Representational Form of Perceptual Average

MyoungAh Kim

The Graduate School

Yonsei University

Graduate Program in Cognitive Science

Representational Form of Perceptual Average

A Master's Thesis

Submitted to the Graduate Program in Cognitive Science

and the Graduate School of Yonsei University

in partial fulfillment of the

requirements for the degree of

Master of Science

MyoungAh Kim

January 2019

This certifies that the master's thesis

of MyoungAh Kim is approved.

Thesis Supervisor: Sang Chul Chong

Min-Shik Kim

Soojin Park

The Graduate School

Yonsei University

July 2018

Table of Contents

List of Figures

List of Tables

Abstract

Representational Form of Perceptual Average

MyoungAh Kim

Graduate Program in Cognitive Science

The Graduate School

Yonsei University

People can accurately represent ensemble properties from a set of multiple items, such as mean size. While much questions have focused on the mechanism of ensemble representation, to our knowledge there was never a discussion about the form of mean representation. The condoned assumption seems to be that mean is represented as a single average, e.g., a single size. However, some evidences contradict this intuitive understanding, one of which is that mean estimation shows large bias in studies that use single item probe to report the mean size. The fact that mean and other various ensemble statistical properties are interrelated also suggests that mean representation is more complex than a single average. The current study explored the form of mean representation by examining how mean size estimation is influenced by the characteristic differences between two comparing ensembles, specifically depending on set size and variance. In each trial, observers were presented with a set of multiple circles. They were asked to report the mean size of the standard display by adjusting the size of a single circle or the overall size of multiple circles in the probe display. We measured percentage error from the actual mean size, as well as the variance of the response. In Experiment 1, we compared mean size estimation performance between using a single probe versus a set probe. Replicating the macro trend across studies, estimation error was greater in the single probe condition than the set probe condition. In Experiment 2, we further divided probe's set size into four levels. Results showed that error becomes systematically smaller as set size disparity decreases between standard and probe displays. In Experiment 3, we checked if this observed set-size disparity effect was possibly due to a difference in sensory memory overlap by examining whether the results change when probe is presented on a different location. Results showed no significant difference between same and different location conditions, ruling out the sensory memory explanation. Finally, Experiment 4 manipulated size variance to see how variance congruency influenced mean size estimation. Error and response variance were always smaller when variance was congruent than when variance was incongruent. All in all, error and response variance of mean estimation were contingent on the characteristics of the probe displays. This supports an idea that mean representation is not represented as a single average, but includes ensemble of statistical properties, such as variance and numerosity.

Keywords: Mean Representation, Mean Size, Statistical Properties, Numerosity, Variance

INTRODUCTION

People can accurately represent ensemble properties from a set of multiple items, such as mean size (Ariely, 2001; Chong & Treisman, 2003). This phenomenon has led to questions trying to understand the mechanism of ensemble perception (Alvarez, 2011; Whitney & Yamanashi Leib, 2018). One of them was how much single-item processing and ensemble coding are related (e.g. Banno & Saiki, 2012; Demeyere, Rzeskiewicz, Fischer & Whitney, 2011; Haberman, Brady & Alvarez, 2015; Haberman, Harp & Whitney, 2009; Haberman & Whitney, 2011; Hochstein, Pavlovskaya, Bonneh & Soroker, 2015; Humphreys & Humphreys, 2008; Parkes, Lund & Angelucci, 2001; Oriet & Corbett 2008; Sweeny, Wurnitsch, Gopnik & Whitney, 2015; Ward, Bear & Scholl, 2016; Yamanashi Leib, Landau, Baek, Chong &Robertson, 2012a; Yamanashi Leib et al., 2012b). Another question in the mean size perception was over how many items are used to compute mean size (Alvarez & Oliva, 2008; Chong, Joo, Emmanouil & Treisman, 2008; Chong & Treisman, 2005; Dakin, 2001; Florey, Clifford, Dakin & Mareschal, 2016; Im & Halberda, 2013; Myczek & Simons, 2008, Maule & Franklin, 2016). More recent discussions are about how items are not weighted equal, but some items have more influence in the mean computation, such as precision or saliency (Albrecht & Scholl, 2010; de Fockert & Marchant, 2008; Kanaya, Hayashi & Whitney, 2018; Mareschal, Morgan & Solomon, 2010). As such, while there was much enthusiasm into figuring out how we are able to extract ensemble properties, to our knowledge, never was there a fundamental discussion about the form of ensemble representation.

Even without much established discussion regarding the form of ensemble representation, it seems that studies naturally assume that perceptual averaging is done by compressing multiple items into a single condensed representation, such as one size or one orientation. This assumption can be found in the fact that the most commonly used method to assess perceptual averaging involves reporting the mean of multiple items (e.g., mean size) using a single item probe (e.g., one circle, one line) (See Table 1 for review). For example, studies that use method-of-adjustment procedure (MOA) show multiple items in a display, followed by a single test probe. The task is to reproduce the mean size of the previous display by adjusting the size of the test probe to approximate the mean (Bauer, 2009; Chong & Treisman, 2003; Li & Yeh, 2017; Marchant, Simons, de Fockert, 2013; Oriet, Hozempa, 2016; to name a few). Another way uses 2-alternative-forced-choice, where a set of multiple items is presented, also followed by a single item probe. The mean size is to be compared against a single probe by reporting whether the mean of the display is larger or smaller than the probe (Allik, Toom, Raidvee, Averin & Kreegipuu, 2013; Ariely, 2001; Bauer, 2017; Joo et al., 2009; Lee, Baek, Chong, 2016). The underlying logic seems to be that if the visual system should summarize multiple items into a representation, then the most obvious strategy is to create an average exemplar that assumes the essential information of the set. For instance, a batch of berries can effectively be reduced into a single berry that has the mean size, mean shape, mean color, and other features in average magnitude of all the berries in the batch. Such premise is also plausible considering that ensemble perception has been suggested as a way of overcoming the limited capacity of the visual system (e.g., Im & Chong, 2014).

Pruning the redundancies into a single average would be a quite efficient way to compress information, given that receptive field properties of cortical neurons are well suited for representing statistical properties in a complex scene (Field, 1987).

Study	Stimulus	Comparison	Task		Result
Bias					
Bauer (2009)	16 horizontal lines	SET vs. 1 probe	MOA	Sixteen horizontal lines were presented, followed by a single line probe below where the stimuli were presented. Task was to adjust the line probe to match the mean length of the set. Feedback was given.	+2.50% bias
Bauer (2017)	9 horizontal lines	SET vs. 1 probe	2AFC	Nine horizontal lines were presented, followed by a single line probe below where the stimuli were presented. Task was to judge whether the probe is larger or smaller than the average length of the set.	+5~46% bias*
Lee, Baek, Chong (2016)	10, 20, 40 circles	SET vs. 1 probe	2AFC	A set of either, 10, 20 or 40 circles was presented, followed by a single probe circle. Task was to judge whether the probe's size was larger or smaller than the mean size of the previous display.	+0% ~ 3% bias (increasing with increasing set size) **
Li & Yeh (2017)	16 circles	SET vs. 1 probe	MOA	Sixteen circles were presented to left and right sides of display, followed by single probe circle. Task was to adjust the size of the probe circle to match the mean size of the set.	+14~33% bias
Marchant, Simons, de Fockert (2013)	4, 8, 16 circles	SET vs. 1 probe	MOA	A set of circles was presented, followed by a single probe circle. Task was to adjust the size of the probe circle to match the mean size of the set.	+14.1~27.7% bias
Oriet, Hozempa (2016)	10-20 colored circles	SET vs. 1 probe	MOA	A set of circles was presented, followed by a single probe circle. Task was to adjust the probe to match the mean size / the largest circle / the smallest circle in the previous display.	Bias depending on variability (8.79px error in high variability, -1.12px error in low variability) .
No Bias					
Oriet & Brand (2013)	lines	SET vs. SET	MOA	Task was to estimate the average length of lines on left side of the display by adjusting the set in the right half that consisted of identical number of items.	No bias***

Note. 'MOA' stands for 'Method-of-adjustment'
** PSE uniformly shifted to about +5~46%*
*** Although the average PSE did not significantly differ from zero, PSE of the mean size increased from 0% to approximately 3% as the number of items in the set increased from 10, 20 to 40 (Inferred from their Figure 3).*
**** Estimation error only varied according to the experimental variable of interest (irrelevant set's mean length disparity) but there was no net shift in bias across conditions (inferred from their Figure 3)*

Table 1. Mean size estimation studies and the reported bias. Whereas bias in mean size estimation is commonly found in 'SET vs. 1 probe' comparisons, no bias is found for 'SET vs. SET' comparison study (refer to the 'Comparison' column).

