. Therefore, there is a need to systematically investigate their neural representation (Whitney & Yamanishi Leib, 2018).

The goal of this proposal is to identify a common measure that can be used to characterize the representation of the average across methodologies and stimulus types, specifically, a measure of sensitivity, or a neurometric function. The central hypothesis examined here states that if there is a neural representation for an
average feature in an ensemble it can be measured using electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) and manipulated by focused attention and the presence of an explicit averaging task. We will show participants groups of ellipses with differing average sizes and orientations while measuring neural activity. Using multivariate pattern analysis (MVPA), we will create neurometric functions by comparing activity to a standard ensemble to ensembles with varying degrees of difference in their mean orientation and size. If successful, this will provide an objective measure of ensemble encoding that is task-irrelevant and can be compared across methodologies and ensemble features. With new theories in the literature regarding the purpose of extracting feature averages, this research will provide a measure of the neural representation that can be used to further our understanding of how the brain can break the bottleneck of attention.

The following sections outline the specific aims of the proposed project followed by a brief review of past and present research on ensemble averaging.

**Aim 1: Identify neural correlates of ensemble averaging using fMRI adaptation.**

First, we used ROI-based fMRI techniques to measure BOLD signals across visual cortex in response to a stimulus paradigm designed to reveal correlates of feature-specific (size or orientation) and feature-general (size and orientation) neural tuning at the level of the ensemble when such ensembles are either task-relevant or task-irrelevant. If successful this aim will reveal tuning characteristics of
neural representations of ensembles and identify candidate regions of visual cortex where such neural populations may exist. We will use this information to both constrain and inform models of ensembles and guide the interpretation of further empirical data collected in Aims 2 and 3.

**Aim 2: Explore the pattern of activity in processing feature averages over time using EEG and MVPA.**

Next, we explored the time course of activity that emerges following ensemble averaging. We collected evoked potentials in response to ensembles with means of various sizes and orientations. We used univariate analysis, split-half correlations, and multivariate pattern analysis to attempt to discover when in time ensemble representations emerge in both task-relevant and task-irrelevant contexts. This can provide evidence for domain-general or domain-specific mechanisms by exploring when in time these representations emerge for extracting averages for different features.

**Aim 3: Identify the effects of attention and task on the representation of the average using EEG and a frequency tagging oddball paradigm.**

Finally, to further explore the neural representation of the average, we examined how changes to the focus of attention affect the sensitivity to representations of the average as a function of task relevance. We did this by presenting two sets of ensembles simultaneously to either hemifield. We measured
effects of spatial attention and averaging task on the representation of the mean size and orientation using a multiplexed frequency tagging oddball paradigm. This can inform models of mean extraction by identifying the neural effects of attention and task on the representation of the mean.

A history of perceiving the average

Work done on investigating the ability to average information dates as far back as William James in his landmark work, The Principles of Psychology. In a section titled The Summation of the Sense-spaces he gives one of the first descriptions of ensemble encoding: ‘Whatever sensible data can be attended to together we locate together. Their several extents seem one extent. The place at which each appears is held to be the same with the place at which the others appear. They become, in short, so many properties of one and the same real thing’ (James, 1890). Another early scientist described a similar phenomenon, which they term constructive combination, or, ‘the combination of different qualities to form a single object’ (Messenger, 1903). Attneave gives us another early insight into this averaging ability. He describes how individuals might perceive a random dot field, in which there is simply too much information for a person to individually represent: ‘It appears, then, that when some portion of the visual field contains a quantity of information grossly in excess of the observer’s perceptual capacity, he treats those components of information which do not have redundant representation somewhat as a statistician treats ‘error variance,’ averaging out particulars and abstracting certain statistical homogeneities’ (Attneave, 1954). In these early discussions, researchers describe how we parse apart perceptual information to make sense of the enormous amount of data we have access to. Although these descriptions may not directly address ensemble encoding, in the way we
have come to think about it today, they do touch upon many key points that define our modern version of ensemble encoding, like identifying the average features and locations of common elements in our perceptual environment.

Although this very early work is more theoretical and does not directly test participant’s ability to extract feature statistics, it does lay an early foundation for later researchers to build upon. Since these early descriptions, researchers have more systematically and empirically studied this phenomenon. This work investigates perceptual averaging using fairly simple paradigms that attempted to find midpoint values of various stimulus features. For example, researchers found that participants are remarkably consistent when judging the midpoint gray when mixing different amounts of white and black paint (Laming & Laming, 1996), or finding the midpoint of a line (Pearson, 1922), and the mid-tempo between two metronomes set at different rates (Wallin, 1912; for a comprehensive review, see: Bauer, 2015). Again, these early researchers do not necessarily refer to this ability as ensemble averaging, but ultimately, these observations strongly support the claim that participants can form basic perceptual averages.

Using these findings, researchers today have delved more deeply into understanding the properties that govern our ability to average and why we might extract these averages. For example, one report, calls this ability intuitive statistics and significantly broadened the scope of the initial midpoint averaging research (Peterson & Beach, 1967). In their review, they discuss how humans need fast and efficient representations of perceptual information that may not be directly available to their consciousness. We need this information as a basis for making decisions and inferences about future events. The researchers go on to describe human performance on a wide variety of descriptive and inferential statistical tasks, such as measures of proportion, central tendency, variance, and
correlations. They conclude, that while human observers generally fail to meet optimally efficient standards, in comparison to an ideal observer or statistician, they do perform relatively well across these types of statistical and predictive tasks. Today, researchers have uncovered a wide variety of functions that ensemble averages might support. Although this proposal will discuss such functions, first, it is important to understand the stimulus characteristics that underlie ensemble encoding.

**Features of ensemble encoding**

Modern research has more thoroughly investigated the range of stimulus features for which human observers can perform statistical operations, although this research has almost exclusively focused on the arithmetic mean. This research may be incomplete, in as much as the scope of statistical operations humans can perform, but it is incredibly important in informing later theories investigating the purpose of extracting averages. For example, it has been known for many years that humans can pool information across certain low level visual or auditory information, like local color, motion, orientation, or auditory tones (for example: Pasternak, Albano, & Harvitt, 1990; Dakin & Watt, 1997). When we think about averaging in perception it should not be surprising that observers are able to perform these types of tasks. After all, the basis of perception is the conjoining of information into separate objects or background and foreground. In order to do these basic perceptual tasks, we must group together stimulus features across space (and time), which must require some kind of averaging or compression of information. For example, when viewing a tree out of the window, we typically do not perceive the tree as consisting of a variety of shades of green, brown, or yellow, we perceive it as one consistent color. Similarly, with shape, we do not perceive the orientations of every leaf and branch. Instead,
we perceive an upright, although slightly tilted, tree. Therefore, it shouldn’t come as a surprise that low level perceptual information can be pooled or averaged. However, the fact that we can also take averages across a wide array of mid and high level information, which do not themselves have distinct neural feature detectors, such as size, facial expression, and lifelikeness, makes this averaging ability more intriguing. This also suggests that in addition to simply extracting these feature averages, we make use of them in ways that go far beyond simple perceptual organization. In the next sections, we describe the wide variety of features that can be extracted and averaged from an ensemble.

**From low- to high- level feature averages**

As stated above, researchers have demonstrated that human observers can efficiently extract averages for low-level perceptual features. One common low-level feature, orientation, has been researched thoroughly, and was one of the first feature dimensions shown to be averaged (Dakin & Watt, 1997; Parkes, Lund, Angelucci, Solomon, & Morgan, 2001; M. Morgan, Chubb, & Solomon, 2008; Dakin, 2001; Miller & Sheldon, 1969). Both average luminance (Bauer, 2009) and hue (Maule, Witzel, & Franklin, 2014; J. Webster, Kay, & Webster, 2014; Demeyere, Rzeskiewicz, Humphreys, & Humphreys, 2008) have been shown to be extracted from ensembles. Lastly, average motion direction (Watamaniuk & McKee, 1998; Watamaniuk, Sekuler, & Williams, 1989) and speed (Watamaniuk & Duchon, 1992) have been demonstrated.

Extracting averages for these low-level feature domains is not surprising as it has been known that these features have specific feature detecting neurons in early
visual cortex. However, many other features can be extracted that do not have specific known feature detectors. These are particularly interesting because, similarly to low-level features, they are extracted quickly and efficiently. This is in spite of the use of multiple stages of processing to identify individuals, let alone ensemble averages. Therefore, it becomes much more difficult to describe a model of averaging these higher level features, like simply pooling across feature detectors.

The most commonly studied mid-level feature is average size (Chong & Treisman, 2003, 2005; Solomon et al., 2018; Allik, Toom, Raidvee, Averin, & Kreegipuu, 2013; Ariely, 2001; Marchant, Simons, & de Fockert, 2013). Size perception is particularly interesting because it can be ambiguous. For example, detecting size also requires computations of depth and context. As size is particularly important to the current proposal we will provide a more formal description of what size is. Although at first glance size perception might seem trivial, the way in which we compute size is dependent on several factors (Irwin, 1969). For example, imagine looking at a car on a black background with no context or other visual information. As there is no other visual information to help you judge the size of the car, it becomes difficult to determine if the percept is of an up close toy car or normal sized car further away. Therefore, every stimulus has a physical or retinal size that represents the actual physical size of the object in space, or the corresponding area on the retina where its reflected light touches. However, this retinal size is often not predictive of its perceived size, which has been shown through various behavioral and neuroimaging studies examining various size
illusions (Fang, Boyaci, Kersten, & Murray, 2008; Murray, Boyaci, & Kersten, 2006). What is interesting about average size perception, is that it has been shown that this computed average size is not based on retinal size, but perceived size (Im & Chong, 2009), meaning that much more computation is required to extract average size than something like average orientation. Related to size, and complicating things further, average depth can also be extracted, leading to questions of how these average representations may interact (Wardle, Bex, Cass, & Alais, 2012).

A wide variety of high-level perceptual information can also be averaged. For example, average facial expression (Haberman & Whitney, 2007, 2009, 2018), animacy (Yamanishi Leib, Kosovicheva, & Whitney, 2016), direction of biological motion (Sweeny, Haroz, & Whitney, 2013), and attractiveness (van Osch, Blanken, Meijs, & van Wolferen, 2015; Walker & Vul, 2014) can be averaged. Similarly to mid-level features these do not have known feature detectors that represent individuals. Or to be more precise, they are represented across a wide array of not-easily defined feature dimensions (Grimaldi, Saleem, & Tsao, 2016). For example, facial expression has many feature dimensions that change between expressions. Therefore, to take averages across these high-level features, there must be high-level mechanisms that pool information across a wide array of feature dimensions.

**Other summary statistics**

The term summary *statistics* implies that there are more than one statistical dimension that can be extracted. Recent evidence shows that we can extract
measures of variance as well. For example, extraction of orientation variance (Dakin & Watt, 1997; M. Morgan et al., 2008; Norman et al., 2018) and size variance has demonstrated (Solomon et al., 2018). In one study, researchers demonstrated that participants can extract the orientation variance from a set of circles with a higher degree of accuracy compared to the mean size. It has also been demonstrated that observers are also sensitive to the variance in emotional expression of a group of faces (Haberman & Whitney, 2018). In our own work, we have investigated potential interactions between representations of mean size and the variance of size across an ensemble (Killebrew, Blair, & Caplovitz, 2014). We have found that in the presence of visual noise simulating sub-optimal viewing conditions observers tend to underestimate the variance of size across an ensemble, whereas mean-size in unaffected. We interpreted this result as suggesting that the mean size of an ensemble can be used to fill-in/substitute information about hard-to-resolve individuals, thereby making them all look more like the average and thereby reducing the overall sense of variability across the ensemble.

Although the nature and accuracy of these variance representations is still debated, it seems fairly clear that we can extract some measure of variance across different features. This is particularly important when thinking about the representation of the average. If variance is also efficiently extracted, it provides some evidence that the representation of the mean is a distribution of responses from a group of items (or a subset of that group), as this distribution would provide measures of both variance and mean. Although simply extracting mean is very
valuable and provides important information about the world, other statistical measures are equally important. For example, as Haberman, Lee, and Whitney describe, extracting the average facial expression of a crowd doesn’t inform you of the reliability of that average (Haberman & Whitney, 2018). One can imagine a mean expression of a crowd being happy or excited. However, just because the average is happy, does not mean that the variance isn’t quite large, meaning there are potentially still quite a few individuals with angry or sad expressions.

One important area of future research on this topic is to identify what types of statistics are being encoded. For example, one might expect that arithmetic mean is calculated. However, research suggests that this is in fact not the case (Utochkin & Vostrikov, 2017). In a clever experiment, researchers showed that observers extracted the numerosity and size of a set of objects separately and independently. This is a surprising result because the arithmetic mean requires the computation of the number of items being averaged. Therefore, it becomes extremely important, when looking for neural representations of ensemble averages, to test across different feature dimensions. This allows the differentiation between general or specific averaging mechanisms depending on the feature being averaged. For example, if the neural mechanism is a simple pooling mechanism, it seems unlikely that there would be only one mechanism that pools information across different feature domains. This means that the representations of averages for different features are also likely to be different. Additionally, because it is likely that through
a distribution of individual responses is how the mean is represented, it is also likely that mean and variance are not extracted separately.

**Stimulus factors influencing ensemble encoding**

Before proceeding, it is important to understand the stimulus dimensions necessary to extract averages from groups. In other words, in order to understand the function or representation of ensemble averages, we must first understand what constitutes an ensemble.

Is an ensemble representation different than that of an individual? In a thorough review of the literature, researchers attempt to define stimulus properties necessary for ensemble encoding (Whitney & Yamanishi Leib, 2018). The most basic of these, is that an ensemble requires at least two items to form an average. There is support for this argument. For example, there is a multitude of studies that show our ability to recognize an individual in an ensemble is much worse than our ability to recognize the average and, in fact, this method is a commonly used measure of ensemble averaging, also known as the membership identity task (for example see: Ariely, 2001; Ward, Bear, & Scholl, 2016; Whitney & Yamanishi Leib, 2018). However, this is somewhat dependent on the number of items within the ensemble and we would, for example, extract the identity of an individual in an ensemble of two much more accurately than in an ensemble of ten.

One area of research that could prove helpful for this discussion examines the capacity limits of both ensembles and individuals, either together or separately.
In other words, how many different representations of ensemble statistics can we extract and hold in mind compared to representations of individual features? One group of researchers explored our ability to extract two ensembles simultaneously (shape and texture). They found that when observers attended to global ensemble elements, changes to one ensemble feature (shape) affected detection performance on another ensemble feature (texture) (Cant, Sun, & Xu, 2015). When attending to local features, or individual items, in an ensemble, participants were not impaired in this task. However, other studies using find a different patterns of result (Luo & Zhao, Under Review; Attarha & Moore, 2015; Haberman, Brady, & Alvarez, 2015; Im & Chong, 2014). For example, studies examining the capacity limits of ensemble perception generally find a limit to the number of ensemble statistics that can be extracted, including extracting separate averages for the same feature across spatially overlapping ensembles and extracting averages of different features from the same ensemble (Luo & Zhao, Under Review; Attarha & Moore, 2015; Im & Chong, 2014; Corbett, 2017). This number tends to be very similar to the number of individuals that can be held in working memory or tracked using spatial attention, or 4+/−1 (Cowan, 2001; Pylyshyn & Storm, 1988). In another study researchers looked at observers ability to process ensembles that require different levels of processing (faces, orientation) (Haberman et al., 2015). They found that participant’s ability to extract an ensemble of one type is not correlated to their ability to extract an ensemble of another type. Additionally, although highly debated in the literature, there is evidence that attention is necessary to extract these
summary averages (Epstein & Emmanouil, 2017; Huang, 2015; although see section on ensembles and attention below).

It seems odd that a neural mechanism that averages information as a distribution of responses would not also respond to a single individual. Additionally, what if we have an ensemble of the exact same object? Shouldn’t this set of objects illicit the same response in an averaging mechanism as one item, albeit a stronger response? If we assume that the representation of the ensemble is just a distribution of responses, then should that representation also hold for an ‘ensemble of one?’

Answers to what the ideal ensemble consists of will require more thorough experimental testing. Neural measures developed throughout this dissertation project will provide a more objective measure of the representation of these averages and allow questions like these to be answered.

Regardless of the number of items in the ensemble, which items are important to the average? One important observation that has been made through multiple studies is the weight of statistical outliers in the ensemble. The consensus seems to be that outliers are at least weighted differently in the average compared to the rest of the ensemble and at most completely discounted (V. Li, Herce Castanon, Solomon, Vandormael, & Summerfield, 2017; Haberman & Whitney, 2010; although see Oriet & Brand, 2013 for conflicting results for size averaging).

Lastly, there is the question of the number of items actually integrated in the average representation, regardless of the number of items present in the ensemble. One way of examining this involves identifying the number of stimuli that are
effectively integrated, usually using an ideal observer analysis (Parkes et al., 2001; Sweeny, Wurnitsch, Gopnik, & Whitney, 2015), and comparing that with the total number of stimuli in the set (Whitney & Yamanishi Leib, 2018). This number varies widely across methodologies, feature types, and individual observers. For example, in studies using high level features, like faces or biological motion, participants efficiently encode between 4-8 items, which is approximately half of the total display (Haberman & Whitney, 2010; Sweeny et al., 2013; Yamanishi Leib et al., 2016). However, in a recent review paper, researchers correlated these two values across a wide range of studies using various methods and features and found that participants tend to encode the square root of the total number of items in the ensemble (Whitney & Yamanishi Leib, 2018).

Attention may also plays a strong role in participant’s ability to extract a subset of the ensemble or the entire ensemble. In other words, do observers spread their attention across the entire array or only to small subsets within the array? There is support for the idea that participants cannot use the entire array due to attentional restrictions to the amount of items that can reasonably be processed (de Fockert & Marchant, 2008; Marchant et al., 2013; Myczek & Simons, 2008; Simons & Myczek, 2008). These studies discuss problems revolving around potential focused attention strategies that participants might employee. For example, participants may be focusing attention on only a small subset of the items or on only one item (de Fockert & Marchant, 2008; Myczek & Simons, 2008). Although these strategies may lower accuracy on a trial-to-trial basis, the mean accuracy would still be large.
However, evidence supporting the idea that we are extracting information from the entire ensemble has also been shown (Chong, Joo, Emmanouil, & Treisman, 2008; Chong & Treisman, 2005a, 2005b; Robitaille & Harris, 2011). For example, it has been shown that participants are better at extracting summary statistics from larger arrays than smaller arrays and that using only one item is not a strategy being employed by observers (Robitaille & Harris, 2011; Chong, Joo, et al., 2008). Additionally, it has been shown that participants can extract separate ensemble averages for multiple groups presented simultaneously, including the same feature from overlapping groups differentiated along another feature dimension, such as color or various gestalt grouping rules (Alvarez, 2011a; Corbett, 2017; Chong & Treisman, 2005b). However, conflicting evidence seems to suggest that, if ensemble encoding is pre-attentive, items cannot be segmented into groups (Ward et al., 2016; Oriet & Brand, 2013).

Ultimately, these debates have merged with the debate surrounding attentional requirements for extracting averages. In other words, the question has become one revolving around our ability to selectively distribute attention to the entire group or subsets within the group, and whether or not attention is even allocated to the ensemble to extract the average (see the attention and ensemble encoding section below for more details). Therefore, it is no longer a question of how many items are used to form an ensemble, but how efficiently and effectively attention can be allocated to a subset of items within the ensemble or the ensemble itself.
The neural representation of the average

What is the neural representation of an ensemble average? To answer this question, researchers have performed behavioral experiments to uncover the neural characteristics of the average. Many proposed functions have been introduced recently that argue for the role of ensemble encoding in cognition and perception. For example, ensembles have been used to describe rapid scene processing (Brady et al., 2017), popout in visual search (Rosenholtz et al., 2012), the richness of conscious experience (Cohen et al., 2016), and as a means to overcome the processing bottleneck that attention and working memory present (Cohen et al., 2016; Im & Chong, 2014; Alvarez & Oliva, 2008). One thing each of these has in common is a shared assumption about the representation of the average. For example, many of these proposed functions state that ensemble averages are quickly and efficiently extracted for areas of visual space not attended to by the observer (Cohen et al., 2016). In other words, ensemble averages are used to fill in our conscious experience, due to our inability to fully attend and process all of the information in the environment. This assumes that ensemble statistics are quickly and efficiently extracted in the absence of attention. However, very little is actually known about the underlying representation of ensemble averages, like their neural representation or attentional and working memory requirements. This section will focus on two questions: how is an ensemble average generated and what is being represented in that average.

Proposed mechanisms in ensemble encoding

Many of the proposed functions of ensemble encoding are to quickly and efficiently facilitate perception by summarizing large, and ultimately unimportant,
areas of space. One proposed mechanism to explain how these summary statistics are computed is the process of spatial pooling (Alvarez, 2011a; Haberman & Whitney, 2012). What seems appealing about this mechanism is that the resulting representation of the mean takes the form of a distribution of responses from lower cortical levels. As discussed in the above sections, a distribution seems to be the ideal way to represent an average, as it can give access to many different statistics in itself. Although details of this process are certainly vague, and more thoroughly explored for some feature domains than others, there is still a fair amount of evidence supporting this type of mechanism.

Generally, this mechanism works as follows. Individuals within an ensemble activate local populations of feature-selective neurons. This then creates a distribution of local responses that are based on the tuning functions of the underlying neurons. Next, a higher-level cortical area samples this population of tuning curves, selecting either a subset of responses from the population or the entire population and creates a global population- response. This global response, based on the summation of local tuning curves, represents the average response. This global response would be based on a tuning curve derived from the local tuning of its input neurons and can thus be used in a variety of ways. For example, altering local tuning functions to match the average through a feedback mechanism or comparing average responses to those of local responses to facilitate visual search (Haberman & Whitney, 2012). Another interesting potential use for this average representation is that of normalization. Normalization or norm-based coding refers
to a neural coding scheme in which stimuli are encoded relative to the norm, which seems strikingly similar to the average (M. A. Webster, 2015).

