American Psychological Association

Journal of Experimental Psychology:
Human Perception and Performance

© 2020 American Psychological Association
ISSN: 0096-1523

http://dx.doi.org/10.1037/xhp0000733

Integration and Segmentation Conflict During Ensemble Coding of Shape

Elric Elias
University of Colorado Denver

Timothy D. Sweeny
University of Denver

During ensemble coding, the visual system extracts summary information from input that has been *integrated*, facilitating gist-level judgments about objects and features that belong together. In contrast, input can be *segmented*, allowing for quick categorical distinctions between objects. Integration and segmentation usually work in parallel but may sometimes conflict in the context of ensemble coding. To investigate this possibility, we examined summary perception of aspect ratio (i.e., "tallness/flatness"). Aspect ratio has a category boundary (e.g., a circle), and individual aspect ratios may be perceptually exaggerated—segmented—away from this boundary. We predicted that summary perception of multiple aspect ratios would be disrupted when, as a set, they spanned the category boundary, since integration and segmentation would then be at odds. We found that when observers reported the average aspect ratio of a set of ellipses, they were less sensitive to the mean of sets that included both tall *and* flat ellipses, compared to sets comprised of tall *or* flat ellipses. Follow-up experiments suggest this occurred because segmentation distorted the appearance of ellipses away from the category boundary, exaggerating set heterogeneity. These experiments advance understanding of how the visual system summarizes information by showing that integration and segmentation can conflict.

Public Significance Statement

The visual system uses two broad strategies to transform input into representations that are concise and useful. Sometimes, it integrates information in order to make generalizations, and other times it segments information to highlight differences. Here, we show that these two fundamental computational methods can conflict. Our findings provide insight into how the solution to some computational problems faced by the visual system can actually constrain, or conflict with, the solution to different problems.

Keywords: integration, segmentation, ensemble coding, categorical perception

Supplemental materials: http://dx.doi.org/10.1037/xhp0000733.supp

Imagine being immersed in an infinite sea of continuous information. Your task is to parse that information into useful, meaningful chunks. How would you decide which bits of information go together and which do not? 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 that constrain visual processing (e.g., attention, Chong & Treisman, 2005; memory, Luck & Vogel, 1997). To distill order and meaning from chaos, the visual system leverages at least two broad strategies: integration and segmentation.

Integration is one of the visual system's basic approaches for solving computational problems. Individual cells integrate information (e.g., Brincat & Connor, 2004; Kastner et al., 2001; Miller, Gochin, & Gross, 1993; Rolls & Tovee, 1995; Sato, 1989; Zoccolan, Cox, & DiCarlo, 2005). So do populations of cells (e.g., Michel, Chen, Geisler, & Seidemann, 2013; Pasupathy & Connor, 2001; Suzuki, 2005). In many cases, information about multiple visual features or objects is integrated via these populations. This happens, for example, in classic processes of visual grouping (see Palmer, 1999; Palmer, 2002; Peterson & Kimchi, 2013; Wagemans et al., 2012), and may result in the perception of texture (Dakin, 2015). 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 Alvarez, 2011; Whitney, Haberman, & Sweeny, 2014; Whitney & Yamanashi Leib, 2018, for reviews). Ensemble coding is a consequence of information integration—one that usually implies the integration of information to acquire summary information (e.g., the mean, variance, etc.) across sets with more than two members (Whitney & Yamanashi Leib, 2018). For example, as you pass a fruit stand full of oranges, your visual system can use ensemble coding to extract the

Elric Elias, Department of Psychology, University of Colorado Denver; Timothy D. Sweeny, Department of Psychology, University of Denver.

MATLAB scripts and all data available through the Center for Open Science at https://osf.io/pv4nz.

Correspondence concerning this article should be addressed to Elric Elias, Department of Psychology, University of Colorado Denver, 1200 Larimer Street, North Classroom, Denver, CO 80217. E-mail: elric.elias@ucdenver.edu

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). Similar operations can be performed for simple features like orientation (Alvarez & Oliva, 2009; Elias, Padama, & Sweeny, 2018; Parkes, Lund, Angelucci, Solomon, & Morgan, 2001; Ross & Burr, 2008) velocity and trajectory (Watamaniuk & Duchon, 1992; Watamaniuk, Sekuler, & Williams, 1989), hue (Maule & Franklin, 2015; Webster, Kay, & Webster, 2014), as well as for high-level visual features like facial expression (Elias, Dyer, & Sweeny, 2017; Haberman, Harp, & Whitney, 2009; Haberman & Whitney, 2007, 2009; Im et al., 2017), gaze (Florey, Clifford, Dakin, & Mareschal, 2016; Sweeny & Whitney, 2014), biological motion (Sweeny, Haroz, & Whitney, 2013), identity (de Fockert & Wolfenstein, 2009; Neumann, Schweinberger, & Burton, 2013; Yamanashi Leib, Landau, Baek, Chong, & Robertson, 2012), viewpoints (Yamanashi Leib et al., 2014), and attractiveness (Walker & Vul, 2014). More complex judgments of race, dominance, social norms, and even group membership itself appear to be rooted in ensemble coding, too (Dannals & Miller, 2017; Goldenberg, Sweeny, Shpigel, & Gross, 2020; Lamer, Sweeny, Dyer, & Weisbuch, 2018; Phillips, Slepian, & Hughes, 2018). Summary judgments even appear possible for more abstract targets, like average lifelikeness of plants, nonhuman animals, and household objects (Yamanashi Leib, Kosovicheva, & Whitney, 2016).

Although ensemble coding helps circumvent computational limitations by compressing information into a "gist" representation that characterizes the group, information about individuals can be lost to conscious access (Allik et al., 2014; Haberman & Whitney, 2007). Indeed, observers sometimes fail to perceive, attend to, and/or recall information about individual set members at all, or do so poorly (e.g., Alvarez & Oliva, 2009; Sweeny et al., 2015). This often occurs because ensemble paradigms typically present images for very brief durations (e.g., Haberman et al., 2009) but has also been shown in the context of neuropsychological deficits that disrupt the perception of individuals (Hochstein, Pavlovskaya, Bonneh, & Soroker, 2015; Robson, Palermo, Jeffery, & Neumann, 2018; Yamanashi Leib, Landau et al., 2012; Yamanashi Leib, Puri et al., 2012), or masking paradigms that limit awareness of constituents (Choo & Franconeri, 2010; Elias et al., 2018; Jacoby, Kamke, & Mattingley, 2013; Ward, Bear, & Scholl, 2016). Yet even in cases like these, information integration and ensemble coding can proceed and make a summary representation available to the perceiver. Thus, ensemble coding is not only fast, efficient, and robust, it is clearly useful. It can provide precise information about the general characteristics of visual information that is outside the focus of attention, or even about forgotten or unperceived individuals in a larger set. This information can then help guide attention in future moments (Alvarez & Oliva, 2009; Alvarez, 2011; Im et al., 2017), aid with detection (or discounting) outliers in the group (Haberman & Whitney, 2010), or even skew the perception of individuals toward the characteristics of the set, overall (Brady & Alvarez, 2011; Corbett, 2017; Khayat & Hochstein, 2018, 2019).

In contrast to information integration, the visual system can *segment* input that does not go together by, for example, exaggerating differences instead of integrating across them. By segmenting information, the visual system is able to avoid making generalizations across objects that do not belong together. 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 a larger, 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 are be perceptually exaggerated to appear different from each other (Morgan, 1999). This is a perceptual consequence of segmentation—objects are perceived to be more distinct than they really are. Similar distortions have been observed for perception of size (Fortenbaugh, Sugarman, Robertson, & Esterman, 2019), shape (Sweeny, D'Abreu, Elias, & Padama, 2017), spatial location (Badcock & Westheimer, 1985; DiGiacomo & Pratt, 2012; Suzuki & Cavanagh, 1997), and direction of motion (Thornton, 2002). Segmentation is likely involved in distinguishing figure (i.e., object) from ground (i.e., background; Grossberg, 1994; Westheimer & Levi, 1987) and can even act at during decision-making (Fritsche & de Lange, 2019; Zamboni, Ledgeway, McGraw, & Schluppeck, 2016).

What sorts of problems might segmentation help the visual system solve? Segmentation may be especially relevant for visual features that have a clear 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 gaze is neither leftward nor rightward). By perceptually exaggerating feature values away from category boundaries, segmentation may allow 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). It may reduce the likelihood of 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). It is important to note that visual categories like these are supported by the organization of the cells that encode a given visual feature (Dickinson, Morgan, Tang, & Badcock, 2017; Kayaert, Biederman, Op de Beeck, & Vogels, 2005; Storrs & Arnold, 2017), but they may also be learned (Kourtzi, 2010).

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 around feature category boundaries (Goldstone & Hendrickson, 2010). Aspect ratio is a feature with such a boundary. 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 (Biederman & Kalocsais, 1997; Treisman & Gormican, 1988). Aspect ratio is encoded by cells in inferotemporal (IT) cortex, separately from simpler visual features like size or curvature (Dickinson et al., 2017; Op de Beeck, Wagemans, & Vogels, 2003; Regan & Hamstra, 1992); these cells are likely organized in a multichannel manner (Dickinson et al., 2017; Storrs & Arnold, 2017). More relevant for our purposes, 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 priority 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 (Dickinson et al., 2017; Suzuki & Cavanagh, 1998; Sweeny, Grabowecky, Kim, & Suzuki, 2011; Sweeny et al., 2017; but also see Fritsche & de Lange, 2019; Kuang, 2019; Zamboni et al., 2016, for discussions about whether observations of repulsion always reflect changes in perception). In addition, 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 (Regan & Hamstra, 1992; Suzuki, 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 et al., 2005). In addition, fewer cells are tuned to values near the null-point, and they respond less strongly than those tuned to extreme values. The perceptual consequences of all this can be surprising. Aside from perceptual exaggeration away from the category boundary, it is easier to mask circles than extreme aspect ratios, likely because the neural representation of circles is relatively weak (Braun & Sweeny, 2019).

