1-1-2019

# Integration and Segmentation Conflict During Ensemble Coding of Aspect Ratio

Elric Matthew Elias
*University of Denver*

# Integration and Segmentation Conflict During Ensemble Coding of Aspect Ratio

## Abstract

The visual system often *integrates* information that "goes together". Once information has been integrated, summary information (e.g., average emotion or average size) can be extracted; this occurs during *ensemble coding*. Integration thus allows for fast and efficient generalizations about sets to be made. In contrast, the visual system sometimes *segments* input that *does not* go together. For example, the perception of objects can be exaggerated away from natural category boundaries (e.g., a perfect circle is a category boundary; it is neither "flat" nor "tall"). Segmentation allows the visual system to make quick categorical distinctions. Much of the time, integration and segmentation work in parallel, and they have most often been studied in isolation. However, investigating how these two processes operate together, and potentially even *conflict*, was the purpose of this dissertation. I examined the ensemble coding of aspect ratio, which is a visual feature roughly equivalent to "tallness/flatness". Aspect ratio has a category boundary (e.g., a circle or square), and the perception of aspect ratio tends to be exaggerated -segmented - away from that boundary. Thus, I predicted that observers' ability to integrate aspect ratio information that spanned the category boundary would be disrupted, since in those instances, integration and segmentation would be at odds. To test this prediction, observers were asked about the average aspect ratio of a set of ellipses. In two experiments, observers were less sensitive to the mean of sets that included both tall *and* flat ellipses, compared to sets that only included tall *or* flat ellipses. A third experiment confirmed that segmentation perceptually distorted the appearance of ellipses near the category boundary *away* from that boundary; shapes were perceived to be more extreme than they actually were. Segmentation thus made sets that included both flat and tall ellipses appear more heterogeneous than they really were, which disrupted ensemble coding. In general, these experiments provide a deeper understanding of how the visual system summarizes large sets of information, by investigating how integration interacts with, and even conflicts with, segmentation.

Document Type
Dissertation

Degree Name
Ph.D.

Department
Psychology

First Advisor
Timothy D. Sweeny, Ph.D.

Keywords
Ensemble coding, Perception, Segmentation, Vision

Subject Categories
Psychology | Quantitative Psychology | Social and Behavioral Sciences

# Integration and Segmentation Conflict during Ensemble Coding of Aspect Ratio

A Dissertation

Presented to:

the Faculty of Social Sciences

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

by

Elric Elias

August 2019

Advisor: Timothy D. Sweeny

Author: Elric Elias
Title: Integration and Segmentation Conflict during Ensemble Coding of Aspect Ratio
Advisor: Timothy D. Sweeny
Degree Date: August 2019

## ABSTRACT

The visual system often *integrates* information that “goes together”. Once information has been integrated, summary information (e.g., average emotion or average size) can be extracted; this occurs during *ensemble coding*. Integration thus allows for fast and efficient generalizations about sets to be made. In contrast, the visual system sometimes *segments* input that *does not* go together. For example, the perception of objects can be exaggerated away from natural category boundaries (e.g., a perfect circle is a category boundary; it is neither “flat” nor “tall”). Segmentation allows the visual system to make quick categorical distinctions. Much of the time, integration and segmentation work in parallel, and they have most often been studied in isolation. However, investigating how these two processes operate together, and potentially even *conflict*, was the purpose of this dissertation. I examined the ensemble coding of aspect ratio, which is a visual feature roughly equivalent to “tallness/flatness”. Aspect ratio has a category boundary (e.g., a circle or square), and the perception of aspect ratio tends to be exaggerated—segmented—away from that boundary. Thus, I predicted that observers’ ability to integrate aspect ratio information that spanned the category boundary would be disrupted, since in those instances, integration and segmentation would be at odds. To test this prediction, observers were asked about the average aspect ratio of a set of ellipses. In two experiments, observers were less sensitive to the mean of sets that included both tall *and* flat ellipses, compared to sets that only included tall *or* flat ellipses. A third experiment confirmed that segmentation perceptually distorted the appearance of ellipses near the category boundary *away* from that boundary; shapes were perceived to be more extreme than they actually were. Segmentation thus made sets that included both flat and tall ellipses appear more heterogeneous than they really were, which disrupted ensemble coding. In general, these experiments provide a deeper understanding of how the visual system summarizes large sets of information, by investigating how integration interacts with, and even conflicts with, segmentation.

# TABLE OF CONTENTS

# LIST OF FIGURES

# INTRODUCTION

Imagine being immersed in an infinite sea of continuous information. Your task is to parse that information into useful, meaningful chunks. Where would you even start? How would you decide which bits of information go together, and which do not? What information is important enough to attend to? The visual system faces, and usually solves, this daunting puzzle every moment light enters the eye. These computational feats occur even in spite of the many bottlenecks in visual processing (e.g. attention, Chong & Treisman, 2005b; memory, Luck & Vogel, 1997). To accomplish them, the visual system leverages at least two strategies to make sense of input. First, it *integrates* information that, in some sense, “goes together”. Gestalt psychologists for example, as well those in the field of perceptual organization, have attempted to understand how the visual system organizes disparate retinal input into the objects and groups of objects we phenomenologically see (see, for example, Wertheimer, 1923; Palmer, 1999; Palmer, 2002; Wagemans et al., 2012, 2012b; Peterson & Kimchi, 2013). For instance, objects that share a common region, move together, or share a common feature like color tend to be grouped together, and grouping of visual elements can inform object recognition (Palmer, 2002). The visual system does not just organize and group input, though, it integrates and extracts summary information from it (see Whitney, Haberman & Sweeny, 2014; Whitney & Leib, 2018, for reviews of ensemble coding; Ross & Burr, 2008;

Alvarez, 2011). Information integration thus helps the visual system reduce redundancy and minimize differences among individual objects, it increases efficiency, it allows objects that go together perceptually to *appear* to be a group, and it facilitates the extraction of summary statistics about that group. Imagine, for example, catching a glimpse of a flock of birds. The individual birds are grouped to form a flock, and by integrating information about individual birds (e.g. motion heading information), the visual system can efficiently extract information about a gestalt object (e.g., the heading of the flock in general), without having to precisely model the heading of each and every bird (and without having to retain access to information about the individual birds).

In contrast to information integration, the visual system also *segments* input that does not go together by, for example, exaggerating differences instead of pooling across them. By segmenting information, the visual system is able to avoid making generalizations across objects that do not belong together. For example, motion information about a flock of birds may not be confused with motion information about a nearby stationary bird perched on a tree; this information has been segmented. Critically, segmentation may be especially relevant for visual features that have a  category boundary—a value that lies exactly between dimensions and belongs to neither (e.g., a perfect circle is neither “tall” nor “flat”; a straight-ahead eye gaze is neither leftward nor rightward). Exaggerating differences around these category boundaries is especially important for the visual system. Segmentation thus allows the visual system to organize objects into one feature category or the other (e.g., is the object tall *or* flat?; Suzuki & Cavanagh, 1998; Sweeny, Grabowecky & Suzuki, 2011), and it reduces the chances of making categorical errors when making noisy perceptual judgments about objects near a boundary (e.g., it would be better to err in perceiving a slightly tall shape as being moderately tall than slightly flat; Kourtzi 2010; Sweeny, Haroz & Whitney, 2012; Wei & Stocker, 2017). Segmentation may thus support perceptual decisions when precision is less important than coarse categorization (Suzuki, 2005). In general, these two broad computational processes–integration and segmentation–help the visual system efficiently handle a vast sea of input in the face of limited computational resources, and it is these two processes that are central to this dissertation.

Although segmentation and integration help the visual system sort input in opposing ways, much of the time they operate in parallel and do not seem to conflict. Perhaps for this reason, these two computational methods have most often been studied in isolation. However, there may be instances in which they interact. For example, imagine laying out your pocket change onto a desk. From your perspective, each coin projects a two-dimensional, flattened ellipse shape on your retinae. Since each coin is aligned in-depth in a similar way (they're all laying on the same surface), extracting the average flatness—and inferring the average orientation-in-depth—should be possible (Biederman & Kaloscai, 1997; Treisman & Gormican, 1988). But now imagine instead a high-speed snapshot of those coins being tossed in the air. Some coins project flat elliptical shapes to the camera, but some are head-on relative to the camera and project a circular aspect ratio, while others project a tall shape by being vertically oriented and rotated relative to the camera. What is the average shape among the set of coins oriented in-depth? In this case, integration and extraction of summary information is more complicated.

On the one hand, the visual system tends to *segment* heterogeneous information like this *away* from the category boundary, exaggerating differences between individual objects. For example, flat objects are made to appear flatter and tall objects are made to appear taller (Suzuki & Cavanagh, 1998, Sweeny, Grabowecky & Suzuki, 2011; Sweeny, D'Abreu, Elias & Padama, 2017). On the other hand, in examples like the one above, the visual system also tends to *integrate* information *across* that same feature boundary to determine what the components have in common as a collective. As mentioned, prior research on integration—and subsequent mean extraction—has left this potential tension largely unexplored. In many prior investigations of integration (e.g., Ariely, 2001; Chong & Treisman, 2005; Haberman & Whitney, 2009; Elias, Dyer & Sweeny, 2017), category boundaries for the relevant visual feature are either ambiguous (e.g. the boundary between "small" and "large" objects), or the impact of those category boundaries was not explicitly considered (e.g., Sweeny & Whitney, 2014). It is important to understand not only how integration helps the visual system solve a host of computational problems, but also how that solution is potentially constrained by the system's solution to *other* problems (e.g., segmentation of categorical visual information). Investigating the relationship—and potential conflict—between these two fundamental computational mechanisms is the primary aim of this dissertation. In doing so, I hoped to add to a more complete understanding of how the visual system computes the properties of sets of objects.

In Experiment 1, I tested the visual system's ability to integrate information about aspect ratio across category boundaries–boundaries around which information segmentation normally operates. I expected integration to operate with reduced efficiency in these cases, and the results clearly support this prediction. Experiment 2 replicated the main results of Experiment 1, while also addressing a few methodological limitations from Experiment 1. Experiments 3a and 3b showed that segmentation repulsed the appearance of ellipses away from the category boundary—the aspect ratios of ellipses appeared to be more extreme than they really were. This distortion introduced exaggerated heterogeneity, which ultimately disrupted integration.

### *Integration in the Form of Ensemble Coding*

Integration is one of the visual system's basic approaches for solving computational problems. Individual cells integrate information (e.g., Miller, Gochin & Gross, 1993; Rolls & Tovee, 1995; Sato, 1989; Kastner, De Weerd, Pinsk, Elizondo, Desimone & Ungerleider, 2001; Brincat & Connor, 2004; Zoccolan, Cox, & DiCarlo, 2005). So do populations of cells (e.g., Pasupathy & Connor, 2001; Suzuki, 2005; Michele, Chen, Geisler & Seidemann, 2013). In many cases, information about individual features or objects are integrated, via these populations. This happens, as suggested already, in classic grouping processes (see Palmer, 1999; Palmer, 2002; Wagemans et al., 2012; Peterson & Kimchi, 2013). When information about multiple objects is integrated such that a summary judgment about the entire group can be made, the process is known as ensemble coding (see Whitney, Haberman & Sweeny, 2014; Whitney & Leib, 2018, for reviews of ensemble coding). Ensemble coding is one consequence of information integration – one that usually implies the pooling of information to acquire summary information (e.g., the mean, variance, etc.) across sets with more than two members (Whitney & Leib, 2018). For example, as you pass a fruit stand full of oranges, your visual system can utilize ensemble coding to extract the average size of the fruits quickly and automatically (Allik, Toom, Raidvee, Averin, Kreegipuu, 2014), without having to sequentially sample each and every fruit (Ariely, 2001; Chong & Treisman, 2003; 2005; Sweeny, Wurnitsch, Gopnik, & Whitney, 2015). Or, perhaps you catch a glimpse of a group of joggers; your visual system may extract information about their average heading in a similar way (Sweeny, Haroz & Whitney, 2013). Similar operations can be performed for simple features like orientation (Parkes, Lund, Angelucci, Solomon & Morgan, 2001; Ross & Burr, 2008; Alvarez & Oliva, 2009; Elias, Padama & Sweeny, 2018) and speed (Watamaniuk & Duchon, 1992; Watamaniuk, Sekuler, & Williams, 1989), as well as for socially relevant information like facial expression and gaze (Haberman, Harp, & Whitney, 2009; Haberman & Whitney, 2007, 2009; Sweeny & Whitney, 2014; Elias, Dyer & Sweeny, 2017).

Importantly, although ensemble coding helps circumvent the computational limitations of the visual system by compressing information about individuals into a "gist" representation that characterizes the group, information about individuals can be lost to conscious access (Haberman & Whitney, 2007; Allik et al., 2014). Indeed, observers sometimes fail to perceive, attend to and/or recall individual set members at all or do so poorly (e.g., Alvarez & Oliva, 2009, Sweeny et al., 2015). This may occur because images are presented for very brief durations (Haberman et al. 2009, Oriet & Corbett 2008), because neuropsychological deficits disrupt the perception of group members (Yamanashi Leib, Landau, Baek, Chong & Robertson, 2012; Yamanashi Leib et al., 2012b; Robson, Palermo, Jeffery & Neumann, 2018), or because set members are masked from awareness in some way (Choo & Franconeri 2010; Jacoby, Kamke & Mattingly, 2013; Ward, Bear & Scholl, 2016; see Elias et al., 2018 for possible limits to pooling of complex visual features that are masked). Yet even in cases like these, information integration and ensemble coding can proceed and make a summary representation available to the perceiver. Ensemble coding even allows people to make judgments about groups that are more precise and accurate than judgments about a single individual seen in isolation (Sweeny, Haroz & Whitney, 2012, 2013; Elias et al., 2017). Interestingly, ensemble coding can also bias perception of an individual's features toward the mean features of the overall group (Brady & Alvarez, 2011). Thus, ensemble coding is not only fast, efficient and robust, it is clearly very useful. It can provide very precise information about the general characteristics of visual information that is outside of focused attention, or even about forgotten or unperceived individuals in a larger set. This information can then help guide focused attention in future moments (Alvarez & Oliva, 2009; Alvarez, 2011; Im et al., 2017), help detect (or ignore) notable outliers in a group (Haberman & Whitney, 2009b; Haberman & Whitney, 2010), or even incorporate group attributes into the perception of individuals (Brady & Alvarez, 2011).

### *Segmentation*

The visual system doesn't always pool information across objects in order to extract summary statistics, though. Sometimes, it instead exaggerates differences, segmenting visual information so that perceivers may make categorical decisions about an object. A striking example of perceptual segmentation is the tilt illusion, in which the orientation of a center patch of parallel lines is perceptually repulsed—segmented—away from the orientation of lines within an adjacent concentric ring (see Clifford, 2014, for a review). Segmentation may also be an important component of the Poggendorff illusion, in which two alternate exterior acute angles may be perceptually exaggerated to appear different from each other (Morgan, 1999; Westheimer, 2008). Segmentation can even act to exaggerate spatial displacement between objects (Baddock & Westheimer, 1985; Suzuki & Cavanagh, 1997), and is likely involved in distinguishing figure (i.e., object) from ground (i.e., background; Westheimer & Levi, 1987; Grossberg, 1994).