This pooling mechanism was originally proposed for, and is most developed for, orientation and motion integration, sometimes called global processing (for orientation: Baldwin, Husk, Meese, & Hess, 2014; Dakin, 2001; Dakin & Watt, 1997; M. Morgan et al., 2008; Parkes et al., 2001; for size: Pasternak et al., 1990; Treue, Hol, & Rauber, 2000; Williams & Sekuler, 1984). Studies examining global motion integration measured sensitivity to a range of directions in a variety of single neurons in area MT, an area shown to be responsive to motion direction (Treue et al., 2000). They then create a population tuning function from these local response functions. Interestingly, they show that the shape of this global tuning function corresponds to subjective reports of global motion direction and number of moving elements. Models of orientation integration have also been proposed. For example, in one model, orientation integration occurs in two stages (Baldwin et al., 2014). First, orientation signals are integrated across small areas of space that are limited by local noise for each area. Then, these integrated local orientation signals are combined depending on task demands and attentional allocation.

The feature dimension of size has also been explored in the context of this pooling model. Researchers have proposed a model called the noise and selection theory (Allik et al., 2013). This model uses the basic pooling framework with two constraints. First, the idea that there is noise inherent in the system, which affects participant’s ability to judge the size of individuals and, as a result, the overall mean
size. Second, only a subset of items are actually sampled and that the size of the subset cannot exceed attention and working memory capacity limitations. This model is able to accurately predict human behavioral responses.

One problem with this type of mechanism is it becomes more difficult to explain higher level ensemble encoding. This is because for many of these high-level features there are no simple locally tuned populations that a pooling mechanism can draw from. This makes the pooling response relatively straightforward. However, similar mechanisms for something like facial expression or even something slightly simpler like size have not yet been identified.

A different mechanism has been proposed for the extraction of numerosity (M. J. Morgan, Raphael, Tibber, & Dakin, 2014). The theory behind this model is as follows. When low level differences of a group of items, like blur, contrast, and total area of the array, are controlled for, participants require a large change in number of items to be detected. Given this, a model for which the observer is making simple spatial contrast judgments performs as well or better than human observers. Although this model elegantly explains average number, it relies on measurements of density and surface area of the ensemble, and therefore cannot be applied to feature ensembles that do not vary along these dimensions, such as orientation or facial expression.
**Attentional requirements of ensemble processing**

Although it seems that the mechanism behind most forms of ensemble encoding is the process of pooling, there are still many questions remaining regarding how this process actually occurs. One of the most important questions is how attention is utilized in ensemble encoding. Many of the proposed models for pooling make assumptions about the use (or lack thereof) of attention to perform the function of pooling or deciding what items should or should not be pooled. Is attention necessary for the extraction of ensemble statistics? This question is far from being resolved, and there is evidence in the literature for both sides.

On one hand there is substantial evidence that ensemble encoding does not require the allocation of attention or is a pre-attentive. Initially, it was thought that ensembles were extracted pre-attentively due to observer’s ability to extract them quickly and efficiently. For example, it has been shown that ensemble averages can be extracted for item sets that far exceed the bottleneck of attention or working memory capacity (4+/-1) (Ariely, 2001; Haberman & Whitney, 2009; Robitaille & Harris, 2011). It has also been shown that averages can be extracted quite quickly, with presentations shown as quickly as 50 ms, which is faster than would be expected if attention was allocated to each item.

While these studies show that attentional allocation to individual items is not required to extract the average, the question of whether the average can be extracted without attention at all is still open. However, this assumption was more formally tested using paradigms with explicit attentional manipulations. Some of
these studies found that the allocation of attention doesn’t affect ensemble encoding accuracy (Alvarez & Oliva, 2008; Epstein & Emmanouil, 2017; Bronfman, Brezis, Jacobson, & Usher, 2014; Hall, Mattingley, & Dux, 2015). In one study, researchers had participants either attend to individual items or the entire display. They showed that withdrawing attention only affected the representation of individual items but not the summary representation (Alvarez & Oliva, 2008). A similar effect was shown for extracting average color from a set of letters (Bronfman et al., 2014) and performing an ensemble averaging task while focusing on individual items for a statistical learning task (Hall et al., 2015). Lastly, one experiment looking at working memory capacity for ensemble features compared to individual features, showed that ensemble representations are maintained despite overloading working memory stores (Epstein & Emmanouil, 2017). Lastly one important study has demonstrated that individual items in an ensemble can be masked using object substitution masking (Jacoby, Kamke, & Mattingley, 2013). Importantly, this affects participants judgments of the average in a systematic way, where the judgments of the average are shifted away from the level of the masked item. This provides more evidence that ensemble statistics can accurately be maintained despite a lack of attentional resources.

This debate is further complicated by the existence of multiple neuropsychological studies that show participants with unilateral spatial neglect (Hochstein, Pavlovskaya, Bonneh, & Soroker, 2015; Yamanashi Leib, Landau, Baek, Chong, & Robertson, 2012), simultagnosia (Demeyere et al., 2008), and
prosopagnosia (Yamanashi Leib, Puri, et al., 2012), are still efficient at recovering averages despite impairment for individual item recognition.

Other studies show a detriment in ensemble encoding accuracy if attention is not focused on the set (Chong & Treisman, 2005a; Huang, 2015; Jackson-Nielsen, Cohen, & Pitts, 2017; McNair, Goodbourn, Shone, & Harris, 2017). One study found that when observers were asked to report the average size of a group of circles while performing a search task on that set, that required serial allocation of attention to individual items, averaging accuracy was reduced (Chong & Treisman, 2005a). Other studies showed that ensemble encoding accuracy was reduced when presenting ensembles during the attentional blink (McNair et al., 2017) and during an inattention blindness paradigm (Jackson-Nielsen et al., 2017). Lastly, one study tested observers ability to extract features from individuals with their ability to extract summary features (Huang, 2015). They showed that participants need as much attention to extract both types of feature information.

Taken together, these studies are inconsistent and inconclusive. Neuroimaging and recording studies are sorely needed to provide a more objective measure of the effects of attention on the representation of ensemble statistics.

**The neural correlates of the average**

To date, a relatively small amount of research has been conducted examining the neural correlates of ensemble averages. One clue to the neural correlates of the mean comes from a study that examined individual differences across observers for their ability to extract high and low level ensembles (Haberman et al., 2015). They found that the ability to
extract either type of average feature information is not well correlated. This data suggests that there is not a unified domain-general averaging mechanism. Additionally, the neuropsychological studies cited in the above section seem to suggest that the mechanism responsible for ensemble encoding are not the same as those responsible for individual item processing (Demeyere et al., 2008; Hochstein et al., 2015; Yamanashi Leib, Landau, et al., 2012; Yamanashi Leib, Puri, et al., 2012).

One area of research that may be promising when studying the neural representation of the mean is redundancy gain (Cohen, Konkle, Rhee, Nakayama, & Alvarez, 2014; Shim, Jiang, & Kanwisher, 2013; Jiang, Kwon, Shim, & Won, 2010; Macevoy & Epstein, 2009). Generally, redundancy gain is an increase in response in early retinotopic cortex due to the presentation of the same objects simultaneously. This is generally different from something like biased competition theory, as the stimuli are presented in separate visual fields meaning their receptive fields are not overlapping. Redundancy gain has been proposed as a result of feedback mechanisms from higher cortical areas that hold a representation of the mean (Shim et al., 2013). This representation may then bias the response of these local populations toward the mean. In other words, it is possible that neurons whose responses are more similar to the mean generate a larger response due to this feedback. This is interesting as it suggests that differences in the response to ensembles are reflected in early levels of visual cortex.

Only a handful of studies have been published that directly examine the neural correlates of ensemble encoding. In a series of studies using fMRI repetition suppression, researchers measured fMRI adaptation to repeated presentation ensembles made up of the
same real world objects (e.g. a group of strawberries or flowers) (Cant & Xu, 2012, 2015, 2017). They measured the response to repeated presentation of either the same object ensemble or to different ensemble (e.g. strawberries to strawberries or strawberries to flowers). Their results indicate that areas in anterior-medial ventral visual cortex, including scene sensitive area PPA, shows suppressed response when shown the same group compared to different groups, indicative of adaptation. Additionally, they report responses in the same areas when observers are presented with textures that reflect similar underlying statistics without the object representations. They conclude that ensemble information is computed in these scene and texture areas.

Many of the functions and mechanisms presented here similarly rely on texture representations and often use these textures interchangeably with ensemble averages. However, we should use caution when making assumptions about the similarities with ensemble averages and textures. While the two are tightly related they are independent of one another, similarly to the relationship between the processing of individual features and object recognition. Object perception is the result of binding features together into a coherent representation. Ensemble statistics are feature averages that are extracted from groups of objects or from textures. A texture is a representation built from feature averages. They contain ensemble information, but are not necessarily an ensemble average in and of themselves. Attneave, describes this with an eloquent example: ‘in observing a cat, however, one does not ordinarily perceive its hairs as individual entities; instead on perceives that the cat is furry. Furriness is a kind of texture; the statistical parameters which characterize it presumably involve averages of shape and
direction, as well as size, of elements’ (Attneave, 1954). This example emphasizes that while the perceived texture might be that of ‘furriness,’ this texture representation is built upon featural averages of direction, size, and shape. Therefore, a major problem with past imaging work is that it often treats ensemble perception as nothing more than texture perception. Because of this, we think it becomes valuable to examine responses to separate ensemble statistics independently. While using real world images is more representative of real world behavior, it does not give a complete picture of ensemble encoding.

Lastly, an fMRI repetition suppression study attempted to show differences between activity to groups of differently sized circles (Chong, Shin, & Jo, 2008). Specifically, they measured adaptation to repeated presentation of circles with the same mean and with different means. Ultimately, they show no significant differences between their conditions. One reason why this might be the case is a lack of control in overall area taken up by the objects within their array. For example, when showing an ensemble with a large mean size you must control for total area by increasing the number of items in the small mean size condition. This is a fairly large confound, as it indicates that the total number of pixels present is also systematically varying along with the overall mean size of the ensembles. In other words, as mean size increases total area stimulated also increases.
**What cognitive resources are needed to extract the average?**

A major question this thesis seeks to answer is *how* the ensemble average is extracted and encoded. First, we will examine the spatial and temporal dynamics of mean extraction using MRI and EEG. In addition to this we will also examine what the role of attention and task are in extracting the ensemble average. Many of the predictions about the spatial and temporal properties of ensemble encoding depend on the role of attention and task relevance. For example, when in time we expect to see this emerge in an evoked potential will vary depending on if this process is automatic and/or preattentive. Similarly, if spatial attention is necessary for this process to emerge, we would not expect to see any evidence of ensemble encoding in early visual areas like V1 or V2. Therefore, we will describe 4 general theories of the role of task and attention in ensemble encoding and the predictions they each make. The first three assume that ensemble encoding happens in parallel but differ in the amount of cognitive resources that are necessary to extract the ensemble. Additionally, the first three theories assume that the representation of the average varies as a function of changes to that average. Hypothetical predicted results are displayed in Figure 1.
Figure 1. Hypothetical predictions in support of ensemble encoding for each proposed hypothesis.

Table displaying which conditions we would predict to see evidence for ensemble encoding given the specific hypothesis. Shown on the x-axis is the given experimental condition and on the y-axis is the given hypothesis. Conditions from left to right are 1) attended task relevant, 2) attended task irrelevant, 3) unattended task relevant, and 4) unattended task irrelevant. The hypotheses are for strong ensemble encoding, weak ensemble encoding 1, and weak ensemble encoding 2 (described above). Check and x marks are representative of evidence in favor of ensemble encoding, in any of the three experiments.

### Strong ensemble encoding

The first theory of ensemble encoding states that neither the allocation of spatial attention nor an explicit averaging task is needed to extract the ensemble.
Many popular theories of ensemble encoding require this ability to be rapid and automatic (Alvarez, 2011b; Brady et al., 2017; Cohen et al., 2016; Rosenholtz et al., 2012; Utochkin, 2015). For example, ensemble averages are used to explain why we have a rich subjective perceptual experience despite our brains severe attention and working memory bottleneck (Cohen et al., 2016). Therefore, this theory predicts that any representation of the mean varies only as a function of changes to that mean, and will not vary as a function of attention or task. Hypothetical data for experiment 1 in support of this hypothesis can be seen in Figure 2.
Figure 2. Hypothetical data for experiment 1 for the strong ensemble encoding hypothesis.

Hypothetical percent signal change in response to each of the four presented feature levels in experiment 1 in support of the strong ensemble encoding hypothesis. Presented here are the hypothetical results for orientation (top row) and size (bottom row) in response to progressively larger differences between the adapting and test stimuli presented in experiment 1. As can be seen, this hypothesis would predict progressively larger responses as we increase the difference. This pattern would be present for both tasks and across visual ROIs, although perhaps to a larger or smaller degree.
**Weak ensemble encoding 1**

The second theory of ensemble encoding states that ensemble encoding requires either the allocation of spatial attention to the ensemble or that the relevant feature be attended to through the presence of an averaging task. This theory still predicts that each element is processed in parallel but that more cognitive resources are needed to initiate this parallel process. Therefore, this theory predicts that any representation of the mean will vary both as a function of changes to the mean and either the focus of spatial attention (weak ensemble encoding 1.1) or the presence or absence of an averaging task (weak ensemble encoding 1.2) that requires the participant to focus on the relevant feature of interest.

**Weak ensemble encoding 2**

The third theory of ensemble encoding states that both the allocation of spatial attention and the presence of an averaging task are necessary to extract the average from an ensemble. Again, like for the first two theories, this theory still predicts that ensemble encoding occurs in parallel. This theory predicts that the representation of the average will vary as a function of changes to the mean only in the presence of spatial attention and an averaging task. Specifically, the representation of the average will vary under one of those conditions. First, when the participant is performing a task and attending to the ensemble compared to when they are performing a task and not attending to the ensemble (weak ensemble
encoding 2.1). Second, when the participant is not attending to the relevant feature (attending to another feature or performing an averaging task on another irrelevant feature) with the allocation of spatial attention compared to when they are not attending to the relevant feature without the allocation of spatial attention (weak ensemble encoding 2.2). Third, when the participant is attending to the relevant feature and allocating spatial attention to the ensemble compared to when they are not attending to the relevant feature and are not allocating spatial attention to the ensemble (weak ensemble encoding 2.3).

**Ensemble encoding as a strategy**

As mentioned in the above sections, there is evidence to support a theory of ensemble encoding as a strategy to rapidly encode information from a large group of items without encoding the feature level of each item in parallel. In other words, this theory states that people aren’t actually extracting the feature value from each object and averaging them together via pooling. Instead, participants might be employing a number of different strategies to accomplish the same goal.

One theory explains ensemble encoding through a random subsampling account (Simons & Myczek, 2008). According to this theory participants may be randomly sampling 3-4 items from the ensemble and computing the average of this subgroup. Although the mean estimation would be worse given this strategy, over many trials the estimated average would regress towards the mean.
Other theories suggest that participants may be subsampling in a more systematic way. For example, one theory suggests that participants may sample outliers of the ensemble in order to get a better estimate of the range of ensemble (Marchant et al., 2013; Maule & Franklin, 2016). For example, by identifying the minimum and maximum sized circles in a ensemble, participants can form an accurate representation of the average.

More recently, researchers have proposed that in addition to sub-sampling, participants may be using other low level visual properties to estimate the mean (Lau & Brady, 2018). For example, it has been demonstrated that participant’s estimates of mean size are biased toward larger estimates when the ensemble is made up of filled circles compared to outlines circles. As a result, participants seem to be using other low-level features as proxies for estimating variability within an ensemble.

**Outstanding questions and proposed experiments**

The above sections have laid a solid groundwork on the properties, both behavioral and neural, of ensemble encoding. As you may have noticed, many of the questions in the literature remain unanswered. Is there a common region that supports the extraction of summary statistics across different feature domains? Is the process automatic or does it require a specific averaging task? Does extracting these statistics require focused attention? To answer these questions we need to better understand the neural correlates and neural representation of the mean.
Given the lack of past studies on this topic, this dissertation project will attempt to systematically test our sensitivity to changes in the mean across two features, orientation and size, while either performing an explicit averaging task or not. We will explore the representation using a variety of different methods, including fMRI and EEG. These techniques will provide a wide array of results including localization of function using fMRI and time coarse analysis using EEG. Additionally, we will use MVPA analysis to try to uncover the distributed pattern of results caused by an ensemble array. In the following sections the specific experiments and analysis are described in detail.

**General methodology and stimulus parameters**

**Introduction**

Ensemble averages are average feature representations that are pooled over space and time (Whitney & Yamanishi Leib, 2018). For example, orientation and size can be averaged across a large number of individuals to a high degree of accuracy over space and time (Dakin & Watt, 1997; Ariely, 2001; Chong & Treisman, 2005, 2003). However, we do not know how these ensembles are represented neurally. In order to characterize the neural representations of these feature averages we will characterize neural sensitivity to changes in the mean. Classically, this has done by defining psychometric or neurometric functions that represent the ability, either behaviorally or neurally, to discriminate between two stimuli. For example, we can
measure accuracy when discriminating between two oriented Gabor patches that differ in the magnitude of the difference between their orientations. Using neural data, we can measure and compare the neural response to stimuli presented at different orientations and compute how discriminable those responses are. Here, we will create such discriminability functions using both neural and behavioral data in response to changes in the mean orientation and size of an ensemble of ellipses, using psychophysics, high-density electroencephalography (hdEEG), and functional magnetic resonance imagining (fMRI).

As relatively little has been done to understand the neural representation of the ensemble mean, we will use a variety of methods to define our neurometric functions. We will collect neural responses using both hdEEG and fMRI and analyze these data using multivariate pattern analyses (MVPA) and frequency domain analyses based on the frequency-tagging EEG technique. In the above sections, we have argued for a mean representation in the form of a distribution of responses from local neuronal populations. Given this representation, MVPA is the ideal analysis technique as it is the best tool to identify the underlying pattern of activity. Lastly, we will match neurometric to behavioral psychometric functions to identify where in the brain and when in time ensemble representations emerge.

**General stimulus parameters**

We will use ensembles comprised of ellipses with different mean sizes and orientations. An ellipse is an ideal stimulus for these experiments because we can
manipulate both the average size and orientation independently. An ellipse also represents an individual object with its own enclosed shape. This allows us to eliminate some potential confounds, such as unwanted grouping or contour formation that can be introduced by using lines or Gabor patches. Importantly, these stimuli also allow us to compare neural activity to ensembles that vary on two dimensions independently. Additionally, it has been shown that participants can extract multiple ensemble averages simultaneously, including extracting the same average feature from multiple ensembles, as well as extracting different features from the same ensemble (Huang, 2015; Im & Chong, 2014).

We will use one reference mean and 4 progressively larger (or more rightward tilted) means that differ from the reference by some increasing amount, which we will refer to as levels of delta (change in mean or $\Delta$; $\Delta_s$ for size and $\Delta_\theta$ for orientation; see Figure 3). For example, one reference size, with an average of 1 DoVA, with 4 progressively larger ensembles or one reference orientation, with an average of $0^\circ$, with 4 progressively rightward tilted ensembles. The amount of change between each level and the reference will be determined using a psychophysical task (see psychophysical experiment section; for an example of individually chosen levels see Figure 4). We will use the same stimuli for each experiment (oriented ellipses), although the way in which they are presented, how many levels will be presented, and how the levels are chosen will be slightly different for each experiment (see details below for specific experiments).
Orientation Levels With Mean Size Constant
\( \Delta_s=0 \) \( \Delta_s=4 \)

Size Levels With Mean Orientation Constant
\( \Delta_\theta=0 \) \( \Delta_\theta=4 \)

Figure 3. Example ensemble stimuli for minimum and maximum size and orientation levels.

Stimulus ensembles made up of ellipses that differ in their mean orientation or mean size. In the first row, mean size is held constant (1 degree of visual angle (DoVA)), whereas the mean orientation is increased (0°-90°). In the second row, mean orientation is held constant (0°), whereas the mean size is increased (1-2 DoVA).
Controlling confounds for overall ensemble configuration

Three studies, which we know of, have been conducted that specifically examine the neural representation of the ensemble average, only one of which was published (Chong, Shin, et al., 2008). These studies have all used fMRI and generally found null results. However, all of these studies have flaws in that they fail to control for crucial stimulus confounds. In an attempt to learn from past mistakes we will pay special attention to these potential confounds. Below is a list of important confounds that will be controlled for.

1. Because we are attempting to identify the representation of the *average feature* we want to design our stimuli in a way that minimizes low level feature information that could increase or decrease the overall response and dampen the response of the average. The theory behind this is as follows. When using fMRI and EEG, increased stimulation will lead to increased neural response. For example, we can reasonably expect the response in visual cortex to be larger when presented with an ensemble of 10 large items than to an ensemble of 10 small items, a problem that we will call the ink on the page problem. Therefore it is critical to compare ensembles for which the mean is different but not the total ink on the page. Controlling for this problem will control for low level confounds and allow us to get a more accurate response to the change in the overall mean. This will be addressed by pseudorandomly varying the number of items in each stimulus array in such a fashion that there will be an equally likely chance that the smallest
mean size be presented with more ‘ink on the page’ as for the largest mean size to be presented with less ‘ink on the page’.

2. Another potential problem in decoding average orientation arises from the oblique effect (Furmanski & Engel, 2000; B. Li, Peterson, & Freeman, 2003). This is a well-documented effect for which detection and discrimination is better for orientations at the horizontal and vertical orientations but not for the diagonals or oblique orientations. In order to control for this confound we will change the standard orientation value for each participant. In other words, while some participants may see stimuli at a standard value of 0 degrees, others will see standards at 45 degrees, for example.