Despite this work described above, there is as of yet very little evidence that aspect ratio can also be integrated by ensemble coding, or any other process (see Oriet & Brand, 2013, for potential aspect ratio integration, though changes in the aspect ratios 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 midlevel visual feature, encoded in intermediate stages of the ventral visual hierarchy (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 integration and ensemble coding is known to act. Thus, our first goal was simply to examine if ensemble coding is capable of acting on aspect ratio.

Our primary goal was to investigate the interaction between integration and segmentation in the context of ensemble coding. We theorized that the distinction between these processes might be especially pronounced around category boundaries, where integration might act to compress information *toward* the boundary and segmentation would exaggerate information *away* from it. Previous work on ensemble coding has left the tension between integration and segmentation relatively unexplored (e.g., Elias et al., 2017; Haberman & Whitney, 2009; Sweeny & Whitney, 2014). Here we examined it directly. We used aspect ratio as our target feature because it is relatively simple and easy to manipulate, and more important, its perception and encoding is clearly organized around a null category boundary (e.g., Braun & Sweeny, 2019; Suzuki, 2005). We predicted that the process of ensemble coding should be particularly efficient for sets of shapes with aspect ratios that fall on one side of the null-point (e.g., only "flat-ish" ellipses), compared to sets with aspect ratios that cross the category boundary (flat *and* tall ellipses). In the former case, 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, integrated information should tend toward the set mean, thus supporting generalizations (or "rapid visual categorizations"; Utochkin, 2015). On the other, segmentation could distort perception of individual shapes 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. Although ensemble coding should be more precise for sets of generally flat objects, and for separate sets of generally tall objects, it should be less precise for sets that span the category boundary, containing both flat *and* tall objects.

We first conceived of and ran a pilot investigation[1] very similar to Experiment 1. After correcting design limitations present in the pilot, and guided by the results of it, we designed and built Experiments 1 and 2. To foreshadow the results of all experiments: in Experiment 1, we confirmed our prediction that ensemble coding operates less efficiently when it is pressed to integrate across the category boundary of aspect ratio, replicating the main results of our pilot investigation (see Footnote 1). Experiment 2 and a reanalysis of Experiment 1 highlight a possible mechanism: segmentation may have repulsed the appearance of ellipses away from the category boundary—the aspect ratios of ellipses may have appeared to be more extreme than they really were. This distortion introduced exaggerated heterogeneity, which ultimately disrupted integration.

Experiment 1

Method

Observers. After institutional review board (IRB) approval, 45 students ($M_{\text{age}} = 19.2$ years; 40 female) from the University of Denver participated in Experiment 1. Observers granted informed consent and had normal or corrected-to-normal visual acuity. The sample size for Experiment 1 was selected based on our observed effect size ($d = 0.6$) for detecting the presence of ensemble coding using a matched-pairs *t* test with 33 observers in a pilot investigation (see Footnote 1) with a nearly identical design and analysis. Assuming this same effect size, we would have needed a sample of 24 observers to obtain power of 0.8 with alpha set at .05 in Experiment 1. However, after correcting experimental design lim-

[1] In a pilot study ($n = 34$), we asked observers to judge the average aspect ratio of sets of four and eight ellipses. Additional trials displayed only one ellipse. Analyses confirmed the presence of ensemble coding, $t(33) = 3.48$, $p < .01$, $d = .6$, for sets of eight ellipses only (by comparing full-range multi-ellipse trials to single ellipse trials; the logic for this comparison is described below). This effect size informed our sample size for Experiment 1. However, some experimental design limitations were present in this pilot (e.g., ellipses were not perfectly equated for surface area). Experiment 1 corrected those limitations, and thus is described as our primary study here. Crucially, all the main results of the pilot study were replicated in Experiment 1. See the online supplemental materials for additional details.

itations present in the pilot, we anticipated a reduced effect size, and thus increased our sample size and set our stop rule at 45.

Stimuli. Our full stimulus set included 27 ellipses (0.2° thick lines) created in Adobe Photoshop CS6 v. 13.0 × 64, 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. In addition, 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 used in the sets. 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. In addition, 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 used in the sets. Thus, the stimulus range included 21 total ellipses. The response stage included those same 21 ellipses, plus six additional ellipses to avoid response compression. 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, 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 = 171, 171, 171; luminance = 41.5 cd/m^2). (Experiments were coded and run using MATLAB The MathWorks, Natick, MA) with the Psychophysics Toolbox (Brainard, 1997).

Stimulus arrays and trial types. 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 eight ellipses arranged in a globally diamond-shaped organization around a central fixation point (see Figure 1). Thus, multi-ellipse trials 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), and the encoding of aspect ratio is relatively insensitive to size (Regan & Hamstra, 1992). 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 (see Figure 1).

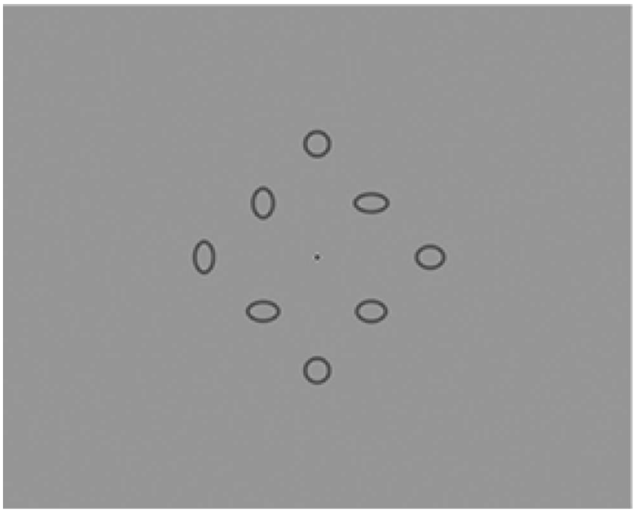

Figure 1. Layout of an eight-ellipse center trial, with fixation.

Multi-ellipse trials contained subconditions including *flat, tall, center, outlier* and *full-range* conditions (see Figure 2). On flat, tall and center trials, ellipses were drawn from a limited range of 11-aspect ratios from within the full set of 21 shapes (excluding the most extreme ellipses present only at the response stage). 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 11 aspect ratios: −0.463, −0.417, −0.371 −0.324, −0.278, −0.232, −0.185, −0.139, −0.093, −0.046, and 0.00 [circle]). Each flat trial was further randomly determined to be either a *low-seed flat trial* or a *high-seed flat trial* (see Figure 2). On low-seed flat trials, two ellipses were randomly selected from the three flattest aspect ratios in the flat range (i.e., −0.463, −0.417, and −0.371). The remaining six ellipses were randomly selected from the entire flat-ellipse range. On high-seed flat trials, two ellipses were randomly selected from the three ellipses with the least-flat aspect ratios (-.093, −0.046, and 0.00). The remaining ellipses were selected from the entire flat-ellipse range. We explain the rationale for this seeding procedure later.

Ellipses from tall trials could have had any of the following 11 aspect ratios: 0.00, 0.046, 0.093, 0.139, 0.185, 0.232, 0.278, 0.324, 0.371, 0.417, and 0.463. Tall trials were constructed similarly to flat trials: low-seed tall trials had two ellipses randomly selected from the three least-tall aspect ratios (0.00, 0.046, and 0.093), whereas high-seed tall trials had two ellipses randomly selected from the three tallest aspect ratios (0.371, 0.417, and 0.463). Note that flat and tall trials could both contain circles, which ensured that the presence of circles was not unique to any condition.

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 with aspect ratios from the center of the stimulus range (aspect ratios: −0.232, −0.185, −0.139, −0.093, −0.046, 0.00, 0.046, 0.093, 0.139, 0.185, and 0.232). Low-seed center trials had two ellipses randomly selected from the flattest three aspect ratios in the center range (−0.232, −0.185, and −0.139), while high-seed center trials had two ellipses randomly selected from the tallest three aspect ratios in this range (0.139, 0.185, and 0.232). These three trial types—flat, tall, and center trials—were our main conditions of interest. Henceforth, flat and tall trials are collectively referred to as "non-boundary" trials, since they never spanned the categorical boundary. Center trials are referred to as such.