Thus, when computing the value of a feature, context matters. Sometimes, making categorical distinctions (e.g., “is the object vertical or not”, “is the object here or there”, “is the object figure or ground”) is what is important. This is especially true for visual features that have category boundaries. Aspect ratio is one such feature (other examples include object taper, skew and convexity; Suzuki, 2005). Aspect ratio can be thought of as a visual object’s “tallness” or “flatness”, and is a 2D visual feature that can provide information about an object’s orientation-in-depth (Beiderman & Kaloscai, 1997; Treisman & Gormican, 1988). Aspect ratio is encoded by cells in the inferotemporal (IT) cortex, separately from simpler visual features like size or curvature (Reagan & Hamstra, 1992; de Beek, Wagemans & Vogels, 2003; Dickinson, Morgan, Tang & Badcock, 2017). Aspect ratio varies around a category boundary or null-point (e.g., perfect circles and squares are equivalently “flat” and “tall”, or equivalently neither), and indeed, being able to discriminate between categorically “tall” and categorically “flat” is a prioritized task for the visual system. For example, at short time scales – and thus perhaps in the face of perceptual uncertainty due to noisy neural representation (Wei & Stocker, 2017) – perceived aspect ratio tends to be exaggerated away from the null-point, toward extreme values (Suzuki & Cavanagh, 1998; Sweeny, Grabowecky & Suzuki, 2011; Dickinson et al., 2017; Sweeny, D’Abreu, Elias & Padama, 2017). Additionally, extremely “tall” or extremely “flat” shapes stand out from a field of perfect circles quite clearly, although the reverse is not true (Treisman & Gormican, 1988). Similarly, observers are especially sensitive to slight changes in aspect ratio around the null-point (Reagan & Hamstra, 1992; Suzuki et al., 2005), supporting the accurate perception of even subtly “flat” or subtly “tall” objects. Perceptual evidence like this is, unsurprisingly, reflected in the way the visual system encodes aspect ratio at the neural level. The majority of cells in IT tuned to aspect ratio respond more strongly to extreme values than to values near the null-point (Kayaert, Biederman, Beeck & Vogels, 2005). Additionally, fewer cells are tuned to values near the null-point, and they respond more weakly than those tuned to extreme values. The perceptual consequences of this can be surprising. It is, for example, easier to mask circles than extreme aspect ratios, likely because the neural representation of circles is relatively weak (Braun & Sweeny, 2019).

What is less clear is exactly how information about aspect ratio is encoded at the neural population level. Multiple encoding schemes could, theoretically, lead to the perceptual segmentation of aspect ratio discussed above. Below, I will briefly discuss two possible encoding schemes. Aspect ratio was once thought to be supported by an *opponent-coding* scheme in which one neural population is broadly tuned to “tall” shapes, and a second is tuned to “flat” shapes, with perceived aspect ratio as the centroid of these two distributions (Regan & Hamstra, 1992; Suzuki, 2003, 2005). Instead, recent work suggests that a *multi-channel* encoding scheme—one in which multiple neural populations are each tuned to “flat”, “tall” or intermediate values—is more likely (Dickinson et al., 2017; Storrs & Arnold, 2017), although as already mentioned, intermediate values do seem to be represented more sparsely and weakly (Kayaert et al., 2005). Even given a multi-channel encoding scheme, it is possible that a greater number of channels are distributed around the null-point compared to the number of channels devoted to extremely “flat” or “tall” values. Irrespective of the precise details, though, categorical judgements are especially important for aspect ratio. The neural coding of aspect ratio supports categorical judgements and segmentation; the perception of aspect ratio does too.

### *Integration and Segmentation of Aspect Ratio*

Although aspect ratio is clearly subject to segmentation, there is as of yet very little evidence that it can be integrated by ensemble coding, or any other process (see Oriet & Brand, 2013, for potential aspect ratio integration, though changes in the aspect ratio of their stimuli were confounded with size and area, which are already known to be easily ensemble coded). Yet it is reasonable to expect that it should be. After all, aspect ratio is a mid-level visual feature, encoded in intermediate stages of vision (e.g., V4; Dumoulin & Hess, 2007), along with other global shape attributes in IT (e.g., Kayaert et al., 2005). Aspect ratio is thus computed between simple features (e.g. orientation) and more complex features (e.g., facial expression) on which pooling is known to act. Further, it is encoded by populations of dedicated IT cells (Kayaert et al., 2005), and pooling operates across dedicated neural populations that encode other features (Pasupathy & Connor, 2004; Suzuki et al., 2005). For these reasons, I expected the integrative process of ensemble coding to operate on aspect ratio. Importantly though, I predicted that this process of ensemble coding should be particularly efficient for sets of aspect ratios that fall on one side of the null-point (e.g., a set of "flat-ish" ellipses), compared to sets that cross the category boundary (flat and tall ellipses). In cases like these, the visual system should be able to leverage information integration without being simultaneously pressed to segment information across a category boundary. In contrast, if "flat" *and* "tall" ellipses are present in a set about which generalizations must be made, the visual system is faced with a dilemma. On the one hand, integration should operate to pool information, "toward" the set mean, thus supporting generalizations. On the other, aspect ratio perception and encoding should segment information "away" from the null-point, thus maximizing perceived differences between set members. So, although summary statistics can be extracted from a wide range of visual features, in the case of aspect ratio, the category boundary should matter. While ensemble coding should operate efficiently for sets of generally-"flat" objects, and for separate sets of generally-"tall" objects, it should operate with less efficiency for sets containing both "flat" *and* "tall" objects – in other words, for sets with a clear category boundary. I reasoned that the presence of a category boundary would diminish the effectiveness of ensemble coding. I began my investigation of the conflict between integration and segmentation there.

# EXPERIMENT 1

## *Method*

***Observers.*** Thirty-four students from the University of Denver participated in Experiment 1. Due to an oversight, demographic information was not collected for Experiment 1. Observers granted informed consent and had normal or corrected-to-normal visual acuity. This sample size was selected based on a previous investigation with a related design, number of trials, and analysis, which had sufficient power to detect and replicate an ensemble-coding effect using different stimuli with an approximately equal number (thirty) of observers (Elias, Dyer & Sweeny, 2017).

***Stimuli.*** The stimulus set included twenty-one ellipses (0.2° thick lines) created in Adobe Photoshop CS6 v. 13.0 x64, each rendered in dark gray (mean luminance: 19 cd/m$^2$). The aspect ratios were symmetrically distributed (in log scale) around the category boundary aspect ratio (i.e., a circle). Flat ellipses included the following log aspect ratios:  −0.419, −0.374, −0.343, −0.311, −0.285, −0.221, −0.176, −0.131, −0.087, −0.043. Tall ellipses included the following aspect ratios: 0.043, 0.087, 0.131, 0.176, 0.221, 0.285, 0.311, 0.343, 0.374, and 0.419. An even (circular) aspect ratio of 0.00 was also included. The edges of each ellipse were blurred in Adobe Photoshop using the Gaussian blur tool with a 2-pixel radius.

Experiments were conducted on a CRT monitor with a refresh rate of 100 Hz at a viewing distance of 55 cm. Stimuli were presented against a uniform gray background

(RGB value = 170, 170, 170; luminance = 41.5 cd/m$^2$). Experiments were coded and run using MATLAB (Version 2014b; The MathWorks, Natick, MA) with the Psychophysics Toolbox (Brainard, 1997).

***Procedure.*** Observers were individually run in a dimly lit room. The experiment consisted of 240 multi-ellipse and single-ellipse trials, counterbalanced. Multi-ellipse trials featured the presentation of either four ellipses arranged in a globally-shaped square around a central fixation point (*four-ellipse trials),* or eight ellipses arranged in a larger diamond shape around fixation (*eight-ellipse trials;* Figure 1.1). Note that the locations of the central ellipses in the eight-ellipse array were the very same locations as those in the four-ellipse array. Thus, both four- and eight-ellipse sets contained a global shape with a null aspect ratio, which is important given that global and local shape perception can interact (e.g., Navon, 1977; Badcock, Whitworth, Badcock, & Lovegrove, 1990). These two different types of trials were included to examine how integration might interact with set size. The centroids of adjacent ellipses were 5.9° away from each other along the horizontal axis and 5.9° away from each other along the vertical axis (Figure 1.1). Set size was, of course, confounded with eccentricity (the larger sets necessarily included more peripherally presented ellipses), although this was necessary in order to equate inter-shape distance in the sets, which was important for preventing visual crowding and maintaining a global square shape.

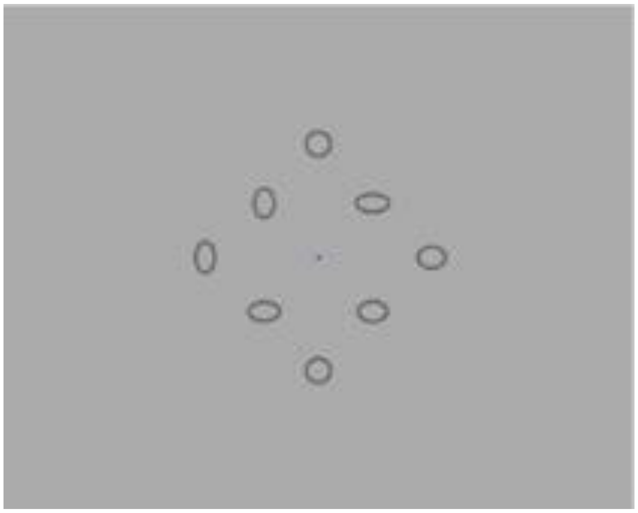

**Figure 1.1.** Layout of an eight-ellipse trial with fixation at center. Note that four-ellipse trials included ellipses only at the four most-central locations.

The primary purpose of this experiment was to investigate how integration of aspect ratio differed as a function of a set's relationship to the category boundary (in the case of aspect ratio, a circle). Multi-ellipse trials thus contained sub-conditions including *flat, tall, center, outlier* or *full-range* conditions (Figure 1.2). On flat, tall and center trials, ellipses were drawn from a limited range of 11-ellipses from within the full set of twenty-one ellipses. On flat trials, all ellipses were drawn from the flat range of the stimulus set; sets from these trials never contained a tall ellipse, and thus the distribution of ellipses never crossed the category boundary (in log units, ellipses from flat trials could have had any of the following eleven aspect ratios: −0.419, −0.374, −0.343, −0.311, −0.285, −0.221, −0.176, −0.131, −0.087, −0.043, and 0.00 (circle)). Note that flat and tall trials could both contain circles, which ensured that the presence of circles was not unique to any condition. Each flat trial was further randomly determined to be either a *low-seed flat trial* or a *high-seed flat trial* (Figure 1.2). On low-seed flat trials, one ellipse (on four-ellipse trials) or two ellipses (on eight-ellipse trials) were randomly selected from the three flattest ellipses (i.e., aspect ratios −0.419, −0.374, and −0.343). The remaining three (on four-ellipse trials) or six (eight-ellipse trials) were randomly selected from the entire flat-ellipse range. On high-seed ellipse trials, one ellipse (on four-ellipse trials) or two (on eight-ellipse trials) was randomly selected from the three ellipses with the least-flat aspect ratio (aspect ratios −.087, −0.043, and 0.00). The remaining ellipses were selected from the entire flat-ellipse range. The rationale for this seeding procedure is explained later. Ellipses from tall trials could have had any of the following eleven aspect ratios: 0.00, 0.043, 0.087, 0.131, 0.176, 0.221, 0.285, 0.311, 0.343, 0.374, and 0.419. Tall trials were constructed similarly to flat trials: low-seed tall trials had one ellipse (two on eight-ellipse trials) randomly selected from the three least-tall ellipses (aspect ratios 0.00, 0.043, and 0.087), while high-seed tall trials had one ellipse (two on eight-ellipse trials) randomly selected from the three tallest ellipses (aspect ratios 0.343, 0.374, and 0.419). Finally, center trials had the same structure, except that they contained both flat *and* tall ellipses, with the category boundary (i.e., circle) in the center of the range from which the ellipses were drawn. Center trials thus included ellipses from the eleven center ellipses in the stimulus range (aspect ratios: −0.221, −0.176, −0.131, −0.087, −0.043, 0.00, 0.043, 0.087, 0.131, 0.176, and 0.221). Low-seed center trials had one ellipse (two on eight-ellipse trials) randomly selected from the flattest three ellipses in this center range (aspect ratios −0.221, −0.176, and −0.131), while high-seed center trials had one ellipse (two on eight-ellipse trials) randomly selected from the tallest three ellipses in this range (aspect ratios 0.131, 0.176, and 0.221).

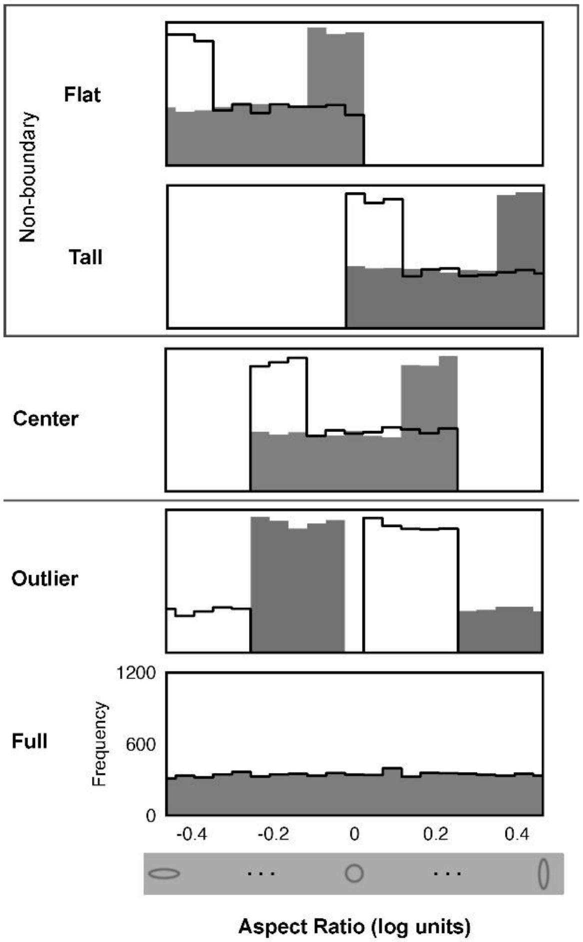

**Figure 1.2.** Multi-ellipse conditions from Experiments 1 and 2. Data and stimuli are from Experiment 2 (eight-ellipse trials), though the basic structure of conditions is the same across both experiments. Histograms denote frequency with which each ellipse from the stimulus range appeared across all flat trials and tall trials (which together formed non-boundary trials), as well as all center trials—the three main conditions of interest. Outlier and full-range trials are also depicted. Open histograms represent frequency data for low-seed trials, while filled gray histograms represent high-seed trials.

Utilizing low- and high-seeds ensured that the distribution of aspect ratios in each set was always skewed. This ensured that if observers simply guessed from the middle of the tall, flat or center ranges on a trial-by-trial basis, their responses would not default to the actual mean of the set and artificially be mistaken for true, perceptual extraction of the mean. Additionally, by employing low- and high-seeds, I could, in theory, determine whether observers were extracting the mean or the median, across trials of any given type, at the data-analysis stage. I verified that low- and high-seed distributions produced dissociated means and medians before beginning the investigation by simulating 100 four-ellipse, low-seed flat trials and computed the average simulated mean and median for each. I then iterated this 100-trial simulation 500 hundred times. Across these iterations, four-ellipse flat-seed trials contained a mean log aspect ratio (*M* = −0.2573, *SD* = 0.006) that was significantly different from the median log aspect ratio (*M* = −0.2691, *SD* = 0.008) t(499) = 79.16, *p* < .01, *d* = 1.67. The means and medians of eight-ellipse flat trials, as well as those of low- and high-seed tall and center trials (all run in separate simulations), differed in a similar way.

In addition to flat, tall and center trials, multi-ellipse trials could also include *outlier* or *full-range* trials. Outlier trials could be either low-seed or high-seed. On four-ellipse low-seed outlier trials, one ellipse (two on eight-ellipse trials) was randomly selected from the five flattest aspect ratios; this ellipse was the “outlier”. The remaining three (or six, on eight-ellipse trials) ellipses were randomly selected from the tall ellipses – specifically, from aspect ratios 0.043, 0.087, 0.131, 0.176, and 0.221. Similarly, on high-seed outlier trials, one ellipse (two on eight-ellipse trials) – the “outlier” – was randomly selected from the tallest five ellipses. The remaining ellipses were randomly selected from the flat range, specifically from aspect ratios −0.221, −0.176, −0.131, −0.087, and −0.043. This approach produced trials in which the majority of ellipses were generally flat, while the outlier(s) was tall, and vice-versa. Finally, on full-range trials, all ellipses (four or eight) were randomly selected from the entire twenty-one ellipse range of the stimulus set (Figure 1.2).