3. Another potential confound arises when attempting to define what is being decoded: each individual feature or the mean feature. In other words, what is being decoded: the differences amongst individuals from one ensemble to the next or the difference in the overall mean. To control for this problem, we will create our ensembles with enough variance to allow for overlap amongst item sizes/orientations between different means. For example, a large ensemble can contain some small individuals and a small ensemble can contain some large individuals. Additionally, for each stimulus presentation, the location of each individual will be jittered in its position and a new size and orientation value will be, while maintaining the mean size and orientation of the overall ensemble.
4. On confounding factor when recording neural data is the difference in peripheral and foveal activation. Stimuli presented at fixation tend to dominate the response. This is particularly important for ensembles, especially considering some unknowns regarding the number of and location of items that are actually encoded from the ensemble. To minimize confounds that may cause the central items to dominate the responses, we will never present items at or near fixation or near the vertical or horizontal meridians.

5. Lastly, we will specifically control the mean values in each of the four quadrants of visual space. One possibility that could confound the results is that the process of ensemble encoding may be lateralized. Therefore one important control that was not taken into account in past studies is mean values across the different subsets within the overall ensemble. Here we will ensure that the mean within each subset in each quadrant is the same.

In order to implement these controls, we will create the ensemble arrays by generating an 8x8 grid for which individual items can appear in any of the cells in the grid within certain constraints (Figure 3). To control for the ink on the page problem, we will randomize the number of items that can appear in the array by adding or subtracting one item in each quadrant. We will present items in a total of 48 or 56 of the total 64 available cells and each condition with more or less ink on the page an equal number of times. In other words, every condition will be compared to the standard with more and less total ink on the page. Additionally,
these controls should not affect the participant’s ability to extract the mean as it has been shown previously that participants can effectively extract averages in both the size dimension and orientation dimension across a wide number of individual items and densities (Chong & Treisman, 2005; Marchant et al., 2013; Whitney & Yamanishi Leib, 2018)

**Psychophysical experiment: Identifying individual’s thresholds for ensemble averaging of two stimulus features**

**Introduction and experimental paradigm**

All participants, for each of the three proposed experiments also first ran in a behavioral psychophysical experiment. In order to find the neural representation of the ensemble mean, we must first establish the behavioral sensitivity for detecting differences in the mean using our stimuli and paradigm. To do so, we implemented a two alternative forced choice paradigm in which participants compared two ensembles and determined which has the larger mean size or more rightward tilted orientation (from 1-2 degrees of visual angle (DoVA) and 0°-90°; see Figure 3). Specifically, we presented participants with 2 consecutive ensembles, one of which is always the reference (0° average orientation and 1 DoVA average size), while the other is pseudo-randomly presented at one of the 5 levels for both size and orientation (25 possible combinations of the 5 levels for size and orientation).
Each ensemble was presented for 300 ms separated by a 1000 ms inter-stimulus interval. Participants were then asked to determine which ensemble was more rightward tilted or had a larger mean size. We presented each of the 25 stimulus combinations, with either the size or orientation averaging task, a total of 4 times. Therefore, a total of 200 trials were presented. The order of presentation of the test and standard was randomized for each trial.

**Data analysis**

First, we segmented trials based on the relevant feature level, thus collapsing across the irrelevant feature, resulting in five times as many trials per feature level (20 trials per level). Next, we calculated the proportion of times each participant reported the test as being either more rightward tilted or larger than the reference for each level of both features for both tasks. This allowed us to create psychometric functions for both tasks (for an example participant see Figure 4). Then, using the psychometric functions for each participant, we can determine the point of subject equality, or the point at which participants are simply guessing, and pick off values that are 10%, 20%, 30%, and 40% greater than the PSE. These values are then used instead of the hard-set values so participant accuracy is not at ceiling, keeping the task challenging and helping ensure participants are attending to the task.
Figure 4. Example participant psychometric functions.

Psychometric functions from one subject for the orientation task (left) and size task (right). Plotted on the y-axis is the proportion of times the participant responded that the test was more rightward (orientation) or larger (size) than the reference for each of the feature levels (x-axis). The red solid lines are the raw percentages for each task while the blue solid lines are the sigmoidal curve fits. The horizontal dashed black line represents chance response rate, for which the participant reported the test as being larger or more rightward tilted than the reference 50% of the time. The vertical dashed lines represent feature level values (x-axis) that fall 10%, 20%, 30%, and 40% above the 50% value (y-axis).

Experiment 1: Identify neural correlates of ensemble averaging using fMRI repetition suppression

Introduction

In order to identify where in the brain ensemble averages for different features are processed we measured neural activity in the form of the BOLD response using fMRI in response to changes in the mean size and orientation of an ensemble of ellipses. Here we measured changes in the BOLD response to varying
degrees of change, or Δ, in the mean of an ensemble across two independent feature
dimensions: orientation and size. Specifically, we will present participants with
blocks of stimuli consisting of alternating sequences of ensembles that differ in the
magnitude of Δ between the standard ensemble and the test ensemble. Then, using
this data we ran a multivariate pattern classifier to identify how similar the pattern
of results are as a function of increased Δ across visual ROI’s.

Methods

Participants

Ten neurotypical adults (9 right-handed) with normal to corrected to normal vision participated in experiment 3 (9 males, aged 23-49). Each participant provided informed written consent as approved by the Institutional Review Board at the University of Nevada, Reno.

Apparatus

The stimulus computer was a 2.53 GHz MacBook Pro with an NVIDIA GeForce 330 M graphics card. Stimuli were controlled using a 2.6Mhz MacMini and presented using PsychToolbox (Brainard, 1997; Pelli, 1997; Kleiner et al., 2007) for MATLAB (MathWorks Inc., Natick, MA). For the retinotopic scans collected at the University of California, Davis, stimuli were presented on a 24 in Cambridge Research System (Kent, UK) LCD BOLD screen display with a 60 Hz refresh rate. The
maximum visual area through the mirror was 19.3° x 12.1°. For the retinotopic and experimental scans collected at Renown Health Hospital, stimuli were presented on a 32 in SensaVue visual display system (Invivo, Inc., Gainesville, FL) with an 85 Hz refresh rate. The maximum visual area through the mirror was 31.5° x 18.9°. In both cases, the screen was positioned outside of the scanner bore and viewed through a tangent mirror attached to the head coil. Stimulus presentation was time-locked to the fMRI data by a trigger sent from the scanner at the start of image acquisition.

**MRI apparatus and scanning procedure**

*Retinotopy and anatomical scans*

Retinotopic data were acquired at one of two facilities. Data for 8 of 10 subjects was acquired at the University of California, Davis Imaging Research Center on a 3T Skyra MRI System (Siemens Healthcare, Erlangen, Germany) with a 64 channel phased-array head coil. The functional images were collected using T2* fast field echo, echo planar functional images (EPIs) sensitive to BOLD contrast (TR = 2.5s, TE=25ms, 32 axial slices, 3.0 mm², matrix size=80x80, 3 mm thickness, interleaved slice acquisition, 0.5 mm gap, FOV=240x240, flip angle=71°). In addition to the functional retinotopic data, high-resolution anatomical scans were collected for reconstruction of each participant’s cortical hemisphere surfaces using the FreeSurfer application from T1 (MPRAGE, TR=2230 ms, TE=4.02 ms, FA=7°, 640x640 matrix, res=0.375x0.375x0.8 mm) and T2 (TR=3 s, TE=304 ms, FA=7°, 640x640 matrix, res=0.375x0.375x0.8 mm) images.
Data for 2 of the 10 subjects was acquired at the Neuroimaging Facility of Renown Health Hospital in Reno, NV on a 3T Philips Ingenia scanner (software version 4.1.1) using a 32 channel digital SENSE head coil (Philips Medical Systems, Best, Netherlands). The functional images were collected using T2* fast field echo, echo planar functional images (EPIs) sensitive to BOLD contrast (TR = 2.5s, TE=40ms, 32 axial slices, 3.0 mm², matrix size=128x128, 3 mm thickness, interleaved slice acquisition, 1 mm gap, FOV=256x256, flip angle=76°). In addition to the functional retinotopic data, high-resolution anatomical scans were collected for reconstruction of each participant’s cortical hemisphere surfaces using the FreeSurfer application obtained using a 3D T1 weighted pulse sequence (TE=4.6 ms, TR=3.0, flip angle=8°, res=1x1x1 mm, matrix size=256x256).

*Experimental scans*

Experiment data were collected at the Neuroimaging Facility of Renown Health Hospital in Reno, NV using a 3T Philips Ingenia scanner using a 32 channel digital SENSE head coil (Philips Medical Systems, Best, Netherlands). Functional data was collected using T2* fast field echo, echo planar functional images (EPIs) sensitive to BOLD contrast (TR=2.0 s, TE=25 ms, 32 axial slices, 3.0 mm², 3.0 mm thickness, 0.5 mm gap, matrix size=80x80, FOV=240x240, Flip angle=71°, interleaved slice acquisition).
Stimuli

As described above, we used ensemble arrays made up of ellipses with different mean sizes and mean orientations. The physical values of each of these were optimized based on behavioral data for each participant collected before the fMRI experiment. Unlike in the last 2 experiments, here we only used 4 feature levels for each feature instead of 5.

Experimental procedure and task design

Participants were asked to perform an average task along one of two feature dimensions while changes to the BOLD signal were recorded using fMRI. Each of the 16 feature combinations were presented in individual blocks throughout the experiment. A typical block consisted of 16 stimulus presentations, 8 reference ensembles interleaved with 8 test ensembles (see Figure 5). The reference ensembles maintained an average of 0° for orientation and 1 DoVA for size throughout the entire experiment. The test ensemble remained constant at the chosen feature level throughout that specific block. Each stimulus was presented for 150 ms followed by an inter-stimulus-interval of 850 ms.

Participants were asked to make judgments between the averages of the test and reference ellipses. After all 16 stimulus presentations, participants viewed a blank response screen for 12 s (inter-block-interval or IBI), in which they were asked to respond to the previous block. Participants responded using a button box with 4 buttons, labeled 1-4, 1 being no change and 4 being maximum change. The
individual position, orientation, and size of single ellipses changed on each stimulus presentation, while maintaining the average orientation and size. Participants were cued half way through the IBI to the task type before the start of each block by a change in the fixation to either an ‘O’ for orientation or an ‘S’ for size. This fixation symbol remained constant throughout the block to avoid any unwanted differences due to stimulation. At the end of the block and after the last stimulus presentation, the fixation changed to a ‘#’ symbol to notify the participant to respond.

Each feature combination was presented a total of 12 times (4 times per run repeated 3 times) for each task. Each block lasted 16 s and each IBI lasted 12 s. Each run consisted of 8 16 s blocks with interleaved 12 s IBIs leaving a total run time of 236 s (3.9 min) per run and a total experiment time of 47.2 min. Blocks were pseudo-randomly divided into groups of 4 runs, with each run containing 4 of the 16 feature level combinations for each of the 2 averaging tasks. This was repeated 3 times leaving a total of 12 runs each containing 8 blocks.
Figure 5. Experiment 1 trial sequence for one block.
Shown here, is an example trial sequence during one block from experiment 1. Participants first see a blank screen for 6 seconds with a small letter (either s or o) in the center of the screen that indicates the task that will be performed for that block. This is then followed by the presentation of the adapting stimulus for 150 ms and a blank screen for 850 ms. Then, the test stimuli appeared for 150 ms followed again by a blank screen for 850 ms. This is repeated 8 times for a total of 16 seconds. Throughout the block the central fixation point remains the cue letter to ensure participants attended to the correct feature. Lastly, the fixation turned from the letter to a ‘#’ to cue participants to respond. Participants were allowed 6 seconds to respond before the start of the next block.

Behavioral data analysis

For each block participants were asked to judge how different, on average, the orientation or size of the test ensemble was from that of the reference ensemble on a scale of 1-4. We averaged each participants response for each feature level collapsed across the levels of the other feature. This resulted in 4 average response values for each of the 4 feature levels for both the orientation and size task (Figure
To ensure that participants could detect differences in the average orientation or size we performed an ANOVA with a linear contrast to detect any positive linear increase in their average ratings.

**fMRI data preprocessing**

Functional MRI data were analyzed using AFNI
(http://afni.nimh.nih.gov/afni/ Cox, 1996), SUMA
(http://afni.nimh.nih.gov/afni/suma), (Saad et al., 2004), FreeSurfer
(http://surfer.nmr.mgh.harvard.edu Dale, et al., 1999; Fischl et al., 1999), and MATLAB (MathWorks Inc., Natick, MA). First, functional scans were slice-time corrected to the first slice of every volume and motion corrected within and between runs (always < 3 mm). Next, the anatomical volume was aligned with the motion-corrected functional volume and the resulting transformation matrix was used to align the surface-based retinotopic ROIs with the volume-based experiment dataset. Functional images were smoothed using a 6-mm Gaussian kernel. Data were normalized to the percent signal change by dividing the each time series by their mean intensity for each run.

**Retinotopic mapping**

In order to localize functional activity to specific areas of cortex, we defined several topographic regions of interest (ROIs) using both AFNI
(http://afni.nimh.nih.gov/afni/ Cox, 1996) and SUMA
(http://afni.nimh.nih.gov/afni/suma). Standard retinotopic mapping procedures were used for each participant (Arcaro et al., 2009; Swisher et al., 2007). Specifically, we presented a color and luminance varying flickering (4 Hz) checkerboard. Participants performed 4 runs of polar angle mapping and 2 runs of eccentricity mapping.

During the polar angle mapping a wedge shaped aperture (45°; spanning 8-13.5° eccentricity) travelled around the checkerboard. Polar angle mapping consisted of eight 40 s stimulus cycles (9° / s) with a 20 s blank period at the beginning of each run. Consecutive runs alternated between directions between clockwise and counter-clockwise. During the eccentricity mapping a moving ring (1.7° width) was the stimulus. During one cycle, the ring moved through the space, either expanding or contracting for consecutive runs, covering the space between 0° and 13.5° eccentricity from fixation. Each eccentricity run contained eight 40 s stimulus cycle (9° / s) with a 10 s blank period between cycles and a 20 s blank period at the beginning of each run. Participants were asked to maintain central fixation at all times. To ensure participants were maintaining attention to the checkerboard, they were asked to press a button when one portion of the wedge changed to gray. Targets appeared every 4.5 s.

Polar angle and eccentricity representations were extracted using standard phase encoding techniques (Bandettini et al., 1993; Engel et al., 1997; Sereno et al. 1995). Borders between topographic areas were defined by reversals in polar angle representations at the vertical meridians. ROIs were drawn for each participant.
using standard definitions (for review see: Wang et al., 2015; Wandell & Winawer, 2011). We defined a total of 12 ROIs for each hemisphere in each participant: V1, V2v, V2d, V3v, V3d, hV4, V01-2, V3a, V3b, LO1-2. Examples of ROIs drawn for one participant in the dorsal and ventral streams can be seen in Figure 6.
Figure 6. Early visual dorsal and ventral regions of interest drawn for one participant.

Visual ROIs overlaid on activity in response to standard retinotopic mapping polar angle presentations. Shown are ROIs for V1, V2v/d, V3v/d, hV4, V3a/b, VO1/2, and LO1/2.
**General linear model**

We used a blocked fMRI repetition suppression paradigm, in which each block consisted of 8 repeated presentations of the adapting stimulus (reference) and test stimulus. We predicted the hemodynamic response to be larger in blocks where the test and reference were maximally different and thus no neural adaptation has occurred. The measured BOLD response for each block (aka, each feature combination) was quantified using a general linear model (GLM; Friston et al., 1995). Square-wave regressors for each of the 16 feature combinations were generated and convolved with a model hemodynamic response function (BLOCK model in AFNI’s 3dDeconvolve function) accounting for the shape and temporal delay of the hemodynamic response. Nuisance regressors were included to account for variance due to baseline drifts across runs, linear and quadratic drifts within each run, and the six-parameter rigid-body head motion estimates. We ran the GLM separately for the orientation and size task, as there was considerable overlap between trials used in the two analysis. This yielded two independent estimates, one for each task. Then using retinotopically defined visual areas, we defined separate masks used to extract the voxels only present within those regions of interest at a significance threshold of 1.
ROI analysis

To quantify ensemble encoding in the fMRI data, we used the GLM generated beta weights, which represent the observed percent signal change in a block. Importantly, we looked at the average change in the BOLD signal over the course of one block, which includes 8 repetitions of adapting stimulus and test stimulus. We predicted the average signal change over the course of a block would be greater when there is less adaptation. In other words, as the test stimulus becomes more different than the adapting stimuli, the BOLD signal will increase. Specifically, we compared the average % signal change across early visual ROIs for each of the 4 feature levels for orientation and size. We collapsed across V2v and V2d and only display average activity for V2 and V3. Additionally, we z-scored the values by subtracting the averaging over the 4 feature levels from each level and then dividing by the standard deviation for each ROI and participant. This normalized the data and emphasized differences between the values.

Evidence of a neural representation of the ensemble average would be evident given a positive linear trend in the signal between levels 1 through 4 in any of the visual ROIs. To test for this we also performed an ANOVA with a linear contrast across the four levels in each ROI for both size and orientation.

As we do see some evidence repetition suppression in early ROIs, we wanted to be sure this activity was originating from voxels that are responsive to either the orientation or size task. To do so, we created an additional mask that only selects
voxels in any ROI that are more active during the orientation task than during the size task (and vice versa). This allowed us to select only those voxels that were responsive during the task of interest. We then repeated the above ROI analysis using this smaller subset of voxels.

Whole brain analysis

There is some evidence that would suggest activity in other regions than just early visual ROIs. For example, if this were a strategy, one might predict stronger activation in more frontal regions. Alternatively, we might expect to see activation in medial temporal regions if there feedback processes that facilitate ensemble encoding. In order to test for this, we performed a whole brain GLM comparing each of the four levels for either task.

Specifically, we first transformed the anatomical and functional data for each person into Talairach space to compare the data across participants. Next, we performed a one sample repeated measures ANOVA between each of the four levels, collapsed across the task irrelevant feature, for both tasks across all voxels. Additionally, we performed two contrasts comparing levels 1 and 2 with 3 and 4 and level 1 with levels 2, 3, and 4. The results one of the analyses are shown in

Results
Behavioral results

Behavioral results were analyzed in a similar way to those from experiments 1 and 2, except that this experiment only had 4 levels per condition instead of 5. The results of the behavioral analysis are plotted in Figure 7. The two left plots show the average behavioral response for each of the 4 feature levels for orientation or size while performing the orientation or size task respectively. Plotted on the two right figures are the average behavioral responses to orientation or size while performing the opposite task. In other words, these right plots look at participants average responses to the unattended feature and serve as a sanity check to ensure participants aren’t using feature information from the irrelevant feature in their judgments.

A positive linear relationship is clearly visible for orientation while doing the orientation task (top left) and size while doing the size task (bottom left). This was confirmed by a repeated measures ANOVA with a linear contrast (orientation: F(1,9) = 57.104, p < 0.001; size: F(1,9) = 97.313, p < 0.001). A linear contrast was also performed for the size values while performing the orientation task and orientation values while performing the size task (orientation while size: F(1,9) = 8.070, p = 0.019; size while orientation: F(1,9) = 2.339; p = 0.161). While the analysis was significant for the orientation values while performing the size task, this is mainly due to very slight increase between the fourth level and first level, which can be seen by the significant post-hoc pairwise comparison. Additionally,
post-hoc pairwise comparisons were performed between each feature level and any significant differences are plotted.

Figure 7. Experiment 1 behavioral results.

Behavioral results from experiment 1 plotted for orientation levels during the orientation task (top left), size during the size task (bottom left), size during the orientation task (top right), and orientation during the size task (bottom right). Plotted are average responses to each of the four levels of either orientation or task. Error bars represent the standard error of the mean. Significance markers represent significance at a threshold of: * - p=0.05, ** - p=0.01, *** - p=0.001.
ROI Analysis

The results from the ROI analysis are plotted in Figure 8 for orientation and Figure 9 for size. The z-scored normalized results from the ROI analysis are plotted in Figure 10 for orientation and Figure 11 for size. Shown here is the % signal change in the BOLD signal for each of the four feature levels across early visual ROIs (V1, V2, V3, hV4, VO1, VO2, LO1, LO2, V3a, and V3b). Linear contrasts were performed comparing each of the four feature level values in each ROI for both orientation (V1: F(1,8)=2.438, p=0.157; V2: F(1,8)=1.320, p=0.284; V3: F(1,8)=1.801, p=0.216; hV4: F(1,8)=1.320, p=0.284; VO1: F(1,8)=2.558, p=0.148; VO2: F(1,8)=2.766, p=0.135; LO1: F(1,8)=0.001, p=0.974; LO2: F(1,8)=0.106, p=0.753; V3a: F(1,8)=1.67, p=0.232; V3b: F(1,8)=0.006, p=0.942) and size (V1: F(1,8)=1.775, p=0.220; V2: F(1,8)=0.002, p=0.962; V3: F(1,8)=0.109, p=0.749; hV4: F(1,8)=0.002, p=0.588; VO1: F(1,8)=0.318, p=0.588; VO2: F(1,8)=0.163, p=0.697; LO1: F(1,8)=0.022, p=0.887; LO2: F(1,8)=0.055, p=0.824; V3a: F(1,8)=0.300, p=0.599; V3b: F(1,8)=0.147, p=0.711). Linear contrasts were also performed for the z-score adjusted values for both orientation (V1: F(1,8)=0.774, p=0.405; V2: F(1,8)=0.117, p=0.741; V3: F(1,8)=0.578, p=0.469; hV4: F(1,8)=0.117, p=0.741; VO1: F(1,8)=1.555, p=0.248; VO2: F(1,8)=3.451, p=0.10; LO1: F(1,8)=0.060, p=0.813; LO2: F(1,8)=0.365, p=0.562; V3a: F(1,8)=1.667, p=0.233; V3b: F(1,8)=0.003, p=0.957) and size (V1: F(1,8)=2.083, p=0.187; V2: F(1,8)=0.204, p=0.663; V3: F(1,8)=0.042, p=0.663; hV4: F(1,8)=0.204, p=0.663; VO1: F(1,8)=0.307,
p=0.595; V02: F(1,8)=0.114, p=0.774; LO1: F(1,8)=0.167, p=0.693; L02: F(1,8)=0.076, p=0.789; V3a: F(1,8)=0.013, p=0.911; V3b: F(1,8)=0.023, p=0.882).