We used this low- and high-seeding system so that the distribution of aspect ratios in each set would always be 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 later be mistaken for true, perceptual extraction of the mean in our analy-

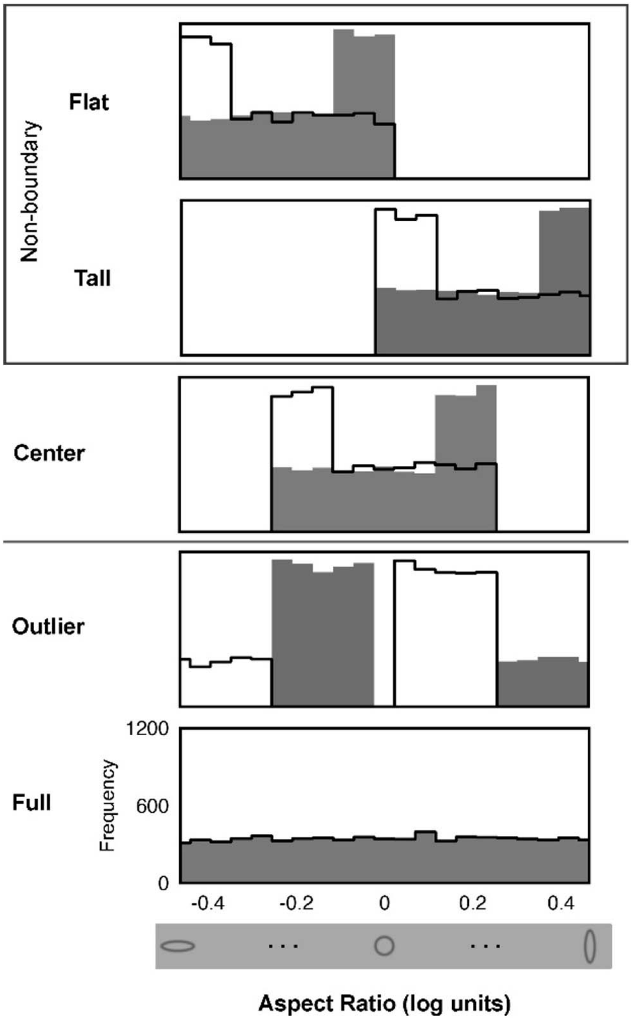

Figure 2. Multi-ellipse conditions from Experiment 1. Histograms denote the frequency with which each ellipse from the stimulus range appeared across the entire experiment, across all flat trials and tall trials (which together formed nonboundary trials), as well as all center trials—the three main conditions 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.

ses. In addition, by using low- and high-seeds, we could, in theory, determine whether observers were extracting the mean or the median, across trials of any given type, at the data-analysis stage (see the online supplemental materials); recent work has shown that people inadvertently encode information about the distribution of a set during ensemble tasks (Chetverikov, Campana, & Kristjánsson, 2017), including mean, skew, and variance (Khayat & Hochstein, 2018; Oriet & Hozempa, 2016), in as little as 100 ms (Michael, De Gardelle, & Summerfield, 2014).

Importantly, because the precision of ensemble coding is known to decrease as the heterogeneity of a set increases, (e.g., Dakin, 2001; Haberman, Lee, & Whitney, 2015; Im & Halberda, 2013; Marchant, Simons, & de Fockert, 2013; Morgan, Chubb, & Solomon, 2008), we wanted to ensure that set heterogeneity was comparable across nonboundary and center trials. Before running the experiment, we confirmed that our design parameters should result in comparable set heterogeneity for nonboundary and center trials by running a 100-trial simulation of Experiment 1, iterated 500 times. This showed that nonboundary trials contained heterogeneity ($M = 0.1439$ log units, $SD = 0.0022$) comparable to center trials ($M = 0.1440$, $SD = 0.002$), $t(499) = .85$, *ns*.

Multi-ellipse trials could also include *outlier* or *full-range* trials. Outlier trials could be either low-seed or high-seed. On low-seed outlier trials, two ellipses were randomly selected from the five flattest aspect ratios; these ellipses were the "outliers". The remaining six ellipses were randomly selected from the tall ellipses—specifically, from aspect ratios 0.046, 0.093, 0.139, 0.185, and 0.232. Similarly, on high-seed outlier trials, two ellipses—the "outliers"—were randomly selected from the tallest five ellipses. The remaining ellipses were randomly selected from the flat range, specifically from aspect ratios −0.232, −0.185, −0.139, −0.093, and −0.046. This approach produced trials in which the majority of ellipses were generally flat, while the outliers were tall, and vice versa. Finally, on full-range trials, all eight ellipses were randomly selected from the entire 21 ellipse range of the stimulus set (see Figure 2).

We reasoned that integration would produce less precise estimates of mean aspect ratio on outlier trials than on full-range trials. First, outliers tend to be discounted during ensemble coding 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, we reasoned that, despite the presence of boundary-crossing in both trial types, outlier trials would show reduced evidence of integration compared to full-range trials. Since it proved difficult to perfectly equate heterogeneity across outlier and full-range trials, we intentionally constructed outlier trials to have slightly less heterogeneity than full-range trials. Thus, if outlier trials did show evidence of reduced integration compared to full-range trials as we predicted, it could not be because they simply had more heterogeneity.

Apart from multi-ellipses trials, we 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 evaluated the aspect ratio of this single ellipse. However, we also randomly selected seven additional ellipses from the full range on each trial, *as if* we were generating a full-range multi-ellipse trial to display. Importantly though, these additional "invisible" ellipses were not displayed. A group mean was nonetheless calculated, and observer error relative to this group mean could be computed. Observer error on these trials, relative to the mean of the entire set (which they could not see), allowed us to quantify the magnitude of error one would expect on the actual multi-ellipse trials if observers simply responded to one random ellipse, without integrating aspect ratio information. 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 and ensemble coding would be present if observer error on true full-range multi-ellipse trials was reduced compared to error on single-ellipse trials. Less important, 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.

Procedure. 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 1200 ms. Next, a multi-ellipse array or a single-ellipse array was displayed for 250 ms. Aspect ratio information can be extracted at extremely brief durations (Sweeny 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 for an additional 500 ms, which prevented the upcoming response screen from being perceptually 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 a set of ellipses that included the full 21 ellipses that could be present in a given trial, plus the three extremely flat and three extremely tall ellipses described above (see Stimuli section). Observers reported their estimate of the set's average aspect ratio (or the individual ellipse's aspect ratio on single-ellipse trials) by moving the mouse left or right, which incrementally adjusted the response ellipse's aspect ratio across the stimulus range. If, for example, the initial response ellipse happened to be a circle, moving the mouse leftward would increase the flatness (and decrease the tallness) of the response ellipse by animating across the stimulus set, one ellipse at a time. After the flattest ellipse in the set was reached, the response ellipse would then begin to smoothly increase in tallness (and decrease in flatness). 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 having been 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 ratios 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), which was intended to further reduce unwanted effects of response compression. When observers reached their desired response, they 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 before the next trial began.

Results

Our primary interest was 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. We began by computing the error of each observer's response relative to the mean aspect ratio of 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 0.046 (too tall relative to the set mean), their error on that trial would be + 0.046. Conversely, negative error values indicated a response that was too flat. For each observer, we compiled these signed-difference scores into separate error distributions, one for each condition (flat, tall, center, outlier, full-range, center and single trials). Next, we calculated the standard deviation of each observer's error distributions, for each condition. Greater sensitivity to mean aspect ratio would produce error distributions with smaller standard deviations. This approach has been used in previous investigations of ensemble coding (e.g., Elias et al., 2017; Haberman & Whitney, 2009; Sweeny et al., 2013; Sweeny & Whitney, 2014). This analysis yielded overall error scores (i.e., the *SD* of an observer's error distribution) for each condition, and for each observer.

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. Crucially, estimates of the set's average aspect ratio were more precise on full-range trials (M SD = .28, SD = .08) than they were on single-ellipse trials (M SD = .3, SD = .04), $t(44) = 2.78$, $p < .01$, $d = 0.42$.

We had no a priori hypothesis regarding performance on flat versus tall trials, so for each observer, we averaged flat and tall *SD* values to yield a measure of their performance on *nonboundary* trials. Critically, observers performed better on nonboundary trials than they did on *center* trials (M SD nonboundary trials = .18, SD = .06; M SD center trials = .21, SD = .06), $t(44) = 4.64$, $p <$ .01, $d = 0.69$. Thus, observers showed evidence of integration and ensemble coding, especially for sets that did not span the category boundary. These main results are summarized in Figure 3.

Secondary results. Of secondary interest was the comparison between outlier trials and full-range trials. Contrary to our prediction, observer performance was better on outlier trials (M SD = .26, SD = .07) than full-range trials (M SD = .28, SD = .08), $t(44) = 2.5$, $p = .02$, $d = .37$. However, as intended, outlier trials did contain slightly less heterogeneity (M SD = .24, SD = .02) than full-range trials (M SD = .27, SD = .05), which could account for some—or all—of this performance advantage. Indeed, we examined the set heterogeneity present on a trial-by-trial basis in relation to the absolute magnitude of observer error, across all trial types. Across observers, the relationship between set heterogeneity and observer error was positively correlated (M slope: .67, SD = .02). We then considered the average amount of heterogeneity on outlier trials (M SD = .24, SD = .02) with the average amount of heterogeneity on full-range trials (M SD = .27, SD = .05). The linear relationship between heterogeneity and error magnitude predicted that, for the increase in heterogeneity from the outlier to the full-range trials, observer error should have increased

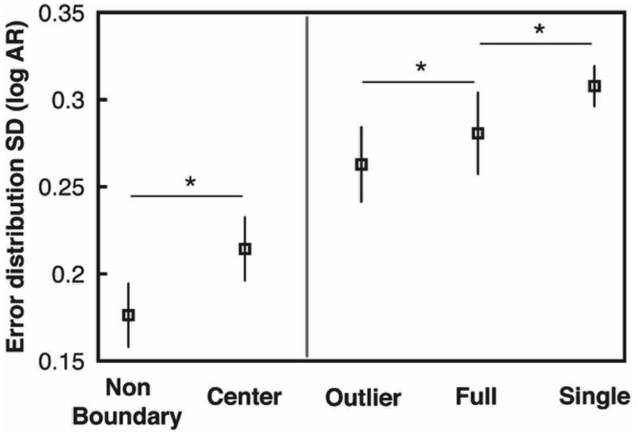

Figure 3. Observers were more sensitive to the average aspect ratio of sets that included only flat or tall ellipses (i.e., nonboundary trials), compared to when sets included flat and tall ellipses (i.e., center trials). Observers were also more sensitive to the average aspect ratio of full-range trials compared to single trials; this indicated the presence of ensemble coding. Finally, observers were more sensitive to the average of outlier trials compared to full-range trials. Error bars in all figures represent 95% confidence intervals. Starred comparisons represent $p < .05$.

by .02 log units. Indeed, this was very close to the observed difference in error magnitude between outlier (*M* error log units = .23, *SD* = .07) and full-range trials (*M* error log units = .25, *SD* = .08). Thus, the difference in performance between the outlier and full-range conditions was likely due simply to an imbalance in heterogeneity.