However, contrary to the intuitive understanding, there are reasons to question the notion that ensemble statistical information is represented as a single average. For one, consistent error is found across studies that require representing the mean size of multiple items by means of a single probe. More specifically, a great majority of them show an overestimation of mean when mean size is estimated using a single probe (around 3% to 40% bias, refer to 'SET vs. 1 probe' comparisons in Table 1). Interestingly, such estimation bias seems to disappear when mean is reported using a probe that consists of multiple items (refer to 'SET vs. SET' comparisons in Table 1). For instance, when observers were asked to adjust the length of multiple lines on the right half of the display to match the average length of the lines on the left half of the display that consisted of identical number of items, net bias was near zero (Oriet & Brand, 2013).

The fact that mean size estimation becomes significantly accurate when they are estimated using multiple items instead of a single item implies that form of mean representation maybe more than a single average. It is claimed that error and reaction time is directly related to compatibility between stimulus and response (Stimulus Response Compatibility: Fitts & Deininger, 1954) or congruency in mapping between stimulus and response elements, such as dimensionality overlap (Kornblum, Hasbroucq & Osman, 1990). If estimation error increases or decreases depending on the kind of probe used to report the mean, it means that the two representations (of the stimulus set and the probe set) are incompatible. Having to report a mean using two incompatible representations must require some form of conversion process which translates the set's representation into a single unit, incurring conversion cost. Leastways, it shows that ensemble size representation includes more information than a single mean size.

Indeed, studies have shown that various ensemble statistical properties, including the mean, are intricately related to each other such that they cannot be independent derivatives of their own. In fact, Lee et al. (2016) have found that observers were able to extract at least two different ensemble properties (numerosity and mean size) at once from the same display, suggesting that the ensemble representation includes various statistical properties. Not only are observers extracting various properties from a set at once, but there is a strong interrelationship between these properties. For example, Lee et al. (2016) showed that there is a strong correlation between the perceived summed area and perceived numerosity ($r = 0.97$). It has also been found that incongruent numerosity interferes with judgements of summed area and incongruent summed area interferes with judgements of numerosity (Hurewitz, Gelman, & Schnitzer, 2006) and that number estimation error increases when the total area of to-be counted objects does not covary with the number of objects (Halberda, Sires, Feigenson, 2006). Similarly, mean is also related to other properties as well, such as numerosity and variance. For example, mean discrimination performance drops when the number of items differs between two comparison displays (Chong & Triesman, 2005). Also, mean size discrimination threshold decreases when variance increases (Corbett, Wurnitsch, Schwartz & Whitney, 2012; Dakin, 2001; Fouriezos, Rubenfeld & Capstick, 2008; Haberman et al., 2015; Im & Halberda, 2013; Morgan, Chubb & Solomon, 2008; Solomon, Morgan & Chubb, 2011) or when the shape of underlying distribution is different (Chong & Treisman, 2003).

Variance is also strongly correlated to the mean in that providing stable mean context over time facilitate variability perception while unstable mean context causes loss in variability sensitivity (Tong, Ji, Chen & Fu, 2015). Such strong interrelationship between various statistical properties suggests that these ensemble properties are not independently computed, but they are a part of one representation that intricately compounds these properties.

There are reasons to believe that various ensemble properties like the mean, numerosity and variance are part of a compound representation. Walsh (2003) proposed a "generalized magnitude system" that processes various magnitudes as time, space, and numerosity. Neuroimaging studies also provide evidence in support this common center. For example, posterior parietal region is related to estimation of these three properties (time: Leon & Shadlen, 2003; Onoe et al., 2001; space: Stein, 1989; numerosity: Sawamura, Shima, & Tanji, 2002) and intraparietal sulcus is related to various magnitude estimations such as size, luminance, and numerosity (Pinel, Piazza, Bihan, & Dehaene, 2004). In this sense, ensemble could be represented in a form of a compound magnitude that includes several statistical properties.

When we can understand that mean representation includes other statistical properties, then it becomes clear why mean estimation error increases when two sets have different number of items. Previous studies have shown that disparity in set size makes it difficult to compare the mean between two ensembles, such that mean size estimation performance drops when number of items differ between the two displays (Albrecht & Scholl, 2010; Chong & Treisman, 2005). In fact, if we look at the degree of error in single probe studies, whereas there is hardly any judgement bias for mean size tasks that asks to average only two items (Chong & Triesman, 2003), the error is generally larger for studies that use larger set size (Bauer, 2017; Li & Yeh, 2017; Marchant, Simons, de Fockert, 2013). The trend of error increasing with greater set size disparity can be found in Marchant et al. (2013) where mean size estimation error increases linearly as set size increases from 4, 8, to 16. Presumably, drastic difference in set size also renders the mean representation very different, which makes it harder to compare the mean. This leads to greater conversion error in the course of trying to compare one representation (that of many items) to the other (one item).

In the same sense, conversion error will also be an issue when comparing ensembles with different variances. In Oriet & Hozempa (2016), observers adjusted size of a single probe to estimate the mean size of circles which had either high size variability or low size variability. While error was relatively small for low variability condition, there was an overestimation of the mean for high variability condition. Considering that a single probe has a variance of zero, low variability set has similar variance to the probe than high variability set. This variance difference between two comparison sets may create different ensembles so that comparing the mean against the probe becomes difficult, leading to conversion error. If the mean was indeed represented in form of a single average size, there is no reason why high variability condition should have larger error than low variability condition, because both ensembles should have same average representation regardless of variance.

The current study investigated the form of mean representation. Specifically, we examined whether mean size is represented as a single size or includes other properties, such as numerosity and variance. For this purpose, we conducted a mean size estimation task and examined how mean size estimation changed depending on the disparity between two comparing ensembles, by varying set size and variance. Observers were presented with a set of multiple circles and were asked to estimate the mean size via adjusting the size of a single circle or the overall size of a set in the probe display. In Experiment 1, we compared mean size estimation performance between using a single probe versus a set probe, to see if the observed overestimation across previous studies (Table 1) could be replicated. In Experiment 2, we further divided probe's set size into four levels to see if error becomes systematically larger as set size disparity increases between the two displays. In Experiment 3, we examined whether this observed difference was possibility due to a difference in sensory memory overlap that also differed according to set size conditions. Finally in Experiment 4, we manipulated size variance to see how variance congruency influences mean size estimation.

EXPERIMENT 1

In Experiment 1, we examined whether the difference in mean size estimation error between a single probe versus a set probe, that is observed as a macro trend across different studies (Table 1), can be replicated within participants. For this purpose, we asked participants to judge the mean size of a set of circles in the standard display by adjusting the size of the test circle in the probe display. We used two types of probe displays. The 'single probe' condition was a typical method-of-adjustment task where one probe circle was shown and participants were to report the mean size of a display by adjusting the size of this one probe circle. The 'set probe' condition showed a set of circles similar to the set in the standard display and participants adjusted the overall size of the circles so that its mean size matches the mean size of the standard display. Based on the trend found across previous studies (Table 1), we hypothesized that there would be an overestimation of the mean size in the 'single probe' condition, whereas there would be none in 'set probe' condition.