None of the linear contrasts reached the significance threshold of p=0.05. Interestingly, although none of the linear contrasts reach significance, in ROIs that approach significance, especially in early visual ROIs (V1-V4), feature level 1 is consistently smaller than the other feature levels. This indicates that more adaptation has occurred in this condition, in which each presentation has the same average, compared to when the mean changes throughout the block. This is evident especially for the orientation condition.
Figure 8. ROI analysis results for orientation.

Averaged percent signal change in response to each of the four feature levels for the orientation condition in early visual ROIs. Data are shown from areas: V1, V2, V3, hV4, VO1 in the first row and VO2, LO1, LO2, V3a, and V3b in the second row. Each consecutive bar represents the percent signal change compared to baseline in response to viewing ensemble arrays at each of the four orientation levels. Shown in the top of each graph are results from an ANOVA with a linear contrast between each of the four levels for each ROI. Error bars represent the standard error of the mean.
Figure 9. ROI analysis results for size.

Averaged percent signal change in response to each of the four feature levels for the size condition in early visual ROIs. Data are shown from areas: V1, V2, V3, hV4, VO1 in the first row and VO2, LO1, LO2, V3a, and V3b in the second row. Each consecutive bar represents the percent signal change compared to baseline in response to viewing ensemble arrays at each of the four size levels. Shown in the top of each graph are results from an ANOVA with a linear contrast between each of the four levels for each ROI. Error bars represent the standard error of the mean.
Figure 10. ROI analysis results z-score normalized for orientation.

Plotted are z-score normalized values from Figure 8. Values are normalized by subtracting the average of each of the four values from each value and then dividing by the standard deviation. Results of a linear contrast performed on each of the z-scored groups are plotted above each graph. Error bars represent standard error of the mean.
Figure 11. ROI analysis results z-score normalized for size.

Plotted are z-score normalized values from Figure 9. Values are normalized by subtracting the average of each of the four values from each value and then dividing by the standard deviation. Results of a linear contrast performed on each of the z-scored groups are plotted above each graph. Error bars represent standard error of the mean.

As we do see some evidence of repetition suppression in early visual ROIs, we wanted to perform the same ROI analysis, but using a more conservative selection method to threshold our voxels. To do so, we thresholded the data using a contrast between the orientation and size tasks (orientation-size for the orientation condition and size-orientation for the size condition).
Data are plotted in Figure 12 for orientation and Figure 13 for size. We
performed a linear contrast in order to check for a ramping pattern of results across
feature levels in each ROI (orientation: V1: F(1,8)=1.076, p=0.330; V2: F(1,8)=3.175,
p=0.113; V3: F(1,8)=3.175, p=0.113; hV4: F(1,8)=0.636, p=0.448; VO1: F(1,8)=1.727,
p=0.225; VO2: F(1,8)=4.440, p=0.068; LO1: F(1,8)=4.258, p=0.073; LO2: F(1,8)=1.703, p=0.228; V3a: F(1,8)=7.497, p=0.026; V3b: F(1,8)=5.690, p=0.044;
size: (V1: F(1,8)=0.408, p=0.541; V2: F(1,8)=1.623, p=0.238; V3: F(1,8)=1.623,
p=0.238; hV4: F(1,8)=0.416, p=0.537; VO1: F(1,8)=8.997, p=0.017; VO2: F(1,8)=0.747, p=0.413; LO1: F(1,8)=2.224, p=0.174; LO2: F(1,8)=4.811, p=0.060; V3a: F(1,8)=3.330, p=0.105; V3b: F(1,8)=10.711, p=0.011). Overall, we do see some
significant linear contrasts, but the overall pattern of results is in the opposite
direction of what we would predict from our hypothesis. In other words, for these
significant linear contrasts, we see the least amount of release from adaptation in
the 4th level and the most in the 1st level.
Figure 12. ROI analysis for orientation thresholded by orientation > size.

ROI analysis only using selected voxels that showed a greater response during the orientation task compared to the size task. Averaged percent signal change in response to each of the four feature levels for the size condition in early visual ROIs. Data are shown from areas: V1, V2, V3, hV4, VO1 in the first row and VO2, LO1, LO2, V3a, and V3b in the second row. Each consecutive bar represents the percent signal change compared to baseline in response to viewing ensemble arrays at each of the four size levels. Shown in the top of each graph are results from an ANOVA with a linear contrast between each of the four levels for each ROI. Error bars represent the standard error of the mean.
Figure 13. ROI analysis for size thresholded by size > orientation.

ROI analysis only using selected voxels that showed a greater response during the size task compared to the orientation task. Averaged percent signal change in response to each of the four feature levels for the size condition in early visual ROIs. Data are shown from areas: V1, V2, V3, hV4, VO1 in the first row and VO2, LO1, LO2, V3a, and V3b in the second row. Each consecutive bar represents the percent signal change compared to baseline in response to viewing ensemble arrays at each of the four size levels. Shown in the top of each graph are results from an ANOVA with a linear contrast between each of the four levels for each ROI. Error bars represent the standard error of the mean.

Whole brain analysis

In addition to the ROI analysis we performed a whole brain analysis.

Specifically, we compared each level for either task using a one-way ANOVA across
all voxels. Additionally, we performed whole brain contrasts looking at the
difference between levels 1 and 2 with 3 and 4 as well as the difference between
level 1 and levels 2, 3, and 4. The data were thresholded at t=0.80.

Data for this analysis are plotted in Figure 14 for orientation and Figure 15
for size. Results from the ANOVA and 1, 2 vs. 3, 4 contrast for both orientation and
size are not plotted as they did not show any significant voxels. Overall, this analysis
confirmed what our ROI analysis showed, in that we have greater activation for
levels 2, 3, and 4 compared to level 1 for both size and orientation in early occipital
areas.
Figure 14. Whole brain contrast for orientation level 1 vs. levels 2, 3, and 4.

Contrast between orientation level 1 and levels 2, 3, and 4 across all voxels (threshold $t=0.8$). Voxels shown in red represent greater activation for levels 2, 3, and 4 compared to level 1. Data are plotted in Talairach space and the green cross hairs are centered on Talairach coordinates 9, 93, and 5.
**Figure 15. Whole brain contrast for size level 1 vs. level 2, 3, and 4.**

Contrast between orientation level 1 and levels 2, 3, and 4 across all voxels (threshold t=0.8). Voxels shown in red represent greater activation for levels 2, 3, and 4 compared to level 1. Data are plotted in Talairach space and the green cross hairs are centered on Talairach coordinates 9, 94, and 9.

**Experiment 2: Exploring the pattern of activity in processing feature averages across time using EEG and MVPA**
Introduction

To initially probe the neural response to an ensemble average, we will measure visually evoked potentials (VEPs) in response to ensemble arrays with different mean sizes and orientations. Then, we will perform MVPA analysis on each millisecond of the time series and across electrodes. This analysis is ideal because it allows us to identify how the representation of the mean evolves over time. In particular, this experiment will attempt to answer questions regarding when in time this process takes place for two different averaging tasks. Specifically, we can identify effects of attention and task relevance. For example, we might expect larger responses early during the VEP if this process does not require attention. Additionally, we might expect non-differentiable responses between the activity evoked by the two tasks, if ensemble encoding is task-irrelevant.

Methods

Participants

Twenty-one, right-handed, neurotypical adults with normal to corrected to normal vision participated in experiment 2 (7 females, aged 19-49). Each participant provided informed written consent as approved by the Institutional Review Board at the University of Nevada, Reno. Two participants were removed due to excessive noise in the EEG data or poor behavioral results indicative of chance performance.
Apparatus

The stimuli were presented on a Mitsubishi Diamond Pro 2070SB CRT monitor (20in, 1024x768) using a 120Hz refresh rate at a viewing distance of 73 cm. Stimuli were controlled using a 2.6Mhz MacMini and presented using PsychToolbox (Brainard, 1997; Pelli, 1997; Kleiner et al., 2007) for MATLAB (MathWorks Inc., Natick, MA).

EEG Data Acquisition

EEG was continuously recorded using a 256 channel HydroCell Geodesic Sensor Net using an EGI Net Amps Bio 300 bioamplifier (Electrical Geodesics Inc., Eugene, OR) at a sampling rate of 1000Hz. The digitized data were then recorded using Netstation 5.1. Impedance values were kept below 100 Ω where possible. Impedances were checked before starting the experiment. Frame-accurate timing of stimulus presentation was tested and validated using a photodiode.

Stimuli

The stimuli used were an 8x8 grid of framed ellipses. As stated above, ellipses were chosen because they possess two easily detectable object features: size and orientation. For each stimulus presentation, we randomized how many items were present in the array (either 36 or 44 total) while maintaining the same mean and number in each of the 4 visual quadrants. We used a reference average size of 1
DoVA and a reference average orientation of 0°. For the individual step sizes for both orientation and size, participant’s individual values were used (See psychophysical experiment for a detailed description; see Figure 4 for an example participant).

**Experimental procedure and task design**

Participants were asked to perform an averaging task in which they were required to discriminate between the average orientation or size of a group of framed ellipses and a single framed reference ellipse (Figure 16). For example, in a typical trial, participants will see the ensemble presented for 300ms followed by a response screen in which a single ellipse appeared at fixation with a size and orientation that matched the reference level on both size and orientation. Participants had between 800-1200ms (pseudo-randomly jittered) to respond in between each stimulus presentation. Participants were asked to judge how much larger or more rightward tilted on average the ensemble was in comparison with the single ellipse on a scale of 1-4, 1 being the same average 4 being maximally different on average.

There were 2 mean feature conditions, size and orientation, each with 5 levels. Here, we used an event related potential design, in which were presented each of the 25 conditions (5 levels or both orientation and size) pseudo randomly throughout the experiments. Trials were presented in blocks of 125, blocked based on the type of averaging task that was required. Each of the 25 feature combinations
was presented 5 times per block. There were a total of 2 runs, each containing 5 blocks, with 10 blocks total. Each averaging task was presented for 5 of the 10 blocks. This leaves total of 1250 trials per participant (625 per run), each lasting, on average, 1.3 seconds, resulting in a run time of 13.5 minutes. Participants were given breaks between blocks. Before participating in the experiment participants were exposed to 10 practice trials of the orientation and size task to familiarize themselves with the stimuli and tasks.

Figure 16. Experiment 2 example trial sequences.

Shown here is an example of the start of one block from experiment 2 along with the first two trials presented. Participants are shown an instruction screen that cues them to the task they will be performing for the duration of the block. After the participant initiates the start of the block with a button press, the first ensemble is displayed for 300 ms. This is then followed by a response screen with a jittered presentation time between 800-1200 ms in which the participant responds with how different, if at all, the ensemble average was from a representative ellipse presented at the center of the screen. This is then repeated for the duration of the block.
**Behavioral data analysis**

The purpose of the experimental design was to ensure participants could quickly and efficiently determine the average of a group of objects. Therefore, we wanted participants to judge the amount of change in the average across a visual feature compared to a standard value (0° for orientation; 1 DoVA for size). To do so we had participants respond to the amount of change present in the average of an ensemble of a scale of 1-4, 1 being no change 4 being maximum change.

First, we looked at participant behavioral performance on this task. Behavioral data were analyzed by first segmenting trials based on condition and task. We segmented trials into groups based on the level of the attended feature, combining trials from all levels of the unattended feature. Then, we averaged the behavioral response (1-4) for all trials of that condition. This resulted in 5 values (see Figure 17) for both tasks.

Our hypothesis would predict a linear increase in the average response that mirrors the linear increase in delta change. In other words, participants average should reflect the level of change between the comparison and test. To measure the statistical significance of this relationship, we performed linear contrasts. To ensure the participant was never responding to the non-attended feature, we also compared the average response when doing one task to the level of the non-attended feature, which allowed us to get a baseline level for comparison.
EEG data preprocessing

EEG data were preprocessed using a combination of FieldTrip toolboxes and custom Matlab scripts. A bandpass filter (0.5 – 30 Hz) was applied to the data so as to remove slow drift over time as well as electrical noise. The data were segmented into 700 ms segments each consisting of 700 time points, including a 100 ms baseline. Segmentation was performed using trigger pulses sent to the acquisition computer by Matlab at the time of stimulus onset. The average temporal offset between stimulus presentation and trigger pulse was measured using a photodiode and corrected for during segmentation. Next, ocular artifacts, including eyeblinks and eye movements were detected using FieldTrip functions. As trials were relatively short, any trial containing an ocular artifact was removed. Bad channels were detected using custom Matlab scripts and visual inspection using a weighted average interpolation from their nearest neighboring electrodes. Trials in which 100 or more channels were found to be bad were removed and channels that were found to be bad in 20 % of trials were removed. Lastly, the data were re-referenced from electrode Cz to the average of all electrodes.

Univariate analysis

First the preprocessed segments were grouped into 10 conditions: 5 orientation levels collapsed across all size levels and 5 size levels collapsed across all orientation levels. These trials were then averaged together for each time point,
electrode, and participant and baseline corrected using the 50ms baseline. A grand average was then calculated across all participants.

As a first pass analysis, to see if there were any consistent differences over time between conditions of varying magnitudes of change from the reference, we calculated the difference at each time point between level 1 of the condition and every other level. The prediction being that as we increase the difference between the presented level and the reference we would see a corresponding increase in the difference in the EEG signals.

To test for time points of significance, we computed paired samples T-Tests at every time point between the corresponding difference value and 0. We plotted the resulting t-values for every electrode as a function of the level of the condition and grouped based on their location on the scalp. Significant time points are shown in higher contrast and were gated based on a p < 0.05 (Figure 18 for orientation; Figure 19 for size).

**Multivariate analysis**

As we are aware of previous studies examining neural correlates of ensemble encoding that have failed to show meaningful results, we wanted to probe the data more thoroughly than is possible with univariate results themselves. Additionally, as we are presenting full field stimuli, differences are much more likely to be found amongst the pattern of activation as opposed to the aggregate activity. In an attempt
to probe for differences that may be hidden in the mean, we next performed various multivariate analysis.

*Split-half correlations*

First, to test whether the VEP contains information differentiating the levels of either feature, we performed a split-half correlation comparing level 1 with every other level for both features. This allows us to measure the similarity of the response between two levels of a feature as a function of the difference magnitude between the two levels. To do this we first divided the trials for each feature level (orientation level 1, 2, 3, 4, 5 collapsed across size levels and size level 1, 2, 3, 4, 5 collapsed across all orientation levels) into two halves pseudo-randomly. We then averaged and baseline corrected these averages using the 100 ms baseline. Next, correlations were performed across all electrodes for the first half of every level and the second half of every other level at every time point. For example, one half of the orientation level one trials were correlated with on half of the orientation level two trials across all electrodes. This resulted in a heat map representing the correlations of the chosen comparison level with every other level over time (Figure 20 for orientation; Figure 21 for size). This analysis was done separately for orientation and size trials. Finally, the topographic maps were averaged over all participants. To test for significance, we took the difference between the within and between level correlations. Specifically we wanted to determine if there were any differences between the VEP of each level for both features. For example, for level 1
correlations, we examine the difference between the correlation of 1 vs 1 and subtracted it from the average correlations between 1 vs 2, 1 vs 3, 1 vs 4, and 1 vs 5. This was done for all levels of both features for all time points. Next, t-tests were taken at each time point and these differences were plotted over time for all level comparisons (Figure 22 for orientation; Figure 23 for size). Error bars represent standard error of the mean.

**Multivariate pattern analysis**

In order to identify differences that occur among the pattern of VEP responses across electrodes and time, we performed multivariate pattern analysis using CosmoMVPA Matlab toolbox (Oosterhof, Connolly & Haxby, 2016). First, in order to increase the signal-to-noise ratio, we performed sub-averages across groups of 5 segments within each group, which halved the number of trials being fed into the classifier but provided much cleaner VEP data for the MVPA classifier. Trials for each sub-average were pseudo-randomly chosen. As the number of trials in each condition was slightly different due to artifact and trial rejection, we threw out any left over trials during the sub-averaging process. For example, if there were 104 trials in one condition, after sub-averaging in groups of 5, the 4 remaining trials were not used. Next, MVPA classification was performed independently for all participants using these sub-averaged trials.

We trained naïve Bayes and LDA classifiers to attempt to discriminate between the pattern of activity evoked by feature level 1 compared with every other
feature level for both features of interest. Classifier performance was evaluated using the k-fold cross validation method. Specifically, for each millisecond in the time course the EEG data were grouped based on the condition of interest for each electrode. Then, we split the data into 10 chunks and randomly assigned each sub-average to 1 of the 10 chunks. Of these 10 chunks 9 were pseudo-randomly chosen as ‘training’ trials and 1 was left out as the ‘test’ set. We repeated this process 10 times, holding out a new trial as the test each time. We evaluated the performance of the classifier by comparing classification performance of the test sets for each of the 10 iterations with chance level (50%). Classification accuracy was then plotted for each time point and one-sample t-tests were performed between the accuracy value and chance (Figure 17 for LDA orientation; Figure 18 for LDA size; Figure 19 for naïve Bayes orientation; Figure 20 for naïve Bayes size).

**Results**

**Behavioral results**

The results of the behavioral analysis are plotted in Figure 10. The two left plots show the average behavioral response for each of the 5 feature levels for orientation or size while performing the orientation or size task respectively. Plotted on the two right figures are the average behavioral responses to orientation or size while performing the opposite task. In other words, these right plots look at participants average responses to the unattended feature and serve as a sanity
check to ensure participants aren’t using feature information from the irrelevant feature in their judgments.

A positive linear relationship is clearly visible for orientation while doing the orientation task (top right) and size while doing the size task (bottom right). This was confirmed by a repeated measures ANOVA with a linear contrast (orientation: \( F(1,18) = 130.317, p < 0.001 \); size: \( F(1,18) = 70.131, p < 0.001 \)). A linear contrast was also performed for the size values while performing the orientation task and orientation values while performing the size task (orientation while size: \( F(1,18) = 16.839, p = 0.001 \); size while orientation: \( F(1,18) = 41.911, p < 0.001 \)). While these analyses were significant, this is mainly due to very slight increase between the last level and first four levels, which can be seen by the post-hoc pairwise comparisons between the first and last level. This difference is expected due to the variance present within each feature in the ensembles. Additionally, post-hoc pairwise comparisons were performed between each feature level and any significant differences are plotted.
Figure 17. Experiment 2 behavioral results.

Behavioral results from experiment 2 plotted for orientation levels during the orientation task (top left), size during the size task (bottom left), size during the orientation task (top right), and orientation during the size task (bottom right). Plotted are average responses to each of the five levels of either orientation or task. Error bars represent the standard error of the mean. Significance markers represent significance at a threshold of: * - p=0.05, ** - p=0.01, *** - p=0.001.

Univariate time course analysis

The first analysis performed, directly compared the visually evoked potentials (VEP) between each of the levels for either feature. We took the preprocessed and averaged VEPs from each of the 5 levels of orientation and size.
and took the difference between the level one and each other level from two to five. This was performed for each electrode at every time point.

As a first pass statistical analysis, we performed one sample t-tests between the amplitude difference and 0 at each time point for every electrode. The resulting test statistics are plotted in the heat maps in Figure 18 for size and Figure 19 for orientation. Warmer colors represent larger amplitudes for the larger levels (levels 2, 3, 4, or 5) whereas cooler colors represent larger amplitudes for level 1. Points that are more saturated represents time points that are significantly greater or less than 0 at an alpha of 0.05. As is clear from visually inspecting the heat maps almost no areas are highlighted for either orientation or size. In other words, no electrodes show a difference significantly greater than 0 for any period of time more than what would be predicted by chance. Additionally, in viewing the areas of significance, no predictable pattern emerges. For example, one prediction that would support a neural representation of the average would be that as we increase the difference between the levels we should see increased significance and this time range of significance should be similar in each comparison. Instead, areas of significance seem to be random. As a result, we cannot detect differences in any potential neural representation of the mean using this analysis.
Figure 18. Experiment 2 test statistics plotted over time for all electrodes for orientation.

Plotted are the resulting heat maps of test-statistics from comparisons of the differences between the average VEP for feature level 1 and every other level with 0 for every time point and electrode for the orientation condition. Graphs from top to bottom are comparisons between orientation levels 1-5, 1-4, 1-3, and 1-2. Electrodes are grouped based on location on the scalp and plotted on the y-axis. The vertical black line represents the end of the baseline period and the start of the stimulus onset. Values are thresholded by an alpha of p=0.05 and significant time points are shown at a higher contrast. Test statistics range from 1 (yellow) to -1 (blue).
Figure 19. Experiment 2 test statistics plotted over time for all electrodes for size.

Plotted are the resulting heat maps of test-statistics from comparisons of the differences between the average VEP for feature level 1 and every other level with 0 for every time point and electrode for the size condition. Graphs from top to bottom are comparisons between size levels 1-5, 1-4, 1-3, and 1-2. Electrodes are grouped based on location on the scalp and plotted on the y-axis. The vertical black line represents the end of the baseline period and the start of the stimulus onset. Values are thresholded by an alpha of p=0.05 and significant time points are shown at a higher contrast. Test statistics range from 1 (yellow) to -1 (blue).

Multivariate analysis
Split-half correlations

To probe this dataset more closely, we next implemented a split-half correlation analysis. This analysis allows us to examine the relationship between the variations in the signal across all electrodes across the two conditions being compared. The results of this analysis are plotted in Figure 20 for orientation correlations and Figure 21 for size correlations. Each plot from top to bottom shows the correlation between one level with every other level. For example, in the first row of the top most heat map, the correlations plotted represent the relationship between the first half of level one with the second half of level one. The first row of the second heat map shows the correlations between the first half of level 2 with the second half of level 1. Warmer values represent positive correlations and cooler values represent negative correlations, ranging from 1 to -1. To test for significance we took the differences between the within category correlation and the average of the between category correlations. The resulting difference waves are plotted in Figure 22 for orientation and Figure 23 for size.