We in addition wanted to determine whether, on nonboundary trials, observers' errors were *systematic.* Although observers were most sensitive to the mean of nonboundary-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, we computed each observer's average signed error separately for flat trials and for tall trials. Across observers, mean estimates of flat trials exaggerated the mean flatness of the set (*M* error = −.05, *SD* = .08), $t(44) = 3.99$, $p < .01$, $d = .6$. Likewise, mean estimates of tall trials exaggerated the mean tallness of the set (*M* error = .05, *SD* = .08), $t(44) = 4.25$, $p < .01$, $d = .63$.

Next, we addressed potential alternative explanations of our primary result. We began by considering whether response compression could account for the observed results, despite our effort (e.g., the extended response range) to mitigate the potential effects of compression from response-stage endpoints. If observers avoided responding with the most extreme ellipses from the response range, then the distributions of responses would have been compressed away from these endpoints, toward the center of the response range. Response compression like this would have narrowed error distributions and thus reduced overall estimates of error. Response compression would have been 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.

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 high-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. We calculated the *SD* of each observer's error distributions for flat trials with a high seed, and separately, tall trials with a low seed. We then averaged these two values for each observer to yield a value that represented performance when response compression was *less* likely. We also calculated performance across flat trials with a low seed and tall trials with a high seed—trials in which potential response compression was *more* likely. Across observers, performance on trials in which response compression was more likely (*M SD* = .16, *SD* = .06) was better than on trials in which it was less likely (*M SD* = .18, *SD* = .07), $t(33) = 2.55$, $p = .02$, $d = 0.38$. Importantly, however, even when we considered only trials in which response compression was less likely (*M SD* = .18, *SD* = .07), observer performance was still better on nonboundary trials than on center trials (*M SD* = .21, *SD* = .06), $t(33) = 4.09$, $p < .01$, $d = 0.61$. This suggests that response compression cannot fully account for our main results.

Next, we examined the possibility that observers were not actually integrating information but were instead simply responding from the midpoint of the relevant stimulus range on each trial. We first 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 low-seed trials, and flat, center, and tall high-seed trials). We then 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) 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 (*M* AR = −.34, *SD* = .09) than on flat high-seed trials (*M* AR = −.22, *SD* = .09), $t(44) = 11.6$, $p < .01$, $d = 1.73$. This pattern persisted for center low-seed trials (*M* AR = −.06, *SD* = .06) and center high-seed trials (*M* AR = .03, *SD* = .05), $t(44) = 7.39$, $p < .01$, $d = 1.1$, as well as tall low-seed trials (*M* AR = .22, *SD* = .07) and tall high-seed trials (*M* AR = .34, *SD* = .09), $t(44) = 10.98$, $p < .01$, $d = 1.64$. Thus, observers were not simply responding from the midpoint of the appropriate range on a trial-by-trial basis.

Finally, we examined the relationship between set heterogeneity and observer error. Recall that prior to Experiment 1, simulations using Experiment 1's parameters predicted a comparable amount of heterogeneity for nonboundary 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 slightly less heterogeneity in nonboundary trials (*M*

SD = .151, *SD* = .003) than in center trials (*M SD* = .153, *SD* = .005), *t*(44) = 2.55, *p* = .02, *d* = .38. Thus, heterogeneity was not perfectly controlled for in Experiment 1, despite careful efforts to do so. To investigate whether this subtle difference in heterogeneity could completely account for our main results, we modeled the relationship between heterogeneity and observer error. On each trial, we recorded the *SD* of all the aspect ratios present in that set. Next, we computed the absolute error magnitude, in log units, between an observer's response and the true trial mean. For each observer, across all trials, we then computed the relationship between heterogeneity and the magnitude of observer error. This linear relationship was positive (*M* slope: .67, *SD* = .02), *t*(44) = 241.14, $p < .01$, *d* = 35.95. We then used this relationship to compute the expected difference in observer-error magnitude between the nonboundary and center trials, given their actual difference in heterogeneity. The relationship between heterogeneity and observer error predicted an increase in error magnitude of .0013 log units between these conditions. However, the observed difference in error magnitude between nonboundary (*M* log units = .16, *SD* = .05) and center trials (*M* log units = .19, *SD* = .05) was more than 18 times greater than predicted. This approach was admittedly post hoc, but it suggests that variation in heterogeneity between trial types cannot fully account for observers' increased sensitivity to the mean on nonboundary trials, relative to center trials.

Discussion

Replicating the main results of our pilot study (see the online supplemental materials), observers in Experiment 1 were more sensitive to the mean aspect ratio of sets that did not span the category boundary compared to sets that did. Observers in Experiment 1 appeared to make their summary judgments using the mean of sets of aspect ratios. They also appeared to be sensitive to skew in distributions of aspect ratio. The sensitivity of ensemble judgments was impacted by the heterogeneity of aspect ratios in our set, but more important, ensemble coding operated best on sets that did not span a category boundary.

There are several reasons to suspect that aspect ratio was the feature integrated in Experiment 1, although it is important to consider alternatives. For example, perhaps after a categorical tall-flat analysis, one-dimensional elongation was summarized instead. This alternative is plausible, and we acknowledge that it may have occurred, but we consider it as being unlikely. First, the visual system has cells devoted to aspect ratio (e.g., Dumoulin & Hess, 2007; Kayaert et al., 2005). If length in the vertical or horizontal direction was integrated in the current investigation, and aspect ratio was not, then that length integration would have to occur *despite* the presumed activity of populations dedicated to aspect ratio. Second, a prior investigation found that height and width information are both integrated by observers automatically, even when the task is to integrate only height or width, separately (Oriet & Brand, 2013). Nonetheless, we acknowledge that we cannot unambiguously confirm that aspect ratio was the integrated feature in Experiment 1. Most importantly, though, our main results—that ensemble coding is impeded when values span the category boundary—are not contingent on doing so.

Experiment 2

We suggested already that segmentation (specifically, perceptual distortion) would be a plausible mechanism for our prediction of less precise ensemble coding when information in a set spans a category boundary. This proposed effect of perceptual distortion should be particularly pronounced for aspect ratios near the category boundary, and less pronounced or absent for more extreme values (Suzuki, 2005; Suzuki & Cavanagh, 1998; Sweeny et al., 2012). Experiment 1 provided some evidence for distortion away from the boundary—observers exaggerated the "flatness" of flat sets and the "tallness" of tall sets. In Experiment 2, we tested this prediction more directly. Below, we describe how distortion around the category boundary could influence ensemble perception.

Consider center trials, in which sets are composed of flat and tall ellipse. Segmentation could distort the appearance of this type of set's constituents, such that slightly flat ellipses would appear flatter and slightly tall ellipses would appear taller. The net effect across these trials would be an increased range of perceived aspect ratio, and thus increased perceptual heterogeneity in these sets. In contrast, on flat trials, segmentation would distort the perception of ellipses near the category boundary toward the flatter set mean, reducing the amount of perceived heterogeneity in the set (a similar pattern, in the opposite direction, would occur for tall trials). Considering that heterogeneity is known to disrupt the integrative process of ensemble coding (e.g., Dakin, 2001; Haberman et al., 2015; Im & Halberda, 2013; Marchant et al., 2013; Morgan et al., 2008), this potential effect of perceptual distortion would account for the disrupted integration on center trials observed in Experiment 1. Looked at the other way, perceptual distortion away from the category boundary may account for the improved performance on nonboundary trials. But first, does this kind of distortion actually occur in sets of ellipses like those used in Experiment 1?

The purpose of Experiment 2 was to examine whether repulsive mechanisms influenced the perception of individual ellipses. Specifically, we presented sets of ellipses or individual ellipses and, using a postcue, asked observers to evaluate the aspect ratio of individual ellipses. We 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. We predicted that errors in aspect ratio judgments would follow an s-shaped pattern (the first derivative of a Gaussian function; Figure 4), with the highest magnitude of repulsive distortion near the category boundary, and a gradual decay of distortion for progressively flatter (or taller) aspect ratios. This pattern of distortion has been observed for other visual features (e.g., Crane, 2012; Sweeny et al., 2012).