Methods

Participants

Six participants (2 male, 4 female, M age = 23 years, age range: 21-26) were recruited from the Yonsei University undergraduate community board in exchange for monetary compensation. All reported normal or corrected to normal vision.

We used an adaptive sampling procedure based on Bayesian inference to determine the sample size. We used the Bayesian analyses of variance in JASP statistical package to assess the evidence for or against our term of interest, which was the effect of probe type (single/set) in yielding different degrees of mean size estimation error. For this, we derived the Bayes Factor (BF_k) for each term, which describes how much more likely one model is likely over the null model. The ANOVA included "participant ID" as a random effect and the factorial term 'Probe Type' (single probe/set probe). We chose to stop data collection once compelling evidence ($BF > 3$) was found either for or against the effect of the probe type, which correlates to 'substantial evidence against the null hypothesis' according to Jeffrey's grades of evidence (Jeffreys, 1961). The evidence for the probe type effect was found quickly ($BF_{10} = 3.301$) after collecting data from six participants, hence stopped data collection. All participants submitted written consent forms prior to participating in experiment and every aspect of the study was reviewed and approved by the Institutional Review Board of Yonsei University.

Apparatus and Stimuli

We presented the stimuli using MATLAB (The MathWorks, Natick, MA) and Psychophysics Toolbox Version 3 (Brainard, 1997; Pelli, 1997). It was presented on a gamma-corrected 21–inch CRT monitor (HP P1230; resolution 1,600 x 1,200 pixels; refresh rate: 85 Hz). The experiment was conducted in a dark room. Participants' heads were fixed by a chin-and-forehead rest at a viewing distance of 60 cm; one pixel subtended 0.0239° at this distance.

Stimuli were presented on a gray background (39.31 cd/m^2). The stimulus display consisted of two sets of twelve outlined circles presented on the left and right halves of the display respectively. The circle was outlined in black (0.04 cd/m^2) on a gray background. The size of twelve circles were randomly selected from a normal distribution in which the mean was between 1.13 ~ 2.26° and had standard deviation one eighth of this mean. The size of the circles was set so that the sizes would not exceed two standard deviations from the mean size. In two thirds of the trials, one of the circles in the set was either a small outlier or one large outlier (five standard deviations smaller or bigger than the set's mean size, but this manipulation was not a variable of interest of this study nor did the manipulation yield any significant result. Each circle was located in a 4-by-3 invisible grid (spanning 12 horizontal x 8 vertical visual degrees) and was randomly jittered within 0.445° without being overlapped. The two grids were 6 visual degrees apart from each other. In the single-probe adjustment condition, only one circle probe appeared in the test screen on either the left or right side, depending on which of the two sets was to be appraised. The probe was located at the center of the previously shown stimulus set. The initial probe diameter was randomly selected from 0.5927~ 3.1105°, and scaled up or down by 1px with each up/down key press. In the set adjustment condition, a set of twelve circles similar to that shown in the stimulus display, appeared on either the left or right side of the test display. The size and the location of the twelve circles in the probe set was determined in the same way as the set used in the stimulus display, except the set's mean size ranged from 0.5927~ 3.1105°. The mean size of the set was scaled by 1px with each up/down key press, and the size as well as the location of each circle was reselected with each key press to avoid gradual scaling of a fixed set, while maintaining the same size variance. All sizes were sampled from distributions based on the Teghtstoonian scale (Teghtsoonian, 1965).

Design and Procedure

Experiment 1 was a one factor within-subject design, seeking to see the mean size judgement error according to two different types of 'probe conditions' (single probe/set probe). The task of Experiment 1 was to estimate the mean size of a set of twelve circles using two different types of probes. The experiment consisted of two sessions. Only one type of probe was used per session and each participant conducted both types (single probe or set probe) across two different days. The order of the session type (single probe-set probe / set probe-single probe) was counterbalanced across participants. The design of Experiment 1 is summarized in Figure 1. Participants were seated in a dark room with their head fixed on a head-and-chin rest. Each trial started with a black fixation cross in the center of the screen and two black rectangle frames on the left and right sides of the screen, which was presented for 500-1000 ms. One of the two frames flashed white for 75 ms, followed by a 25 ms fixation. This flashing frame was an experimental manipulation which did not yield any significant impact, nor was it the experimental condition of interest. Then the standard display was presented for 100 ms with two sets of circles presented on the left and right sides of the center fixation cross respectively. After a 500 ms blank screen, the probe display appeared with the probe on either the left or right side of the display. Two different types of probe displays were shown depending on the probe condition. During single probe sessions, a single circle was presented as the probe either on the left side of the screen or the right side of the screen. If the probe appeared on the left side, participants had to estimate the mean size of the set that appeared on the left side in the previous standard display, whereas if the probe appeared on the right side, they had to estimate the mean size of the set that appeared on the right side. Participants adjusted the size of this probe to match the mean size of the set presented in the standard display by using the up/down arrow keys on the keyboard. During set probe sessions, twelve circles were presented as the probe either on the left side or the right side of the screen, also indicating which side set to report the mean size of. Participants pressed the up/down arrow keys to adjust the overall size of the circles in the probe set to match the mean size of the set presented in the standard display. They pressed 'space' key to confirm their response, after which next trial began immediately. Participants conducted 18 practice trials before each session. Only one type of probe was used per session, which consisted of 216 trials. There were five breaks between blocks per session where participants could choose to continue at their own discretion by pressing the space key on the keyboard. Each participant conducted all two sessions across two different days. Each session took 20 minutes – 25 minutes to complete.

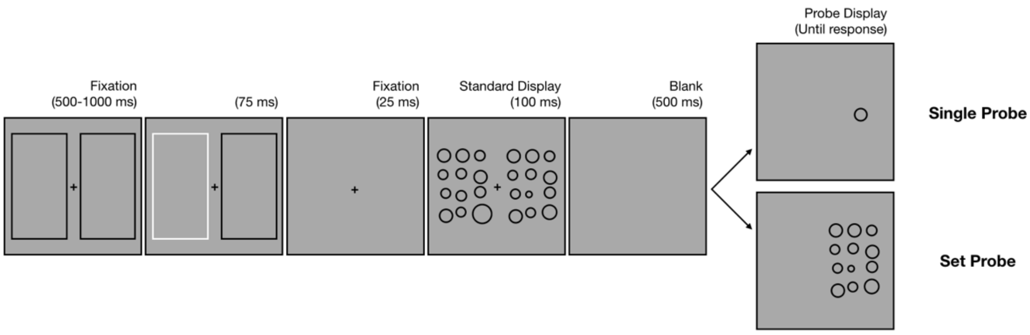

Figure 1. Stimuli and trial procedure of Experiment 1. Observers were instructed to judge the perceived mean size of the standard display by adjusting the size of the probe/probe set in the probe display.

Results

Data Analysis

To see how mean size estimation error changed depending on different probe conditions, we calculated the Relative Mean Size Difference (hereafter, RMSD), defined as the difference between the estimated mean size and the actual mean size of the stimulus display divided by the actual mean size. As for the single probe condition, the estimated mean size was the size of the final probe circle. As for set probe condition, the estimated mean size was the averaged size of the twelve circles in the final probe set.

$$\text{Relative Mean Size Difference (RMSD)} = \frac{\text{Estimated mean size - Actual mean size}}{\text{Actual mean size}}$$

For each participant, we created the error distribution for each probe condition respectively, with RMSD on the y-axis (Figures 2A and 3A). We looked at two measures in this distribution. First was the average RMSD for each adjustment condition, taken as the mean point of the RMSD distribution (Figure 2A arrows). If this RMSD is positive, it indicates that there was an overestimation of the mean size (e.g., 0.4 indicates that observers overestimated the size to be 40% greater than the actual size). On the other hand, if RMSD is negative, it indicates there was an underestimation of the mean size (e.g., -0.4 indicates that observers underestimated the size to be 40% smaller than the actual size). RMSD closer to 0 indicates accurate perception of the mean. Second, we also looked at the variance of error (variance of responses), taken as the standard deviation of the RMSD distribution (Figure 3A dotted arrows). While the overall deviation of distribution reflects systematically biased perception, the variability in responses may reflect observer's uncertainty, whether it be the precision of representation. Thus, larger response variance can be interpreted as bigger uncertainty (or less precision) of the mean size representation when using the respective probe. Each of these two measures (relative mean size difference, variance of error) was averaged across participants then compared between the two different probe conditions.