On initial inspection, correlations are approximately 0 during the pre-stimulus baseline and for the first 100 ms. This was expected as nothing in the VEP should be visible before the information has had an opportunity to reach visual cortex. However, after this initial period, the correlations become consistently positive across all conditions. Importantly, the correlations between the first and second halves of each feature of the same level (e.g. level 1 half 1 with level 1 half 2) are close to 1 for the entire stimulus presentation, which indicates that the data are
clean and the signal to noise ratio is low. This shows that each level is showing clean VEPs but that the differences between these neural signatures are either too small to detect using this method or are non-existent.

Figure 20. Orientation split-half correlations.

Pearson’s correlation coefficients are plotted along the x-axis for every time point. Each graph compares the average of half of the data for one level with the average of the other half of every other level for the orientation condition using all channels. Specifically, plotted along the y-axis are the r-values for each time point, comparing the average of half of the trials of levels 1, 2, 3, 4, and 5 with the average of the other half of levels 1 (first row), 2 (second row), 3 (third row), 4 (fourth row), and 5 (fifth row). The vertical black line represents the end of the baseline period and the start of the stimulus onset. Correlation values range from 1 (yellow) to -1 (blue).
Figure 21. Size split half correlations.

Pearson’s correlation coefficients are plotted along the x-axis for every time point. Each graph compares the average of half of the data for one level with the average of the other half of every other level for the size condition using all channels. Specifically, plotted along the y-axis are the r-values for each time point, comparing the average of half of the trials of levels 1, 2, 3, 4, and 5 with the average of the other half of levels 1 (first row), 2 (second row), 3 (third row), 4 (fourth row), and 5 (fifth row). The vertical black line represents the end of the baseline period and the start of the stimulus onset. Correlation values range from 1 (yellow) to -1 (blue).
Figure 22. Differences between within and between conditions split half correlations for orientation.

Plotted are difference waves calculated by taking the difference between the correlation of one half of each feature level with the other half of the same feature level (within condition correlation) and the average correlation of one half of each feature with the other half of all other feature levels (between condition correlation). Plotted on the y-axis are correlation statistics differences. Error bars represent standard error of the mean. The vertical black line represents the end of the baseline period and the start of the stimulus onset.
Figure 23. Differences between within and between conditions split half correlations for size.

Plotted are difference waves calculated by taking the difference between the correlation of one half of each feature level with the other half of the same feature level (within condition correlation) and the average correlation of one half of each feature with the other half of all other feature levels (between condition correlation). Plotted on the y-axis are correlation statistics differences. Error bars represent standard error of the mean. The vertical black line represents the end of the baseline period and the start of the stimulus onset.

Lastly, as we used all electrodes in the correlations, we z-score transformed the VEP data for each electrode across time. Specifically, we z-scored the average VEP data from each half for each of the 10 conditions by subtracting out the average
of all time points from each time point for all electrodes. We then repeated the split half correlation analysis described above. The data are plotted in the heat maps in Figure 24 for orientation and Figure 25 for size. As can be seen here, the correlations are very small and, like the above analysis, are very similar to one another.

![Heat Maps](image)

**Figure 24. Split half correlations of the z-scored VEPs for the orientation task.**

Pearson's correlation coefficients are plotted along the x-axis for every time point, comparing the average of half of the data for each feature with the other half of every other feature for the z-scored orientation condition. Plotted along the y-axis are the correlations between the second half of VEPs for orientation levels 1, 2, 3, 4, and 5 with the first half of levels 1 (first row), 2 (second row), 3 (third row), 4
(fourth row), and 5 (fifth row). The vertical black line represents the end of the baseline period and the start of the stimulus onset. Correlation values range from .25 (yellow) to -.25 (blue).

**Figure 25.** Split half correlations of the z-scored VEPs for the size task.

Pearson’s correlation coefficients are plotted along the x-axis for every time point, comparing the average of half of the data for each feature with the other half of every other feature for the z-scored size condition. Plotted along the y-axis are the correlations between the second half of VEPs for orientation levels 1, 2, 3, 4, and 5 with the first half of levels 1 (first row), 2 (second row), 3 (third row), 4 (fourth row), and 5 (fifth row). The vertical black line represents the end of the baseline period and the start of the stimulus onset. Correlation values range from .25 (yellow) to -.25 (blue).
*Multivariate pattern analysis*

Although the first multivariate analysis did not show significant differences between conditions, there may still be small underlying differences not detectable by the split-half correlation analysis. To further explore this dataset, we trained a multivariate pattern analysis (MVPA) classifier to look for differences between level 1 compared with each other level for both size and orientation. Importantly, we ran separate MVPA classifiers for each comparison level separately so chance performance remains at 50% as there were only every two options for the classifier to chose from. To ensure that the classification accuracy wasn’t solely dependent on the type of classifier used, we trained two: a naïve Bayes and an LDA classifier.

LDA classification accuracy for all time points is plotted in Figure 26 for orientation and Figure 27 for size while naïve Bayes classification accuracy is plotted in Figure 28 for orientation and Figure 29 for size. Similarly to the split-half correlation analysis, we see no evidence of increased classification performance as the two feature levels become more different for size or orientation in either of the classifiers used.
Figure 26. Orientation LDA classification accuracy.

LDA classification accuracy for the orientation condition for all time points for each of 4 comparisons made between feature level 1 and feature levels 2-5 (rows 1-4). Shaded regions around the curve represent standard error of the mean. The horizontal black line represents the 50% chance performance mark. The vertical black line represents the end of the baseline period and the start of the stimulus onset.
Figure 27. Size LDA classification accuracy.

LDA classification accuracy for the size condition for all time points for each of 4 comparisons made between feature level 1 and feature levels 2-5 (rows 1-4). Shaded regions around the curve represent standard error of the mean. The horizontal black line represents the 50% chance performance mark. The vertical black line represents the end of the baseline period and the start of the stimulus onset.
Figure 28. Orientation Naïve Bayes classification accuracy.

Naïve Bayes classification accuracy for the orientation condition for all time points for each of 4 comparisons made between feature level 1 and feature levels 2-5 (rows 1-4). Shaded regions around the curve represent standard error of the mean. The horizontal black line represents the 50% chance performance mark. The vertical black line represents the end of the baseline period and the start of the stimulus onset.
Figure 29. Size Naïve Bayes classification accuracy.
Naïve Bayes classification accuracy for the size condition for all time points for each of 4 comparisons made between feature level 1 and feature levels 2-5 (rows 1-4). Shaded regions around the curve represent standard error of the mean. The horizontal black line represents the 50% chance performance mark. The vertical black line represents the end of the baseline period and the start of the stimulus onset.

Experiment 3: Identify the effects of attention and task on the representation of the average using EEG and a frequency tagging oddball paradigm.
Introduction

Experiment 3 investigates the influence of attention and task on the neurometric function of ensemble encoding. Specifically, we used EEG and a multiplexed frequency tagging oddball paradigm (Guillaume, Mejias, Rossion, Dzhelyova, & Schiltz, 2018; Liu-Shuang, Norcia, & Rossion, 2014). Using this design allows us to measure the potential representation of the mean while manipulating both the portion of the ensemble that is attend to (spatial attention) and the type of averaging task the participant is performing. Specifically, we will look at the neural signature of the average while the ensemble is inside and outside the focus of attention while the participant is either actively trying to extract the mean feature or not for each feature and each of the 5 levels of those features.

After initial piloting of the experimental task, we found that the stimulus levels chosen through the psychophysical task were too difficult to use in the current paradigm. Therefore, in order to make the task easier, we implemented a minimum difference between the levels of each feature. We ensured a minimum difference of 10° for orientation and 0.1 DoVA for size. These new levels resulted in better overall performance on the task.

Methods
Participants

Twenty neurotypical adults (19 right-handed) with normal to corrected to normal vision participated in experiment 2 (11 males, aged 21-39). Each participant provided informed written consent as approved by the Institutional Review Board at the University of Nevada, Reno.

Apparatus

The stimuli were presented on a Mitsubishi Diamond Pro 2070SB CRT monitor (20in, 1024x768) using a 120Hz refresh rate at a viewing distance of 73 cm. Stimuli were controlled using a 2.6Mhz MacMini and presented using PsychToolbox (Brainard, 1997; Pelli, 1997; Kleiner et al., 2007) for MATLAB (MathWorks Inc., Natick, MA).

EEG Data Acquisition

EEG was continuously recorded using a 256 channel HydroCell Geodesic Sensor Net using an EGI Net Amps Bio 300 bioamplifier (Electrical Geodesics Inc., Eugene, OR) at a sampling rate of 1000Hz. The digitized data were then recorded using Netstation 5.1. Impedance values were kept below 100 Ω where possible. As each run was relatively long, impedances were checked before starting the experiment and half-way through the experiment. Frame-accurate timing of stimulus presentation was tested and validated using a photodiode.
Stimuli

The stimuli used here were the same experiment 1, except in the case where the minimum difference between the feature levels did not meet the minimum threshold (10° for orientation; 0.1 DoVA for size). In other words, if the difference between orientation level 1 and orientation level 2 was less than 10° then the new orientation level 2 becomes orientation level 1 plus 10°. Similarly, if the difference between the new orientation level 2 and orientation level 3 was less than 10° then the new orientation level 3 becomes orientation level 2 plus 10°. This rule was enforced for each level for both size and orientation.

Experimental procedure and task design

Participants were asked to perform an averaging task that controls for the stimulus that is actively being attended to and averaged (Figure 30). First, before the start of the trial, we cued participants to which feature they should be actively averaging. In addition to attending to only one feature, participants were instructed to spatially attend to and make judgments on only one side of visual space. During each trial, ensembles appeared and disappeared on the left and right side of fixation at either 3 or 5 Hz (pseudo-randomly chosen for each trial). During each new presentation (each stimulus onset) the position, orientation, and size of each individual ellipse changed while still maintaining the overall average. Throughout the trial, oddball feature changes occurred at specific underlying frequencies (0.6 and 0.75 Hz oddball frequencies for the 3 Hz carrier frequency and 0.8 and 2 Hz
oddball frequencies for the 5 Hz carrier frequency). Then, after each trial, participants were probed by asking how different the oddballs were from the reference presentations for the specific feature dimension on a scale from 1-4, 1 being no change and 4 being maximum change.

Each of the 25 feature combinations was presented 8 times, twice for each task and attended hemifield, leaving a total of 200, 20 second trials. We chose 20 seconds to maximize resolution in the frequency domain and so we had enough resolution to detect the oddball frequencies. As each trial lasted 20 seconds and participants were encouraged to take breaks between trials, the total run time was fairly long. As a result, the experiment was split into two halves, each run on separate days. The focus of spatial attention was changed for either run (pseudo-randomly chosen left/right for first/second run). Participants were cued before each block and instructed to attend to either size or orientation. The averaging task was specifically directed to the chosen hemifield. Importantly, the non-relevant feature also changed in the attended hemifield at the other oddball rate, and both the relevant and non-relevant feature were changed in the unattended hemifield. This allowed us to examine the frequency tags to changes in the oddball for both the task-relevant and task-irrelevant features in the attended and unattended side of visual space.
Figure 30. Experiment 3 frequency tagging trial sequence.

Shown here is an example of the start of one 20 second trial from experiment 3 along with the first two 2 seconds of the stimulus presentation. Participants are first shown an instruction screen that cues them to the task they will be performing for the duration of the block (not shown). Then participants are shown a cue at the start of each trial indicating which side of the screen they should be attending. This is followed by 20 seconds of continuous flickering of the ensemble. Each half of the ensemble is presented at a different flicker frequency. Shown in the example here, on the right is 3 Hz and left is 5 Hz. Periodically, oddball stimuli are presented in which either the orientation or size average changes. Shown here, orientation is changing at a rate of .6 Hz and size at a rate of .75 Hz on the right and 0.8 Hz for orientation and 2 Hz for size on the left. The trial lasts for an additional 18 seconds (20 seconds total). After the trial finishes the participant is presented with a response screen in which they have unlimited time to response using the 1-4 keys, indicating how different the oddball averages were from the standard presentations.

The multiplexed frequency tagging oddball paradigm

In a frequency tagging paradigm stimuli that are presented at a periodic rate will produce a corresponding periodic neural signature that can be detected using EEG (Norcia, Appelbaum, Ales, Cottereau, & Rossion, 2015; Regan, 1966) (See Figure 31). For example, presenting a stimulus that is flickering at 6 Hz will produce a large
6 Hz component in the EEG signal, also known as a frequency tag. This frequency tag is the result of overlapping transient evoked potentials in response to each individual stimulus presentation and can be detected in the EEG signal using Fourier analysis. More recently, a modified version of this technique has been developed that uses frequency tagging along with the presentation of an oddball stimulus in the flickering stream at different periodic rates (Liu-Shuang et al., 2014). For example, in an experiment looking at neural markers of numerosity, researchers presented groups of objects to participants that differed across a variety features such as item density, individual item size, and item shape, at a periodic rate of 10 Hz (carrier frequency) (Guillaume et al., 2018). Importantly, embedded within this periodic stimulus presentation was another slower periodic presentation (oddball frequency) in which the numerosity of the item set would change. In other words, while some of the features of the set changed 10 times per second, the numerosity of the set changed at a rate of 1.25 Hz, or every 8 stimulus presentations. This then produces a 1.25 Hz component in the corresponding EEG. Additionally, we can multiplex this design by using two oddball frequencies inlaid within the baseline frequency, one in which the size changes and the other in which the orientation changes. Lastly, we can present two separate streams to the left and right side of fixation, with carrier frequencies at 3 and 5 Hz, each with their own two oddball rates, 0.6 and 0.75 Hz in the 3 Hz stream and 0.8 and 2 Hz in the 5 Hz stream.
Figure 31. Multiplexed frequency tagging oddball paradigm.

An example of a frequency tagging paradigm. In this example the stimuli is flickering at 6 Hz, meaning it turns on and off 6 times. Then the preprocessed and averaged EEG signal is transformed using a fast Fourier transform. This then transforms the data into the frequency domain. Plotted in the third row are example frequency tags for the above waveform, with larger amplitudes for the 1st, 2nd, and 3rd harmonics of the presentation frequency.

Behavioral data analysis

Similarly to experiment 1, participants were asked to judge the difference of the test ensemble average compared to the reference ensemble average on a scale of 1-4. Trials were collapsed across the non-relevant feature leaving 10 trials per feature level per task per attended hemifield. Responses to each feature level were
averaged together and plotted in Figure 32. To look for a positive linear pattern among the responses, we performed an ANOVA with a linear contrast.

**EEG data preprocessing**

EEG data were preprocessed using a combination of FieldTrip toolboxes and custom Matlab scripts. First, a bandpass filter (0.5 – 59 Hz) was applied to the data so as to remove slow drift over time as well as electrical noise. The data was segmented for each trial, each segment consisting of 20,000 time points. Segmentation was performed using trigger pulses sent to the acquisition computer by Matlab at the time of stimulus onset. The average temporal offset between stimulus presentation and trigger pulse was measured using a photodiode and corrected for during segmentation. Next, ocular artifacts, including eyeblinks and eye movements were detected using FieldTrip functions. As trials were relatively long, trials that contained more than 10 ocular artifacts were removed. Lastly, bad channels were detected using custom Matlab scripts and visual inspection using a weighted average interpolation from their nearest neighboring electrodes. Trials in which 100 channels were found to be bad were removed and channels that were found to be bad in 20% of trials were removed.

**Carrier frequency analysis**

Before examining the frequency tags for the oddball stimuli we examined the carrier frequency tags (3 and 5 Hz). In order to be confident in the oddball results, it
is integral to perform some quality control baseline level analysis. To do so we will examine the carrier frequency tags. One reason for this is that these tags should be fairly high powered as every trial has a stimulus being presented at those rates. Additionally, the carrier frequencies should reflect some general changes based on the state of the participant but conditions not related to experimental conditions. For example we should be larger frequency tags to attended stimuli compared to unattended stimuli. After preprocessing, segments were grouped based on carrier frequency, attended hemifield, and task.

Sanity check: Are participants awake?

Before any other analysis, we want to ensure the reliability of our frequency tags. In order to do so, we examined how variable the phase of each frequency tag was within frequency tags of the same condition. First, we segmented trials in groups by carrier frequency, attended hemifield, and task. Importantly, we collapsed across the various feature levels. So, for example, one group contained all trials in which 3 Hz was presented on the left, the left was the attended hemifield, and the participant was performing the orientation task. This was done for 5 and 3 Hz, orientation and size task, and both for the attended and unattended hemifield. Then we performed a fast Fourier transform (FFT) on each trial in each group, translating the data from the time to the frequency domain. Next, we picked of the frequencies that related to our carriers (first three harmonics of each carrier). This was done for all electrodes and for all participants. Using this raw output from the FFT we
calculated the phase angle for each frequency tag using a $T_{\text{circ}}^2$ analysis (Victor & Mast, 1991). The $T_{\text{circ}}^2$ statistic is a statistical measure of how variable the phase angle is between multiple frequency components (Victor & Mast, 1991). We then plotted each of these $T_{\text{circ}}^2$ test statistics for all subjects thresholded by a p-value of 0.01 (See Figure 33 for orientation and Figure 34 for size in an example participant).

**Sanity check: Are participants paying attention?**

Next, we wanted to ensure participants were attending to the required hemifield. To do so we compared attended frequency tags with unattended frequency tags. We grouped trials in the same way as for the $t_{\text{circ}}$ analysis. This resulted in frequency tags for 3 and 5 Hz, while attending left and right or not attending left and right, while doing the orientation or size task. Next we created index values comparing the attended and unattended tags, but dividing their difference by their sum:

$$\text{Index}_{f_1} = \frac{\text{Attended}_{f_1} - \text{Unattended}_{f_1}}{\text{Attended}_{f_1} + \text{Unattended}_{f_1}},$$

where $f_1$ represents one of the carrier frequencies. This produced index values comparing attended and unattended frequency tags for 3 or 5 Hz stimuli presented on the left or right for the orientation or size task. These index values ranged from 1 to -1, where positive values correspond with larger attended tags. Next we averaged the indices across participants and performed a paired samples $t$-test for each electrode and 0, thresholded at an alpha of $p = 0.01$. The resulting test statistics
were then plotted in topographic heat maps (Figure 35). Electrodes that reached significance are bolded.

**Oddball frequency analysis**

*Feature levels vs Δ levels*

This experiment left relatively few trials per condition. Overall, there were 10 trials per feature level per task per attended hemifield and further divided based on carrier and oddball frequency location. For example, although there were 10 trials for each condition, only a portion of those trials were 5 Hz left and 3 Hz right and an even smaller number of trials had 0.8 Hz orientation oddball. Therefore in order to increase power, we created delta levels using our feature level frequency tags.

On one hand, using the feature level frequency tags should show a predictable pattern consistent with our hypothesis. This is because the chosen paradigm takes advantage of classic visual adaptation. In other words, we present oddball stimuli, in which the mean changes relative to the continuous reference mean seen at the carrier frequency. As a result, we should see larger oddball frequency tags for stimuli that are more different from the reference.

However, even though we should see linear differences between the frequency tags as we increase the difference between the test and reference, we can also increase power by creating what we call delta levels. To make delta levels we
take the average of differences between the frequency tags of subsequently larger step sizes between each feature level. So, for a delta level of 1, we took the average of the differences between feature level 1 and 2, 2 and 3, 3 and 4, and 4 and 5. For a delta level of 2 we took the average of the differences between feature level 1 and 3, 2 and 4, and 3 and 5. This was done for delta levels 1-4 for each condition. For the remaining analyses we will report both the feature level and delta level results.

_Frequency tag vs. index value analysis_

One problem with comparing frequency tags directly, is that tags for different frequencies can often times be larger or smaller regardless of any effects due to stimulation. For example, frequency tags from 1-3 Hz tend to be larger than those from 20-24 Hz. Therefore, we created index values using the attended and unattended frequency tags. For example, for the orientation tags presented on the left, we took the attended left and unattended left tags, for each oddball frequency, and divided their difference by their sum:

\[
Index_{f_1} = \frac{\text{Attended Left}_{f_1} - \text{Untended Left}_{f_1}}{\text{Attended Left}_{f_1} + \text{Untended Left}_{f_1}}.
\]

This results in one value that ranges between -1 to 1, where positive values represent larger attended tags and negative values represent larger unattended tags. For each analysis that follows that contrasts two conditions (e.g., task relevant vs. task irrelevant, attended vs. unattended) we analyze both the frequency tags for both conditions as well as the index values made from these frequency tags.
Four types of ensemble encoding

As discussed in the introduction, our hypothesis can lead to 3 specific predictions about how the average is represented in the neural signal as well as how attention and task can change this representation. For example, if ensemble encoding is pre-attentive and task general, we would expect the frequency tags from each consecutive level to be consecutively larger in a linear fashion. Importantly, we would predict no difference in the frequency tags as a function of task or focus of attention. In other words, if this process is truly pre-attentive and task general, we would predict the attended and unattended oddball frequencies to be statistically the same. Similarly, if this process requires attention, we would predict a linear increase in the size of the attended tags and no increase or change in the unattended tags. Here, we show analysis designed to test each of the three predictions.

Strong ensemble encoding

Prediction 1 states that ensemble encoding is a pre-attentive stimulus general process. In other words, the process of extracting the average feature happens in parallel automatically and with no effort. To test this prediction, we first grouped our segmented and preprocessed trials only using the feature level for either the size or orientation oddballs. In other words, we collapsed across all other conditions, including attended hemifield and task. This resulted in five groups of trials for both size and orientation for each of the four oddball frequencies. Next, we averaged together all trials from each group and performed a fast Fourier transform
on each average. This transformed the data from the time to the frequency domain. We then picked off the corresponding frequency tags that matched the oddball rate for that group. We then averaged across oddball frequency tag and participant. Lastly, we examined the relationship between the frequency tags of each level. To do so, we performed an ANOVA with a linear contrast to test if the frequency tags increased in a linear fashion. We plotted the resulting f-statistics for each electrode in a topographic heat map and bolded significant electrodes by a threshold of \( p = 0.05 \) (Figures Figure 36Figure 37Figure 38Figure 39). FDR q values are also calculated and reported.