Method

Observers. In Experiment 1, observers did integrate information, as indicated by the comparison between *full-range* and *single-ellipse* trials. Experiment 1 and Experiment 2 were conceived of and created at the same time. The task in Experiment 2 required observers to make judgments about individual objects in crowds, and previous work suggests that this can be difficult or even impossible (e.g., Allik et al., 2014; Haberman & Whitney, 2007). So, based on the same power analysis used in Experiment 1, we

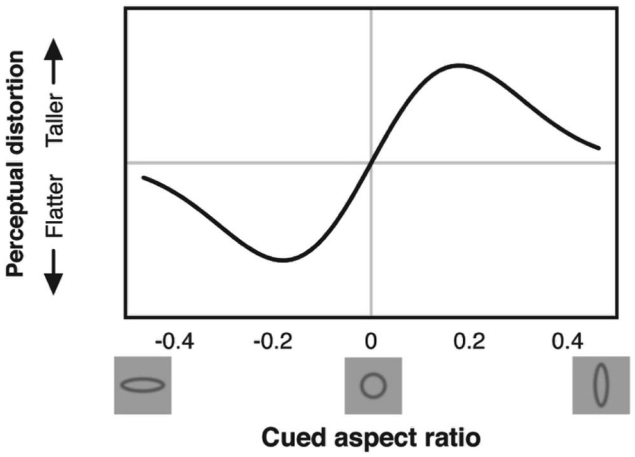

Figure 4. Idealized depiction of perceptual distortion around a category boundary. Ellipse aspect ratio is plotted on the *x*-axis from flattest to tallest, and systematic error in aspect ratio perception (i.e., distortion relative to the cued shape's actual aspect ratio) is plotted on the y-axis. In this example, flat ellipses near the category boundary are perceived to be even flatter, and 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.

increased the sample size in anticipation of what we expected to be a difficult task and recruited 45 students (M_{age} = 19.2 years; 34 female) from the University of Denver, with IRB approval.

Stimuli. Stimuli in Experiment 2 were identical to those used in Experiment 1.

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 Experiment 1.

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°) appeared 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 Experiment 1. Observers were then presented with the same method-of-adjustment response screen used in Experiment 1, except in Experiment 2 they were instructed to "indicate the cued tallness or flatness." Observers were allowed as many practice trials as they wished beforehand, and the experimenter confirmed that observers understood the instructions.

Results

We began by quantifying how much repulsion, if any, occurred for the cued ellipse on each 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, because 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 would have occurred. In contrast, if an observer responded with aspect ratio 0.185, 0.046 log units of perceptual repulsion would have occurred, since the slightly tall ellipse would have been 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, we computed an average repulsion index for each cued aspect ratio. Across observers, we 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 taller than they actually were, whereas extremely tall aspect ratios were rated as flatter than they actually were (Figure 5A). Overall these errors produced a linear pattern with a negative slope, not the s-curved shape of repulsion as we predicted. These results do not reflect a perceptual effect of attraction toward the middle of each aspect ratio category (Brady & Alvarez, 2011; Corbett, 2017), since this guessing pattern persisted, even when we isolated flat and tall trials separately. Rather, we suggest that this pattern is consistent with observers guessing during a very difficult task. In fact, simulations from previous work illustrate how guessing would produce this same pattern (Sweeny et al., 2012). If, for example, a very flat aspect ratio was cued, and an observer 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, 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 parsimoniously account for the pattern of results seen across multi-ellipse trials. To be clear, we did not predict this, although in hindsight it makes sense, especially if observers were unable to 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., Allik et al., 2014; Haberman & Whitney, 2007).

In contrast to multi-ellipse trials, the overall distribution of repulsion indices for single aspect ratios around the category boundary did conform to our hypotheses, following an s-shaped curve (Figure 5B). 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 first derivative of a Gaussian function ($R^2 = .86$, $p < .01$; Figure 5B). A corrected Akaike information criterion analysis confirmed with 99.99% certainty that the fit for the first derivative of a Gaussian characterized the pattern of data better than a linear fit ($R^2 = .33$, *ns*). The first derivative of a Gaussian function was given by $y = xawce^{-(wx)2}$, where *a* determined the amplitude of the curve peaks, *w* scaled the curve width, and *c* was a constant. The values for *a*, *w*, and *c* were .1162, 15.58, and .3206, respectively.

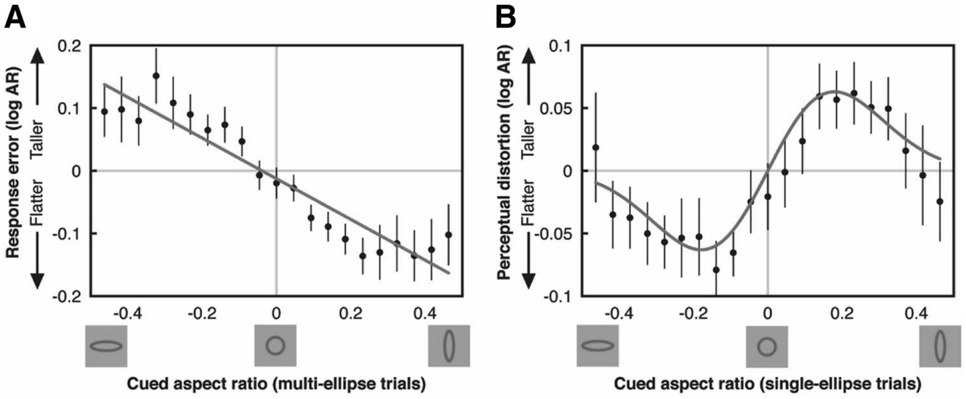

Figure 5. Cued ellipse response error on multi-ellipse trials (A). Ellipse aspect ratio is plotted on the X-axis from flattest to tallest, and systematic error in observer response is plotted on the Y-axis. In response to extremely flat ellipses, observers erred with tall values; vice versa in response to extremely tall ellipses. Data points linearly fit (straight line), with a negative slope indicative of guessing. Perceptual distortion around the category boundary on single-ellipse trials (B). Flat ellipses near the category boundary were perceived to be even flatter, whereas tall ellipses near the boundary were perceived to be taller than they actually were. Data points were fit with the first derivative of a Gaussian (curved line). Error bars in both figures represent 95% confidence intervals.

Discussion

Overall, Experiment 2 suggests that on single-ellipse trials, distortion did occur, and it occurred for some ellipses more strongly than others. Slightly flat ellipses were reported to be flatter than they actually were, and slightly tall ellipses were reported as taller than they actually were.

Interestingly, distortion was not evident on multi-ellipse trials, in contrast with our prediction. As a result, we acknowledge that it is possible that no such distortion occurred on these trials. On the other hand, we suggest that the absence of distortion on multi-ellipse trials is consistent with a common finding whereby observers 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., Alvarez & Oliva, 2008; Haberman & Whitney, 2007; Neumann, Ng, Rhodes, & Palermo, 2018). In fact, computing group-level summaries has been shown to detract from learning about individual set members (Zhao, Ngo, McKendrick, & Turk-Browne, 2011) unless perceivers are encouraged to pay attention to individuals when making summary judgments (Hall, Mattingley, & Dux, 2015). Integration 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 individual aspect ratios in multi-ellipse sets were encoded to some degree, otherwise integration (such as that observed in Experiment 1) 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 in Experiment 1.

Reanalysis of Experiment 1

If the perception of individual aspect ratios in multi-ellipse trials in Experiment 1 was distorted in a manner compatible with the pattern observed on single-ellipse trials in Experiment 2, then the error of observers' ensemble estimates from Experiment 1 should be reduced when calculated relative to set means based on *distorted* values from Experiment 2 rather than the aspect ratios that were physically present.

Recall that in Experiment 1, observer error was computed relative to the mean of the actual aspect ratios displayed. However, Experiment 2 suggests that observers may not have encoded (or perceived, in the case of single trials) those aspect ratios veridically; they may have been distorted. Thus, if we retrospectively relabeled the *displayed* aspect ratios in Experiment 1 with the *distorted* aspect ratio values from Experiment 2 (taken from the first derivative of a Gaussian fit from single-ellipse trials), then estimates of observer error in Experiment 1 should be systematically reduced. Investigating this possibility was the aim of our reanalysis of Experiment 1.

We predicted that observer error would decrease overall, when error values from Experiment 1 were recalculated relative to the aspect ratios that observers likely perceived and integrated. More specifically, we expected that recalculated observer accuracy would improve the most for transformed nonboundary trials, compared to transformed center trials. Our reasoning for this prediction was 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 should not only change the values of individual aspect ratios in nonboundary sets, it should also significantly shift the set mean of nonboundary trials. If observers were basing their mean judgments on distorted values on these nonboundary 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 (recalculation) 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 nonboundary trials. This leads to the second prediction for our reanalysis of Experiment 1. We expected encoded heterogeneity (as opposed to the actual heterogeneity) to increase for center trials more than for nonboundary trials, for the reasons described above. In sum, after transformation, we expected observers' mean estimates to become especially more accurate when extracting the mean of nonboundary trials. In addition, we expected encoded heterogeneity to increase more for center trials than for nonboundary trials.

Method

We transformed the data from Experiment 1 guided by the s-shaped curve obtained in Experiment 2. For each trial, we replaced the actual aspect ratio of each ellipse in a set with its corresponding value from the fit in Figure 5B. Set means from every trial in Experiment 1 were then recalculated using these transformed values, and new error values were calculated relative to these transformed means.

For example, say that one of the eight aspect ratios on a given trial was −.139, somewhat flat. That same aspect ratio yielded a value of −.0597 log units of distortion in Experiment 2. Thus, we presume that aspect ratio −.139 was likely to have been perceived as an aspect ratio with a value of −.19871, flatter than it actually was. We would then replace the original aspect ratio (−.139) with this new, distorted aspect ratio (−.19871). We repeated this process for every aspect ratio present in every set in the recorded data from Experiment 1, which yielded a new, transformed Experiment 1 data set, referred to as such henceforth. Note that we did not transform observer responses in this way, since observers were free to deploy focused attention to the response ellipses, for an unrestricted amount of time (*M* reaction time [RT] = 2.28 s, *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 1's original, untransformed trial data. In other words, did observer accuracy improve, once the distortion described in Experiment 2 was taken into account? Note that, in this case, we considered mean absolute error to be a more appropriate measure of observer performance, as opposed to the *SD* of a signed error distribution. This is because mean absolute observer is capable of capturing shifts in a distribution's mean when overall *SD* remains constant, and such shifts were a predicted outcome after transformation. In addition to simply assessing whether accuracy increased, it was critical to determine if that increase was greater for nonboundary trials than for center trials.