For both measures, a 1 factor (probe type: single probe, set probe) paired samples *t*-test was conducted.

Relative Mean Size Difference

The mean RMSD for the two probe conditions are presented in Figure 2. There was a significant main effect of the probe condition, $t(5) = 3.10$, $p = .027$, $d = 1.27$, $BF_{10} = 3.30$; the mean RMSD was larger in the single probe condition (27.20% overestimation) than set probe condition (3.50% overestimation). We conducted one sample *t*-test for each probe type to see whether the mean RMSD significantly differed from zero. The mean RMSD in the single probe condition was significantly different from zero, $t(5) = 3.45$, $p = .018$, $d = 1.41$, $BF_{10} = 4.41$, while the mean RMSD in the set probe condition did not significantly differ from zero, $t(5) = 0.93$, $p = .397$, $d = 0.38$.

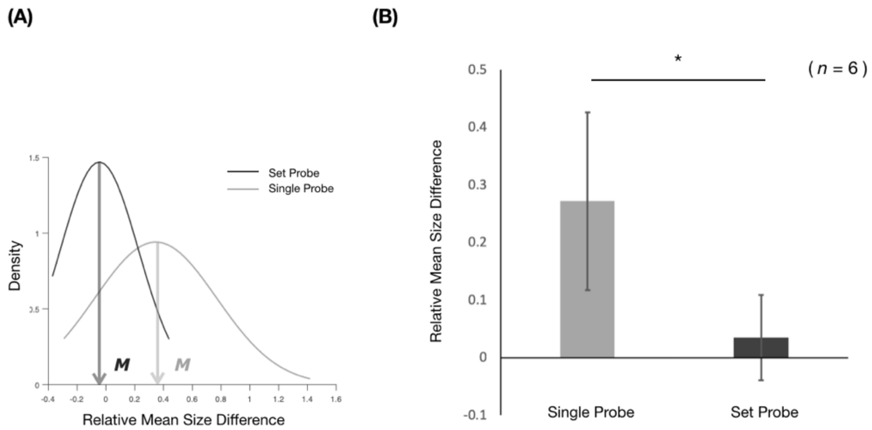

Figure 2. Mean RMSD results of Experiment 1. (A) Example distribution of Relative Mean Size Difference (RMSD) of a representative participant. The average of the RMSD distribution (as marked by the arrows) was averaged per condition across participants as shown in graph B. (B) The mean RMSD for the two probe conditions. Error bars indicate the 95% confidence interval. RMSD is significantly bigger for single probe condition than set probe condition, which indicates that observers overestimated the mean size in the single probe condition, but not in the set probe condition.

Variance of Error

The variance of error for the two probe conditions are presented in Figure 3. There was a significant main effect of the probe condition, $t(5) = 4.79$, $p = .005$, $d = 1.96$, $BF_{10} = 11.82$; the variance of error was bigger in the single probe condition (0.39) than the set probe condition (0.29), suggesting that using set probe reduced estimation uncertainty (or increased precision of mean representation).

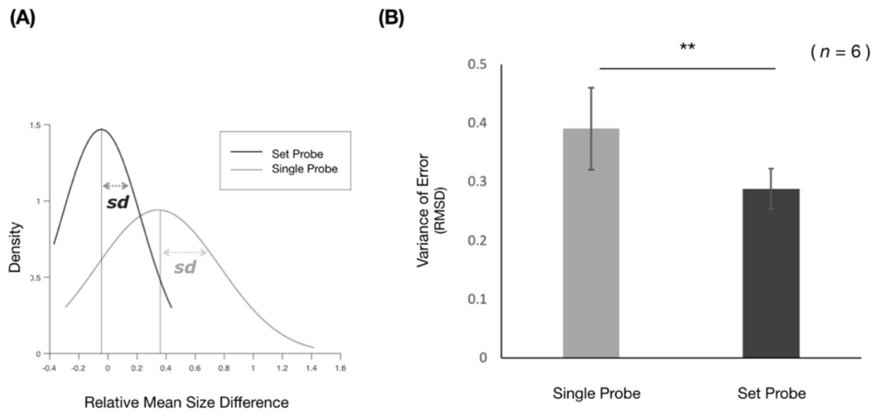

Figure 3. Variance of error results of Experiment 1. (A) Example distribution of RMSD of a representative participant. The standard deviation of the RMSD distribution (as marked by the dotted line arrows) was averaged per condition across participants as shown in graph B. (B) Variance of error for the two probe conditions. Error bars indicate the 95% confidence interval. Variance of error is significantly larger for single probe condition than set probe condition.

Discussion

In Experiment 1, we tested whether overestimation of size is present in the ‘single probe’ condition, but not in the ‘set probe’ condition. As hypothesized, participants overestimated the mean size of the standard display when using a ‘single probe’ to report the mean size whereas the estimation was accurate when the mean size was matched by a ‘set probe’, which was a similarly looking set. This successfully replicates the overarching trend shown in previous studies that using a single probe to report the mean size of many items incurs large estimation error, predominantly in a positive direction.

The fact that mean estimation bias is larger for the single probe condition as opposed to the set probe condition suggests that representing the mean of multiple items in terms of a single average unit is more difficult, or leastways, not the most natural way that that size information is summarized in the visual system. When viewing multiple items, the representation of the items’ average size is not actually reduced to the form of a single size. So when the task actually requires the mean information to be reduced down to a single unit, the original representation should go through a conversion process resulting in estimation bias as the conversion cost. This comparison, on the other hand, is easier when using a similar set, where no information has to be reduced and converted. In addition, the fact that ‘variance of error’ is larger in the single probe condition as opposed to the set probe condition insinuates that representing the mean size in terms of a single size unit has more uncertainty compared to the set probe condition.

If the estimation error is indeed due to some form of representational conversion, then we can expect to see that mean size estimation error decrease as two comparison sets become more similar. We tested this across two experiments. In Experiment 2, we tested whether error becomes systematically smaller as the number of items become more identical between two sets by manipulating the set size disparity between two displays into four levels. In Experiment 4, we tested whether error becomes smaller when two sets have the same variance as opposed to different variance by manipulating variance congruency.

EXPERIMENT 2

In Experiment 1, we observed that participants overestimated the mean size in the 'single probe' condition, but not in the 'set probe' condition. This shows that the difference in perception bias found for different probe tasks across multiple studies is actually replicated within participants. If such overestimation in single probe condition is due to representational discordance which requires converting the original representation into a single item representation, we can expect perception bias to become smaller as the probe becomes more similar to the stimulus display. To test this hypothesis, in Experiment 2, we manipulated the set size disparity between the stimulus display (remained the same: 16 items) and the probe display (1, 4, 9, and 16 items) into four conditions: 16:1, 16:4, 16:9, and 16:16. If this bias in size perception is influenced by disparity in the number of items between the two comparison displays, we predicted that error will become systematically larger as the set size disparity increases between the standard and probe displays.

Methods

Participants

Eight participants (4 male, 4 female, M age = 24.125 years, age range: 20-28) were recruited from the Yonsei University undergraduate community board in exchange for monetary compensation. All reported normal or corrected to normal vision. As in Experiment 1, we used the same adaptive sampling procedure based on Bayesian inference to determine the sample size. The ANOVA included “participant ID” as a random effect and the factorial term ‘set size disparity’ (16:1, 16:4, 16:9, 16:16). We chose to stop data collection once compelling evidence ($BF > 3$) was found either for or against ($BF < 0.33$) the effect of probe type. The evidence for the probe type effect was found quickly ($BF_{10} > 413.74$) after collecting data from eight participants, hence stopped data collection. All participants submitted written consent forms and every aspect of the study was reviewed and approved by the Institutional Review Board of Yonsei University.