I reasoned that pooling would produce less precise estimates of mean aspect ratio on outlier trials than on full-range trials. First, outliers tend to be discounted during pooling of color (Michael, De Gardelle, & Summerfield, 2014) and facial expression (Haberman & Whitney, 2010). If outliers are also excluded from summary judgments of aspect ratio, then judgments on trials with outliers should be less close to the mean of their sets than those without outliers. Second, because perception of extreme aspect ratios tends to be exaggerated away from the null-point (Suzuki & Cavanagh, 1998), the consistent presence of a *categorical* outlier might heighten the conflict between segmentation and integration, and thus further disrupt integration. Full-range trials, in contrast, were less likely to have a minority of ellipses that were categorically different from the rest. Thus, I reasoned that, despite the presence of boundary-crossing in both trial types, outlier trials would show reduced evidence of pooling, compared to full-range trials.

Importantly, because the precision of ensemble integration is known to decrease as the heterogeneity of a set increases, (e.g., Dakin, 2001; Morgan, Chubb & Solomon, 2008; Marchant, et al., 2013; Im & Halberda, 2013; Haberman, Lee & Whitney, 2015), I wanted to ensure that set heterogeneity was comparable across outlier and full-range trials. I confirmed this by running a 100-trial simulation with outlier and full-range trials, iterated 500 times. Across iterations, with these design parameters, four-ellipse outlier trials had slightly *less* heterogeneity (*mean SD* = 0.2443 log units, *SD* = 0.002) than four ellipse full-range trials (*mean SD* = 0.2479, *SD* = 0.005), *t*(499) = 8.94, *p* < .01, *d* = 0.95. This simulation yielded the same pattern for eight-ellipse trials. Thus, if outlier trials did show evidence of reduced pooling, compared to full-range trials, it could not be because they simply had more heterogeneity.

Apart from multi-ellipses trials, I also included *single-ellipse* trials. On single-ellipse trials, one ellipse was randomly selected from the full range, and was displayed at a random location; observers could only base their response on this single ellipse. However, I also randomly selected three (or seven) additional ellipses from the full range, as if I were generating a full-range multi ellipse trial to display. Importantly though, these additional "invisible" ellipses were *not* displayed. Although these additional ellipses were invisible, a group mean was nonetheless calculated, and observer error relative to this group mean could be computed. Thus, single-ellipse trials were crucial. Observer error on these trials, relative to the mean of the entire set (which they could not see), allowed me to quantify the magnitude of error one would expect if observers simply responded to one random ellipse, without integrating aspect ratio information, on *true* multi-ellipse full-range trials. If observers did integrate information from multiple visible ellipses on true multi-ellipse trials, their estimates should approach that trial's true average, since in those cases multiple visible ellipses were available for integration. Thus, convincing evidence of integration (i.e., ensemble coding) and mean extraction would be present if observer error on true full-range multi-ellipse trials was reduced, compared to error on single-ellipse trials. Less centrally, single-ellipse trials also served as a measure of sensitivity to peripherally-viewed aspect ratio when error was computed relative to the actual aspect ratio of the single visible ellipse.

For every observer, the experiment began with the central display of the following instructions: *Estimate the average shape. Maintain your gaze on fixation at all times. Move mouse L or R to adjust response. Spacebar to begin.* Each trial began with a central fixation point displayed for a random duration between 800 and 1200ms. Next, a four or eight multi-ellipse array, or a single-ellipse array, was displayed for 250ms. Aspect ratio information can be extracted at extremely brief durations (D’Abreu et al., 2017), and others have used similar display durations for sets of relatively simple static stimuli (e.g., Chong & Treisman, 2005; Oriet & Brand, 2013). This duration also prevented multiple fixations, and serial scanning of individual set members. Next, fixation was displayed again for 500ms, which prevented the upcoming response screen from being incorporated into the stimulus set. Finally, a response ellipse appeared in the center of the screen. The aspect ratio of the initial response ellipse was randomly selected from the full stimulus range. Observers reported their estimate of the set average (or the individual ellipse on single-ellipse trials) by moving the mouse left or right, which smoothly adjusted the response ellipse across the stimulus range. If, for example, the initial response ellipse happened to be a circle, moving the mouse leftward would increase the “flatness” of the response ellipse by animating across the stimulus set, in discrete steps, one ellipse at a time. After the “flattest” ellipse in the set was reached, the response ellipse would then begin to increase in “tallness”. Moving the mouse rightward had the opposite effect. If the observer continuously moved the mouse left or right, eventually, after a cycle that included all the ellipses in the stimulus set being displayed at least once, they would encounter an endpoint (i.e., a point at which further left or right movement did not further change the response ellipse). Observers could then move the mouse in the opposite direction to continue response adjustment. Importantly, the aspect ratio of these endpoints (if they were encountered at all) were randomized across trials – they did not systematically correspond to the actual endpoints of the stimulus set (i.e., they did not systematically correspond to the "flattest" and "tallest" ellipses). When observers reached their desired response, they simply clicked the mouse to finalize their choice. Finally, to prevent any effect of an afterimage from the response ellipse on the next trial, a backward mask composed of a scrambled circle from the stimulus set was displayed for 250 ms at the center of the screen, then the next trial began. Observers were allowed as many practice trials as they wished before responses were recorded.

### *Results*

My primary interest was whether observers were more sensitive to a set of ellipses' average aspect ratio when that set's members did not span the category boundary, compared to when they did. I began by computing the error of each observer's response relative to the mean aspect ratio the set, on a trial-by-trial basis. For example, if on one trial a set of ellipses had a null (0.00) aspect ratio on average, and the observer responded with aspect ratio 0.043 , their error on that trial would be +0.043. In this case, their response ellipse was too tall relative to the set mean. Negative error values indicated a response that was too flat. For each observer, I compiled these signed-difference scores into separate error distributions, one for each condition (flat, tall, center, outlier, full-range, center and single trials). Next, I calculated the standard deviation of each observer's error distributions, for each condition. Greater sensitivity to mean aspect ratio was expected to produce error distributions with smaller standard deviations. This approach has been used in previous investigations of ensemble coding (e.g., Haberman & Whitney, 2009; Sweeny, Haroz, & Whitney, 2013; Sweeny & Whitney, 2014; Elias et al., 2017). This analysis yielded overall error scores (i.e., the SD of an observer's error distribution) for each condition.

***Main Results.*** A repeated measures 6 (trial type: flat, tall, center, outlier, full-range, single trials) x 2 (set size: four-ellipse, eight-ellipses) analysis of variance (ANOVA) revealed main effects of trial type, $F(5, 29) = 85.44$, $p < .01$, $\eta_p^2 = 0.94$, and set size, $F(1, 33) = 130.61$, $p < .01$, $\eta_p^2 = 0.8$. The interaction between trial type and set size was significant, $F(5, 29) = 6.82$, $p < .01$, $\eta_p^2 = 0.54$. The main indicator of ensemble coding was the comparison between full-range trials and single trials. Crucially, estimates of average aspect ratio were more precise on full-range trials (*mean SD* = .21, *SD* = .05) than they were on single-ellipse trials (*mean SD* = .24, *SD* = .04), when eight ellipses were displayed, $t(33) = 3.48$, $p < .01$, $d = 0.6$. However, this comparison was not significant for four-ellipse trials, $t(33) = .98$, *n.s.*, although it trended in the right direction. Thus, although performance was better on full-range trials in both conditions, the best evidence for ensemble coding of aspect ratio was present on eight-ellipse trials.

Performance on flat and tall trials did not differ, on either four-ellipse, $t(33) =0.09$, *n.s*, or eight-ellipse trials $t(33) = 1.24$, *n.s*. Since I had no a-priori hypotheses regarding performance on flat trials versus tall trials, and the category boundary was not crossed in either condition, I combined both flat and tall trials into one error distribution, for each observer. I refer to these trials simply as *non-boundary-spanning*, or simply *non-boundary* trials (for a schematic representation of non-boundary trials, see Figure 1.2). To do so, I calculated performance across flat and tall trials, separately and for each observer. I then took the average of these two values, per observer, to yield a measure of each observer's performance on non-boundary trials. This approach was superior to collapsing data from these conditions into one, super distribution. This alternative approach could have, hypothetically, allowed for two narrow distributions with means biased away from zero, in opposite directions, to produce a super-distribution with an inflated SD, which I obviously did not want. Critically, observers performed better on non-boundary trials than they did on center trials, when both four (*non-boundary mean SD* = .11, *SD* = .03; *center mean SD* = .15, *SD* = .05) $t(33) = 5.49$, $p < .01$, $d = 0.94$, and eight ellipses (*non-boundary mean SD* = .12, *SD* = .03; *center mean SD* = .17, *SD* = .04) $t(33) = 5.26$, $p < .01$, $d = 0.9$, were displayed. Thus, regardless of whether observers were employing ensemble coding (as is likely for eight-ellipse trials), or leveraged some other method (as is possible for four-ellipse trials), sensitivity to a set's average aspect ratio was greatest when ellipses did not span aspect ratio's category boundary. These main results are summarized in Figure 1.3. Since the strongest evidence for ensemble coding occurred on eight-ellipse trials, all following analyses were carried out on those trials only.

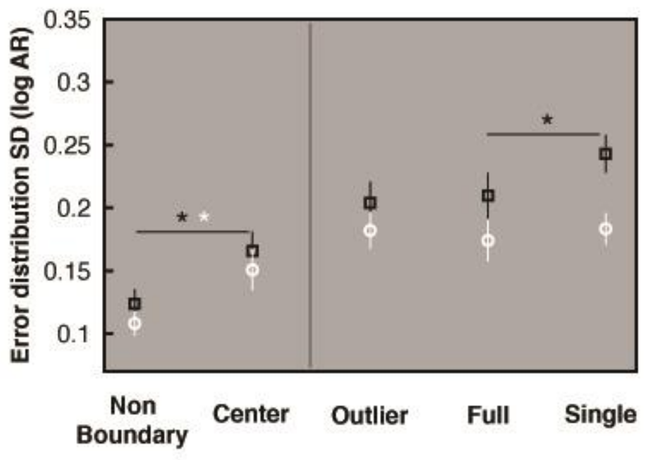

**Figure 1.3.** On both four (white data points)- and eight-ellipse (black data points) trials, observers were more sensitive to the average aspect ratio in sets that included only "flat" or "tall" ellipses (i.e., non-boundary trials), compared to when sets included "flat" and "tall" ellipses (i.e., center trials). The comparison between single and full-range trials—the main indicator of ensemble coding—was significant for sets of eight ellipses. Error bars in all figures represent 95% confidence intervals, all starred comparisons p < .01.

***Secondary Results.*** Of secondary interest was the comparison between outlier trials and full-range trials. Observer performance was comparable for both outlier (*mean SD* = .20, *SD* = .05) and full-range trials (*mean SD* = .21, *SD* = .05) *t*(33) = .85, *n.s* (Figure 1.3). Interestingly, observer performance was comparable for outlier and full-range trials even though outlier trials did contain less heterogeneity (*mean SD* = .24, *SD* = .03) than full-range trials (*mean SD* = .25, *SD* = .07), as pre-experimental simulations predicted they would. It is possible that the reduced heterogeneity in outlier trials did not compensate for the disruptive effects of the presence of ellipses that spanned the category boundary, although only two ellipses were categorically different from the rest, on eight-ellipse trials.

Irrespective of these secondary comparisons, observers were capable of integrating aspect ratio information, but were especially sensitive to the mean of sets that did not cross the category boundary. This is the main result of Experiment 1.

***Addressing alternative explanations.*** Next, I addressed three alternative explanations of Experiment 1's primary result. First, I considered response compression. Then, I investigated whether observers were extracting the mean of the sets, the median of the sets, simply averaging the two most extreme ellipses present in a given set, or simply responding from the middle of the relevant range on a trial-by-trial basis. Finally, I considered the relationship between set variability (i.e., heterogeneity) and response error magnitude.

I began by considering whether response compression could account for the observed results. The stimulus set for Experiment 1 contained "endpoints". That is, there was a "flattest" and a "tallest" ellipse. These endpoints were present both in the sets of ellipses and in the response stage. If observers avoided responding with these extreme ellipses, then distributions of responses would have been compressed away from endpoints, toward the center of the stimulus range. Response compression like this would narrow error distributions and thus lower overall estimates of error. Response compression could be particularly relevant for trials in which the mean of the set was especially "flat" or "tall"—the very trials that did indeed display reduced observer error.

Although steps were taken to reduce the impact of endpoints and potential response compression at the response stage (see *Procedure)*, I conducted further analyses to determine if response compression was responsible for the above results. Consider flat trials with a low seed: these trials had, on average, the “flattest” mean aspect ratio. In contrast, tall trials with a high seed had, on average, the “tallest” mean aspect ratio. Thus, endpoints were most relevant on flat low-seed and tall hi-seed trials, since the means of those trials were nearest the stimulus set’s endpoints; response compression was especially likely on these trials. In contrast, endpoints were less relevant for flat trials with a *high* seed and tall trials with a *low* seed since the means of these trials were nearer the category boundary (and “farther” from the potential sources of response compression). I calculated the SD of each observer’s error distributions for flat trials with a high seed, and separately, tall trials with a low seed. I then averaged these two values for each observer to yield a value that represented performance across flat and tall trials in which response compression would be relatively *less* relevant. I also calculated performance on flat trials with a low seed and tall trials with a high seed. Response compression, if present, would be relatively *more* relevant on these trials. Observer performance on trials in which response compression would be more relevant (*mean SD* = .11, *SD* = .04) was better than on trials in which it would be less relevant (*mean SD* = .13, *SD* = .04) $t(33) = 2.33$, $p = .03$, $d = 0.4$. Thus, response compression may have impacted our results. Importantly, however, even when I considered only trials for which response compression would be less relevant (*mean SD* = .13, *SD* = .04), performance was better on trials that did not span the category boundary than on center trials that did (*mean SD* = .17, *SD* = .04) $t(33) = 4.75$, $p < .01$, $d = 0.82$. This suggests that response compression cannot fully account for the main pattern of results presented above.

Next, I investigated whether observers were indeed extracting the mean of sets, the median of sets, or were simply responding by selecting the midpoint of the appropriate stimulus range on a trial-by-trial basis (e.g., responding with the middle-flat option for all flat trials). In addition to mean and median extraction though, another form of integration was possible: observers may have selectively averaged the two most extreme aspect ratios present in the set. For example, if a trial contained six circles, one slightly flat ellipse and one very tall ellipse, the observer may have simply averaged the slightly flat and very tall ellipse, and ignored the intermediaries. To disentangle these three possibilities (mean or median extraction, and averaging-of-extremes), I first isolated low-seed trials for flat ellipses. Across these trials, I calculated each observer's error distribution using error computed relative to the veridical trial mean. I then did the same for low-seed center and tall trials. I isolated high-seed error distributions from flat, center and tall trials in the same manner. The average of these six values provided a single measure of each observer's sensitivity across my primary three trial types, relative to the mean. I repeated this entire process for error relative to each trial's *median* and relative to the average of each trial's flattest and tallest ellipse. As a result, I obtained three final values: average sensitivity across my main three trial types relative to the mean, median, and average-of-extremes. Observer performance relative to the mean (*M* = .13, *SD* = .03) was better than performance relative to the median (*M* = .14, *SD* = .03) *t*(33) = 22.31, *p* < .01, *d* = 3.83. However, it was worse than performance relative to the average-of-extremes (*M* = .11, *SD* = .02) *t*(33) = 3.84, *p* < .01, *d* = 0.66. Thus, it is possible that observers were simply averaging the two most extreme ellipses on a given flat, tall or center trial.