**Weak ensemble encoding 1**

Prediction 2 states that in order to extract the average we need to be actively trying to, either by attending to the area in which the ensemble is presented or attending the average feature itself. To test the first part of the weak ensemble encoding prediction, we grouped our segmented and preprocessed trials based on attended hemifield and feature level, ignoring the task condition. Next, we averaged together all trials from each group and performed a fast Fourier transform on each average. This transformed the data from the time to the frequency domain. We then picked off the corresponding frequency tags that matched the oddball rate for that group. For example, when looking at the oddball frequency tags for each level of orientation, we would group trials according to the orientation oddball frequency presented in either the attended or unattended hemifield for that trial, regardless of
the task the participant is asked to perform. This leaves 2 groups: 1 for the attend and unattended hemifields, for each oddball frequency and for each feature level.

Like for prediction one above, we examined the relationship between the frequency tags of each level. Using these frequency tags (or index values), we performed an ANOVA with a linear contrast to test if the tags increased in a linear fashion, with level 1 having the smallest tag and level 5 the largest. We plotted the resulting f-statistics for each electrode in a topographic heat map and bolded significant electrodes using a threshold of \( p = 0.05 \) (Figures 40, 41, 42, 43, 44, 45, 46, and 47). FDR q values are also calculated and reported.

Part 2 of prediction 2 predicts that ensemble encoding requires an explicit averaging task, but not attention, in order to extract the average. In other words, this prediction states that if participants are actively trying to extract the average orientation, they can do so regardless of where they focus their spatial attention. To test this, we grouped our segmented and preprocessed trials based on the task the participant was asked to perform and feature level, ignoring the attended hemifield. Next, we averaged together all trials from each group and performed a fast Fourier transform on each average. This transformed the data from the time to the frequency domain. We then picked off the corresponding frequency tags that matched the oddball rate for that group. For example, when looking at the oddball frequency tags for each level of orientation, we grouped trials according to the orientation oddball frequency presented while the participant is performing the
orientation or size task for that trial, regardless of the attended hemifield. This leaves 2 groups: 1 for the size and orientation tasks, for each oddball frequency and for each feature level. Index values and linear contrasts were performed the same way as in part 1 of prediction 2, except the two conditions of interest were relevant task vs non-relevant task instead of attend vs. unattended hemifield. Frequency tags and index values are plotted in Figures Figure 48, Figure 49, Figure 50, Figure 51, Figure 52, Figure 53, Figure 54, and Figure 55.

**Weak ensemble encoding 2**

For the last set of analysis, we tested the prediction that ensemble encoding is only found when the participant is actively trying to extract the feature of interest and attending to the ensemble (non-automatic and post-attentive). To test this prediction we performed 3 analysis in a similar way to those presented for prediction two. We grouped our trials based on the presence or absence of an averaging task and the presence or absence of attention. Specifically, we grouped the trials as a function of: no task attended vs no task unattended, task attended vs task unattended, and task attended vs no task unattended.

Part 1 of prediction 3 examines the relationship between spatial attention and task on our ability to extract the average of an ensemble. It predicts that ensemble encoding requires participants to either: attend while performing an averaging task or perform an averaging task while attending. To test these predictions, we grouped our segmented and preprocessed trials by attended
hemifield while the participant was performing an averaging task or not. Next, we averaged together all trials from each group and performed a fast Fourier transform on each average. This transformed the data from the time to the frequency domain. We then picked off the corresponding frequency tags that matched the oddball rate for that group. This resulted in four groups of trials: task present attended hemifield, task present unattended hemifield, task absent attended hemifield, and task absent unattended hemifield. Index values and linear contrasts were performed the same way as in prediction 2. However, the two comparisons made were between task present attended vs task present unattended and task absent attended vs task absent unattended (see Figures Figure 56, Figure 57, Figure 58, Figure 59, Figure 60, Figure 61, Figure 62, and Figure 63 for task absent attended vs. task absent unattended and Figures Figure 64, Figure 65, Figure 66, Figure 67, Figure 68, Figure 69, Figure 70, and Figure 71 for task present attended vs. task present unattended).

Part 2 of prediction 3 more directly examines the relationship between task and attention, by comparing conditions when the participant was required to perform an averaging task while attending vs when the participant was not attending nor performing an averaging task. To test this prediction, we separated our segmented and preprocessed trials into two groups one for tags that were attended and task relevant and one for tags that were unattended and task irrelevant. Next, we averaged together all trials from both groups and performed a fast Fourier transform on each average. This transformed the data from the time to
the frequency domain. We then picked off the corresponding frequency tags that matched the oddball rate for that group. Index values and linear contrasts were performed the same way as in prediction 2 (see Figures Figure 72, Figure 73, Figure 74, Figure 75, Figure 76, Figure 77, Figure 78, and Figure 79).

Results

Behavioral results

Behavioral results were analyzed in a similar way to those from experiment 1 and 2. However, this experiment had the added condition of attended hemifield. For this behavioral analysis we collapsed these conditions together however the pattern shown in the collapsed analysis strongly match the results of each analysis done separately.

The results of the behavioral analysis are plotted in Figure 32. The two left plots show the average behavioral response for each of the 5 feature levels for orientation or size while performing the orientation or size task respectively. Plotted on the two right figures are the average behavioral responses to orientation or size while performing the opposite task. In other words, these right plots look at participants average responses to the unattended feature and serve as a sanity check to ensure participants aren’t using feature information from the irrelevant feature in their judgments.
Similarly to experiment 1, a positive linear relationship is clearly visible for orientation while doing the orientation task (top right) and size while doing the size task (bottom right). This was confirmed by a repeated measures ANOVA with a linear contrast (orientation: F(1,19) = 173.837, p < 0.001; size: F(1,19) = 469.108, p = p < 0.001). A linear contrast was also performed for the size values while performing the orientation task and orientation values while performing the size task (orientation while size: F(1,19) = 1.665, p = 0.212; size while orientation: F(1,19) = 11.738; p = 0.003). While the analysis was significant for the size values while performing the orientation task, this is mainly due to very slight increase between the fourth level and first level, which can be seen by the significant post-hoc pairwise comparison. Additionally, post-hoc pairwise comparisons were performed between each feature level and any significant differences are plotted.
Figure 32. Experiment 3 behavioral results.

Behavioral results from experiment 3 plotted for orientation levels during the orientation task (top left), size during the size task (bottom left), size during the orientation task (top right), and orientation during the size task (bottom right). Plotted are average responses to each of the five levels of either orientation or task. Error bars represent the standard error of the mean. Significance markers represent significance at a threshold of: * - p=0.05, ** - p=0.01, *** - p=0.001.

Carrier frequency analysis
$T_{circ^2}$ analysis

First, using the raw output from the Fourier transformed trials for each condition we extracted the phase. Using these phase values, we performed a $T_{circ^2}$ analysis (Victor & Mast, 1991). This statistical analysis takes the phase angle of the carrier frequency components in all trials in a given condition and determines if they are significantly different from one another. This was performed for each participant and for tags presented in either hemifield, for both features. The resulting test statistics for one example participant are plotted in Figure 33 for orientation and Figure 34 for size. Lighter values in the topographic map represent larger test statistics and darker values represent smaller test statistics. Electrodes that reached at an alpha value of $p = 0.01$ are shown in bold. As can be seen the large majority of electrodes are significant at this lowered threshold (Number of significant electrodes on average for orientation: 3 Hz left-188, 3 Hz right-173, 5 Hz left-189, 5 Hz right-189; for size: 3 Hz left-169, 3 Hz right-169, 5 Hz left-182, 5 Hz right-183) indicating that our frequency tags are reliable across trials.
Figure 33. Example $T_{\text{circ}}^2$ for one participant for orientation.

Topographic heat maps showing the $T_{\text{circ}}^2$ statistic for both carrier frequencies for the orientation condition (top and bottom rows) when presented in the left and right sides of visual space (left and right columns). Significant electrodes at a threshold of $p=0.01$ are bolded. Test statistics range from 0-5 (yellow-red).
Figure 34. Example $T_{circ}^2$ for one participant for size.

Topographic heat maps showing the $T_{circ}^2$ statistic for both carrier frequencies for the size condition (top and bottom rows) when presented in the left and right sides of visual space (left and right columns). Significant electrodes at a threshold of $p=0.01$ are bolded. Test statistics range from 0-5 (yellow-red).

Are participants paying attention?

After establishing that the frequency tags are consistent across trials, we want to establish that the participant is attended to the correct visual field. To do this, we compared attended and unattended carrier frequency tags. The results are plotted in Figure 35. The top row of figures displays the topographic maps for
orientation levels 1-5 and the bottom row displays size levels 1-5. Darker colors represent more significant electrodes. Electrodes that reached significance at an FDR corrected alpha threshold are bolded. This indicates that the participants were attending to the correct side of visual space during the experiment.
Figure 35. Carrier frequencies: Attended-Unattended.

Topographic heat maps showing test statistics for every electrode comparing carrier frequency tags representing attended stimuli vs. unattended stimuli. Test statistics
are thresholded at an FDR corrected p-value of 0.0008. Significant electrodes are plotted in black. Comparisons for the orientation condition are plotted in the left column and size in the right column for each of the five levels (rows 1-5). Yellow values represent smaller test statistics and red values represent larger test statistics (ranging form 0-10).

**Oddball analysis**

**Strong ensemble encoding**

The first analysis we ran tested the prediction that ensemble encoding is rapid, automatic, and pre-attentive. For this analysis we grouped the data only based on the feature and the feature level and collapsed across attended hemifield and averaging task. Based on this prediction, we would expect there to be a positive linear relationship between the tags of increased levels for both features, regardless of the task the participant is performing or where they are attending.

**Feature level analysis**

First, we compared each of the five levels for both tasks. Because the raw frequency tend to be very large, it can be difficult to detect differences by simply viewing these topographic maps. Therefore we performed an ANOVA with a linear contrast for each electrode across all participants. The resulting f-statistics are plotted in Figure 36 for orientation and Figure 37 for size with electrodes that reached significance at an alpha of p = 0.05 are bolded. Values are FDR corrected and the q-values are reported below.
As can be seen, the majority of electrodes did not reach significance in any of the three harmonics for size or orientation. There were 6, 0, and 3 significant electrodes in the orientation condition for the 1\textsuperscript{st}, 2\textsuperscript{nd}, and 3\textsuperscript{rd} harmonic (FDR q>0.99 for all harmonics) and 5, 6, and 16 significant electrodes in the size condition (FDR q>0.99, q>0.99, q=0.812 respectively). This indicates that some level of attention is required to extract the ensemble, be it spatial attention or attention to the feature of interest.
Figure 36. Feature level linear contrast – Orientation.

Topographic heat maps representing F-statistics from linear contrasts comparing the frequency tags across the five feature levels for the orientation condition for all electrodes. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of p=0.05 and the corresponding FDR q-values and F-statistics are shown at the bottom of each map.
Figure 37. Feature level linear contrast – Size.

Topographic heat maps representing F-statistics from linear contrasts comparing the frequency tags across the five feature levels for the size condition for all electrodes. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of $p=0.05$ and the corresponding FDR $q$-values and F-statistics are shown at the bottom of each map.

Δ level analysis

Next, in order to increase the power of our signal, we computed four delta levels from our 5 feature levels, which look at the averaged difference between feature levels of increasing difference (See methods section). Using the averaged
values of each delta level for all participants we again performed a linear contrasts. Similarly to the feature level analysis, we plotted the resulting f-statistics for each electrode with electrodes that reached significance at an alpha of $p = 0.05$ bolded in black (Figures Figure 38 and Figure 39). Values are FDR corrected and the q-values are reported below.

Again, despite the increased power, very few electrodes are significant, even at the liberal threshold. There were 10, 5, and 9 significant electrodes in the orientation condition for the 1st, 2nd, and 3rd harmonic (FDR $q=0.812$, $q>0.99$, $q>0.99$ respectively) and 8, 19, and 20 significant electrodes in the size condition (FDR $q>1.00$, $q=0.52$, $q=0.361$ respectively).
Figure 38. ∆ level linear contrast – Orientation.

Topographic heat maps representing F-statistics from linear contrasts comparing the frequency tags across the four ∆ levels for the orientation condition for all electrodes. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of p=0.05 and the corresponding FDR q-values and F-statistics are shown at the bottom of each map.
**Figure 39. Δ level linear contrast – Size.**

Topographic heat maps representing F-statistics from linear contrasts comparing the frequency tags across the four Δ levels for the size condition for all electrodes. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of p=0.05 and the corresponding FDR q-values and F-statistics are shown at the bottom of each map.

**Weak ensemble encoding 1**

The second series of analysis we ran tested the prediction that ensemble encoding requires either spatially focused attention or the presence of an explicit averaging task. For this prediction, we performed two series of analysis, differing in
how we segmented the trials into groups. For the first analysis, we segmented trials as a function of which side of visual space was attended and the feature level, while collapsing across the task. Based on this prediction, we would expect a positive linear relationship between the tags of each level for both features only when those frequency tags are being attended and regardless of what task the participant is doing.

For the second analysis, we segmented trials as a function of the task the participant was asked to perform and the feature level, while collapsing across the attended hemifield. Based on this prediction, we would expect a positive linear relationship between the tags of each level for both features only when the relevant task is being performed regardless of where spatial attention is focused. In other words, we would predict the orientation frequency tags to show a strong linear relationship even in the unattended hemifield as long as the participant is actively attending to the orientation feature.

As mentioned in the methods section above, in addition to examining the average frequency tags, we also created an index between the average frequency tags of two conditions of interest. The reasoning for this is fairly straightforward. While the index value is a better way to visualize and quantify any difference between the frequency tags, there are times when it can mask potentially important results. For example, it could be the case that both conditions being contrasted show a positive linear relationship. In other words, the frequency tags are increasing as a function of the feature level, but are doing so in both conditions (e.g., attended vs.
unattended, task vs. no task, etc.). In this case an index value would hide this result as the difference between any level in both conditions would be the same (or similar) across all levels. This would lead to no significant linear relationship in the index. For rest of the presented analysis, we performed linear contrasts on both the averaged frequency tags and the index value created with these tags.

**Is the allocation of spatial attention needed to extract the ensemble, regardless of attend feature?**

*Feature level analysis*

The resulting f-statistics are plotted for the frequency tag analysis in Figure 40 for orientation and Figure 41 for size while the f-statistics for the index value analysis are plotted in Figure 42 for orientation and Figure 43 for size. Lighter colors represent smaller values while darker colors represent larger values. Electrodes that reached significance at an alpha of $p = 0.05$ are bolded in black. Values are FDR corrected and the $q$-values are reported below.

As can be seen, the majority of electrodes did not reach significance for the attended or unattended conditions in any of the three harmonics for the index and frequency tag analysis for size or orientation. For the frequency tag analysis we found 14, 8, and 18 significant electrodes in the orientation attended condition (FDR $q=2.167$, $q=1.182$, $q=0.310$) and 29, 9, and 0 significant electrodes in the orientation unattended condition (FDR $q=0.433$, $q=1.30$, $q=\text{NaN}$) for the 1\textsuperscript{st}, 2\textsuperscript{nd}, and 3\textsuperscript{rd}
harmonic. There were 14, 8, and 17 significant electrodes in the size attended condition (FDR $q=0.929$, $q=1.625$, $q=0.722$) and 5, 5, and 22 significant electrodes in the size unattended condition (FDR $q=2.60$, $q=2.60$, $q=0.591$) for the 1$^{st}$, 2$^{nd}$, and 3$^{rd}$ harmonic.

For the index analysis we found 10, 7, and 6 significant electrodes in the orientation condition (FDR $q=1.30$, $q=1.857$, $q=0.2.167$) and 29, 9, and 0 significant electrodes in the size condition (FDR $q=0.433$, $q=1.30$, $q=\text{NaN}$) for the 1$^{st}$, 2$^{nd}$, and 3$^{rd}$ harmonic. This is indicative of no representation of the ensemble in the frequency tags when the participant is attending or not attending, at least if we assume that the averaging task is not necessary for extracting the ensemble.
Figure 40. Attended vs. unattended – Frequency tags – Feature levels – Orientation.

Topographic heat maps representing F-statistics from linear contrasts comparing the frequency tags across the five feature levels for the orientation condition for all electrodes. The two rows represent the attended (left) and unattended (right) conditions. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of p=0.05 and the corresponding FDR q-values and F-statistics are shown at the bottom of each map.
Figure 41. Attended vs. unattended – Frequency tags - Feature levels – Size.

Topographic heat maps representing F-statistics from linear contrasts comparing the frequency tags across the five feature levels for the size condition for all electrodes. The two rows represent the attended (left) and unattended (right) conditions. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of $p=0.05$ and the corresponding FDR q-values and F-statistics are shown at the bottom of each map.
Figure 42. Attended vs. unattended – Index values - Feature levels – Orientation.

Topographic heat maps representing F-statistics from linear contrasts comparing the index values across the five feature levels for the orientation condition for all electrodes. Each index value represents a normalized difference between the attended and unattended conditions. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of p=0.05 and the corresponding FDR q-values and F-statistics are shown at the bottom of each map.
Figure 43. Attended vs. unattended – Index values – Feature levels – Size.

Topographic heat maps representing F-statistics from linear contrasts comparing the index values across the five feature levels for the size condition for all electrodes. Each index value represents a normalized difference between the attended and unattended conditions. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of $p=0.05$ and the corresponding FDR q-values and F-statistics are shown at the bottom of each map.

$\Delta$ level analysis

Just like the feature level analysis we examined the delta levels. We plotted the f-statistics for the frequency tag and index value analysis in 44 and 46 for
orientation and 45 and 47 for size respectively. Lighter colors represent smaller values while darker colors represent larger values. Electrodes that reached significance at an alpha of $p = 0.05$ are bolded in black. Values are FDR corrected and the q-values are reported below.

Like for the feature level analysis, the majority of electrodes did not reach significance for the attended or unattended conditions in any of the three harmonics for the index and frequency tag analysis for size or orientation. For the frequency tag analysis we found 3, 17, and 54 significant electrodes in the orientation attended condition (FDR $q=4.333$, $q=0.765$, $q=0.232$) and 19, 7, and 1 significant electrodes in the orientation unattended condition (FDR $q=0.684$, $q=1.857$, $q=13.00$) for the 1$^{\text{st}}$, 2$^{\text{nd}}$, and 3$^{\text{rd}}$ harmonic. There were 9, 6, and 14 significant electrodes in the size attended condition (FDR $q=1.444$, $q=2.167$, $q=0.929$) and 2, 4, and 19 significant electrodes in the size unattended condition (FDR $q=6.50$, $q=3.25$, $q=0.684$) for the 1$^{\text{st}}$, 2$^{\text{nd}}$, and 3$^{\text{rd}}$ harmonic. For the index analysis we found 26, 5, and 10 significant electrodes in the orientation condition (FDR $q=0.50$, $q=2.60$, $q=1.30$) and 26, 6, and 21 significant electrodes in the size condition (FDR $q=0.50$, $q=2.167$, $q=0.619$) for the 1$^{\text{st}}$, 2$^{\text{nd}}$, and 3$^{\text{rd}}$ harmonic.

Given these results, even with the increased power provided by the delta levels, it seems that attention has no effect on the representation of the ensemble. In other words, the representation of the ensemble does not change as a function of attended hemifield, when we do not consider the task being performed.
Topographic heat maps representing F-statistics from linear contrasts comparing the index values across the four $\Delta$ levels for the orientation condition for all electrodes. The two rows represent the attended (left) and unattended (right) conditions. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of $p=0.05$ and the corresponding FDR q-values and F-statistics are shown at the bottom of each map.
45. Attended vs. unattended – Frequency tags - Δ levels – Size.

Topographic heat maps representing F-statistics from linear contrasts comparing the index values across the four Δ levels for the size condition for all electrodes. The two rows represent the attended (left) and unattended (right) conditions. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of p=0.05 and the corresponding FDR q-values and F-statistics are shown at the bottom of each map.
46. Attended vs. unattended – Index values - Δ levels – Orientation.

Topographic heat maps representing F-statistics from linear contrasts comparing the index values across the four Δ levels for the orientation condition for all electrodes. Each index value represents a normalized difference between the attended and unattended conditions. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of p=0.05 and the corresponding FDR q-values and F-statistics are shown at the bottom of each map.
47. Attended vs. unattended – Index values - Δ levels – Size.

Topographic heat maps representing F-statistics from linear contrasts comparing the index values across the four Δ levels for the size condition for all electrodes. Each index value represents a normalized difference between the attended and unattended conditions. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of p=0.05 and the corresponding FDR q-values and F-statistics are shown at the bottom of each map.

Is the presence of an explicit averaging task necessary to extract the ensemble, regardless of spatial attention?
Feature level analysis

Next, we compared the tags and indices when participants were performing a relevant averaging task to when they were not, regardless of the focus of spatial attention. The resulting f-statistics are plotted for the frequency tag analysis in Figure 48 for orientation and Figure 49 for size while the f-statistics for the index value analysis are plotted in Figure 50 for orientation and Figure 51 for size. Lighter colors represent smaller values while darker colors represent larger values. Electrodes that reached significance at an alpha of $p = 0.05$ are bolded in black. Values are FDR corrected and the q-values are reported below.

As can be seen, the majority of electrodes did not reach significance for the task relevant or task irrelevant conditions in any of the three harmonics for the index and frequency tag analysis for size or orientation. For the frequency tag analysis we found 13, 12, and 12 significant electrodes in the orientation task relevant condition (FDR $q=1.00$, $q=1.083$, $q=1.083$) and 10, 8, and 14 significant electrodes in the orientation task irrelevant condition (FDR $q=1.30$, $q=1.625$, $q=0.929$) for the 1st, 2nd, and 3rd harmonic. There were 16, 29, and 13 significant electrodes in the size task relevant condition (FDR $q=0.812$, $q=0.448$, $q=1.00$) and 4, 8, and 6 significant electrodes in the size task irrelevant condition (FDR $q=3.25$, $q=1.625$, $q=2.167$) for the 1st, 2nd, and 3rd harmonic. For the index analysis we found 18, 18, and 24 significant electrodes in the orientation condition (FDR $q=0.722$, $q=0.722$, $q=0.542$) and 15, 22, and 10 significant electrodes in the size condition (FDR $q=0.867$, $q=0.591$, $q=1.30$) for the 1st, 2nd, and 3rd harmonic. This is indicative
of no representation of the ensemble in the frequency tags when the participant is attending or not attending to the relevant feature, at least if we assume that spatial attention is not necessary for extracting the ensemble.