Apparatus and Stimuli

All experimental apparatuses and stimuli were identical to Experiment 1. The size of the sixteen circles in the standard display were determined in the same way as in Experiment 1, except the set’s mean size ranged from 1.13 ~ 2.26°. Each circle was located in each square of a 4-by-4 invisible grid (each inner square’s side length was 4°)

and was randomly jittered within 0.35° without being overlapped. The grid was located at the center of the screen. For 16:1 condition, only one probe circle appeared in the probe display. The location of the probe was randomly chosen among the sixteen possible inner squares of the 4-by-4 invisible outer grid, spanning the same size as the one used in the standard display. For 16:4 condition, four circles were located in a 2-by-2 invisible grid (each inner square spanned 4° vertically and horizontally), and this 2-by-2 grid was randomly presented among the nine possible locations on the 4-by-4 outer grid. For 16:9 condition, nine circles were located in a 3-by-3 invisible grid and this inner grid was randomly presented among the four possible locations on the outer grid. For 16:16 condition, sixteen circles were located in the 4-by-4 invisible grid and was centered on the screen. The sizes of the probe were determined in the same way as the set used in the stimulus display, except the set's mean size ranged from 0.9352~ 2.6174°.

Design and Procedure

The design and the procedure of Experiment 2 are summarized in Figure 4. Experiment 2 was a one factor design that examined the difference in mean size judgement across the four levels of 'set-size disparity' (16:1, 16:4, 16:9, 16:16). The general task of Experiment 2 was identical to Experiment 1, but with two major differences. First, only one set was presented in the standard display, which consisted of sixteen outlined circles and was located at the center of the display. Second, the probe type was subdivided into four conditions (according to set size disparity: 16:1, 16:4, 16:9, and 16:16), and they were blocked.

Each trial started with a black fixation cross presented for 500-1000 ms. In the standard display, sixteen circles were presented for 100 ms then disappeared. After a 500 ms blank screen, the probe display was shown. Four different probe displays were presented depending on the set-size disparity. The participants pressed the up/down arrow keys to adjust the overall size of circles in the probe display (or the size of the circle in case of 1 probe condition) to match the mean size of the set presented in the standard display. They pressed the space key to confirm their response, after which the next trial begun immediately. Participants conducted 12 practice trials (3 trials of each probe type) before conducting the main experiment. The main trials consisted of 4 blocks and only one set-size disparity condition was used per block. The order of block was determined according to Latin-square and was assigned to each participant according to order of arrival. Each block consisted of 216 trials and breaks were assigned in between the blocks in which participants could choose to continue at their own discretion by pressing the space key on the keyboard. The entire experiment took 20 minutes – 25 minutes to complete.

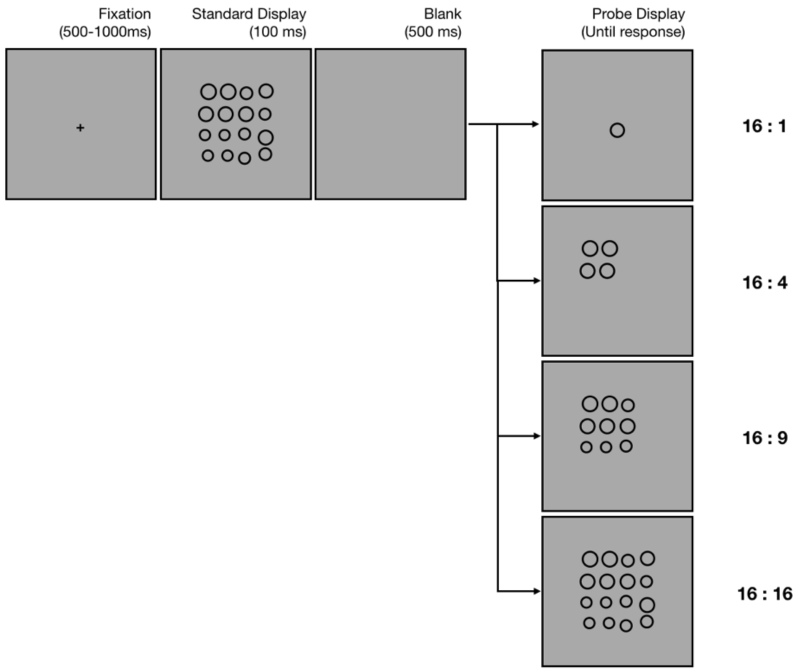

Figure 4. Stimuli and trial procedure of Experiment 2. Observers were instructed to judge the perceived mean size of the standard display by adjusting the probes' mean size in the probe display.

Results

Data Analysis

Data was analyzed in the same manner as in Experiment 1. Mean RMSD and variance of error was derived for each set size disparity condition. A one-way (set size disparity: 16:1, 16:4, 16:9, 16:16) repeated-measures ANOVA was conducted for both mean RMSD and variance of error.

Relative Mean Size Difference

The mean RMSD for each set size disparity is presented in Figure 5A. There was a significant main effect of set size disparity, $F(3, 21) = 11.60$, $p < .001$, $\eta^2_p = .62$, $BF_{10} = 413.74$; the mean RMSD was the largest in the 16:1 condition (13.80% overestimation) followed by 16:4 (8.5% overestimation), 16:9 (0.00%), and 16:16 (-.03%) respectively. The mean RMSD was significantly larger when set size disparity increased from 16:4 to 16:9, $t(7) = 2.35$, $p = .026$, $d = .83$, $BF_{+0} = 3.58$, while no significant difference was found between 16:1 vs. 16:4 conditions, nor 16:9 vs. 16:16 conditions (both *p*s > .06). This means that mean size perception bias occurs beyond the disparity ratio of 16:4. We conducted one sample *t*-test for each set size disparity to see whether the mean RMSD significantly differed from zero. The mean RMSD for 16:1 ($t(7) = 3.73$, $p = .007$, $d = 1.32$, $BF_{10} = 8.35$) and 16:4 ($t(7) = 3.63$, $p = .008$, $d = 1.28$, $BF_{10} = 7.47$) conditions were significantly different from zero, while 16:9 and 16:16 conditions did not (both *p*s > .37).

Variance of Error

The variance of error for each set size disparity is presented in Figure 5B. There was a significant main effect of set size disparity, $F(3, 21) = 4.94$, $p = .010$, $\eta^2_p = .41$, $BF_{10} = 8.70$. The variability of error was the largest in the 16:1 condition, and was significantly larger than the 16:4 condition, $t(7) = 2.06$, $p = .039$, $d = .73$, $BF_{10} = 1.37$, with no significant difference thereafter (no difference among 16:4, 16:9, 16:16 conditions, all *p*s > .07).