To further explore this possibility, I combined data from my primary three trial types: flat, tall and center trials, and I disregarded seed type. For each observer, I then computed a correlation coefficient between that observer's responses and the true means of all their trials. This yielded a value representing the strength of the relationship between observer response and the true mean of a trial, across my main three trial types. I repeated this process for these trials' medians and average-of-extremes. Across observers, the correlation between response and the true mean (*mean R* = .79, *SD* = .09) was stronger than the correlation between response and trial median (*mean R* = .77, *SD* = .09) *t*(33) = 3.47, *p* = .04, *d* = 0.59, or trial average-of-extremes (*mean R* = .76, *SD* = .09) *t*(33) = 4.8, *p* = .02, *d* = 0.82. Thus, the picture is somewhat murky. On the one hand, when using error distributions as the outcome, observer sensitivity to summary information was highest when measured relative to the average of a set's two most extreme ellipses, and slightly reduced relative to a set's mean. However, the relationship between observers' responses and a set's true mean was stronger than that between observer response and a set's true median or average-of-extremes. There are several reasons to doubt that observers were only averaging the two most extreme aspect ratios, which I will discuss shortly.

Next, I investigated whether observers were not actually integrating information, but were instead simply employing a response strategy. On flat trials, for example, observers may have simply selected a response from the middle of the flat range. They may have done likewise for center and tall trials, as well. To determine if observers employed this response strategy, I again sorted each observer's data by trial type (isolating flat, center and tall trials), as well as by seed (low-seed and high-seed). This resulted in six trial types (*trial type*: flat, center, and tall X *seed*: low and high). This time, however, I computed each observer's *average chosen aspect ratio*, in each of the six trial types (rather than error). Note that the center of the range for both flat low-seed and flat high-seed trials is identical (this is also true of center and tall trials). If observers were simply picking from the center of the appropriate range on a trial-by trial basis, their responses would not depend on that trial's seed. If observers were truly integrating information, however, their responses should vary, depending on the presence of a low or high seed.

A repeated measures 3 (trial type: flat, center, tall) × 2 (seed: low, high) analysis of variance (ANOVA) revealed main effects of trial type, $F(2, 32) = 460.43$, $p < .01$, $\eta_p^2 = 0.97$, and seed, $F(1, 33) = 119.23$, $\eta_p^2 < .01$, $d = 0.78$. The interaction between trial type and seed was significant, $F(2, 32) = 7.52$, $p < .01$, $\eta_p^2 = 0.32$. Planned comparisons revealed that observers chose a flatter aspect ratio (AR) on flat low-seed trials (*mean AR* = -.26, *SD* = .06) than on flat hi-seed trials (*mean AR* = -.19, *SD* = .07) $t(33) = 5.7$, $p = .02$, $d = 0.98$. This pattern persisted for center low-seed trials (*mean AR* = -.1, *SD* = .06) and center high-seed trials (*mean AR* = .03, *SD* = .07) $t(33) = 9.22$, $p < .01$, $d = 1.58$. Center low-seed trials yielded a particularly flat aspect ratio; the likely source of the interaction effect. Nonetheless, the pattern persisted for tall low-seed trials (*mean AR* = .15, *SD* = .06) and tall high-seed trials (*mean AR* = .25, *SD* = .08) $t(33) = 7.0$, $p < .01$, $d =$

1.2 as well (Figure 1.4)**.** In other words, observer response in all of my main trial types depended on whether a trial was low- or high-seed. This is strong evidence against a simple center-of-the-range response strategy.

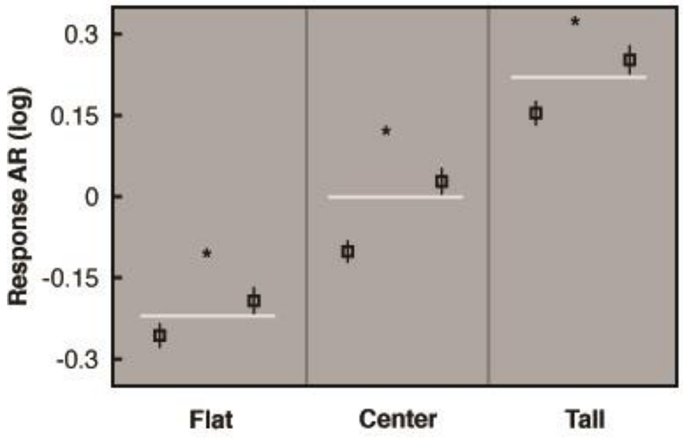

**Figure 1.4.** Across all three main conditions, low-seed trials (left data point in each pair) were estimated to have a flatter mean than high-seed trials (right data point in each pair). White horizontal lines represent the midpoint of the flat, center and tall trial ranges. If observers were simply selecting from the midpoint of flat, center and tall ranges, each pair of data points would align with the horizontal white line. Instead, observers were sensitive low and high seeds, indicating they were not simply choosing from the midpoint of the relevant range, on a trial-by-trial basis.

Finally, I investigated the relationship between set heterogeneity (i.e., how much aspect ratio variation was present in a set) and observer error. In general, I expected increased set heterogeneity to predict increased error. Crucially, there was less heterogeneity in non-boundary trials (*mean SD* = .14, *SD* = .004) than in center trials (*mean SD* = .15, *SD* = .006) t(33) = 6.02, *p* < .01, *d* = 1.03. Although increased set variability is known to impede integration, it was important to investigate whether this subtle difference in heterogeneity between trial types could account for superior observer mean sensitivity on non-boundary trials. In order to do so, I modeled the relationship between heterogeneity and observer performance, and then used that model to predict the amount of error I might expect if heterogeneity fully accounted for observer performance. First, I quantified set heterogeneity in a manner similar to other ensemble coding research (e.g., Elias et al., 2016). On each trial, I simply took the SD of all the aspect ratios present in that set. Next, I computed the absolute error magnitude between an observer's response and the true trial mean. For each observer, across all trials, I then computed the slope of the linear relationship between heterogeneity and the magnitude of observer error. These slopes were positive (*mean slope:* .36, *SD* = .31) t(33) = 6.78, *p* < .01, *d* = 1.16. Finally, I used this relationship to predict the amount that observer error magnitude would be expected to increase, given the actual difference in heterogeneity between non-boundary and center trials. The linear relationship between heterogeneity and observer error predicted an increase in error magnitude of .0027 log units given the 0.01 difference in heterogeneity between the two conditions. However, the observed difference in error magnitude between non-boundary (*mean log units* = .106, *SD* = .03) and center trials (*mean log units* = .146, *SD* = .04) was more than fifteen times greater than predicted. This approach was less-than-ideal, as it was done post-hoc and lacked an objective threshold for assessing heterogeneity's ability to account for the main pattern of results described above. Still, the approach does suggest that variation in heterogeneity between trial types cannot fully account for observers' increased sensitivity to the mean on non-boundary trials, relative to center trials.

### *Experiment 1 Discussion*

Experiment 1 produced several noteworthy results. First, ensemble coding—information integration—can indeed operate on aspect ratio. Secondly, integration operates best when it does not have to contend with a category boundary (i.e. when all ellipses in the group are "flat" *or* "tall", and the group does not span the category boundary). This is an important result, since no ensemble coding research has explored this potential constraint on the process of integration, at least to the best of my knowledge. Observers did not appear to rely on the median to make their summary judgments. Nor did observers simply pick from the midpoint of the relevant range, on a trial-by-trial basis. Notably, observer performance on full-range trials was relatively imprecise. This result is hard to explain if observers were simply averaging the two most extreme ellipses in the set. After all, on any given full-range trial, extreme ellipses (both flat and tall) were likely to be present. Averaging the flattest and tallest ellipse should be especially *easy* on such trials, given that extreme ellipses tend to stand-out (Treisman & Gormican, 1988). Instead, observers performed worse on full-range trials than any other multi-ellipse condition. Additionally, across many full-range trials, the average of the two most extreme ellipses is simply the category boundary circle (indeed, across all eight-ellipse full-range trials, the average of the two most extreme aspect ratios was -.0003). However, observers seemed to show no preference for the category boundary on full-range trials (Figure 1.5)**.** Strictly speaking, even if observers really were averaging the two most extreme aspect ratios from a given set, this would constitute ensemble coding

(Whitney & Leib, 2018). Still, on balance, the evidence suggests that observers were likely sensitive to the mean of the set, not the average of the two most extreme ellipses.

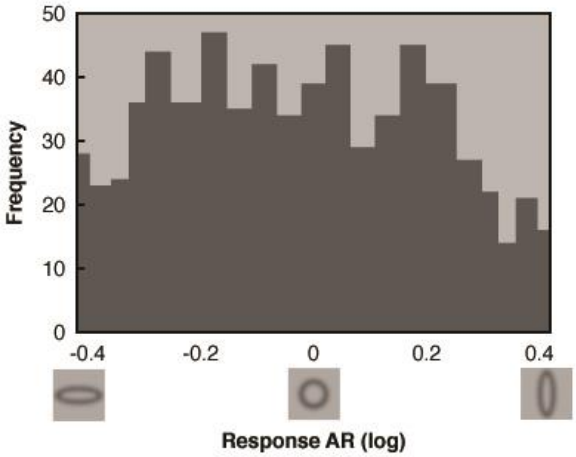

**Figure 1.5.** Across all eight-ellipse full-range trials, the average of the two most extreme aspect ratios present was a circle. However, observers in Experiment 1 did not seem to show a preference for responding with a circle on those trials. This is one reason it is unlikely that observers were simply averaging the two most extreme ellipses present in a given set.

It may be that the presence of a category boundary introduces a conflicting computational strategy in the visual system—segmentation. Segmentation, then, may be responsible for impeding integration, when a visual category boundary is present in a set. Before I investigated this possibility, though, I conducted an additional experiment to address a few limitations from Experiment 1.

# EXPERIMENT 2

In Experiment 1, the endpoints of the stimulus set (i.e., the “flattest” and “tallest” ellipses) were present in the stimuli arrays as well as in the response stage. Even though the response compression that likely that resulted from this issue probably did not meaningfully impact the main findings, it was still less-than-ideal. In Experiment 2, the response range again included all the ellipses that could be present on a given trial. Additionally though, three extremely flat and three extremely tall ellipses, ellipses that were never present on any trial, were included in the response ellipse range. By effectively “extending” the endpoints of the stimulus set at the response stage only, the potential for response compression should be reduced.

Additionally, in Experiment 1, the increase (or decrease) in aspect ratio between each ellipse in the stimulus set was not perfectly equivalent. For example, the difference in aspect ratio (in log units) between the flattest ellipse and the second-flattest ellipse was close to, but not the same as, the difference between the fourth-flattest and fifth-flattest ellipses. This subtle discrepancy accounted for the difference in heterogeneity between non-boundary and center trials, even though both had a range of 11 ellipses. Although the analyses above suggest this discrepancy cannot fully account for the main results, I wanted to address the limitation in a more direct way. Finally, the ellipses in Experiment 1 were not perfectly equated in terms of area. Integration of size certainly occurs (e.g., Allik et al., 2014), and if trial types systematically differed in terms of average size, it would be difficult to attribute observer sensitivity to *aspect ratio* integration, as opposed to *size* integration, with certainty. Experiment 2 addressed both of these issues. In Experiment 2, I used a new stimulus set for which increases (and decreases) in aspect ratio were equated between each step in the stimulus set, and all ellipses were precisely equated in terms of area.

### ***Method***

***Observers.*** Forty-five students (*mean age* = 19.2 years; 89% female) from the University of Denver participated in Experiment 2. Observers granted informed consent and had normal or corrected-to-normal visual acuity. I conducted a power analysis for the test of the presence of ensemble coding in Experiment 1 (i.e., full-range versus single-ellipse trials). Assuming the same large effect size (*d* = .6), I determined that a sample of thirty-three would be necessary to obtain power of 0.8. I anticipated that some of the modifications made to Experiment 2 would result in a smaller effect size; I therefore increased the sample size to forty-five.

***Stimuli.*** The new stimulus set included twenty-seven ellipses (0.2° thick lines) created in Adobe Photoshop CS6 v. 13.0 x64, each rendered in dark gray (luminance: 19 cd/m$^2$). The aspect ratios were symmetrically distributed (in log scale) around the category-boundary aspect ratio (i.e., circle). Flat ellipses present in set displays included the following aspect ratios: −0.463, −0.417, −0.371 −0.324, −0.278, −0.232, −0.185, −0.139, −0.093, and −0.046. Additionally, at the response stage only, three extremely flat ellipses (−.602, −.556, and −.510) were available as response options in addition to the rest of the flat ellipses. Tall ellipses present in set displays included the following aspect ratios: 0.046, 0.093, 0.139, 0.185, 0.232, 0.278, 0.324, 0.371, 0.417, and 0.463. Additionally, at the response stage, three extremely tall ellipses (.510, .556, and .602) were available as response options in addition to the rest of the tall ellipses. Note that the appearance of unequal changes in aspect ratio across the stimulus range in the lists above is due to rounding error. The incremental change between adjacent aspect ratios across the stimulus set was equated, in log units, past the tenth decimal. The areas of all ellipses were equated to the second decimal, and the edges of each ellipse were blurred in Adobe Photoshop using the Gaussian blur tool with a 2-pixel radius.

Experiments were conducted on a CRT monitor with a refresh rate of 100 Hz at a viewing distance of 55 cm. Stimuli were presented against a uniform gray background (RGB value = 170, 170, 170; luminance = 41.5 cd/m$^2$). Experiments were coded and run using MATLAB (Release 2014b; The MathWorks, Natick, MA) with the Psychophysics Toolbox (Brainard, 1997).

***Procedure.*** The procedure for Experiment 2 was nearly identical to that of Experiment 1. The experiment consisted of 240 multi-ellipse and single-ellipse trials, counterbalanced. Since evidence of ensemble coding was strongest for eight-ellipse trials in Experiment 1, all multi-ellipse trials in Experiment 2 consisted of eight ellipses arranged in a diamond shape around fixation (Figure 1.1). The centroids of adjacent ellipses were 5.9° away from each other along the horizontal axis and 5.9° away from each other along the vertical axis.

Before running the experiment, I confirmed that my design parameters should result in comparable set heterogeneity for non-boundary and center trials. I confirmed this by again running a 100-trial simulation of Experiment 2, iterated 500 times. This showed that, with the altered aspect ratios in Experiment 2, non-boundary trials contained comparable heterogeneity (*M* = 0.1439 log units, *SD* = 0.0022) to center trials (*M* = 0.1440, *SD* = 0.002), *t*(499) = .85, *n.s*.

The response stage was nearly identical to that of Experiment 1, but with one important difference. The aspect ratio of the initial response ellipse was randomly selected from a set of ellipses that included the full twenty-one ellipses that could be present in a given trial, plus the three extremely flat and three extremely tall ellipses described above (see *Stimuli)*. Observers then adjusted the aspect ratio of the response ellipse by moving the mouse left or right, which cycled the response ellipse across Experiment 2's extended stimulus range one ellipse at a time.

### *Results*

My primary interest was again whether observers were more sensitive to a set of ellipses' average aspect ratio when that set did not span aspect ratio's category boundary, compared to when a set did span aspect ratio's category boundary. I began by computing each observer's error in the same manner as Experiment 1. This yielded overall error scores (i.e., the SD of an observer's error distribution) for each observer, and for each condition. Notably, observer reaction times on non-boundary trials (*mean RT in seconds*

= 2.22, *SD* = .82) and reaction times on center trials (*mean RT in seconds* = 2.21, *SD* = .8) did not differ *t*(44) = .17, *n.s.*

***Main Results.*** A repeated measures one-way (trial type: flat, tall, center, outlier, full-range, single trials) analysis of variance (ANOVA) revealed a main effect of trial type, *F*(5, 220) = 94.81, *p* < .01, $\eta_p^2 = 0.68$. The main indicator of ensemble coding was the comparison between full-range trials and single trials. As in Experiment 1, estimates of the set's average aspect ratio were more precise on full-range trials (*mean SD* = .28, *SD* = .08) than they were on single-ellipse trials (*mean SD* = .3, *SD* = .04) *t*(44) = 2.78, *p* < .01, *d* = 0.42.