We began by computing average error magnitude for each of our main three trial types (flat, tall and center trials). Averaging across these transformed trial types yielded an average transformed error magnitude score for each observer. We repeated this process on Experiment 1's original, untransformed main three trial types. Next, for each observer, we subtracted the transformed average error magnitude score from their untransformed error magnitude score; this difference score reflected the amount of change in error magnitude (averaged across all three of our main trial types) between the transformed and untransformed data sets. We 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 (*M* log units = .006, *SD* = .007), $t(44) = 5.36$, $p < .01$, $d = .8$ (Figure 6A). We 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 nonboundary trials. We did the same for transformed center trials. Across observers, as predicted, observers evidenced a larger error magnitude change index for nonboundary trials (*M* log units = .0078, *SD* = .0098) than for center trials (*M* log units = .0022, *SD* = .0049), $t(44) = 4.45$, $p < .01$, $d = .66$ (Figure 6A).

Next, we investigated a crucial prediction of Experiment 2. We hypothesized that encoded (as opposed to veridical) heterogeneity would increase for center trials more than for nonboundary trials, once distortion was taken into account. To examine this, we began by computing the heterogeneity present on every trial (i.e., we computed the *SD* of every trial's eight ellipses), for both the transformed and original, untransformed Experiment 1 data set. For each observer, we then computed the average heterogeneity present for flat, tall and center trials. We did this for both data sets. Mirroring the logic described immediately above, we subtracted each observer's average untransformed flat trial (e.g.) 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 untransformed and trans-

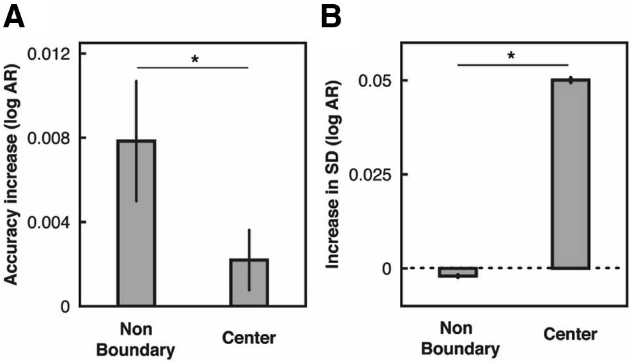

Figure 6. After transforming data from Experiment 1 in a manner retrodicted by Experiment 2's derivative of a Gaussian fit and reanalyzing that data relative to the resulting transformed trial means, observer accuracy improved, particularly on nonboundary trials (A). Additionally, after transformation, average set heterogeneity increased for center trials, and very slightly decreased for nonboundary trials (B). Error bars in both figures represent 95% confidence intervals. Starred comparisons represent $p < .01$.

formed data sets, for each observer. We repeated this process for tall and center trials, and called this difference score the "change in heterogeneity index". In general, a positive (or negative) index represented an 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 nonboundary trials, specifically.

Across observers, the change in heterogeneity index for nonboundary trials was slightly negative (M log units $= -.002$, $SD =$ .001), reflecting a slight decrease in heterogeneity when distortion was taken into account (Figure 6B). In contrast, the change in heterogeneity index for center trials was positive (M log units $=$ .05, $SD = .001$), reflecting a large increase in heterogeneity for center trials. The two indices were significantly different from one another, $t(44) = 172.83$, $p < .01$, $d = 25.76$ (Figure 6B).

Discussion

Our reanalysis of Experiment 1 is compatible with two conclusions. First, as predicted, observers' accuracy was significantly improved once the distortion described in Experiment 2 was taken into account. This was particularly true for nonboundary trials. Thus, in Experiment 1, observers may have been basing their judgments on distorted, rather than veridical, aspect ratios on both single- and multi-ellipse trials. Further, our reanalysis suggests that the distortion described by Experiment 2 (Figure 5B) is a reasonable model of what observers encoded in Experiment 1. Nonetheless, we acknowledge that this is indirect evidence of the quantity and quality of distortion that may have occurred on multi-ellipse trials and should be treated with caution.

Although the model of distortion used in our reanalysis is reasonable, it is also likely imperfect. It may not, for instance, account for late-stage ensemble noise (Baek & Chong, 2019; Solomon, Morgan, & Chubb, 2011). It is perhaps more surprising, then, that such an imperfect model yielded an increase in observer accuracy in our reanalysis of Experiment 1.

Importantly, distortion may be responsible for observers' disrupted ability to integrate information that spans the category boundary. By incorporating the distortion observed in Experiment 2, our reanalysis of Experiment 1 suggests that encoded (as opposed to veridical) heterogeneity increased precisely for the sets that observers had difficulty integrating. As we stated already, heterogeneity is known to disrupt integration and ensemble coding. Together, Experiment 2 and the reanalysis of Experiment 1 suggest that heterogeneity need not be veridical in order to disrupt integration; it can potentially be a result of the distortion—the segmentation—that occurs near a visual feature's category boundary.

General Discussion

This investigation produced several important, novel findings. First, integration in the form of ensemble coding operated on sets of shapes with different aspect ratios, but this integration was disrupted if set members included values that spanned the category boundary. In other words, observers were most sensitive to the mean of groups that were *either* flat *or* tall, not both. Second, in line with previous work, values of single aspect ratios near the category boundary were distorted—segmented—away from that boundary when seen for a brief duration. Flat ellipses appeared flatter, and tall ellipses appeared taller (although we acknowledge that distortions around category boundaries may not always be perceptual; Fritsche & de Lange, 2019; Kuang, 2019; Zamboni et al., 2016). The distortion observed for individual ellipses may also have occurred for the perception of multiple ellipses seen in a set. Even though observers were unable to accurately report on the aspect ratios of individual shapes within a set, our reanalysis is compatible with the contention that distorted values were still encoded and integrated. Finally, encoded heterogeneity was potentially exaggerated for groups that included both flat and tall ellipses. This exaggerated heterogeneity may have contributed to the visual system's disrupted ability to integrate information across category boundaries. These results deepen the field's understanding of integration by suggesting that it can interact with, and even conflict with, another fundamental computational method in the visual system—segmentation.

Although our results highlight the role of heterogeneity in guiding the precision of summary perception, we also acknowledge a potential role of visual grouping in the context of category boundaries. For example, spatial proximity alone can unite objects into coherent groups (see, e.g., Chong & Treisman, 2003; Palmer, Brooks, & Nelson, 2003), and summary statistics can then be extracted from those groups (Chong & Treisman, 2003). But grouping via spatial proximity can also aid ensemble coding; mean statistics can be extracted from spatially proximate groups with particular accuracy (Im & Chong, 2014; similarly, grouping and integration can aid visual search by making oddballs more salient; Utochkin & Yurevich, 2016). Interestingly, grouping can also sometimes produce less accurate summary judgments, perhaps by creating gestalt objects across which integration does not act (Cha & Chong, 2018). Indeed, people seem to use the underlying distribution of feature values across a set as a cue either to integrate information into a single set or to form separate summaries of distinct subsets (Utochkin, 2015; Utochkin, Khvostov, & Stakina, 2018). In general, grouping seems to "gate" integration; it helps the visual system determine which information to average across, and which information to exclude from ensemble computations. Integration seems to operate particularly well on grouped information, perhaps regardless of which grouping principle (e.g., common region) facilitated that grouping. For instance, it is not just spatially proximate objects that are grouped and integrated. Mean statistics can be computed for groups that are segregated by midlevel features like color (Brady & Alvarez, 2011; Chong & Treisman, 2005; Corbett, 2017). 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), perhaps via grouping cues like common fate (Palmer, 1999) or the principle of synchrony (Palmer, 2002). So, grouping may also contribute to the efficient and useful 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 here is compatible with the contention that segmentation interfered with integration via encoded heterogeneity. But, information that spans a feature category boundary may also be resistant to perceptual grouping. 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; Peterson & Kimchi, 2013; Wagemans et al., 2012). 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 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, and subsequently received additional attentional resources devoted to refined processing of the gestalt's constituent parts (in this work, the ellipses themselves; Flevaris, Martínez, & Hillyard, 2013). This possibility seems likely, because grouping can facilitate averaging of remembered items in a set (e.g., Corbett, 2017). Regardless, though, the evidence presented here suggests that segmentation can 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.

As disruptive as segmentation may be to the mechanisms of ensemble coding, in this investigation, it likely did not prevent them 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 we 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, we suspect that segmentation acted first. After all, in this investigation, as in others (e.g., Alvarez & Oliva, 2008; Haberman & Whitney, 2007; Neumann et al., 2018), observers may have encoded and then integrated distorted 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.

Some prior work on summary judgments of size suggests that approximately three to five items are sampled from a set (Gorea, Belkoura, & Solomon, 2014; Im & Halberda, 2013), or as a more general rule, the square root of the set size (Dakin, 2001, see Whitney & Yamanashi Leib, 2018). Here, though, 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. From one perspective, 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. Given the positive relationship between heterogeneity and observer error described in the present work, this perspective suggests that it is unlikely that observers integrated information from all eight ellipses in each multi-ellipse set. Alternatively, it is possible that observers encoded some information, or degraded 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, as well as late-stage noise associated with the ensemble itself, may have contributed the pattern of observer error reported here (Alvarez, 2011; Baek & Chong, 2019; Brady, Konkle, & Alvarez, 2011; Neumann et al., 2018; Solomon et al., 2011). If this second perspective 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 (both early noise and later, ensemble-related noise). Discriminating between these two perspectives is beyond the scope of this investigation.