Figure 48. Task vs. no task – Frequency tags – Feature levels – Orientation.

Topographic heat maps representing F-statistics from linear contrasts comparing the frequency tags across the five feature levels for the orientation condition for all electrodes. The two rows represent the task relevant (left) and task irrelevant (right) conditions. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of
p=0.05 and the corresponding FDR q-values and F-statistics are shown at the bottom of each map.

Figure 49. Task vs. no task – Frequency tags - Feature levels – Size.

Topographic heat maps representing F-statistics from linear contrasts comparing the frequency tags across the five feature levels for the size condition for all electrodes. The two rows represent the task relevant (left) and task irrelevant (right) conditions. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of p=0.05 and the corresponding FDR q-values and F-statistics are shown at the bottom of each map.
Figure 50. Task vs. no task – Index values - Feature levels – Orientation.

Topographic heat maps representing F-statistics from linear contrasts comparing the index values across the five feature levels for the orientation condition for all electrodes. Each index value represents a normalized difference between the task relevant and task irrelevant conditions. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of p=0.05 and the corresponding FDR q-values and F-statistics are shown at the bottom of each map.
Figure 51. Task vs. no task – Index values – Feature levels – Size.

Topographic heat maps representing F-statistics from linear contrasts comparing the index values across the five feature levels for the size condition for all electrodes. Each index value represents a normalized difference between the task relevant and task irrelevant conditions. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of p=0.05 and the corresponding FDR q-values and F-statistics are shown at the bottom of each map.

Δ level analysis

Like in part 1 we also examine the delta levels. We plotted the f-statistics for the frequency tag and index value analysis in Figure 52 and Figure 54 for
orientation and Figure 53 and Figure 55 for size respectively. Lighter colors represent smaller values while darker colors represent larger values. Electrodes that reached significance at an alpha of $p = 0.05$ are bolded in black. Values are FDR corrected and the q-values are reported below. Like for the feature level analysis, the majority of electrodes did not reach significance for the attended or unattended conditions in any of the three harmonics for the index and frequency tag analysis for size or orientation.

For the frequency tag analysis we found 12, 7, and 8 significant electrodes in the orientation task relevant condition (FDR $q=1.083, q=1.857, q=1.625$) and 7, 3, and 6 significant electrodes in the orientation task irrelevant condition (FDR $q=1.857, q=4.333, q=2.167$) for the 1st, 2nd, and 3rd harmonic. There were 4, 4, and 7 significant electrodes in the size task relevant condition (FDR $q=3.25, q=3.25, q=1.857$) and 12, 31, and 22 significant electrodes in the size task irrelevant condition (FDR $q=1.083, q=0.419, q=0.591$) for the 1st, 2nd, and 3rd harmonic. For the index analysis we found 13, 6, and 12 significant electrodes in the orientation condition (FDR $q=1.00, q=2.167, q=1.083$) and 19, 27, and 16 significant electrodes in the size condition (FDR $q=0.684, q=0.481, q=0.812$) for the 1st, 2nd, and 3rd harmonic. Given these results, even with the increased power provided by the delta levels, it seems that task has no effect on the representation of the ensemble. In other words, the representation of the ensemble does not change as a function of the feature that is currently being attended to.
Figure 52. Task vs. no task – Frequency tags - ∆ levels – Orientation.

Topographic heat maps representing F-statistics from linear contrasts comparing the index values across the four ∆ levels for the orientation condition for all electrodes. The two rows represent the task relevant (left) and task irrelevant (right) conditions. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of p=0.05 and the corresponding FDR q-values and F-statistics are shown at the bottom of each map.
Figure 53. Task vs. no task – Frequency tags - Δ levels – Size.

Topographic heat maps representing F-statistics from linear contrasts comparing the index values across the four Δ levels for the size condition for all electrodes. The two rows represent the task relevant (left) and task irrelevant (right) conditions. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of p=0.05 and the corresponding FDR q-values and F-statistics are shown at the bottom of each map.
Figure 54. Task vs. no task – Index values - Δ levels – Orientation.

Topographic heat maps representing F-statistics from linear contrasts comparing the index values across the four Δ levels for the orientation condition for all electrodes. Each index value represents a normalized difference between the task relevant and task irrelevant conditions. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of p=0.05 and the corresponding FDR q-values and F-statistics are shown at the bottom of each map.
Figure 55. Task vs. no task – Index values - Δ levels – Size.

Topographic heat maps representing F-statistics from linear contrasts comparing the index values across the four Δ levels for the size condition for all electrodes. Each index value represents a normalized difference between the task relevant and task irrelevant conditions. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of p=0.05 and the corresponding FDR q-values and F-statistics are shown at the bottom of each map.

Weak ensemble encoding 2

This represents the ‘weakest’ form of ensemble encoding, in that it predicts that ensemble averages are only encoded if the participant is actively attending to
the feature of interest in the spatially attended hemifield. For this analysis, we examine the effects of spatial attention as a function of task. Specifically, we compare the representation of the average when participants are performing a relevant task while attending or not attending, when participants are performing an irrelevant task while attending or not attending, and when participants are performing a relevant task while attending and when they are performing an irrelevant task while not attending.

**No task attended vs. no task unattended**

*Feature level analysis*

First, we compared the tags and indices when participants were performing an irrelevant averaging task and attending to when they were performing an irrelevant averaging task and not attending. Importantly, this analysis will allow us to examine the effects of attention on the representation of the ensemble when the participant is not attending the feature of interest. The resulting f-statistics are plotted for the frequency tag analysis in Figure 56 for orientation and Figure 57 for size while the f-statistics for the index value analysis are plotted in Figure 58 for orientation and Figure 59 for size. Lighter colors represent smaller values while darker colors represent larger values. Electrodes that reached significance at an alpha of \( p = 0.05 \) are bolded in black. Values are FDR corrected and the q-values are reported below.
As can be seen, the majority of electrodes did not reach significance for the task relevant or task irrelevant conditions in any of the three harmonics for the index and frequency tag analysis for size or orientation. For the frequency tag analysis we found 3, 17, and 19 significant electrodes in the orientation task irrelevant attended condition (FDR q=4.33, q=0.765, q=0.684) and 11, 27, and 19 significant electrodes in the orientation task irrelevant unattended condition (FDR q=1.182, q=0.481, q=0.684) for the 1st, 2nd, and 3rd harmonic. There were 10, 4, and 6 significant electrodes in the size task irrelevant attended condition (FDR q=1.30, q=3.25, q=2.167) and 26, 7, and 11 significant electrodes in the size task irrelevant unattended condition (FDR q=0.50, q=1.857, q=1.182) for the 1st, 2nd, and 3rd harmonic. For the index analysis we found 3, 1, and 8 significant electrodes in the orientation condition (FDR q=4.33, q=13.0, q=1.625) and 13, 10, and 17 significant electrodes in the size condition (FDR q=1.00, q=1.30, q=0.684) for the 1st, 2nd, and 3rd harmonic. Like for the previous analysis these results are indicative of no representation of the ensemble that is detectable in the frequency tag data. In other words, if the participant is performing an irrelevant averaging task, then there is not an effect of attention on the representation of the ensemble.
Figure 56. No task attended vs. no task unattended – Frequency tags – Feature levels – Orientation.

Topographic heat maps representing F-statistics from linear contrasts comparing the frequency tags across the five feature levels for the orientation condition for all electrodes. The two rows represent the task irrelevant attended (left) and task irrelevant unattended (right) conditions. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of $p=0.05$ and the corresponding FDR q-values and F-statistics are shown at the bottom of each map.
Figure 57. No task attended vs. no task unattended – Frequency tags - Feature levels - Size.

Topographic heat maps representing F-statistics from linear contrasts comparing the frequency tags across the five feature levels for the size condition for all electrodes. The two rows represent the task irrelevant attended (left) and task irrelevant unattended (right) conditions. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of p=0.05 and the corresponding FDR q-values and F-statistics are shown at the bottom of each map.
Figure 58. No task attended vs. no task unattended – Index values - Feature levels – Orientation.

Topographic heat maps representing F-statistics from linear contrasts comparing the index values across the five feature levels for the orientation condition for all electrodes. Each index value represents a normalized difference between the task irrelevant attended and task irrelevant unattended conditions. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of p=0.05 and the corresponding FDR q-values and F-statistics are shown at the bottom of each map.
Figure 59. No task attended vs. no task unattended – Index values – Feature levels – Size.

Topographic heat maps representing F-statistics from linear contrasts comparing the index values across the five feature levels for the size condition for all electrodes. Each index value represents a normalized difference between the task irrelevant attended and task irrelevant unattended conditions. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of p=0.05 and the corresponding FDR q-values and F-statistics are shown at the bottom of each map.
\( \Delta \) level analysis

Like in part 1 we also examine the delta levels. We plotted the f-statistics for the frequency tag and index value analysis in Figure 60 and Figure 62 for orientation and Figure 61 and Figure 63 for size respectively. Lighter colors represent smaller values while darker colors represent larger values. Electrodes that reached significance at an alpha of \( p = 0.05 \) are bolded in black. Values are FDR corrected and the q-values are reported below. The majority of electrodes did not reach significance for either conditions in any of the three harmonics for the index and frequency tag analysis for size or orientation.

For the frequency tag analysis we found 2, 6, and 13 significant electrodes in the orientation task irrelevant attended condition (FDR q=6.50, q=2.167, q=1.00) and 2, 10, and 1 significant electrodes in the orientation task irrelevant unattended condition (FDR q=6.50, q=1.30, q=13.0) for the 1\(^{\text{st}}\), 2\(^{\text{nd}}\), and 3\(^{\text{rd}}\) harmonic. There were 9, 10, and 10 significant electrodes in the size task irrelevant attended condition (FDR q=1.44, q=1.30, q=1.30) and 24, 14, and 24 significant electrodes in the size task irrelevant unattended condition (FDR q=0.542, q=0.929, q=0.542) for the 1\(^{\text{st}}\), 2\(^{\text{nd}}\), and 3\(^{\text{rd}}\) harmonic. For the index analysis we found 15, 8, and 6 significant electrodes in the orientation condition (FDR q=0.867, q=1.625, q=2.167) and 31, 30, and 6 significant electrodes in the size condition (FDR q=0.419, q=0.433, q=2.167) for the 1\(^{\text{st}}\), 2\(^{\text{nd}}\), and 3\(^{\text{rd}}\) harmonic. Given these results, even with the increased power provided by the delta levels, it seems that when the task does not require the
participant to attended to the relevant feature, attention has not effect on the representation of the ensemble.

Figure 60. No task attended vs. no task unattended – Frequency tags - Δ levels – Orientation.

Topographic heat maps representing F-statistics from linear contrasts comparing the frequency tags across the four Δ levels for the orientation condition for all electrodes. The two rows represent the task irrelevant attended (left) and task irrelevant unattended (right) conditions. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of p=0.05 and the corresponding FDR q-values and F-statistics are shown at the bottom of each map.
Figure 61. No task attended vs. no task unattended – Frequency tags - Δ levels – Size.

Topographic heat maps representing F-statistics from linear contrasts comparing the frequency tags across the four Δ levels for the size condition for all electrodes. The two rows represent the task irrelevant attended (left) and task irrelevant unattended (right) conditions. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of p=0.05 and the corresponding FDR q-values and F-statistics are shown at the bottom of each map.
Figure 62. No task attended vs. no task unattended – Index values - Δ levels – Orientation.

Topographic heat maps representing F-statistics from linear contrasts comparing the index values across the four Δ levels for the orientation condition for all electrodes. Each index value represents a normalized difference between the task irrelevant attended and task irrelevant unattended conditions. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of p=0.05 and the corresponding FDR q-values and F-statistics are shown at the bottom of each map.
Figure 63. No task attended vs. no task unattended – Index values - ∆ levels – Size.

Topographic heat maps representing F-statistics from linear contrasts comparing the index values across the four ∆ levels for the size condition for all electrodes. Each index value represents a normalized difference between the task irrelevant attended and task irrelevant unattended conditions. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of p=0.05 and the corresponding FDR q-values and F-statistics are shown at the bottom of each map.

Task attended vs. Task unattended
**Feature level analysis**

Next, we compared the tags and indices when participants were performing a relevant averaging task and attending to when they were performing a relevant averaging task and not attending. Like the irrelevant attended compared to irrelevant unattended analysis, this will allow us to examine the effects of task and attention on the representation of the average simultaneously. The resulting f-statistics are plotted for the frequency tag analysis in Figure 64 for orientation and Figure 65 for size while the f-statistics for the index value analysis are plotted in Figure 66 for orientation and Figure 67 for size. Lighter colors represent smaller values while darker colors represent larger values. Electrodes that reached significance at an alpha of $p = 0.05$ are bolded in black. Values are FDR corrected and the $q$-values are reported below.

Again, the majority of electrodes did not reach significance for either condition in any of the three harmonics for the index and frequency tag analysis for size or orientation. For the frequency tag analysis we found 14, 12, and 37 significant electrodes in the orientation task relevant attended condition (FDR $q=0.929, q=1.083, q=0.342$) and 2, 6, and 12 significant electrodes in the orientation task relevant unattended condition (FDR $q=6.50, q=2.167, q=1.083$) for the 1$^{st}$, 2$^{nd}$, and 3$^{rd}$ harmonic. There were 7, 3, and 0 significant electrodes in the size task relevant attended condition (FDR $q=1.857, q=4.33, q=\text{NaN}$) and 13, 17, and 0 significant electrodes in the size task relevant unattended condition (FDR $q=1.00, q=0.765, q=\text{NaN}$) for the 1$^{st}$, 2$^{nd}$, and 3$^{rd}$ harmonic. For the index analysis we found
9, 7, and 28 significant electrodes in the orientation condition (FDR q=1.44, q=1.857, q=0.464) and 7, 18, and 3 significant electrodes in the size condition (FDR q=1.857, q=0.722, q=4.33) for the 1st, 2nd, and 3rd harmonic.

Figure 64. Task attended vs. task unattended – Frequency tags – Feature levels – Orientation.

Topographic heat maps representing F-statistics from linear contrasts comparing the frequency tags across the five feature levels for the orientation condition for all electrodes. The two rows represent the task relevant attended (left) and task relevant unattended (right) conditions. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of
p=0.05 and the corresponding FDR q-values and F-statistics are shown at the bottom of each map.

Figure 65. Task attended vs. task unattended – Frequency tags – Feature levels – Size.

Topographic heat maps representing F-statistics from linear contrasts comparing the frequency tags across the five feature levels for the size condition for all electrodes. The two rows represent the task relevant attended (left) and task relevant unattended (right) conditions. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electodes that reached significance are plotted in black. Values are thresholded at an alpha of p=0.05 and the corresponding FDR q-values and F-statistics are shown at the bottom of each map.
Figure 66. Task attended vs. task unattended – Index values – Feature levels – Orientation.

Topographic heat maps representing F-statistics from linear contrasts comparing the index values across the five feature levels for the orientation condition for all electrodes. Each index value represents a normalized difference between the task relevant attended and task relevant unattended conditions. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of p=0.05 and the corresponding FDR q-values and F-statistics are shown at the bottom of each map.
Figure 67. Task attended vs. task unattended – Index values – Feature levels – Size.

Topographic heat maps representing F-statistics from linear contrasts comparing the index values across the five feature levels for the size condition for all electrodes. Each index value represents a normalized difference between the task relevant attended and task relevant unattended conditions. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of $p=0.05$ and the corresponding FDR $q$-values and F-statistics are shown at the bottom of each map.
\textit{\Delta level analysis}

We also examine the delta levels. The f-statistics for the frequency tag and index value analysis are plotted in Figure 68 and Figure 70 for orientation and Figure 69 and Figure 71 for size respectively. Lighter colors represent smaller values while darker colors represent larger values. Electrodes that reached significance at an alpha of $p = 0.05$ are bolded in black. Values are FDR corrected and the q-values are reported below. The majority of electrodes did not reach significance for either conditions in any of the three harmonics for the index and frequency tag analysis for size or orientation.

For the frequency tag analysis we found 3, 13, and 9 significant electrodes in the orientation task relevant attended condition (FDR $q=4.33$, $q=1.00$, $q=1.30$) and 3, 3, and 4 significant electrodes in the orientation task relevant unattended condition (FDR $q=4.33$, $q=4.33$, $q=3.25$) for the $1^{\text{st}}$, $2^{\text{nd}}$, and $3^{\text{rd}}$ harmonic. There were 2, 4, and 5 significant electrodes in the size task relevant attended condition (FDR $q=6.50$, $q=3.25$, $q=2.60$) and 9, 4, and 4 significant electrodes in the size task relevant unattended condition (FDR $q=1.44$, $q=3.25$, $q=3.25$) for the $1^{\text{st}}$, $2^{\text{nd}}$, and $3^{\text{rd}}$ harmonic. For the index analysis we found 9, 4, and 8 significant electrodes in the orientation condition (FDR $q=1.44$, $q=3.25$, $q=1.625$) and 42, 2, and 40 significant electrodes in the size condition (FDR $q=0.310$, $q=6.50$, $q=0.317$) for the $1^{\text{st}}$, $2^{\text{nd}}$, and $3^{\text{rd}}$ harmonic. Taken together with the above analysis, it seems there is no change in the representation of the ensemble if the participant is attending or not attending while performing a relevant or irrelevant task.
Figure 68. Task attended vs. task unattended – Frequency tags – $\Delta$ levels – Orientation.

Topographic heat maps representing F-statistics from linear contrasts comparing the frequency tags across the four $\Delta$ levels for the orientation condition for all electrodes. The two rows represent the task relevant attended (left) and task relevant unattended (right) conditions. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of $p=0.05$ and the corresponding FDR q-values and F-statistics are shown at the bottom of each map.
Figure 69. Task attended vs. task unattended – Frequency tags – Δ levels – Size.

Topographic heat maps representing F-statistics from linear contrasts comparing the frequency tags across the four Δ levels for the size condition for all electrodes. The two rows represent the task relevant attended (left) and task relevant unattended (right) conditions. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of p=0.05 and the corresponding FDR q-values and F-statistics are shown at the bottom of each map.
Figure 70. Task attended vs. task unattended – Index values – Δ levels – Orientation.

Topographic heat maps representing F-statistics from linear contrasts comparing the index values across the four Δ levels for the orientation condition for all electrodes. Each index value represents a normalized difference between the task relevant attended and task relevant unattended conditions. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of \( p=0.05 \) and the corresponding FDR q-values and F-statistics are shown at the bottom of each map.
Figure 71. Task attended vs. task unattended – Index values – Δ levels – Size.

Topographic heat maps representing F-statistics from linear contrasts comparing the index values across the four Δ levels for the size condition for all electrodes. Each index value represents a normalized difference between the task relevant attended and task relevant unattended conditions. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of p=0.05 and the corresponding FDR q-values and F-statistics are shown at the bottom of each map.

Task attended vs No task unattended
**Feature level analysis**

Next, we directly compared the tags and indices when participants were performing a relevant averaging task and attending to when they were performing an irrelevant averaging task and not attending. This allows us to examine the relationship between the two conditions directly. The resulting f-statistics are plotted for the frequency tag analysis in Figure 72 for orientation and Figure 73 for size while the f-statistics for the index value analysis are plotted in Figure 74 for orientation and Figure 75 for size. Lighter colors represent smaller values while darker colors represent larger values. Electrodes that reached significance at an alpha of \( p = 0.05 \) are bolded in black. Values are FDR corrected and the q-values are reported below.

Again, the majority of electrodes did not reach significance for either condition in any of the three harmonics for the index and frequency tag analysis for size or orientation. For the frequency tag analysis we found 14, 12, and 37 significant electrodes in the orientation task relevant attended condition (FDR \( q=0.929, q=1.083, q=0.342 \) ) and 22, 6, and 2 significant electrodes in the orientation task irrelevant unattended condition (FDR \( q=0.591, q=2.167, q=6.50 \) ) for the 1\(^{st}\), 2\(^{nd}\), and 3\(^{rd}\) harmonic. There were 7, 3, and 0 significant electrodes in the size task relevant attended condition (FDR \( q=1.857, q=4.33, q=\text{NaN} \) ) and 3, 13, and 8 significant electrodes in the size task irrelevant unattended condition (FDR \( q=4.33, q=1.00, q=1.625 \) ) for the 1\(^{st}\), 2\(^{nd}\), and 3\(^{rd}\) harmonic. For the index analysis we found 3, 15, and 70 significant electrodes in the orientation condition (FDR \( q=4.33, \)
q=0.867, q=0.188) and 8, 9, and 7 significant electrodes in the size condition (FDR q=1.625, q=1.44, q=1.857) for the 1\textsuperscript{st}, 2\textsuperscript{nd}, and 3\textsuperscript{rd} harmonic.

**Figure 72. Task attended vs. no task unattended – Frequency tags – Feature levels – Orientation.**

Topographic heat maps representing F-statistics from linear contrasts comparing the frequency tags across the five feature levels for the orientation condition for all electrodes. The two rows represent the task relevant attended (left) and task irrelevant unattended (right) conditions. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of p=0.05 and the corresponding FDR q-values and F-statistics are shown at the bottom of each map.
Figure 73. Task attended vs. no task unattended – Frequency tags – Feature levels – Size.

Topographic heat maps representing F-statistics from linear contrasts comparing the frequency tags across the five feature levels for the size condition for all electrodes. The two rows represent the task relevant attended (left) and task irrelevant unattended (right) conditions. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of p=0.05 and the corresponding FDR q-values and F-statistics are shown at the bottom of each map.
Figure 74. Task attended vs. no task unattended – Index values – Feature levels – Orientation.