As in Experiment 1, I separately computed the SDs of each observer's error distributions on flat and tall trials. For each observer, the average of these two values yielded a measure of their performance on *non-boundary* trials. Critically, observers performed better on non-boundary trials than they did on *center* trials (*mean SD non-boundary trials* = .18, *SD* = .06; *mean SD center trials* = .21, *SD* = .06) *t*(44) = 4.64, *p* < .01, *d* = 0.69. Thus, Experiment 2's results mirrored those of Experiment 1; observers showed evidence of integration, especially for sets that did not span the category boundary. These main results are summarized in Figure 2.1.

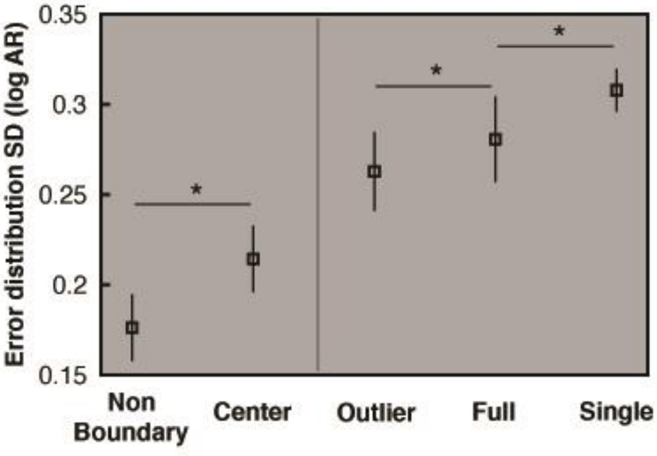

**Figure 2.1.** Again in Experiment 2, observers were more sensitive to the average aspect ratio of sets that included only "flat" or "tall" ellipses (i.e., non-boundary trials), compared to when sets included "flat" and "tall" ellipses (i.e., center trials).

***Secondary Results.*** Of secondary interest was the comparison between outlier trials and full-range trials. In contrast to Experiment 1, observer performance was better on outlier trials (*mean SD* = .26, *SD* = .07) than full-range trials (*mean SD* = .28, *SD* = .08) *t*(44) = 2.5, *p* = .02, *d* = .37. However, as intended, outlier trials did contain less heterogeneity (*mean SD* = .24, *SD* = .02) than full-range trials (*mean SD* = .27, *SD* = .05), which could account for some—or all—of this performance advantage. Indeed, I again examined the set heterogeneity present on a trial-by-trial basis against the absolute magnitude of observer error, across all trial types. Across observers, the relationship between set heterogeneity and observer error was positively correlated (*mean slope*: .67, *SD* = .02). I then compared the average amount of heterogeneity on outlier trials (*mean SD* = .24, *SD* = .02) with the average amount of heterogeneity on full-range trials (*mean SD* = .27, *SD* = .05). The relationship between heterogeneity and error magnitude predicted that, for this amount of increase in heterogeneity, observer error should increase by .02 log units. This was very close to the observed difference between observer error magnitude on outlier (*mean error log units* = .23, *SD* = .07) compared to full-range (*mean error log units* = .25, *SD* = .08) trials. Thus, performance on outlier and full-range trials likely did not differ once the effect of heterogeneity was taken into account.

I additionally wanted to know whether, on non-boundary trials, observers' errors were *systematic.* In other words, although observers were most sensitive to the mean of non-boundary-spanning sets, they were of course not error-free. Did those errors tend to exaggerate or alternatively underestimate the flatness or tallness of the set? To answer this question, I simply computed each observer's average signed error for flat trials and separately for tall trials. Across observers, mean estimates of flat trials exaggerated the mean flatness of the set (*mean error* = -.05, *SD* = .08) *t*(44) = 3.99, *p* < .01, *d* = .6. Likewise, across observers, mean estimates of tall trials exaggerated the mean tallness of the set (*mean error* = .05, *SD* = .08) *t*(44) = 4.25, *p* < .01, *d* = .63.

***Addressing alternative explanations.*** I again wanted to address three potential alternative explanations of my primary result. I began by considering whether response compression could account for the observed results, in the same manner as Experiment 1. Recall that endpoints were most relevant on flat low-seed and tall hi-seed trials, since the means of those trials were nearest the stimulus set's endpoints; response compression was especially likely on these trials.

To investigate this, I again calculated the SD of each observer's error distribution on flat trials with a high seed, and separately, their performance on tall trials with a low seed. I then averaged these two values for each observer, to yield a value that represented performance across trials in which response compression would be relatively *less* relevant. I then calculated the SD of each observers' error distribution on flat trials with a low seed and tall trials with a high seed. Across observers, performance on trials in which response compression would be more relevant (*mean SD* = .16, *SD* = .06) was better than on trials in which it would be less relevant (*mean SD* = .18, *SD* = .07) *t*(33) = 2.55, *p* = .02, *d* = 0.38. Importantly, however, even when I considered only trials in which response compression would be less relevant (*mean SD* = .18, *SD* = .07), observer performance was better on non-boundary trials than on center trials (*mean SD* = .21, *SD* = .06) *t*(33) = 4.09, *p* < .01, *d* = 0.61. This suggests that response compression cannot fully account for the main pattern of results presented above, consistent with Experiment 1. Thus, this result replicated even with the extended response range—and presumably attenuated response compression—present in Experiment 2.

Next, I again wanted to investigate whether observers were indeed extracting the mean of sets, the median of sets, averaging the two most extreme aspect ratios of each set, or were simply responding from the midpoint of appropriate stimulus range on a trial-by-trial basis. I began by investigating whether mean or median extraction occurred, or if observers simply averaged the two most extreme aspect ratios ("averaging-of-extremes"). To disentangle these possibilities, across observers, I calculated the average SD of error distributions relative to the veridical mean for flat, center and tall trials (exactly as in Experiment 1). I repeated this entire process for error relative to each trial's *median* and relative to the average of each trial's flattest and tallest ellipse. As a result, I obtained three final values: average SD of error distributions relative to the mean, median, and average-of-extremes. Observer performance relative to the mean (*M* = .18, *SD* = .06) was better than performance relative to the median (*M* = .19, *SD* = .05) *t*(33) = 2.6, *p* = .01, *d* =.38. Additionally, performance relative to the mean was better than relative to the average-of-extremes (*M* = .19, *SD* = .05) *t*(44) = 5.13, *p* < .01, *d* = 0.76. In addition to these comparisons, I combined data from my primary three trial types: flat, tall and center trials, and I disregarded seed labels. For each observer, I then computed a correlation coefficient between that observer's responses and the true means of these trials. This yielded a value representing the strength of the relationship between observer response and the true mean of a trial, across my main three trial types. I repeated this process for these trials' medians and average-of-extremes. Across observers, the correlation between response and the true mean (*mean R* = .76, *SD* = .16) was stronger than the correlation between response and trial median (*mean R* = .75, *SD* = .16) *t*(44) = 5.19, *p* < .01, *d* = 0.77, and average-of-extremes (*mean R* = .74, *SD* = .16) *t*(44) = 10.37, *p* < .01, *d* = 1.55. Thus, the picture in Experiment 2 was clearer than from the picture in Experiment 1. Observers were most sensitive to the mean aspect ratio of a set, and were likely extracting mean, as opposed to median, values.

Next, I again wanted to investigate whether observers were not actually integrating information, but were instead simply responding from the midpoint of the relevant stimulus range. I again sorted each observer's data by trial type (isolating flat, center and tall trials), as well as by seed (low-seed and high-seed). This resulted in six trial types (flat, center, and tall lo-seed trials, and flat, center, and tall high-seed trials). I recorded each observer's *average chosen aspect ratio*, in each of the six trial types. If observers were simply picking from the center of the appropriate range on a trial-by trial basis, their responses would not depend on that trial's seed. If observers were truly integrating information, however, their responses should vary, depending on the presence of a low or high seed. A repeated measures 3 (trial type: flat, center, tall) × 2 (seed: low, high) analysis of variance (ANOVA) revealed main effects of trial type, $F(2, 43) = 581.62$, $p < .01$, $\eta_p^2 = 0.96$, and seed, $F(1, 44) = 215.77$, $p < .01$, $\eta_p^2 = 0.83$. The interaction between trial type and seed was not significant, $F(2, 43) = 3.02$, $p = .054$, $\eta_p^2 = 0.12$. Planned comparisons revealed that observers chose a flatter aspect ratio (AR) on flat low-seed trials (*mean AR* = -.34, *SD* = .09) than on flat hi-seed trials (*mean AR* = -.22, *SD* = .09) $t(44) = 11.6$, $p < .01$, $d = 1.73$. This pattern persisted for center low-seed trials (*mean AR* = -.06, *SD* = .06) and center high-seed trials (*mean AR* = .03, *SD* = .05) $t(44) = 7.39$, $p < .01$, $d = 1.1$, as well as tall low-seed trials (*mean AR* = .22, *SD* = .07) and tall high-seed trials (*mean AR* = .34, *SD* = .09) $t(44) = 10.98$, $p < .01$, $d = 1.64$**.**

Finally, I again investigated the relationship between set heterogeneity (i.e., how much aspect ratio variation was present in a set) and observer error. Recall that prior to Experiment 2, simulations using Experiment 2's parameters predicted a comparable amount of heterogeneity for non-boundary and center trials. There was also no reason to expect a difference in heterogeneity, since aspect ratios increased/decreased linearly across the stimulus range, and flat, tall and center trials all spanned an equal number of aspect ratios. Nonetheless, perhaps due to random chance or rounding error resulting from representing aspect ratios to three decimals, there was less heterogeneity in non-boundary trials (*mean SD* = .151, *SD* = .003) than in center trials (*mean SD* = .153, *SD* = .005) *t*(44) = 2.55, *p* = .02, *d* = .38. Thus, heterogeneity was not perfectly controlled for in Experiment 2, despite careful efforts to do so. In order to investigate whether heterogeneity could completely account for the pattern of results described above, I again modeled the relationship between heterogeneity and observer error. As in Experiment 1, on each trial, I simply took the SD of all the aspect ratios present in that set. Next, I computed the absolute error magnitude, in log units, between an observer's response and the true trial mean. For each observer, across all trials, I then computed the relationship between heterogeneity and the magnitude of observer error. This linear relationship was positive (*mean slope:* .67, *SD* = .02) *t*(44) = 241.14, *p* < .01, *d* = 35.95. Finally, I used this relationship to predict the amount that observer error magnitude would be expected to increase, given the difference in heterogeneity between non-boundary and center trials. The linear relationship between heterogeneity and observer error predicted an increase in error magnitude of .0013 log units. However, the observed difference in error magnitude between non-boundary (*mean log units* = .16, *SD* = .05) and center trials (*mean log units* = .19, *SD* = .05) was more than eighteen times greater than predicted. Although this approach was again not ideal (for same reasons it was not ideal in Experiment 1), it does suggest that variation in heterogeneity between trial types thus cannot fully account for observer's increased sensitivity to the mean on non-boundary trials, relative to center trials, in Experiment 1 or 2.

### *Experiment 2 Discussion*

Experiment 2 replicated the main findings of Experiment 1. Observers were again more sensitive to the mean aspect ratio of sets that did not span the category boundary compared to sets that did. Experiment 2 addressed multiple methodological limitations from Experiment 1. Namely, in Experiment 2 all aspect ratios in the stimulus set increased/decreased in aspect ratio linearly, and were equated for area. Additionally, the potential for response compression was addressed by extending the response range. Still, the main results held.

Observers in Experiment 2 seem to have been sensitive to the mean of sets of aspect ratios, as opposed to the median or the average-of-the-extremes. Observers again performed poorly on full-range trials; this is hard to explain if they were simply averaging the two most extreme ellipses. Additionally, Experiment 2 indicated both higher sensitivity (i.e. smaller error distribution SD's) relative to the means of sets, and a stronger correlation between observer response and the true means of sets, compared to set medians or average-of-the-extremes. Observers did not simply select a response from the center of the appropriate range. Experiment 2 again demonstrated this by showing that observers were sensitive to skew in the distributions of aspect ratio, which owed to the presence of high and low seeds—something that would not have occurred if observers were simply selecting from the midpoint of the relevant range, on a trial-by-trial basis.

# EXPERIMENT 3a

I have suggested already that segmentation (specifically, perceptual distortion) is a plausible mechanism underlying ensemble coding's reduced sensitivity when information in a set spans a category boundary. This perceptual distortion may have been particularly pronounced for aspect ratios near the category boundary, and less pronounced or absent for more extreme values (Suzuki & Cavanagh, 1998; Suzuki, 2005; Sweeny et al., 2012). Some circumstantial evidence for distortion away from the boundary was observed in Experiment 2—observers exaggerated the "flatness" of flat trials and the "tallness" of tall trials. Experiment 3a more directly tested the hypothesis that distortion away from the category boundary occurred during ellipse perception, particularly for aspect ratio values around the category boundary (Figure 3a.1).

On center trials, slightly flat ellipses were hypothesized to appear flatter and slightly tall ellipses to appear taller; the net effect across trials would be an increased range of perceived aspect ratio, and thus increased perceptual heterogeneity. In contrast, on a flat trial (for example), repulsion would push the perception of ellipses near the category boundary toward the flatter set mean, potentially *reducing* the amount of perceived heterogeneity in the set. Considering that heterogeneity is known to disrupt the integrative process of ensemble coding (e.g., Dakin, 2001; Morgan, Chubb & Solomon, 2008; Marchant, et al., 2013; Im & Halberda, 2014; Haberman, Lee & Whitney, 2015), this potential effect of perceptual distortion would elegantly account for the disrupted integration on center trials observed in Experiments 1 and 2. Looked at the other way, perceptual distortion away from the category boundary may account for the improved performance on non-boundary trials. But first, does such distortion actually occur?

The purpose of Experiment 3a was to examine whether repulsive mechanisms influenced the perception of individual ellipses. Specifically, I presented sets of ellipses or individual ellipses and, using a post-cue, asked observers to evaluate the aspect ratio of individual ellipses. I then evaluated the extent to which the perception of an ellipse's aspect ratio was systematically distorted as a function of its proximity to the category boundary. I predicted that errors in aspect ratio judgments would follow an s-curve shaped pattern (the derivative of a Gaussian function; Figure 3a.1), with the highest magnitude of repulsive distortion near the category boundary, and a gradual decay of distortion as cued aspect ratios become progressively flatter (or taller) relative to the boundary. This distortion pattern has been observed for other visual features (e.g., Crane, 2012; Sweeny et al., 2012).

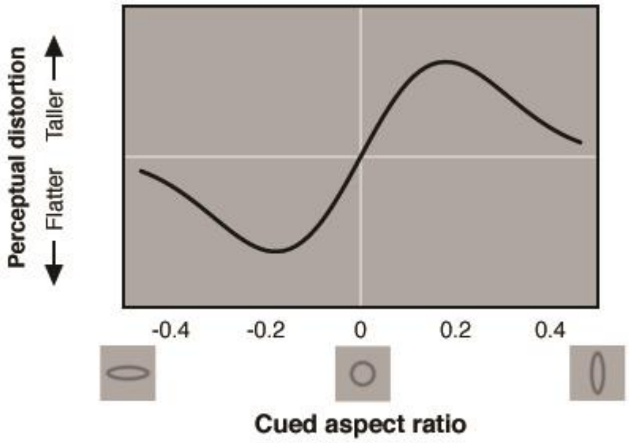

**Figure 3a.1.** Idealized hypothetical model of perceptual distortion around the category boundary. Ellipse aspect ratio is plotted on the X-axis from flattest to tallest, and systematic error in aspect ratio perception (i.e., distortion) is plotted on the Y-axis. In this example, flat ellipses near the category boundary are perceived to be even flatter, while tall ellipses near the boundary are perceived to be taller than they actually were. Perception with no distortion would be represented by a flat line with no slope. This effect is hypothesized to have occurred in Experiment 1 and 2; Experiment 3a tests for the presence of this effect.