In addition, in the present work, it is not possible to say precisely which ellipses were sampled, assuming that only a subset was sampled. For example, it is possible that observers tended to sample more extreme ellipses (Kanaya, Hayashi, & Whitney, 2018, but see, e.g., Haberman & Whitney, 2010). If true, this tendency may have been most pronounced on nonboundary trials in the current investigation. This is compatible with observers' tendency to 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 "oversampling" account of group mean exaggeration, and the account that relies on the distortion of individual ellipses (described above), are not incompatible. While we acknowledge that the oversampling of extreme shapes may have contributed to our results, we do not think it fully accounts for them. After all, in Experiment 1, we found no evidence of reduced mean sensitivity on the trials most likely to contain extreme ellipses (see the online supplemental materials for more details).

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 category boundaries 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 et al., 2011). And, in general, perceptual bias (e.g., distortion) may be a function of how noisily a feature is encoded in the first place—the noisier the representation, the more distortion is needed to avoid random perturbations across the category boundary (Wei & Stocker, 2017). Thus, it is possible 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 straightforward candidate feature to investigate this interaction. Future ensemble coding work, and integration work in general, may benefit from considering the conflict between segmentation and integration demonstrated here.

The first sentences of an untold number of vision science papers, this one among them, run 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 processes that occur “under the hood.” 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—that is, 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 perception 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. http://dx.doi.org/10.1016/j.visres.2014.05.003

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

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

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,* 7345–7350. http://dx.doi.org/10.1073/pnas.0808981106

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

Badcock, D. R., & Westheimer, G. (1985). Spatial location and hyperacuity: The centre/surround localization contribution function has two substrates. *Vision Research, 25,* 1259–1267. http://dx.doi.org/10.1016/0042-6989(85)90041-0

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,* 617–629. http://dx.doi.org/10.1068/p190617

Baek, J., & Chong, S. C. (2019). Distributed attention model of perceptual averaging. *Attention, Perception & Psychophysics, 82,* 63–79.

Biederman, I., & Kalocsais, P. (1997). Neurocomputational bases of object and face recognition. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences, 352,* 1203–1219. http://dx.doi.org/10.1098/rstb.1997.0103

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science, 22,* 384–392. http://dx.doi.org/10.1177/0956797610397956

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.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10,* 433–436. http://dx.doi.org/10.1163/156856897X00357

Braun, A., & Sweeny, T. D. (2019). Anisotropic visual awareness of shapes. *Vision Research, 156,* 17–27. http://dx.doi.org/10.1016/j.visres.2019.01.002

Brincat, S. L., & Connor, C. E. (2004). Underlying principles of visual shape selectivity in posterior inferotemporal cortex. *Nature Neuroscience, 7,* 880–886. http://dx.doi.org/10.1038/nn1278

Cha, O., & Chong, S. C. (2018). Perceived average orientation reflects effective gist of the surface. *Psychological Science, 29,* 319–327. http://dx.doi.org/10.1177/0956797617735533

Chetverikov, A., Campana, G., & Kristjánsson, Á. (2017). Representing color ensembles. *Psychological Science, 28,* 1510–1517. http://dx.doi.org/10.1177/0956797617713787

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research, 43,* 393–404. http://dx.doi.org/10.1016/S0042-6989(02)00596-5

Chong, S. C., & Treisman, A. (2005). Attentional spread in the statistical processing of visual displays. *Perception & Psychophysics, 67,* 1–13. http://dx.doi.org/10.3758/BF03195009

Choo, H., & Franconeri, S. L. (2010). Objects with reduced visibility still contribute to size averaging. *Attention, Perception & Psychophysics, 72,* 86–99. http://dx.doi.org/10.3758/APP.72.1.86

Clifford, C. W. (2014). The tilt illusion: Phenomenology and functional implications. *Vision Research, 104,* 3–11. http://dx.doi.org/10.1016/j.visres.2014.06.009

Corbett, J. E. (2017). The whole warps the sum of its parts: Gestalt-defined-group mean size biases memory for individual objects. *Psychological Science, 28,* 12–22. http://dx.doi.org/10.1177/0956797616671524

Crane, B. T. (2012). Direction specific biases in human visual and vestibular heading perception. *PLoS ONE, 7*(12), e51383. http://dx.doi.org/10.1371/journal.pone.0051383

Dakin, S. C. (2001). Information limit on the spatial integration of local orientation signals. *Journal of the Optical Society of America, A, Optics, Image Science, and Vision, 18,* 1016–1026. http://dx.doi.org/10.1364/JOSAA.18.001016

Dakin, S. (2015). Seeing statistical regularities: Texture and pattern perception. In J. Wagemans (Ed.), *The Oxford handbook of perceptual organization* (pp. 150–167). Oxford, UK: Oxford University Press.

Dannals, J. E., & Miller, D. T. (2017). Social norm perception in groups with outliers. *Journal of Experimental Psychology: General, 146,* 1342–1359. http://dx.doi.org/10.1037/xge0000336

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

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.

DiGiacomo, A., & Pratt, J. (2012). Misperceiving space following shifts of attention: Determining the locus of the attentional repulsion effect. *Vision Research, 64,* 35–41. http://dx.doi.org/10.1016/j.visres.2012.05.009

Dumoulin, S. O., & Hess, R. F. (2007). Cortical specialization for concentric shape processing. *Vision Research, 47,* 1608–1613. http://dx.doi.org/10.1016/j.visres.2007.01.031

Elias, E., Dyer, M., & Sweeny, T. D. (2017). Ensemble perception of dynamic emotional groups. *Psychological Science, 28,* 193–203. http://dx.doi.org/10.1177/0956797616678188

Elias, E., Padama, L., & Sweeny, T. D. (2018). Perceptual averaging of facial expressions requires visual awareness and attention. *Consciousness and Cognition: An International Journal, 62,* 110–126. http://dx.doi.org/10.1016/j.concog.2018.03.005

Flevaris, A. V., Martínez, A., & Hillyard, S. A. (2013). Neural substrates of perceptual integration during bistable object perception. *Journal of Vision, 13*(13), 17. http://dx.doi.org/10.1167/13.13.17

Florey, J., Clifford, C. W. G., Dakin, S., & Mareschal, I. (2016). Spatial limitations in averaging social cues. *Scientific Reports, 6,* 32210. http://dx.doi.org/10.1038/srep32210

Fortenbaugh, F. C., Sugarman, A., Robertson, L. C., & Esterman, M. (2019). The attentional repulsion effect and relative size judgments. *Attention, Perception, & Psychophysics, 81,* 442–461. http://dx.doi.org/10.3758/s13414-018-1612-x

Fritsche, M., & de Lange, F. P. (2019). Reference repulsion is not a perceptual illusion. *Cognition, 184,* 107–118. http://dx.doi.org/10.1016/j.cognition.2018.12.010

Goldenberg, A., Sweeny, T. D., Shpigel, E., & Gross, J. J. (2020). Is this my group or not? The role of ensemble coding of emotional expressions in group categorization. *Journal of Experimental Psychology: General, 149,* 445–460.

Goldstone, R. L., & Hendrickson, A. T. (2010). Categorical perception. *WIREs Cognitive Science, 1,* 69–78. http://dx.doi.org/10.1002/wcs.26

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

Grossberg, S. (1994). 3-D vision and figure-ground separation by visual cortex. *Perception & Psychophysics, 55,* 48–121. http://dx.doi.org/10.3758/BF03206880

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

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

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology, 17,* R751–R753. http://dx.doi.org/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. http://dx.doi.org/10.1037/a0013899

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

Hall, M. G., Mattingley, J. B., & Dux, P. E. (2015). Distinct contributions of attention and working memory to visual statistical learning and ensemble processing. *Journal of Experimental Psychology: Human Perception and Performance, 41,* 1112–1123. http://dx.doi.org/10.1037/xhp0000069

Hochstein, S., Pavlovskaya, M., Bonneh, Y. S., & Soroker, N. (2015). Global statistics are not neglected. *Journal of Vision, 15*(4), 7.

Im, H. Y., Albohn, D. N., Steiner, T. G., Cushing, C. A., Adams, R. B., Jr., & Kveraga, K. (2017). Differential hemispheric and visual stream contributions to ensemble coding of crowd emotion. *Nature Human Behaviour, 1,* 828–842. http://dx.doi.org/10.1038/s41562-017-0225-z

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

Im, H. Y., & Halberda, J. (2013). The effects of sampling and internal noise on the representation of ensemble average size. *Attention, Perception & Psychophysics, 75,* 278–286. http://dx.doi.org/10.3758/s13414-012-0399-4

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,* 233–244. http://dx.doi.org/10.1037/a0028762

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,* 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,* 1398–1411. http://dx.doi.org/10.1152/jn.2001.86.3.1398

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

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

Khayat, N., & Hochstein, S. (2019). Relating categorization to set summary statistics perception. *Attention, Perception & Psychophysics, 81,* 2850–2872. http://dx.doi.org/10.3758/s13414-019-01792-7

Kourtzi, Z. (2010). Visual learning for perceptual and categorical decisions in the human brain. *Vision Research, 50,* 433–440. http://dx.doi.org/10.1016/j.visres.2009.09.025

Kuang, S. (2019). Dissociating sensory and cognitive biases in human perceptual decision-making: A re-evaluation of evidence from reference repulsion. *Frontiers in Human Neuroscience, 13,* 409. http://dx.doi.org/10.3389/fnhum.2019.00409

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,* 683–701. http://dx.doi.org/10.1037/xge0000443

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390,* 279–281. http://dx.doi.org/10.1038/36846

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,* 245–250. http://dx.doi.org/10.1016/j.actpsy.2012.11.002

Maule, J., & Franklin, A. (2015). Effects of ensemble complexity and perceptual similarity on rapid averaging of hue. *Journal of Vision, 15*(4), 6.