Topographic heat maps representing F-statistics from linear contrasts comparing the index values across the five feature levels for the orientation condition for all electrodes. Each index value represents a normalized difference between the task relevant attended and task irrelevant unattended conditions. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of p=0.05 and the corresponding FDR q-values and F-statistics are shown at the bottom of each map.
**Figure 75. Task attended vs. no task unattended – Index values – Feature levels – Size.**

Topographic heat maps representing F-statistics from linear contrasts comparing the index values across the five feature levels for the size condition for all electrodes. Each index value represents a normalized difference between the task relevant attended and task irrelevant unattended conditions. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of p=0.05 and the corresponding FDR q-values and F-statistics are shown at the bottom of each map.
Δ level analysis

We also examine the delta levels. The f-statistics for the frequency tag and index value analysis are plotted in Figure 76 and Figure 78 for orientation and Figure 77 and Figure 79 for size respectively. Lighter colors represent smaller values while darker colors represent larger values. Electrodes that reached significance at an alpha of \( p = 0.05 \) are bolded in black. Values are FDR corrected and the q-values are reported below. The majority of electrodes did not reach significance for either conditions in any of the three harmonics for the index and frequency tag analysis for size or orientation.

For the frequency tag analysis we found 5, 4, and 2 significant electrodes in the orientation task relevant attended condition (FDR q=2.60, q=3.25, q=6.50) and 64, 41, and 43 significant electrodes in the orientation task irrelevant unattended condition (FDR q=0.206, q=0.325, q=0.310) for the 1st, 2nd, and 3rd harmonic. There were 4, 2, and 5 significant electrodes in the size task relevant attended condition (FDR q=2.60, q=6.50, q=3.25) and 5, 51, and 8 significant electrodes in the size task irrelevant unattended condition (FDR q=2.60, q=0.260, q=1.625) for the 1st, 2nd, and 3rd harmonic. For the index analysis we found 7, 1, and 6 significant electrodes in the orientation condition (FDR q=1.857, q=13.0, q=2.167) and 3, 4, and 4 significant electrodes in the size condition (FDR q=4.33, q=3.25, q=3.25) for the 1st, 2nd, and 3rd harmonic. Taken together with the two above analysis, it seems there is no change in the representation of the ensemble if the participant is attending or not attending while performing a relevant or irrelevant task. Notice, however, that when the two
conditions are fully separated there is a pattern of significance in the no task unattended condition compared to the task attended condition, particularly in the delta level analysis for orientation harmonics 1-3 and size harmonic 2. However, as this pattern is not consistent with other comparisons we should be cautious in our interpretation.

**Figure 76. Task attended vs. no task unattended – Frequency tags – Δ Levels – Orientation.**

Topographic heat maps representing F-statistics from linear contrasts comparing the frequency tags across the four Δ levels for the orientation condition for all electrodes. The two rows represent the task relevant attended (left) and task
irrelevant unattended (right) conditions. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of $p=0.05$ and the corresponding FDR $q$-values and F-statistics are shown at the bottom of each map.

**Figure 77. Task attended vs. no task unattended – Frequency tags – Δ Levels – Size.**

Topographic heat maps representing F-statistics from linear contrasts comparing the frequency tags across the four Δ levels for the size condition for all electrodes. The two rows represent the task relevant attended (left) and task irrelevant unattended (right) conditions. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of
p=0.05 and the corresponding FDR q-values and F-statistics are shown at the bottom of each map.

**Figure 78. Task attended vs. no task unattended – Index values – Δ Levels – Orientation.**

Topographic heat maps representing F-statistics from linear contrasts comparing the index values across the four Δ levels for the orientation condition for all electrodes. Each index value represents a normalized difference between the task relevant attended and task irrelevant unattended conditions. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of p=0.05 and the corresponding FDR q-values and F-statistics are shown at the bottom of each map.
Figure 79. Task attended vs no task unattended – Index values – Δ Levels – Size.

Topographic heat maps representing F-statistics from linear contrasts comparing the index values across the four Δ levels for the size condition for all electrodes. Each index value represents a normalized difference between the task relevant attended and task irrelevant unattended conditions. Each row represents the first, second, and third harmonic. F-statistics are represented by color values ranging from yellow to red, yellow representing small values and red representing large values. Electrodes that reached significance are plotted in black. Values are thresholded at an alpha of p=0.05 and the corresponding FDR q-values and F-statistics are shown at the bottom of each map.
Experiment 3 summary

There were 6 separate analysis presented in the above oddball analysis section. It seems appropriate to summarize the analysis here. First, we tested the prediction that ensemble encoding does not require the allocation of spatial attention or an explicit averaging task. We found no significant differences between the levels of either feature in the frequency tags, collapsed across attended hemifield and relevant task. This indicates that either spatial attention or an averaging task is required to extract the average. Alternatively, this is also evidence for theory three, that ensemble encoding is a strategy.

Next, we tested the prediction that ensemble encoding requires either spatial attention or a relevant averaging task to focus participant’s attention to the relevant feature. To test this we compared the neural response to when participants were attending compared to not attending as well as to when participants were performing a relevant task to when they were performing an irrelevant task. For both of these analyses we collapsed across the other condition (collapsed across task for attention comparison and attention for task comparison). We found no significant differences between either of these comparisons. This indicates that both attention and task are needed to see any representation of the average. This is also evidence in support of theory four, that ensemble encoding is a strategy as opposed to a representation of the average.

Lastly, we tested the prediction that ensemble encoding requires both spatial attention and a relevant averaging task. To test this prediction we performed three
analyses. First, we compared the neural response to when participants were performing an irrelevant task and attending to the correct side of visual space with when participants were performing an irrelevant task and not attending. Second, we compared the neural response to when participants were performing a relevant task and attending to when they were performing a relevant task and not attending. Third, we compared the neural response to when participants were performing a relevant task and attending to when they were performing an irrelevant task and not attending. For the large majority of these analyses done for each of these comparisons, we didn't find a significant linear contrast in more electrodes than you would expect by chance. However, in the relevant task attended vs. the irrelevant task unattended comparisons, there is some evidence of ensemble averaging in the irrelevant task not attending condition. This is potentially indicative of ensemble encoding to unattended information, which would support many theories of ensemble encoding. This should be interpreted with caution however, considering the lack of any consistent findings in any of the other comparisons. Again, this is also evidence in support of the fourth theory of ensemble encoding as a strategy. Overall, although a small number comparisons contained many significant electrodes, this pattern was not consistent across comparisons. Therefore it is most likely that the 4th theory of ensemble encoding best fits our data.
General Discussion

This thesis attempted to identify the neural dynamics of ensemble encoding. As almost nothing has been published on this topic to date, we took a fairly broad strokes approach, in which we used fMRI to elucidate where in the brain this occurs and EEG to elucidate when in time these processes emerge. In addition to questions of where and when in time this happens we also probed the underlying processes of ensemble encoding by asking questions about the involvement of spatial attention and task relevance. We hypothesized that if there is a neural representation of the average we can detect it using EEG and fMRI by systematically manipulating the average feature level of the ensemble, the focus of spatial attention, and the relevant feature being averaged. Specifically we proposed three aims to focus our investigation: (1) investigate the neural correlates of ensemble averaging using fMRI repetition suppression; (2) explore the pattern of activity in processing feature averages over time using EEG and MVPA; and (3) identify the effects of attention and task on the representation of the average using EEG and frequency tagging.

Checks for quality control

First, it is important to establish that the data we are using in the analysis are clean and not overly noisy. To do so we can look at several measures for each experiment. In the VEP data, we see evidence of clean VEP signal in the split half correlation analysis. Specifically, in the split half correlations the expected pattern of results would show little to no significant correlations during the pre-stimulus
baseline and for ~50 ms after stimulus onset, as there should not be significant correlations before the signal has had a chance to reach visual cortex. For the frequency tagging data we performed a $T_{circ}^2$ analysis on the carrier frequencies to ensure the source of the signal was consistent across presentations. The overwhelming majority of electrodes were significant for both carrier frequencies in this analysis. In order to ensure that participants are performing the task at the attended location, we also looked at the difference between the carrier frequencies while participants were either attending or not attending to those frequencies. Again, we found that the large majority of electrodes showed greater frequency tags for the attended than unattended condition, which indicates that they are attending to the correct side of visual space. As a result, we can be fairly confident in the reliability of the data collected for these two experiments.

**The representation of the ensemble**

Initially, we set out with the goal of finding a neural representation of the ensemble. However, it seems that the ensemble is not represented in the way many people hypothesize that it might be. For example, one proposal of ensemble encoding, is to facilitate rapid scene perception, of ‘gist’ perception (Brady et al., 2017). In other words, we may extract low-level ensemble averages, like orientation and spatial frequency statistics, from scenes in order to encode the gist of the scene (the forest for the trees). As another example, one theory proposes that ensemble encoding is used to fill in information that is not actively attended to, especially in
the periphery (Cohen et al., 2016). In order to create a seemingly rich visual experience we rapidly extract summary statistics across a range of features in areas of visual space that are unattended. We then use these averages to fill in these areas of unattended visual space. The result is a much more rich subjective experience, which matches our actual perceptual experience. Put simply, many theories of ensemble encoding rely on it being a rapid and parallel process. However, as our data demonstrate, this is most likely not the case.

**The neural correlates of the average**

In experiment 1, we attempted to probe the neural correlates of the ensemble average. We did so by comparing the average BOLD signal between levels of two ensemble features (orientation and size) across all early visual regions of interest (ROIs). Specifically, we looked for evidence of repetition suppression during repeated exposure to an adapting ensemble followed by ensembles of various levels of difference in the relevant feature compared to the adapting stimulus. However, we found no significant differences, in the form of a positive linear relationship, in any of the ROIs. Interestingly, we found that the response for the level one stimuli for orientation was consistently smaller than the other three features in V1-V3, which is reflected by a linear contrast that approaches significance. This is indicative of some amount of repetition suppression for the condition in which the average does not change, which we would predict. This wasn’t the case for the other three feature levels however. The pattern of results was maintained even when the data
was z-score normalized. These results reflect a pattern that is not consistent with our hypothesis. However, there does seem to be a consistent pattern present in early visual ROIs for smaller 1st level responses and maximum response in the second level. This is followed by smaller responses for the remaining two levels. This pattern is consistent in the orientation condition across all ROIs. Although the size condition shows some hints of this pattern, the results for size are much less consistent across ROIs and feature levels.

One potential explanation for these results is the interaction of attention and task difficulty. For example, data here are averaged across a block of 16 repetitions of adapting and test stimuli. However, participants may detect the level of the change during the first one or two presentations. This may lead them to completely ignore the rest of the presentations in the trial. If this were the case, we might expect to see greater activation for the second level compared to the first, but smaller activity for the later levels. Additionally, it is apparent that the pattern of results across the two features is drastically different. For orientation, as mentioned above, we see the largest activation for level 2 and the smallest activation for level 1 across ROIs. For size, we see more inconsistent and somewhat random results. As each size level is presented during every orientation level his may be indicative of separate representations for each of these two features. However, these results are certainly inconsistent and therefore, no firm conclusions can be made about areas in the brain that contribute to ensemble encoding.
Temporal dynamics of ensemble encoding

In experiment 2, we attempted to probe the temporal dynamics of ensemble encoding using EEG. We did so by comparing the visually evoked potential (VEP) in response to presentations of ensembles with standard feature averages to ensembles with progressively larger or more rightward titled averages. We predicted that the signal would be differentiable during some portion of the VEP either early in time if, for example, attention is not required to extract the ensemble or later in time if attention is required. We also predicted that these patterns of differentiability might be different for the two types of features being averaged. However, we did not find any evidence of differentiability in the VEP for the varying feature levels that matched our predictions. Instead, we observed VEPs that were highly correlated among themselves and each other. This indicates 2 things: first the VEP data we are seeing is clean signal. The pattern of the split half correlations show very low correlations between levels pre-stimulus baseline and up to ~50 ms before stimulus onset, which is what would be expected. Second, during the time period where you would expect to see differentiability among the signals of different levels, we instead see high correlations among all the levels. This indicates that within the neural signal there are either no measurable differences or that the brain is representing each level similarly.

We also performed two multivariate analysis, split-half correlations and MVPA, which again showed no consistent or reliable differentiable responses during any time points of the VEP. Again, not only do these results not reveal the temporal
dynamics of ensemble encoding, but they are indicative of no differentiable response among varied levels of the average feature. Therefore, as with the neural correlate, we cannot make a firm conclusion about when in time the average emerges.

**Is ensemble encoding automatic and pre-attentive?**

In addition to questions about the neural correlates and temporal dynamics of ensemble encoding, we also made predictions of the effects of attention and task on the representation of the average. We tested three predictions of the role of attention and task relevance on the representation of the ensemble. First, we tested the prediction that ensemble averages are extracted regardless of where the focus of spatial attention is or whether or not participants are actively trying to extract the mean. Next, we tested whether spatial attention or a relevant task are needed to extract the average. Lastly, we tested the prediction that ensemble encoding is reliant on both spatial attention and a relevant task. We found largely inconsistent results for each of these analyses.

We found very few significant electrodes for any of the analysis comparing each of the five levels collapsed across attended hemifield and task, indicating that ensemble encoding relies on either the allocation of spatial attention or the presence of a task that forces them to attend to the particular feature. Additionally, we found few electrodes that reached significance in any of the analysis done for prediction two. Specifically, comparing attended and unattended conditions
collapsed across task and comparing relevant task with irrelevant task collapsed across attended hemifield. One exception to this can be found in the attended vs. unattended comparison for orientation in the third harmonic for both the delta and feature levels (Figure 40 and 44 bottom left). This analysis shows a number of significant electrodes in the attended condition despite no significant electrodes in the unattended condition. Overall, however, these data indicate that ensemble encoding requires some amount of allocated attention either spatially or to the specific feature of interest.

For the third prediction we again found inconsistent results. For the majority of comparisons we found very few significant electrodes. However there were some analyses that showed many significant electrodes. In the task attended compared with task unattended condition (Figure 72), we see more significant electrodes in the task attended condition than the task unattended condition. This is also reflected in the index value analysis (Figure 74), which shows a large number of significant electrodes. This is supportive of theory 3 in which ensemble encoding requires the allocation of attention and a relevant task. At the same time, however, we also see stark differences in the number of significant electrodes between the Δ levels for the orientation task attended and no task unattended conditions (Figure 76). Specifically, we see more significant electrodes in the irrelevant task unattended condition than in the task attended condition. This would be evidence that suggests that ensemble encoding may only emerge after attention has been engaged elsewhere. Unexpectedly, after further investigation of the underlying
topographic maps of the Δ level values used in the linear contrasts, we see a pattern of activation opposite to what we would predict (Figure 80). Specifically, we see the greatest activation for the first level and smaller activation as the level of the oddball increases relative to the standard ensembles. These results are conflicting and counter-intuitive to what you might predict if the brain was encoding a representation of the average.

We also see a dramatically different pattern of responses between the size and orientation conditions. Across every comparison we see a different pattern of significant electrodes between the two features of interest. This is interesting because each level contains data from every other level of the opposite feature. More work is needed to further parse apart the role of attention and task on ensemble encoding.
Figure 80. Individual Δ levels for the first three harmonics of the frequency tags in the orientation no task unattended condition.

Topographic heat maps of the four averaged Δ levels (columns) for the first three harmonics (rows) of the orientation task irrelevant unattended condition. Values range from 0-400 μV. Larger amplitudes are represented by red and smaller values are represented by yellow on the heat maps.

**Ensemble encoding as a strategy**

If ensemble encoding were not a purely parallel and pre-attentive process, it would still be efficient to condense large amounts of redundant perceptual information, even if by some strategy. For example, imagine giving a speech in front
of a large audience. It would behoove the speaker to follow the general mood and focus of attention of the audience. To do this efficiently the speaker could extract the average expression and average eye gaze of the crowd. The resulting averages extracted would still be extremely valuable to the individual, even though the speaker might have to pay attention to the crowd of faces and actively try to see the average facial expression or gaze direction. Additionally, it wouldn’t matter to the speaker if he were doing this in parallel without attention or by using some kind of efficient strategy to do so.

As mentioned previously throughout the dissertation, one theory of ensemble encoding states that participants are simply using a strategy to extract the average (Lau & Brady, 2018; Simons & Myczek, 2008). For example, by using the largest and smallest items (or least and most rightward tilted items) participants can make a relatively accurate judgment of the ensemble (Simons & Myczek, 2008). This type of strategy may not be recognizable within the neural signal as it is drowned out by the visual response to the rest of the items on the screen, similarly to if we were to attempt to differentiate between the neural response of two ensemble of exactly the same objects where two to three of the objects differed along one of the relevant feature dimensions.

One interesting finding that supports this theory is that participants are still able to perform this task behaviorally, despite there being no neural difference. As can be seen in Figure 7, Figure 17, and Figure 32, participants are performing the task for both orientation and size with relatively high accuracy. Therefore,
participants must be extracting some form of the representation of the average, even if it is not differentiable in the neural response. These results seem to lend support to ensemble encoding as a strategy instead of a pre-attentive parallel process.

**What is the neural representation of the average and why do we extract it?**

Ensemble encoding serves a specific purpose: to compress or summarize redundant or repetitive sensory information. This has been demonstrated behaviorally many times for many features, paradigms, and in support of a variety of different cognitive and perceptual abilities. But what is an ensemble average and why do we use neural resources to extract it?

Some theories propose that ensemble averages are extracted to enhance memory representations, facilitate rapid visual search, facilitate gist perception, and provide us with a rich perceptual experience (Alvarez, 2011b; Alvarez & Oliva, 2008; Brady et al., 2017; Cohen et al., 2016; Rosenholtz et al., 2012; Utochkin, 2015). Although it may be tempting to attribute one perceptual representation to such a wide array of perceptual and cognitive functions, it seems unlikely that ensemble encoding is responsible or involved with them all. One reason is that each of these abilities has a different set of requirements that must be met to extract the ensemble. For example, in order to fill in large areas of visual space to facilitate rich perception or gist perception, participants must have a relatively unlimited capacity for encoding, be able to rapidly extract the average without explicitly attending to
them, and do this for multiple features simultaneously. On the other hand, if we assume that ensemble encoding is not pre-attentive or automatic and is simply a way to represent a large number of items that differ along some feature dimension, it can still be an extremely useful ability and fit with a variety of different theories. For example, if we are searching for one item amongst many items, then the average can be used to facilitate this process by providing a template to compare individual items with. The results of the current set of experiments seem to support a more limited capacity form of ensemble encoding. So, although it is theoretically important to understand how an average is extracted, the fact remains that we know that participants can extract some representation of the underlying statistical structure for a variety of features within a group of objects and this information can be used to make seemingly difficult tasks easier.

An important question still remains however: what is being represented by the ensemble average. So far the overwhelming majority of research studies have examined the extraction of ensemble statistics through the proxy of the mean. However, these data, along with a variety of other studies have suggested that other statistics are extracted. For example, if we assume that, instead of individual features being extracted in parallel, the range of the feature is being extracted, then this indicates that a distribution of feature levels is instead encoded. Using this distribution, a variety of useful statistical information can be extracted, like the variance, mean, median or skew. Additionally, outlier items become much easier to identify. This also fits nicely with both theories of parallel processing of ensembles
or of ensemble encoding as a strategy. It would also provide a link between the representations of individual item features with that of the average feature, which can provide more insights into the nature of the average representation. By considering ensemble encoding in this light, it may be easier to design and execute future studies that attempt to find neural representations of ensemble encoding. At the end of the day though, we must continue to probe the underlying nature of this perceptual representation because regardless of how it represented or extracted, it is clearly a useful source of information in a variety of perceptual and cognitive functions.

**Limitations**

One important potential limitation with the current series of studies is task difficulty and the resulting allocation of attention. One issue that may arise in analyzing neuroimaging or recording data with long trial times is whether or not participants are actually allocating attention in the same way and to the same degree for each level in the task. For example, if it was too easy for the participant to perform the task and they detect the difference in the ensemble early in the trial (or block) we might expect the participant to stop attending to the ensembles. As a result, we would predict that the neural signal will be smaller than if they were consistently paying attention for the duration of the trial. As we use participant specific levels that are based on the participant’s behavioral performance, we should not have this problem and the task should remain relatively difficult for each level.
However, the psychophysical experiment was different in terms of difficulty compared to the tasks of the other experiments. One potential way to solve this is to have separate behavioral tasks for each of the experiments for which you can more accurately judge participant's individual levels. Another way to solve this is to have a separate task during the trial that ensures participants are focusing their attention throughout the trial.

Although we implemented as many stimulus controls as possible, these experiments are still limited by inherent problems associated with the ensemble stimuli. For example, to control for potential strategies, like encoding minimum and maximum values, we would have to manipulate the variance among each level so that minimum and maximum values from the smallest and largest items in the smallest and largest ensembles are overlapping. As a result the variance would be too great to accurately perform the task. Additionally, the purpose of this experiment was to identify a differentiable signal representative of the average in the neural data and not necessarily to rule out potential strategies used. So performing this manipulation would be counter-intuitive to the goal of the current series of experiments.

Lastly, this experiment suffers from problems inherent with neuroimaging and recording using larger full field stimuli. If we assume that there are neural representations of ensemble averages to be detected they may simply be to subtle to be detected in the large amount of activation evoked by these large stimuli. These
problems are not easy to overcome and more work will need to be done in an attempt to draw out any representation of the average.

Conclusions

In conclusion, we did not find consistent differentiable representations of ensemble averaging in the EEG or fMRI data for orientation or size averages. Unsurprisingly, we also found little evidence of any effects of attention or task on the neural representation of the average. Our data seem to be most consistent with a theory of ensemble encoding that serves more as a strategy for extracting feature information from a subset of items as opposed to a system that extracts information from items within the set in parallel. Understanding the neural processes underlying ensemble encoding is an important task to help in understanding how humans represent feature information and more work will need to be done to get more of a complete picture.
References


Norman, L. J., Department of Psychology, D. U., Durham, United Kingdom, liam.norman@dur.ac.uk, Heywood, C. A., Department of Psychology, D. U., Durham, United Kingdom, c.a.heywood@durham.ac.uk, …