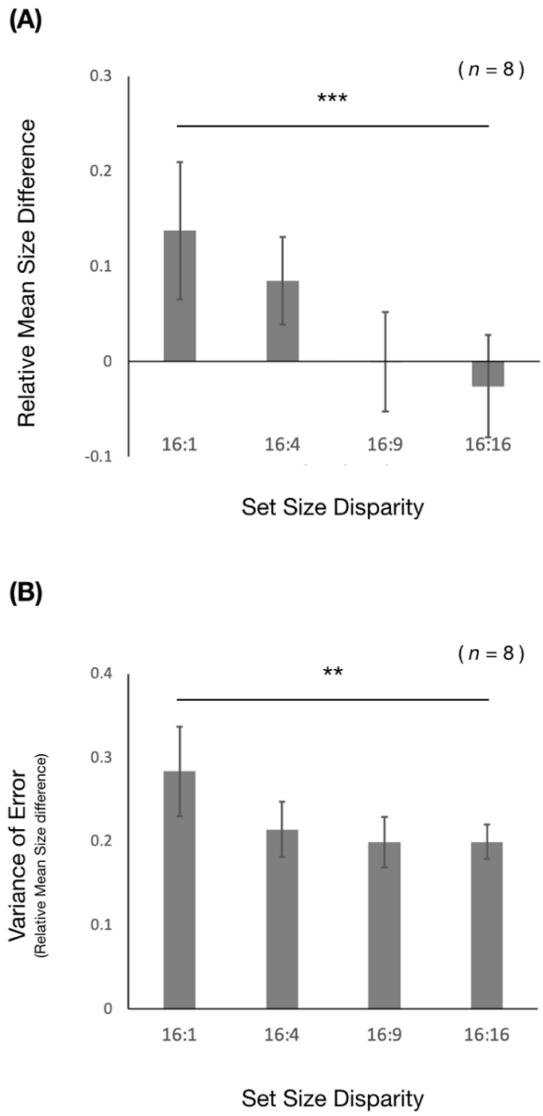

Figure 5. Results of Experiment 2. (A) The mean RMSD according to set size disparity between standard and probe displays (standard:probe). The degree of RMSD systematically reduces as set size disparity decreases between the two comparison displays. (B) Variance of error for each set size disparity condition. Variance of error is the largest for 16:1 (single probe) condition, where set size disparity is the largest. However, there is no difference among the set conditions (16:4, 16:9, 16:16). All error bars indicate the 95% confidence interval.

Discussion

In Experiment 2, we tested if mean size estimation bias reduces as the number of items become similar between standard and probe displays. As hypothesized, mean size estimation error did become systematically smaller as set size disparity reduced, suggesting that mean size comparison became easier when the number of items were similar between the two comparison displays. This means that mean representation includes numerosity property, that when number of items differ between two sets, conversion is necessary on the representation to translate it into a comparable unit. When the mean size representation is in a form of one size, there is no reason why mean estimation should become systematically more accurate as two sets become more identical in its number of items.

One thing to consider is why response variance (proxy of observer certainty) remains more or less the same after four items. In order words, if participants are similarly certain of their mean judgement beyond four item probes, yet their mean size overestimation continues to decrease nevertheless, it might indicate that bias is not solely a reflection of representation conversion noise, but a systematic bias in the mean representations themselves. We discuss this further in the general discussion section in relation to large size saliency-weighted average hypothesis (Kanaya, Hayashi, & Whitney, 2018).

EXPERIMENT 3

Experiments 1 and 2 showed that observers' mean size estimation bias reduces as the set size of two comparison sets become similar to each other. However, one could argue that the decreased estimation error for larger set sizes found in studies above is not due to representation differences, but because using probe of similar set size is easier to match up on a sensory memory level. In fact, in Experiment 1, the set probe condition had more location overlap between the standard display and probe display compared to the single probe condition, which only had $1/12^{th}$ overlap with the standard display. In Experiment 2, larger set size discrepancies yielded smaller location overlap and more varied set location. In this regard, participants could have done the task without using any size representation but using sensory information such as the items' positions and boundaries to match the set probe to the standard display. In such case, having smaller set size disparity with more location overlap will make the task easier to match up. For this purpose, in Experiment 3, we controlled for this locational overlap to check if this location overlap was the cause of the difference found in the former experiments. Specifically, we manipulated the location (left side, right side) where the stimuli sets appears on the screen and compared when standard display sets and probe display sets appear on the same side versus different side. When standard display and probe display sets appear on different sides, we can safely preclude the influence of sensory memory as participants will have to depend solely on size representation to perform the task.

Methods

Participants

Eight participants (4 male, 4 female, M age = 24.13 years, age range: 21-27) were recruited from the Yonsei University undergraduate community website. All reported normal or corrected to normal vision. Same as in Experiments 1 and 2, we used the same adaptive sampling procedure based on Bayesian inference to determine sample size. The ANOVA included "participant ID" as a random effect and the factorial term 'side congruency' (same side, different side) and 'set size disparity' (9:4, 9:9, 9:16). We decided to stop data collection once compelling evidence was found either for ($BF > 3$) or against ($BF < 0.33$) side congruency effect. The evidence against side congruency was found quickly ($BF_{10} = 0.296$) after collecting data from eight participants, hence stopped data collection. As in the previous experiments, every aspect of the study was reviewed and approved by the Institutional Review Board of Yonsei University.

Apparatus and Stimuli

All experimental apparatus was identical to previous experiments. The standard displays always consisted of nine outlined circles presented in a 3-by-3 invisible grid on either the left or the right side of the display. The probe display on the other hand, had three types of set sizes: four, nine, sixteen, which also could appear either on the left or the right side. The location of the grid containing the circles was shifted 9.5° to the left or the right from the center of display, for left side and right side presentation respectively. The size of the circles both for the standard and probe displays were selected in the same manner as in Experiment 2.

Design and Procedure

Experiment 3 was a 2 (side congruency: same side, different side) x 3 (set size disparity: 9:4, 9:9, 9:16) within-subjects design. The experimental conditions are summarized in Figure 6. The task and the experimental procedure were the same as in Experiment 2 with the only difference being the location of where the set appeared in the standard and probe displays, which varied according to the condition (left side/right side). The standard displays always consisted of nine circles. The probe display presented either 4, 9 or 16 circles according to set size disparity condition (9:4, 9:9, 9:16 respectively). Both side congruency and set size disparity condition were intermixed within blocks. Participants conducted 12 practice trials before conducting the main experiment. The main trials consisted of 4 blocks, 75 trials each. The entire experiment took 15 minutes – 25 minutes to complete.

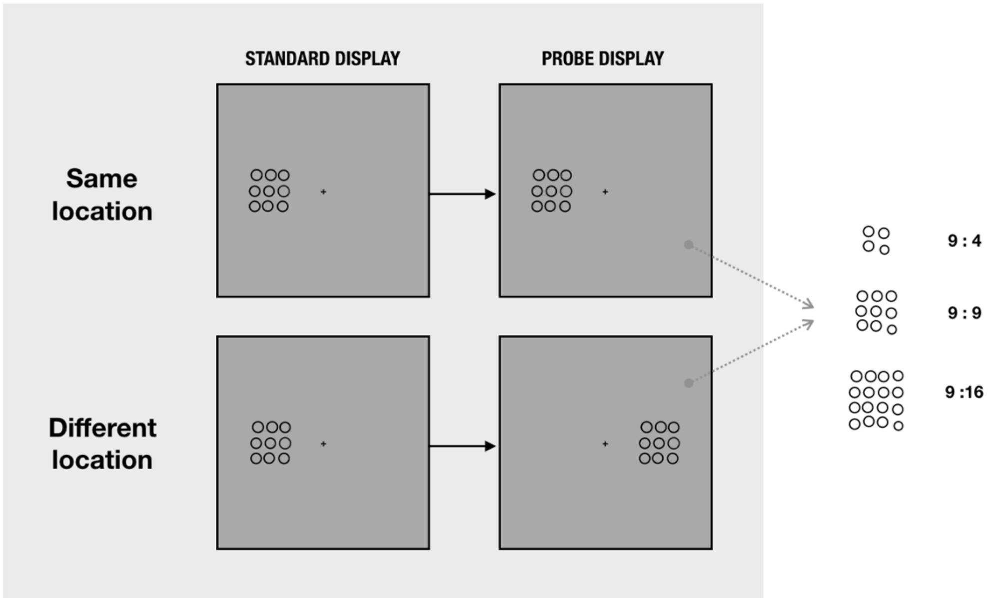

Figure 6. The stimulus conditions of Experiment 3. The standard displays always consisted of 9 circles and could be shown on either the left or right side of the display. The probe display consisted of either 4, 9, or 16 circles depending on the set size disparity condition (9:4, 9:9, 9:16 respectively). In the same location condition, the probe display set was presented on the same side as where the standard display set was presented. In the different location condition, the probe display set was presented on different side of where the standard display set was presented.