Michael, E., de Gardelle, V., & Summerfield, C. (2014). Priming by the variability of visual information. *Proceedings of the National Academy of Sciences of the United States of America, 111,* 7873–7878. http://dx.doi.org/10.1073/pnas.1308674111

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,* 1477–1483. http://dx.doi.org/10.1038/nn.3517

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*(1–2), 25–29. http://dx.doi.org/10.1016/0006-8993(93)90187-R

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,* 2361–2380. http://dx.doi.org/10.1016/S0042-6989(98)00243-0

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,* 353–383. http://dx.doi.org/10.1016/0010-0285(77)90012-3

Neumann, M. F., Ng, R., Rhodes, G., & Palermo, R. (2018). Ensemble coding of face identity is not independent of the coding of individual identity. *Quarterly Journal of Experimental Psychology, 71,* 1357–1366. http://dx.doi.org/10.1080/17470218.2017.1318409

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

Op de Beeck, H., 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,* 491–511. http://dx.doi.org/10.1037/0096-3445.132.4.491

Oriet, C., & Brand, J. (2013). Size averaging of irrelevant stimuli cannot be prevented. *Vision Research, 79,* 8–16. http://dx.doi.org/10.1016/j.visres.2012.12.004

Oriet, C., & Hozempa, K. (2016). Incidental statistical summary representation over time. *Journal of Vision, 16*(3), 3.

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

Palmer, S. E. (2002). Perceptual organization in vision. In H. Pashler (Ed.), *Stevens' handbook of experimental psychology* (pp. 177–234). New York, NY: Wiley. http://dx.doi.org/10.1002/0471214426.pas0105

Palmer, S. E., Brooks, J. L., & Nelson, R. (2003). When does grouping happen? *Acta Psychologica, 114,* 311–330. http://dx.doi.org/10.1016/j.actpsy.2003.06.003

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. http://dx.doi.org/10.1038/89532

Pasupathy, A., & Connor, C. E. (2001). Shape representation in area V4: Position-specific tuning for boundary conformation. *Journal of Neurophysiology, 86,* 2505–2519. http://dx.doi.org/10.1152/jn.2001.86.5.2505

Peterson, M. A., & Kimchi, R. (2013). Perceptual organization in vision. In D. Reisburg (Ed.), *The Oxford Handbook of Cognitive Psychology* (pp. 2–28). Oxford, UK: Oxford University Press. http://dx.doi.org/10.1093/oxfordhb/9780195376746.013.0002

Phillips, L. T., Slepian, M. L., & Hughes, B. L. (2018). Perceiving groups: The people perception of diversity and hierarchy. *Journal of Personality and Social Psychology, 114,* 766–785. http://dx.doi.org/10.1037/pspi0000120

Regan, D., & Hamstra, S. J. (1992). Shape discrimination and the judgement of perfect symmetry: Dissociation of shape from size. *Vision Research, 32,* 1845–1864. http://dx.doi.org/10.1016/0042-6989(92)90046-L

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. http://dx.doi.org/10.1016/j.neuropsychologia.2018.02.019

Rolls, E. T., & Tovee, M. J. (1995). The responses of single neurons in the temporal visual cortical areas of the macaque when more than one stimulus is present in the receptive field. *Experimental Brain Research, 103,* 409–420. http://dx.doi.org/10.1007/BF00241500

Ross, J., & Burr, D. (2008). The knowing visual self. *Trends in Cognitive Sciences, 12,* 363–364. http://dx.doi.org/10.1016/j.tics.2008.06.007

Sato, T. (1989). Interactions of visual stimuli in the receptive fields of inferior temporal neurons in awake macaques. *Experimental Brain Research, 77,* 23–30. http://dx.doi.org/10.1007/BF00250563

Solomon, J. A., Morgan, M., & Chubb, C. (2011). Efficiencies for the statistics of size discrimination. *Journal of Vision, 11*(12), 13.

Storrs, K. R., & Arnold, D. H. (2017). Shape adaptation exaggerates shape differences. *Journal of Experimental Psychology: Human Perception and Performance, 43,* 181–191. http://dx.doi.org/10.1037/xhp0000292

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). New York, NY: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780198529699.003.0006

Suzuki, S., & Cavanagh, P. (1997). Focused attention distorts visual space: An attentional repulsion effect. *Journal of Experimental Psychology: Human Perception and Performance, 23,* 443–463. http://dx.doi.org/10.1037/0096-1523.23.2.443

Suzuki, S., & Cavanagh, P. (1998). A shape-contrast effect for briefly presented stimuli. *Journal of Experimental Psychology: Human Perception and Performance, 24,* 1315–1341. http://dx.doi.org/10.1037/0096-1523.24.5.1315

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,* 2179–2189. http://dx.doi.org/10.3758/s13414-017-1381-y

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,* 1236–1257. http://dx.doi.org/10.1152/jn.00061.2010

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. http://dx.doi.org/10.1016/j.visres.2012.05.008

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. http://dx.doi.org/10.1037/a0028712

Sweeny, T. D., & Whitney, D. (2014). Perceiving crowd attention: Ensemble perception of a crowd's gaze. *Psychological Science, 25,* 1903–1913. http://dx.doi.org/10.1177/0956797614544510

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

Thornton, I. M. (2002). The onset repulsion effect. *Spatial Vision, 15,* 219–243. http://dx.doi.org/10.1163/15685680252875183

Treisman, A., & Gormican, S. (1988). Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review, 95,* 15–48. http://dx.doi.org/10.1037/0033-295X.95.1.15

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

Utochkin, I. S., Khvostov, V. A., & Stakina, Y. M. (2018). Continuous to discrete: Ensemble-based segmentation in the perception of multiple feature conjunctions. *Cognition, 179,* 178–191. http://dx.doi.org/10.1016/j.cognition.2018.06.016

Utochkin, I. S., & Yurevich, M. A. (2016). Similarity and heterogeneity effects in visual search are mediated by "segmentability." *Journal of Experimental Psychology: Human Perception and Performance, 42,* 995–1007. http://dx.doi.org/10.1037/xhp0000203

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,* 1172–1217. http://dx.doi.org/10.1037/a0029333

Walker, D., & Vul, E. (2014). Hierarchical encoding makes individuals in a group seem more attractive. *Psychological Science, 25,* 230–235. http://dx.doi.org/10.1177/0956797613497969

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. http://dx.doi.org/10.1016/j.cognition.2016.01.010

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

Watamaniuk, S. N., Sekuler, R., & Williams, D. W. (1989). Direction perception in complex dynamic displays: The integration of direction information. *Vision Research, 29,* 47–59. http://dx.doi.org/10.1016/0042-6989(89)90173-9

Webster, J., Kay, P., & Webster, M. A. (2014). Perceiving the average hue of color arrays. *Journal of the Optical Society of America A, Optics, Image Science, and Vision, 31,* A283–A292. http://dx.doi.org/10.1364/JOSAA.31.00A283

Wei, X. X., & Stocker, A. A. (2017). Lawful relation between perceptual bias and discriminability. *Proceedings of the National Academy of Sciences of the United States of America, 114,* 10244–10249. http://dx.doi.org/10.1073/pnas.1619153114

Westheimer, G., & Levi, D. M. (1987). Depth attraction and repulsion of disparate foveal stimuli. *Vision Research, 27,* 1361–1368. http://dx.doi.org/10.1016/0042-6989(87)90212-4

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). Cambridge, MA: MIT Press.

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

Yamanashi Leib, A., Fischer, J., Liu, Y., Qiu, S., Robertson, L., & Whitney, D. (2014). Ensemble crowd perception: A viewpoint-invariant mechanism to represent average crowd identity. *Journal of Vision, 14*(8), 26.

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

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. http://dx.doi.org/10.3389/fnhum.2012.00267

Yamanashi Leib, A. Y., Puri, A. M., Fischer, J., Bentin, S., Whitney, D., & Robertson, L. (2012). Crowd perception in prosopagnosia. *Neuropsychologia, 50,* 1698–1707. http://dx.doi.org/10.1016/j.neuropsychologia.2012.03.026

Zamboni, E., Ledgeway, T., McGraw, P. V., & Schluppeck, D. (2016). Do perceptual biases emerge early or late in visual processing? Decision-biases in motion perception. *Proceedings of the Royal Society B: Biological Sciences, 283,* 1–8.

Zhang, W., & Luck, S. J. (2009). Feature-based attention modulates feedforward visual processing. *Nature Neuroscience, 12,* 24–25. http://dx.doi.org/10.1038/nn.2223

Zhao, J., Ngo, N., McKendrick, R., & Turk-Browne, N. B. (2011). Mutual interference between statistical summary perception and statistical learning. *Psychological Science, 22,* 1212–1219. http://dx.doi.org/10.1177/0956797611419304

Zoccolan, D., Cox, D. D., & DiCarlo, J. J. (2005). Multiple object response normalization in monkey inferotemporal cortex. *The Journal of Neuroscience, 25,* 8150–8164. http://dx.doi.org/10.1523/JNEUROSCI.2058-05.2005

Received October 29, 2019
Revision received December 22, 2019
Accepted January 29, 2020 ■