### *Method*

***Observers.*** In Experiment 1, observers did integrate information, as indicated by the comparison between *full-range* and *single-ellipse* trials. This primary comparison from Experiment 1 was used to estimate the necessary sample size for both Experiment 2 and 3, since Experiment 2 and Experiment 3a were conceived of and created at the same time. Assuming the same large effect size (*d* = .6), I determined that a sample of thirty-three would be necessary to obtain a power of 0.8. The task in Experiment 3a required observers to make judgments about individual objects in crowds, and previous work suggests that this can be very difficult or even impossible (e.g., Haberman & Whitney, 2007; Allik et al., 2014). So, I increased the sample size for Experiment 3a by approximately 50%, to compensate for what I expected to be a difficult task. For Experiment 3a, forty-five students (*mean age* = 19.2 years; 76% female) from the University of Denver were recruited.

***Stimuli.*** Stimuli in Experiment 3a were identical to those used in Experiment 2.

***Procedure.*** The experiment consisted of 240 multi-ellipse (flat, tall, center, outlier and full-range trials), as well as single-ellipse trials, counterbalanced. All arrays were constructed and displayed in the same manner as they were in Experiments 1 and 2.

After each multi-ellipse or single-ellipse array was displayed, a blank screen was displayed for 100 ms. Next, a solid black circular cue (.78$^o$) appear at the centroid of the location of one randomly selected ellipse (or at the location of the visible ellipse on single-ellipse trials), for 250 ms. The cue was followed by another 400-ms blank screen. In this way, disregarding the cue display time, the amount of time between the offset of the array of aspect ratios and the onset of the response screen (500 ms) was held constant, relative to Experiments 1 and 2. Observers were then presented with the same response screen used in Experiment 2, except in Experiment 3a they were instructed to "indicate the cued tallness or flatness". Observers were also allowed as many practice trials as they wished beforehand, and the experimenter confirmed that observers understood that the task was to replicate the aspect ratio of the cued ellipse.

### *Results*

I began by quantifying how much repulsion, if any, occurred for the cued ellipse on every trial. For example, if the cued ellipse happened to have a slightly "tall" aspect ratio (e.g., 0.139), and an observer responded with aspect ratio 0.00, then no repulsion occurred, since the observer's response was not exaggerated away from the category boundary but was instead attracted to it. In this example, −0.139 log units of perceptual *attraction* occurred. In contrast, if an observer responded with aspect ratio 0.185, 0.046 log units of perceptual repulsion occurred, since the slightly tall ellipse was perceived to be exaggerated away from the category boundary (i.e., it was perceived to be taller than it really was). Across trials, this analysis yielded an average repulsion index, for every observer, and more importantly, for every cued aspect ratio. Thus for every observer, I computed an average repulsion index for each cued aspect ratio. Across observers, I then had a measure of repulsion for every aspect ratio in the stimulus set.

Across all multi-ellipse (i.e., all flat, tall, center, outlier and full-range) trials, the aspect ratios of ellipses were consistently underestimated. For example, extremely flat aspect ratios were rated as much taller than they truly were, while extremely tall aspect ratios were rated as much flatter than they actually were (Figure 3a.2A). Overall these errors produced a linear pattern with a negative slope, not the s-curved shape of repulsion as I predicted**.** Rather than reflecting a perceptual effect of attraction, this pattern instead suggests that observers were guessing during a very difficult task (see Sweeny, Haroz & Whitney, 2012 for simulations illustrating how guessing could produce this pattern; see Brady, Schurgin & Wixted, 2019, for a potential distinction between subjective and objective guessing). If, for example, a very flat aspect ratio was cued, and the observer simply responded randomly, the majority of random responses would, necessarily, be *less flat* (i.e., taller) than the cued aspect ratio. Similarly, if a very tall aspect ratio was cued, random guessing would be most likely to produce a response that was *less tall* (i.e. flatter) than the cued aspect ratio. The magnitude of these errors would, of course diminish as the cued ellipse approached the center of the response range. Thus, guessing could plausibly account for the pattern of results seen across multi-ellipse trials. I did not predict this, although in hindsight it makes sense, especially if observers did not retain conscious access to the cued ellipse. This is reasonable since prior ensemble-coding research has illustrated that access to individual objects can be severely diminished when they are viewed in the context of a crowd (e.g., Haberman & Whitney, 2007; Allik et al., 2014).

In contrast to multi-ellipse trials, the overall distribution of repulsion indices for single aspect ratios around the category boundary did conform to my hypotheses, following an s-shaped curve (Figure 3a.2B)**.** Slightly flat ellipses were perceived as flatter than they actually were, while slightly tall ellipses were perceived to be taller than they actually were. Across all aspect ratios, the absolute magnitude of repulsion indices was greater than zero (*mean log units* = .06, *SD* = .03) $t(44) = 15.12$, $p < .01$, $d = 2.25$. Across all observers, the pattern of repulsion indices on single-ellipse trials was well fit by a derivative of a Gaussian function ($R^2 = .86$, $p < .01$) (Figure 3a.2B). An AICc analysis confirmed with 99.99% certainty that the fit for the derivative of a Gaussian characterized the pattern of data better than a linear fit ($R^2 = .33$, *n.s.*). The derivative of a Gaussian function used had three parameters and the mathematical constant, *e*: $f(\mathrm{x}) = \mathrm{x}*P_1*P_2*P_3*e(-(P_2*\mathrm{x}^2))$. The values for $P_1$, $P_2$, and $P_3$ were .1162, 15.58, and .3206, respectively.

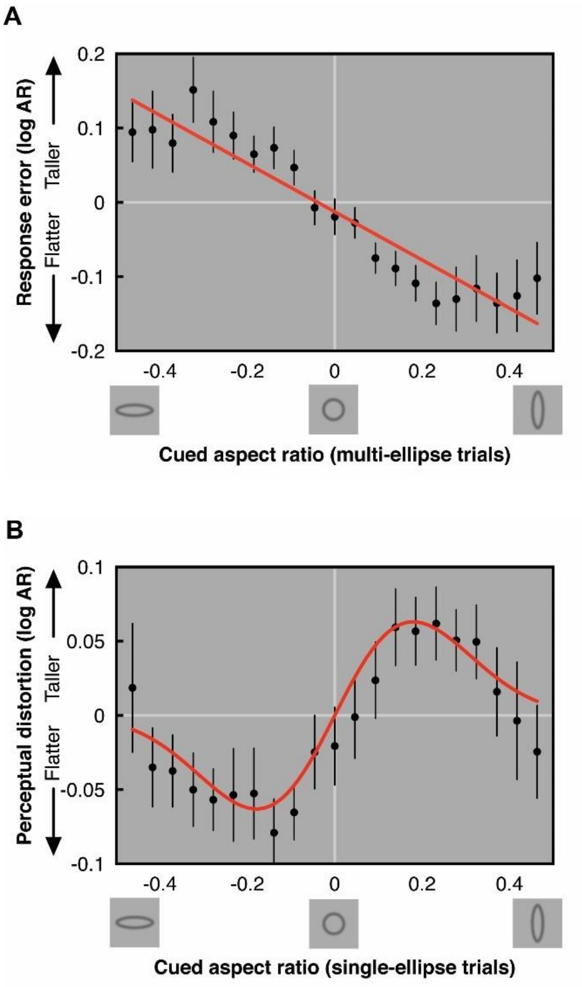

**Figure 3a.2.** Cued ellipse response error on multi-ellipse trials (A). Ellipse aspect ratio is plotted on the X-axis from flattest to tallest, and systematic error in observer response is plotted on the Y-axis. In response to extremely flat ellipses, observers erred with tall values; vice-versa in response to extremely tall ellipses. Data points linearly fit (red line); negative slope is likely indicative of guessing. Perceptual distortion around the category boundary on single-ellipse trials (B). Systematic error in aspect ratio perception (i.e., distortion) is plotted on the Y-axis. Flat ellipses near the category boundary are perceived to be even flatter, while tall ellipses near the boundary are perceived to be taller than they actually were. Data points are fit with the derivative of a Gaussian (red curve).

### *Experiment 3a Discussion*

Overall, Experiment 3a suggests that on single-ellipse trials, perceptual distortion did occur, and it occurred for some ellipses more than others. Slightly flat ellipses were perceived to be even flatter than they really were, and slightly tall ellipses were perceived as taller. Interestingly, this pattern was not apparent across multi-ellipse trials, though I hypothesized it would be. This is perhaps not too surprising, as observers often lose the ability to report on individual members of a set even while ensemble information about the gist of the set remains nonetheless accessible to their conscious report (e.g., Haberman & Whitney, 2007; Alvarez & Oliva, 2008; Neumann, Ng, Rhodes & Palermo, 2018). Pooling of visual information may even temporally precede awareness of individual objects (Allik et al., 2014). Although unavailable to conscious report, it seems reasonable to assume that the aspect ratios of multi-ellipse sets were encoded to some degree – otherwise integration (such as that observed in Experiments 1 and 2) could not have occurred. And importantly, it is possible that observers could have encoded and integrated *distorted* rather than veridical representations on multi-ellipse trials or Experiments 1 and 2.

# EXPERIMENT 3b

If the representations of individual aspect ratios in multi-ellipse trials in Experiment 2 were distorted in a manner compatible with the pattern observed on single-ellipse trials in Experiment 3, then the error of observers' estimates from Experiment 2 should be reduced when calculated relative to set means based on *distorted* values rather than the aspect ratios that were physically present.

Recall that in Experiment 2, observer error was computed relative to the actual aspect ratios displayed. However, Experiment 3 suggests that observers may not have encoded (or perceived, in the case of single trials) those aspect ratios veridically; they were distorted. Thus, if I retrospectively re-labeled the *displayed* aspect ratios in Experiment 2 with the *distorted* aspect ratio values from Experiment 3 (taken from the derivative of a Gaussian fit), then estimates of observer error in Experiment 2 should be systematically reduced. Investigating these possibilities was the aim of Experiment 3b.

I predicted that observer error would decrease overall, when error values were recalculated relative to the aspect ratios that observers likely perceived and integrated. More specifically, I expected re-calculated observer accuracy to improve the most for transformed non-boundary trials, compared to transformed center trials. My reasoning for this prediction is as follows: imagine a flat trial with a mean somewhere near the center of the flat range. On this hypothetical flat trial, some aspect ratios, especially those that are somewhat flat, would be perceived as even flatter. In contrast, aspect ratios that are extremely flat would be distorted less or not at all. The net effect would be that all the flat aspect ratios on this hypothetical trial would become, on average, *more* similar to one another. The same would be true for tall trials. As a result, transformation will not only change the values of individual aspect ratios in non-boundary sets, it should also significantly *shift the mean* of non-boundary trials. If observers were basing their mean judgments on distorted values on these non-boundary trials, then computing their error relative to those shifted means should significantly decrease the magnitude of observer error. In contrast, imagine a center-spanning trial with a mean of zero (a circle). In this case, the perception of slightly flat or slightly tall ellipses would be distorted *away* from the mean of the set; they would appear *less* similar to one another—effectively *increasing* heterogeneity! Crucially, across many center trials, the transformation (re-calculation) of aspect ratios based on perceived values should shift the mean less or not at all, since the distortion of slightly flat and slightly tall ellipses away from the category boundary should effectively cancel each other out. Thus, observer error relative to transformed values on center trials should improve less, or not at all relative to non-boundary trials. This brings me to the second prediction for Experiment 3b. I expected the *encoded* heterogeneity (as opposed to the actual heterogeneity) to increase for center trials more than for non-boundary trials, for the exact reasons described above. In sum, after transformation, I expected observers' mean estimates to become *especially* more accurate when extracting the mean of non-boundary trials. Additionally, I expected *encoded* heterogeneity to increase more for center trials than for non-boundary trials.

### *Method*

I transformed the data from Experiment 2, using the s-shaped curve obtained in Experiment 3a to precisely guide these transformations. For example, if a set of ellipses from a given trial in Experiment 2 contained an ellipse with aspect ratio −0.185 (the fourth flattest ellipse from the circle value), its distorted aspect ratio according to the fit in Figure 3a.2B would be used to calculate the set mean, not its actual aspect ratio. All the data for each and every trial in Experiment 2 were transformed in this way.

For example, say that one of the eight aspect ratios on some trial was -.139 – somewhat flat. Applying the s-shaped function to that aspect ratio yielded a value of -.0597 log units of distortion. Thus, the aspect ratio -.139 was potentially perceived as an aspect ratio with a value of -.19871 – flatter than it really was. I then replaced the original aspect ratio (-.139) with this new, distorted aspect ratio (-.19871). I repeated this process for every aspect ratio present in every set in the recorded data from Experiment 2, which yielded a new, transformed Experiment 2 data set, referred to as such henceforth. Note that I did not transform observer responses in this way, since observers were free to deployed focused attention to the response ellipses, for an unrestricted amount of time (*mean reaction time in seconds* = 2.28, *SD* = 1.79).

### ***Results***

To begin, it was critical to determine if the magnitude of observer error—the error relative to the mean of *transformed* trials—was reduced, compared to error relative to Experiment 2's original, *un-transformed* trial data. In other words, did observer accuracy improve, once the distortion described in Experiment 3a was taken into account? Note that, in this case, *sensitivity* (i.e. error distribution SD) would have been an inappropriate measure to analyze. Distortion should shift the entire distribution of errors across all trials, systematically increasing/decreasing the average magnitude of observer error, but leaving the SD of an observer's error *distribution* unaffected. Thus, absolute error magnitude was analyzed instead of SD. In addition to simply assessing whether observer accuracy increased, it was critical to determine if that increase was greater for non-boundary trials than for center trials.

I began by computing observer error relative to the new, transformed Experiment 2 data set. I computed the absolute magnitude of each observer's error, relative to the transformed trial mean, for every trial. For each observer, I was then able to compute average error magnitude, for all of my main three trial types (flat, tall and center trials). Averaging across all these transformed trial types yielded an average transformed error magnitude score for each observer. I repeated this process on Experiment 2's original, un-transformed main three trial types. Next, for each observer, I subtracted the transformed average error magnitude score from their un-transformed error magnitude score; this difference score reflected the amount of change in error magnitude (averaged across all three of my main trial types) between the transformed and un-transformed data sets. I refer to this difference score as the "error magnitude change index". Positive error magnitude change indices represented an improvement in observer accuracy, after transformation. Across observers, the error magnitude change index was indeed positive (*mean log units* = .006, *SD* = .007) *t*(44) = 5.36, *p* < .01, *d* = .8 (Figure 10A)**.** This indicates not just that observer error accuracy improved once perceptual distortion was taken into account, but also by how much. I then computed this index separately for flat and tall trials separately; the average of the two, computed for each observer, provided a measure of accuracy improvement for transformed non-boundary trials. I did the same for transformed center trials. Across observers, as predicted, observers evidenced a larger error magnitude change index for non-boundary trials (*mean log units* = .0078, *SD* = .0098) than for center trials (*mean log units* = .0022, *SD* = .0049) *t*(44) = 4.45, *p* < .01, *d* = .66 (Figure 3b.1A)**.**

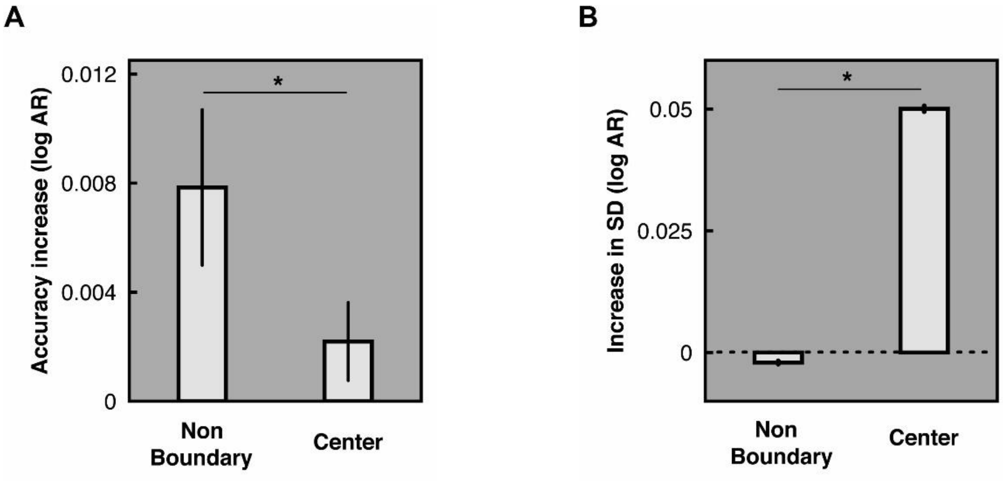

**Figure 3b.1.** After transforming data from Experiment 2 in a manner retrodicted by Experiment 3a's derivative of a Gaussian fit, and re-analyzing that data relative to the resulting transformed trial means, observer accuracy improved, particularly on non-boundary trials (A). Additionally, after transformation, average set heterogeneity increased for center trials, and very slightly decreased for non-boundary trials (B).