Results

Data Analysis

The mean RMSD and the variance of error was computed for the six conditions. A 2 (side congruency: same side, different side) x 3 (set size disparity: 9:4, 9:9, 9:16) repeated-measures ANOVA was conducted for both mean RMSD and variance of error.

Relative Mean Size Difference

The mean RMSD for each condition is presented in Figure 7A. As expected, there was no significant difference of side congruency $F(1, 7) = 0.11$, $p = .747$. However, a significant interaction was found between side congruency and set size disparity, $F(2, 14) = 9.55$, $p = .002$, $\eta^2_p = .577$, $BF_{10} = 3.13$. Contrary to previous experiments, the effect of set size disparity was not replicated in the 'same side condition', $F(2, 14) = 0.29$, $p = .753$. However, in the 'different side condition', a significant set size effect was found, $F(2, 14) = 10.26$, $p = .002$. The RMSD of '9:4' condition increase significantly when probe set was presented on different side as opposed to the same side, $F(1, 7) = 7.63$, $p = .028$. However, there was no side congruency effect found for '9:9' ($F(1, 7) = 4.28$, $p = .077$) and '9:16' ($F(1, 7) = 2.04$, $p = .196$) conditions. None of the six conditions significantly differ from zero (all ps > .11).

Variance of Error

The variance of error for each condition is presented in Figure 7B. As expected, error variance did not differ depending on side congruency, $F(1, 7) = 1.04$, $p = .342$. Unlike previous experiments, error variance did not differ depending on set size disparity condition either, $F(2, 14) = 1.05$, $p = .378$.

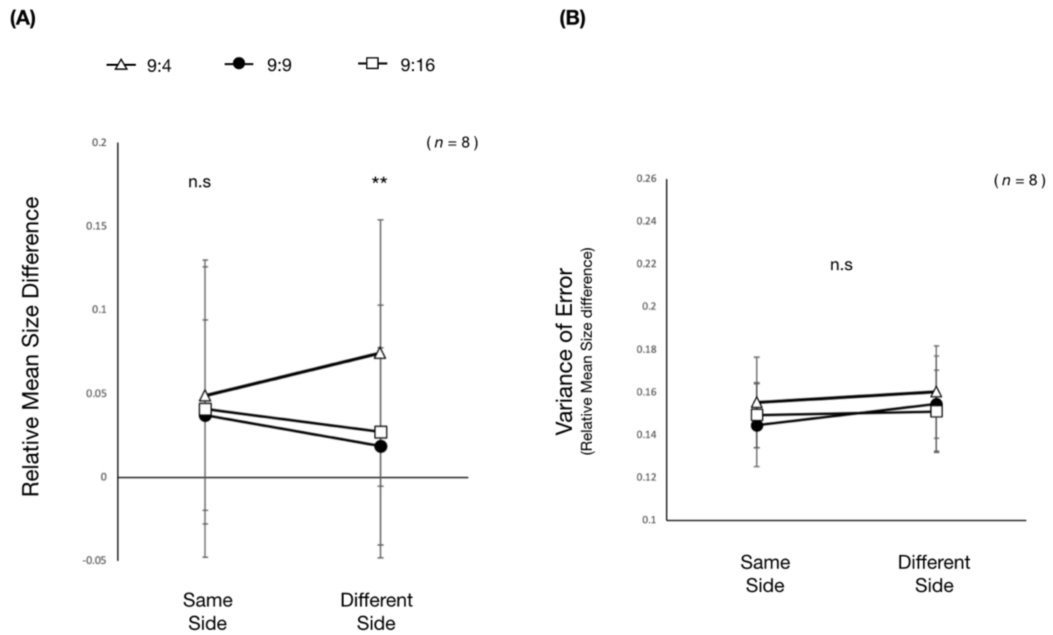

Figure 7. Results of Experiment 3. (A) The mean RMSD according to side congruency & set size disparity. There was no significant RMSD difference of side congruency, but interaction was found between side congruency and set size disparity (B) Variance of error according to side congruency & set size disparity. Error variance did not differ depending on side congruency. All error bars indicate the 95% confidence interval.

Discussion

In Experiment 3, we compared mean estimation error and response variance when stimulus and probe displays appear on the same side versus different side to check if sensory memory overlap was the cause of any difference in mean size perception. In Experiment 2, 16:16 condition, or the condition when set size disparity was smallest was also when there was a larger spatial overlap between the standard and probe displays. This large overlap could have made the task easier by enabling observers to match up the stimuli in terms of sensory memory without referring to the size representation. If it was true that spatial overlap influenced error differences in the previous experiments, then error should be larger when probe display set is presented on the different side of the standard display set. Contrary to the concern, there was no significant bias difference between the same side and different side conditions. However, in this particular setting of the experiment, set side disparity effect was not replicated even in the ‘same side’ condition (previous experiments could be seen analogous to the same side condition). Interestingly, only the bias in the 9:4 disparity condition increased significantly when probe display was presented on different side. The reason behind this interaction needs further analysis. However, the reason why set size effect was not replicated even in the ‘same side’ condition could have been due to differences in the experimental setting from former experiments: Unlike Experiment 2 where sets were presented at the center of the display (near fovea), sets in Experiment 3 were presented on the periphery as opposed to the center. This reduction in resolution could have rendered perceptual difference between the sets smaller. Also, the set size disparities used in this experiment (9:4) maybe perceptually smaller than set size disparities of Experiment 2 (16:4), which could be another reason why set size effect was not replicated. The bottom line is, the fact that set size effect is more evident in the different side condition than the same side condition suggests that set size differences found in previous experiments are still evident when there is no spatial overlap differences. Thus, it is safe to assume that the differences were not influenced by sensory memory but due to ensemble representation.

EXPERIMENT 4

Disparity between two sets may be rendered not only through numerosity differences, but variance differences as well. In Experiment 4, we manipulated variance congruency between the standard display and the probe display to examine whether having different size variance between the two comparison displays makes mean size comparison more difficult. In fact, in a study using single probe adjustment task, mean size overestimation was found for high variability condition but not for low variability condition (Oriet & Hozempa, 2016). Considering that single item probe has same variance as homogenous or low variability sets, this variance difference between two comparison sets may have created different ensemble representations, such that comparing the mean becomes difficult, leading to conversion error. If the ensemble's size representation was indeed represented in a form of a single average, there is no reason why high variability condition should yield larger errors than low variability condition. Because regardless of variance, both ensembles would have similar representations when same number of items are included in the set. So in Experiment 4, we manipulated the similarity between two ensembles in terms of variance congruency (same variance, different variance). There were two kinds of size variance (low variance, high variance). The standard display and the probe display could either have same size variance (low & low, high & high) or different variance (low & high, high & low). If size perception bias is influenced by variance disparity between the two comparison displays, we predicted that error should be larger in the different variance conditions than in the same variance conditions.

Methods

Participants

Six participants (3 male, 3 female, M age = 28.33 years, age range: 26-37) were recruited from the Yonsei University graduate school of psychology. All reported normal or corrected to normal vision. The same adaptive sampling procedure based on Bayesian inference was used to determine the sample size. The ANOVA included "participant ID" as a random effect and the factorial term 'variance congruency' (congruent, incongruent). We chose to stop data collection once compelling evidence ($BF > 3$) was found either for or against ($BF < 0.33$) the variance congruency effect. The evidence for variance type was found quickly ($BF_{10} = 136.17$) after collecting data from six participants, hence stopped data collection. As in the previous experiments, every aspect of the study was reviewed and approved by the Institutional Review Board of Yonsei University.

Apparatus and Stimuli

All experimental apparatuses were identical to Experiments 1 and 2. The standard display consisted of a set of sixteen outlined circles presented on the center of the display. The size of twelve circles were randomly selected from a normal distribution in which the mean was between 0.7 ~ 1.4°. There were two types of size variance, low variance and high variance. The standard deviation of distribution was set to $1/8^{th}$ of the selected mean for low variance sets and $1/2^{th}$ of the selected mean for high variance sets. As in the previous experiments, the size of the circles was set so that it would not exceed 2 standard deviations from the mean size. The location of circles was determined in the same way as the standard display with sixteen circles in Experiment 2 and was randomly jittered within 0.9169° without being overlapped. The sizes of the probe display set were determined in the same way as the stimulus display, except the mean of the sampling pool distribution ranged from 0.1652~ 2.1662°.