Next, I wanted to investigate perhaps the most important prediction of Experiment 3b. I hypothesized that *encoded* (as opposed to veridical) heterogeneity would increase for center trials more than for non-boundary trials, once distortion was taken into account. To investigate this, I began by computing the heterogeneity present on every trial (i.e., I computed the SD of every trial's eight ellipses), for both the transformed and original, un-transformed Experiment 2 data set. For each observer, I then computed the average heterogeneity present for flat, tall and center trials. I did this for both data sets. Mirroring the logic described immediately above, I subtracted each observer's average un-transformed flat trial (for example) heterogeneity from their transformed flat trial heterogeneity. This yielded a difference score that reflected the change in set heterogeneity, averaged across all flat trials, between un-transformed and transformed data sets, for each observer. I repeated this process for tall and center trials. I called this difference score the "change in heterogeneity index". In general, a positive (or negative) index represented a *increase* (or *decrease*) in average *transformed* heterogeneity across trials of a given type once data had been transformed. Averaging the change in heterogeneity index across flat and tall trials provided an index for non-boundary trials, specifically.

Across observers, the change in heterogeneity index for non-boundary trials was slightly negative (*mean log units* = -.002, *SD* = .001), reflecting a slight *decrease* in heterogeneity when distortion was taken into account (Figure 3b.1B). In contrast, the heterogeneity index for center trials was positive (*mean log units* = .05, *SD* = .001), reflecting a large *increase* in heterogeneity for center trials. The two indices were significantly different from one another $t(44) = 172.83$, $p < .01$, $d = 25.76$.

### *Experiment 3b Discussion*

Experiment 3b supports two important conclusions. First, as predicted, observer accuracy was significantly improved, once the distortion described in Experiment 3 was taken into account. This was particularly true for non-boundary trials. This suggests that, in Experiment 2, observers were basing their judgments on *distorted*, rather than veridical aspect ratios. Further, it suggests that the distortion described by Experiment 3a is a good model of the distortion observers likely encoded in Experiment 2.

Further, this distortion may be *responsible* for observers' disrupted ability to integrate information that spans the category boundary. By taking distortion into account, Experiment 3b showed that *encoded* (as opposed to veridical) heterogeneity increased precisely for the sets that observers had difficulty integrating. As stated many times already, heterogeneity is known to disrupt pooling and integration. Experiments 3a and 3b suggest that heterogeneity need not be veridical in order to disrupt integration, it can be a result of the distortion—the segmentation—that occurs near a visual feature's category boundary.

# GENERAL DISCUSSION

The results reported above suggest several important, novel conclusions. First, integration in the form of ensemble coding can operate on perceived aspect ratio, but this integration is disrupted if set members include values that span the category boundary. In other words, integration works best on groups that are *either* flat *or* tall, not both. Second, in line with previous work, values of single aspect ratios near the category boundary are distorted—segmented—away from that boundary when seen for a brief duration. Flat ellipses appear flatter, and tall ellipses appear taller. The distortion observed for individual ellipses may also occur for the perception of multiple ellipses seen in a set. Even though observers were unable to report a randomly cued aspect ratio within a set, those distorted values were likely still encoded. After all, once distortion was taken into account, people became even *more accurate* when extracting the mean of sets of aspect ratios that do not span the category boundary. The inability to report on individuals in a set, while still retaining access to summary information about that set, has been reported before for features other than aspect ratio (e.g., Haberman & Whitney, 2007; Alvarez & Oliva, 2008; Neumann et al., 2018). Importantly, observers also exaggerated the "flatness" of flat trials and the "tallness" of tall trials—further suggesting that multiple simultaneously presented ellipses were indeed distorted.  Finally, e*ncoded heterogeneity* was exaggerated for groups that include both flat and tall ellipses. I propose that this exaggerated heterogeneity contributes to the visual system's disrupted ability to integrate information across category boundaries. These results deepen the field's understanding of integration by showing how it can interact with, and even conflict with, another fundamental computational method in the visual system—segmentation.

It is important to note that heterogeneity may not entirely account for the disruption of integration on sets that spanned the category boundary. Visual grouping may also have contributed. For example, spatial proximity alone can unite objects into coherent groups (see Palmer, 1999; Chong & Treisman, 2003), and summary statistics can then be extracted from those groups (Chong & Treisman, 2003). But grouping via spatial proximity also *aids* integration; mean statistics can be extracted from spatially proximate groups with particular accuracy (Im & Chong, 2014). This suggests that grouping can "gate" integration: it can help the visual system determine which information to average across, and which information to exclude. Integration seems to operate particularly well on grouped information, perhaps regardless of which grouping principle (e.g., common region) facilitated grouping. Indeed, it is not just spatially proximate objects that are grouped and integrated. Mean statistics can be computed for groups that are segregated by mid-level features like color (Brady & Alvarez, 2011). Sets of faces that *behave* together over time are also treated as particularly "group-like" by the visual system, and indeed, the integrative process of ensemble coding is particularly sensitive to such sets (Elias et al., 2017); importantly, this is true even when heterogeneity is controlled for. The faces that behaved together may have been grouped by common fate (Palmer, 1999) or the principle of synchrony (Palmer, 2002). Similarly, when presented with a set of complex visual objects that includes spatially interspersed sub-sets (e.g., a single crowd of faces that includes faces of both black and white men), observers do seem to be sensitive to the means of those sub-sets (Lamer, Sweeny, Dyer & Weisbuch, 2018). Perhaps those sub-sets were grouped by classic grouping principles, or perhaps the grouping of those sub-sets was informed by top-down processes relevant to social categorization, including past experience (see Wagemans et al., 2012; Peterson & Kimchi, 2013; for brief reviews of the influence of past experience on grouping). So grouping (via common region, common color, or even high-level information like shared emotion expression or racial category) may also contribute to the efficient and accurate integration of visual information. It may, in fact, gate the precision of ensemble coding.

Thus, in the present work, a series of factors may have been acting to disrupt the integration of sets that spanned the category boundary. Yes, the evidence strongly suggests that segmentation interfered with integration via encoded heterogeneity. But, information that spans a feature category boundary may also be resistant to perceptual grouping. Interestingly, the pooling of height *and* width information occurs automatically, even if the task is just to pool height *or* width information (Oriet & Brand, 2013). This suggests that both flat and tall objects can be automatically grouped together prior to integration. As noted before, though, Oriet & Brand (2013) did not equate their stimuli for size, so it is unclear whether aspect-ratio information was automatically grouped and integrated. In the present work, it is difficult to say how much grouping contributed to the effects observed, or to speak meaningfully about the time-course of grouping's potential contribution, especially since grouping can act at multiple stages of the visual processing hierarchy (either early or relatively late; see Palmer, 2002; Wagemans et al., 2012; Peterson & Kimchi, 2013). Grouping may have been disrupted first. For example, it is possible that sets that spanned the category boundary were not grouped as efficiently or quickly or even at all, and this disruption to grouping disrupted integration *above and beyond* the effects of increased heterogeneity from segmentation. This disruption to grouping could have, speculatively, been a result of the very early (within 100 ms) modulation of feature-based attention (Zhang & Luck, 2009). Similarly, it is also possible that groups of similar objects (e.g., ones that do not span a category boundary), were treated as a gestalt object by the visual system (perhaps as early as striate cortex) and subsequently received additional attentional resources, devoted to refined processing of the gestalt object's constituent parts (in this work, the ellipses themselves) (Flevaris, Martinez & Hillyard, 2013). Or, perhaps grouping was disrupted only *after* segmentation acted; indeed, perhaps segmentation caused grouping to be disrupted. Regardless of the details, though, the evidence presented here strongly suggests that segmentation did interfere with integration. Disrupted grouping may have had an additional effect, though this cannot be directly evaluated based on the current data. Untangling the interaction between the disruptive effects of segmentation and those of disrupted grouping is a promising direction for future research.

In the current investigations, the category boundary was a perfect circle. As already discussed, that category boundary may be a result of the way neurons dedicated to aspect ratio are organized at the population level (Kayaert et al., 2005; Dickinson et al., 2017; Storrs & Arnold, 2017). Still, it is possible that the precise value of aspect ratio's category boundary is somewhat malleable – it may shift with learning or past experience. If, for example, an observer adapted to many trials composed of exclusively flat ellipses, the perceived category boundary might shift to become slightly taller. If that is indeed possible, an open question is whether or not distortion can occur around the new, perceptually shifted category boundary.

It is also important to note that segmentation likely did not prevent the neural mechanisms that underlie ensemble coding from operating full-stop. Instead, those mechanisms proceeded, but acted on a distorted and more heterogeneous set of information. It is in this sense that I have used the word "conflict" throughout these investigations. The *functions* of segmentation and integration can be at odds, and can lead to the disrupted operation of integrative processes, even if strictly speaking, the two processes still unfolded serially (i.e., even if the conflict between the two processes was not winner-takes-all). If segmentation and integration did unfold serially in this investigation, I suspect that segmentation acted first. After all, in this investigation, as in others (e.g., Haberman & Whitney, 2007; Alvarez & Oliva, 2008; Neumann, Ng, Rhodes & Palermo, 2018), observers likely encoded and then integrated individual objects in the set, even if they could not consciously report those individuals later. Thus, it seems likely that in the current investigation, segmentation acted first, and integration then operated on a subset of distorted ellipses. Still, a more direct test of the temporal relationship between segmentation and integration in sets that include a category boundary is open for future research.

Prior work on summary judgments of size suggests that approximately three to five items are sampled from a set (Im & Halberda, 2013; Gorea, Belkoura & Solomon, 2014), or as a more general rule, the square root of the set size (Dakin, 2001, see Whitney & Leib, 2018). This brings me to a limitation of the present work. It is not possible to confirm (or even estimate) how many ellipses observers sampled when integrating aspect ratio information in the current investigation. It could have been anywhere from two to eight ellipses. However, the estimates described immediately above certainly seem reasonable, for the following reason. Theoretically, heterogeneity should impact the precision of mean estimation less as more objects are sampled from a set (Marchant et al., 2013). If observers sample every item from a set, heterogeneity should be irrelevant. Imagine randomly four a subset of ellipses from a set of eight with zero heterogeneity. Neglecting perceptual and decision noise, this random sample will, of course, yield a perfect estimate of the overall set mean—zero error. However, now imagine randomly sampling four ellipses from a set of eight with a large amount of heterogeneity. Error will, in that case, be increased, since the ellipses you sampled were less representative of the set mean. Given the positive relationship between heterogeneity and observer error described in the present work, it thus seems unlikely that observers integrated information from all eight ellipses in each multi-ellipse set. Still, it is possible that observers encoded information about all eight ellipses in each multi-ellipse set (Robitaille & Harris, 2011). If all eight ellipses were encoded on a given trial, then the *noisiness* of those representations may have contributed the pattern of observer error reported here (Alvarez, 2011; see Brady, Konkle & Alvarez, 2011; Neumann et al., 2018; Schurgin, Wixted & Brady, 2019; Brady, Schurgin & Wixted, 2019). If this account is correct, then although all eight ellipses were encoded on each multi-ellipse trial, the ellipses contained in center trials would have been encoded with less fidelity and more noise. This sort of explanation requires an additional assumption regarding noisier representation around the category boundary, and is thus less parsimonious than a simpler model in which a subset of ellipses are used to estimate the mean. Regardless of the details though, it is clear that in the current investigation, observers integrated information from between two and eight ellipses. A more precise estimate would require additional experimentation. Most importantly, this uncertainty does not change the main results of this investigation.

Additionally, in the present work, it is not possible to say precisely *which* ellipses were sampled, assuming that only a subset were sampled. For example, it is possible that observers tended to sample more extreme ellipses (Kanaya, Hanashi & Whitney, 2018). If true, this tendency may have been most pronounced on non-boundary trials in the current investigation. This is compatible with observers' tendency to judge exaggerate the "flatness" or "tallness" of flat and tall sets respectively. By systematically sampling more extreme ellipses, group mean judgments would be exaggerated away from the category boundary. Importantly, this account of group mean exaggeration, and the account that relies on the distortion of individual ellipses (described above), are not mutually incompatible. In fact, since the accuracy of mean estimates increased once the distortion of individual ellipses was taken into account, biased sampling is unlikely to completely explain group mean exaggeration. Instead, the distortion of individual ellipses in a set likely contributed to the exaggeration of group means away from the category boundary.

Taking a broad view, it is possible that the results of the present investigation apply to visual features with a category boundary in general, not just to aspect ratio. There is good reason to think this may be the case. After all, distortion around the category boundary has been observed for relatively high-level visual features like 3-D depth (Grossberg, 1994) and biological motion (Sweeny et al., 2012) in addition to simpler visual information like curvature (Sweeny, Grabowecky, Kim & Suzuki, 2011). And, in general, perceptual bias (e.g., distortion) is likely a function of how noisily a feature is encoded in the first place—the more noisy the representation, the more distortion is needed to avoid crossing the category boundary (Wei & Stocker, 2017). Thus, it is likely that distortion away from the category boundary helps the visual system avoid categorical errors, given imperfectly encoded feature information. This distortion, although useful in many circumstances, may also sometimes interfere with integration by introducing exaggerated heterogeneity, regardless of what feature is being integrated. Thus, the tension between segmentation and integration may be ubiquitous. In fact, this investigation was never intended as an examination of aspect ratio, per se, but rather of the interaction between category boundaries, segmentation and integration more generally. Aspect ratio simply provided a good candidate visual feature to investigate this interaction for the first time. Future ensemble coding work, and integration work in general, may benefit from considering conflict between segmentation and integration demonstrated here. It is also possible that the conflict between segmentation and integration could ultimately have behavioral consequences. For example, people may be slower to respond to the gist of sets in which segmentation conflicts with integration, or may experience extracting the gist from such sets to be more difficult or effortful (although, notably, the reaction times between boundary-spanning and center trials did not differ here).

The first sentences of an untold number of vision science papers, this one among them, runs something like this: “Think about how complex a computation your everyday vision is. It’s remarkable that we can see at all! Don’t take it for granted!” This opener is a good one – our mundane, everyday visual experience really does belie the sophisticated computational machinations that occur “under the hood”. However, it is important to remember that ultimately, the visual system almost certainly does not operate the way it does *so that* you can enjoy a visual experience. Rather, the visual system must solve a number of difficult computational problems *so that* an organism can *act* effectively in the world in which it finds itself (Gibson, 2014). The problems that the visual system must solve are many, and the methods used to solve them are varied. Often, the system operates smoothly. At the very least, the system’s methods do not produce incompatible solutions or potentially incompatible recommendations for action. However, this is apparently not *always* the case. Our visual systems—i.e., our conscious and nonconscious visual “minds”—have evolved to solve computational problems in a somewhat modular way. Sometimes, the methods used to solve one problem (e.g., “what’s the gist of this clump of stuff?”) can conflict with, or be constrained by, the methods used to solve another (e.g., “is this stuff *this* or *that*?”). By investigating the visual system *at these sites of conflict*, we stand to gain a fuller, richer, more nuanced understanding of the visual system in general

# REFERENCES

Allik, J., Toom, M., Raidvee, A., Averin, K., & Kreegipuu, K. (2014). Obligatory averaging in mean size perception. *Vision Research*, *101*, 34-40.

Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science*, *19*(4), 392-398.

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

Alvarez, G. A., & Oliva, A. (2009). Spatial ensemble statistics are efficient codes that can be represented with reduced attention. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(18), 7345-7350.

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

Awh, E., Barton, B., & Vogel, E.K. (2007). Visual working memory represents a fixed number of items regardless of complexity. *Psychological Science*, *18*, 622-628.

Badcock, D. R., & Westheimer, G. (1985). Spatial location and hyperacuity: The centre/surround localization contribution function has two substrates. *Vision Research*, *25*(9), 1259-1267.

Badcock, J. C., Whitworth, F. A., Badcock, D. R., & Lovegrove, W. J. (1990). Low-frequency filtering and the processing of local—global stimuli. *Perception*, *19*(5), 617-629.

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

Brady, T. F., Konkle, T., & Alvarez, G. A. (2011). A review of visual memory capacity: Beyond individual items and toward structured representations. *Journal of Vision*, *11*(5): 4.

Brady, T. F., Schurgin, M., Wixted, J. The Target Confusability Competition Model and the importance of distinguishing between subjective and objective guessing in visual working memory. Poster presented at: Nineteenth annual meeting of the Vision Sciences Society; 2019 May 17—22; St. Pete Beach, FL.

Braun, A., & Sweeny, T. D. (2019). Anisotropic visual awareness of shapes. *Vision research*, *156*, 17-27.

Biederman, I., & Kalocsai, P. (1997). Neurocomputational bases of object and face recognition. *Philosophical Transactions: Biological Sciences*, *352*, 1203-1219.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10*, 433–436.

Brincat, S. L., & Connor, C. E. (2004). Underlying principles of visual shape selectivity in posterior inferotemporal cortex. *Nature Neuroscience*, *7*(8), 880-886.

Buhmann, J. M., Malik, J., & Perona, P. (1999). Image recognition: Visual grouping, recognition, and learning. *Proceedings of the National Academy of Sciences*, *96*(25), 14203-14204.

Chong, S. C., & Treisman, A. (2005). Statistical processing: Computing the average size in perceptual groups. *Vision Research*, *45*(7), 891-900.

Chong, S. C., & Treisman, A. (2005b). Attentional spread in the statistical processing of visual displays. *Perception & Psychophysics*, *67*(1), 1-13.

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research*, *43*(4), 393-404.

Choo, H., & Franconeri, S. L. (2010). Objects with reduced visibility still contribute to size averaging. *Attention, Perception, & Psychophysics*, *72*(1), 86-99.

Clifford, C. W. (2014). The tilt illusion: Phenomenology and functional implications. *Vision Research*, *104*, 3-11.

Crane, B. T. (2012). Direction specific biases in human visual and vestibular heading perception. *PLoS One*, *7*(12), e51383.

Dakin, S. C. (2001). Information limit on the spatial integration of local orientation signals. *Journal of the Optical Society of America A*, *18*(5), 1016-1026.

de Beeck, H. O., Wagemans, J., & Vogels, R. (2003). The effect of category learning on the representation of shape: dimensions can be biased but not differentiated. *Journal of Experimental Psychology: General*, *132*(4), 491-511.

Dickinson, J. E., Morgan, S. K., Tang, M. F., & Badcock, D. R. (2017). Separate banks of information channels encode size and aspect ratio. *Journal of Vision*, *17*(3): 27.

Dumoulin, S. O., & Hess, R. F. (2007). Cortical specialization for concentric shape processing. *Vision Research*, *47*, 1608–1613.

Elias, E., Dyer, M., & Sweeny, T. D. (2017). Ensemble perception of dynamic emotional groups. *Psychological Science*, *28*(2), 193-203.

Elias, E., Padama, L., & Sweeny, T. D. (2018). Perceptual averaging of facial expressions requires visual awareness and attention. *Consciousness and Cognition*, *62*, 110-126.

Flevaris, A. V., Martinez, A., & Hillyard, S. A. (2013). Neural substrates of perceptual integration during bistable object perception. *Journal of Vision*, *13*(13): 17.

Gibson, J. J. (2014). *The Ecological Approach to Visual Perception: Classic Edition*. Psychology Press.

Grossberg, S. (1994). 3-D vision and figure-ground separation by visual cortex. *Perception & Psychophysics*, *55*(1), 48-121.

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

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

Haberman, J., & Whitney, D. (2009). Seeing the mean: Ensemble coding for sets of faces. *Journal of Experimental Psychology: Human Perception and Performance, 35*, 718-734. doi:10.1037/a0013899

Haberman, J., & Whitney, D. (2009b). The visual system ignores outliers when extracting a summary representation. *Journal of Vision*, *9*(8): 804.

Haberman, J., & Whitney, D. (2010). The visual system discounts emotional deviants when extracting average expression. *Attention, Perception, & Psychophysics*, *72*(7), 1825-1838.

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

Haberman, J., Lee, P., & Whitney, D. (2015). Mixed emotions: Sensitivity to facial variance in a crowd of faces. *Journal of Vision*, *15*(4), 16-16.

Im, H. Y., & Chong, S. C. (2014). Mean size as a unit of visual working memory. *Perception*, *43*(7), 663-676.

Im, H. Y., & Halberda, J. (2013). The effects of sampling and internal noise on the representation of ensemble average size. *Attention, Perception, & Psychophysics*, *75*(2), 278-286.

Im, H. Y., Albohn, D. N., Steiner, T. G., Cushing, C. A., Adams, R. B., & Kveraga, K. (2017). Differential hemispheric and visual stream contributions to ensemble coding of crowd emotion. *Nature Human Behaviour*, *1*(11), 828-842.

Jacoby, O., Kamke, M. R., & Mattingley, J. B. (2013). Is the whole really more than the sum of its parts? Estimates of average size and orientation are susceptible to object substitution masking. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(1), 233-244.

Kanaya, S., Hayashi, M. J., & Whitney, D. (2018). Exaggerated groups: amplification in ensemble coding of temporal and spatial features. *Proceedings of the Royal Society B: Biological Sciences*, *285*(1879), 20172770.

Kastner, S., De Weerd, P., Pinsk, M. A., Elizondo, M. I., Desimone, R., & Ungerleider, L. G. (2001). Modulation of sensory suppression: Implications for receptive field sizes in the human visual cortex. *Journal of Neurophysiology, 86*(3), 1398-1411.

Kayaert, G., Biederman, I., Beeck, H. P. Op De, & Vogels, R. (2005). Tuning for shape dimensions in macaque inferior temporal cortex. *European Journal of Neuroscience*, *22*, 212-224. http://doi.org/10.1111/j.1460-9568.2005.04202.x

Kourtzi, Z. (2010). Visual learning for perceptual and categorical decisions in the human brain. *Vision Research*, *50*(4), 433-440.

Lamer, S. A., Sweeny, T. D., Dyer, M. L., & Weisbuch, M. (2018). Rapid visual perception of interracial crowds: Racial category learning from emotional segregation. *Journal of Experimental Psychology: General*, *147*(5), 683-701.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279-281.

Marchant, A. P., Simons, D. J., & de Fockert, J. W. (2013). Ensemble representations: Effects of set size and item heterogeneity on average size perception. *Acta Psychologica*, *142*(2), 245-250.

Michel, M. M., Chen, Y., Geisler, W. S., & Seidemann, E. (2013). An illusion predicted by V1 population activity implicates cortical topography in shape perception. *Nature neuroscience*, *16*(10), 1477-1483.

Miller, E. K., Gochin, P. M., & Gross, C. G. (1993). Suppression of visual responses of neurons in inferior temporal cortex of the awake macaque by addition of a second stimulus. *Brain Research, 616*(1), 25-29.

Morgan, M. J. (1999). The Poggendorff illusion: a bias in the estimation of the orientation of virtual lines by second-stage filters. *Vision Research*, *39*(14), 2361-2380.

Morgan, M., Chubb, C., & Solomon, J. A. (2008). A 'dipper' function for texture discrimination based on orientation variance. *Journal of vision*, *8*(11): 9.

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive psychology*, *9*(3), 353-383.

Neumann, M. F., Ng, R., Rhodes, G., & Palermo, R. (2017). Ensemble coding of face identity is not independent of the coding of individual identity. *The Quarterly Journal of Experimental Psychology*, (just-accepted), 1-27.

Oriet C., Corbett, J,E. (2008). Evidence for rapid extraction of average size in RSVP displays of circles. *Journal of Vision, 8*(6): 13.

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

Palmer, S. E. (1999). Vision science: Photons to phenomenology. Cambridge: MIT Press.

Palmer, S. E. (2002). Perceptual organization in vision. In: Pashler, H. (Ed.). (2002). *Stevens' Handbook of Experimental Psychology*, 177—234. John Wiley & Sons.

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

Pasupathy, A., & Connor, C. E. (2001). Shape representation in area V4: position-specific tuning for boundary conformation. *Journal of Neurophysiology*, *86*(5), 2505-2519.

Peterson, M. A., & Kimchi, R. (2013). Perceptual organization in vision. In: Reisburg, D. (Ed.). (2013). *The Oxford Handbook of Cognitive Psychology*, 2-28. Oxford University Press.

Regan, D., & Hamstra, S. J. (1992). Shape discrimination and the judgement of perfect symmetry: Dissociation of shape from size. *Vision Research*, *32*(10), 1845-1864.

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

Robson, M. K., Palermo, R., Jeffery, L., & Neumann, M. F. (2018). Ensemble coding of face identity is present but weaker in congenital prosopagnosia. *Neuropsychologia*, *111*, 377-386.

Rolls, E. T., & Tovee, M. J. (1994). Processing speed in the cerebral cortex and the neurophysiology of visual masking. *Proceedings of the Royal Society of London-B-Biological Sciences*, *257*(1348), 9-16

Ross, J., & Burr, D. (2008). The knowing visual self. *Trends in Cognitive Sciences, 12*, 363-364.

Sato, T. (1989). Interactions of visual stimuli in the receptive fields of inferior temporal neurons in awake macaques. *Experimental Brain Research*, *77*(1), 23-30.

Schurgin, M., Wixted, J., Brady, T. F. The Target Confusability Competition Model: Unambiguous evidence in favor of a signal detection model of visual working memory. Poster presented at: Nineteenth annual meeting of the Vision Sciences Society; 2019 May 17—22; St. Pete Beach, FL.

Storrs, K. R., & Arnold, D. H. (2017). Shape adaptation exaggerates shape differences. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(1), 181-191.

Suzuki, S. (2003). Attentional selection of overlapped shapes: a study using brief shape aftereffects. *Vision Research*, *43*, 549-561.

Suzuki, S., & Cavanagh, P. (1997). Focused attention distorts visual space: an attentional repulsion effect. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(2), 443-463.

Suzuki, S., & Cavanagh, P. (1998). A shape-contrast effect for briefly presented stimuli. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(5), 1315-1341.

Sweeny, T. D., Grabowecky, M., Kim, Y. J., & Suzuki, S. (2011). Internal curvature signal and noise in low-and high-level vision. *Journal of Neurophysiology*, *105*(3), 1236-1257.

Sweeny, T. D., Grabowecky, M., & Suzuki, S. (2011). Simultaneous shape repulsion and global assimilation in the perception of aspect ratio. *Journal of Vision*, *11*(1): 16.

Sweeny, T. D., Haroz, S., & Whitney, D. (2012). Reference repulsion in the categorical perception of biological motion. *Vision Research*, *64*, 26-34.

Sweeny, T. D., Haroz, S., & Whitney, D. (2013). Perceiving group behavior: Sensitive ensemble coding mechanisms for biological motion of human crowds. *Journal of Experimental Psychology: Human Perception and Performance*, *39*, 329–337.

Sweeny, T. D., & Whitney, D. (2014). Perceiving crowd attention: Ensemble perception of a crowd's gaze. *Psychological Science*, *25*(10), 1903-1913.

Sweeny, T. D., Haroz, S., & Whitney, D. (2013). Perceiving group behavior: Sensitive ensemble coding mechanisms for biological motion of human crowds. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(2), 329-361.

Sweeny, T. D., Wurnitsch, N., Gopnik, A., & Whitney, D. (2015). Ensemble perception of size in 4–5-year-old children. *Developmental science*, *18*(4), 556-568.

Sweeny, T. D., D'Abreu, L. C., Elias, E., & Padama, L. (2017). Object-substitution masking weakens but does not eliminate shape interactions. *Attention, Perception, & Psychophysics*, *79*(7), 2179-2189.

Suzuki, S., & Cavanagh, P. (1997). Focused attention distorts visual space: an attentional repulsion effect. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(2), 443.Suzuki, S., & Cavanagh, P. (1998). A shape-contrast effect for briefly presented stimuli. *Journal of Experimental Psychology. Human Perception and Performance*, *24*(5), 1315-1341.

Suzuki, S. (2005). High-level pattern coding revealed by brief shape aftereffects. In C. W. Clifford, & G. Rhodes (Vol. Eds.), *Fitting the mind to the world: Adaptation and after-effects in high-level vision: vol. 2*, (pp. 135-172). Oxford University Press.

Treisman, A., & Gormican, S. (1988). Feature analysis in early vision: evidence from search asymmetries. *Psychological Review*, *95*(1), 15-48.

Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., & von der Heydt, R. (2012). A century of Gestalt psychology in visual perception: I.

Perceptual grouping and figure–ground organization. *Psychological Bulletin*, *138*(6), 1172-1217.

Wagemans, J., Feldman, J., Gepshtein, S., Kimchi, R., Pomerantz, J. R., Van der Helm, P. A., & Van Leeuwen, C. (2012b). A century of Gestalt psychology in visual perception: II. Conceptual and theoretical foundations. *Psychological Bulletin*, *138*(6), 1218-1252.

Ward, E. J., Bear, A., & Scholl, B. J. (2016). Can you perceive ensembles without perceiving individuals?: The role of statistical perception in determining whether awareness overflows access. *Cognition*, *152*, 78-86.

Watamaniuk, S. N., & Duchon, A. (1992). The human visual system averages speed information. *Vision Research, 32,* 931-941.

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

Wei, X. X., & Stocker, A. A. (2017). Lawful relation between perceptual bias and discriminability. *Proceedings of the National Academy of Sciences*, *114*(38), 10244-10249.

Wertheimer M. (1923). Untersuchungen zur Lehre von der Gestalt. II. *Psychologische Forschung*, *4*, 301—350. [Partial translation in W. D. Ellis (Ed.) (1950). *A Sourcebook of Gestalt psychology*. New York: Humanities Press.]

Westheimer, G., & Levi, D. M. (1987). Depth attraction and repulsion of disparate foveal stimuli. *Vision Research*, *27*(8), 1361-1368.

Whitney, D., Haberman, J., & Sweeny, T. D. (2014). From textures to crowds: Multiple levels of summary statistical perception. In J. S. Werner, & L. M. Chalupa (Eds.). *The New Visual Neurosciences* (pp. 695-710). MIT Press.

Whitney, D., & Yamanashi Leib, A. (2018). Ensemble Perception. *Annual Review of Psychology*, *69*, 105-129.

Yamanashi Leib, A., Landau, A. N., Baek, Y., Chong, S. C., & Robertson, L. (2012). Extracting the mean size across the visual field in patients with mild, chronic unilateral neglect. *Frontiers in Human Neuroscience*, *6*, 267-278.

Yamanashi Leib, A., Puri, A. M., Fischer, J., Bentin, S., Whitney, D., & Robertson, L. (2012b). Crowd perception in prosopagnosia. *Neuropsychologia*, *50*(7), 1698-1707.

Zhang, W., & Luck, S. J. (2009). Feature-based attention modulates feedforward visual processing. *Nature Neuroscience*, *12*(1), 24-25.

Zoccolan, D., Cox, D. D., & DiCarlo, J. J. (2005). Multiple object response normalization in monkey inferotemporal cortex. *Journal of Neuroscience*, *25*(36), 8150-8164.