Design and Procedure

Experiment 4 was a 2 (variance congruency: same variance, different variance) x 2 (type of standard display: low variance standard display, high variance standard display) within-subjects design. The experimental conditions are summarized in Figure 8. All experimental procedures were same as Experiment 2 with the only difference being the type of standard display and probe display used. Both the standard and probe displays consisted of 16 circles and could either have low size variance or high size variance. In the same variance condition, the standard display and the probe display both had either low variance or high variance (low & low, high & high). In the different variance condition, variance of the standard display and the probe display was different (low & high, high & low). We sought to see how mean size estimation differ when variance between standard and probe displays are the same versus when they are different. We also looked into how mean size estimation differ according to the type of variance used in the standard display (low or high). The four standard-probe variance pairs were conducted 50 trials each, thus 200 trials in total. The conditions were intermixed across 4 blocks. Participants conducted 12 practice trials (3 trials of each condition) before conducting the main experiment. The entire experimental session took 20 minutes – 25 minutes to complete.

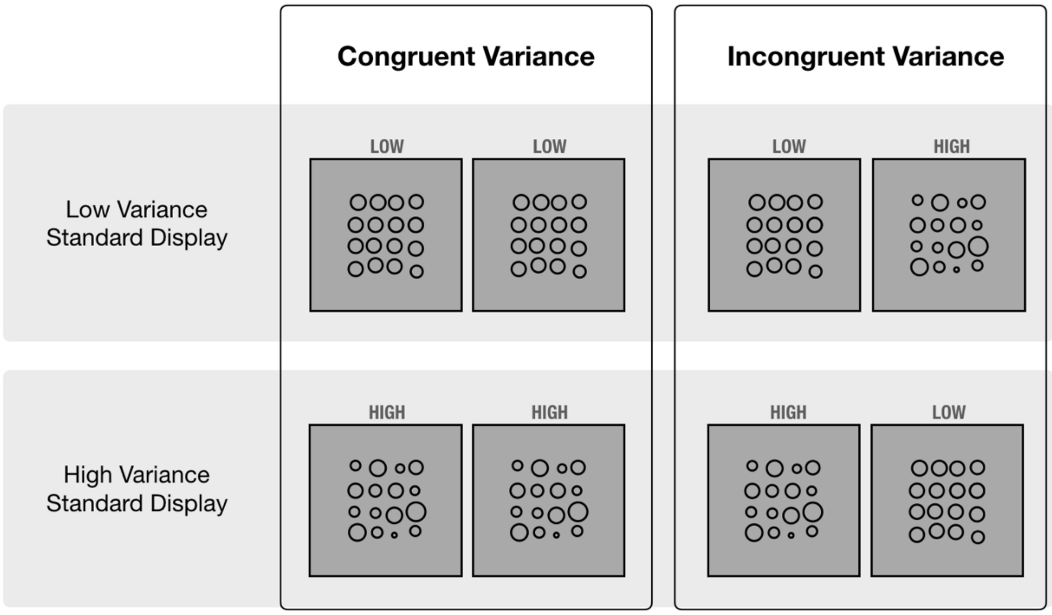

Figure 8. The four standard-probe variance pairs of Experiment 4. Both the standard and probe displays consisted of 16 circles and could have either low size variance or high size variance. In the congruent variance condition, the standard and probe displays both had the same variance type (low & low, high & high). In the incongruent variance condition, the standard and probe displays had different variance type (low & high, high & low).

Results

Data Analysis

The mean RMSD and variance of error was derived for the four types of standard-probe variance pairs. A 2 (variance congruency: same variance, different variance) x 2 (standard display type: low variance standard display, high variance standard display) repeated-measures ANOVA was conducted for both mean RMSD and variance of error.

Relative Mean Size Difference

The mean RMSD for each condition is presented in Figure 9A. There was a significant interaction between variance congruency and standard display type, $F(1, 5) = 32.98$, $p = .002$, $\eta^2_p = .868$, $BF_{10} = 2083.89$. There was no significant difference between the standard display types in the congruent condition of 'low & low' ($M = .019$, $SD = .029$) and 'high & high' ($M = .056$, $SD = .076$), $F(1, 5) = 1.70$, $p = .249$, neither did these two conditions significantly differ from zero (ps $> .129$). However, there was a significant difference between the two standard display types in the incongruent condition of 'low & high' ($M = -.014$, $SD = .029$) and 'high & low' ($M = .135$, $SD = .071$), $F(1, 5) = 58.82$, $p < .001$. While the 'low & high' showed negative RMSD indicating underestimation, the value did not differ significantly from zero, $t(5) = -1.19$, $p = .286$. On the other land, the RMSD for 'high & low' condition was significantly larger than zero (13.50% overestimation), $t(5) = 4.64$, $p = .006$, $d = 1.89$, $BF_{10} = 10.64$.

Variance of Error

The variance of error for each condition is presented in Figure 9B. There was a significant main effect of standard display type, $F(1, 5) = 55.66$, $p < .001$, $\eta^2_p = .92$, $BF_{10} = 42.68$. Error variance was significantly larger when standard display had high variance regardless of the probe display's variance type. There was a marginal effect of variance congruency, $F(1, 5) = 6.01$, $p = .058$, $\eta^2_p = .55$, $BF_{10} = 1.35$. Error variance was larger when variance was incongruent between the two displays.

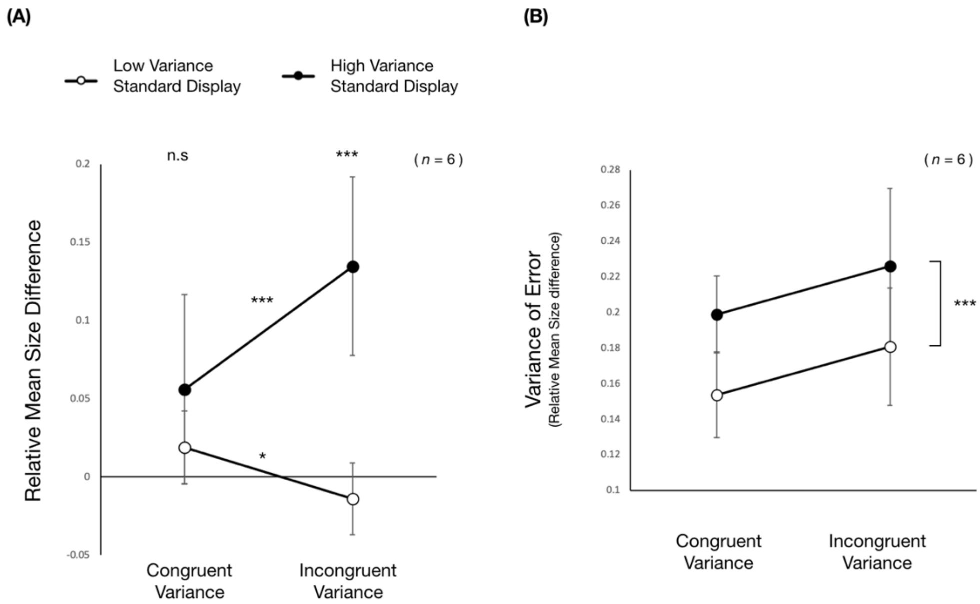

Figure 9. Results of Experiment 4. (A) The mean RMSD depending on variance congruency and standard display type. If standard display has high variance, using a same high variance probe decreases RMSD (B) Variance of error depending on variance congruency and standard display type. Error variance was a marginally larger when variance was incongruent. It was significantly larger when standard display had high variance regardless of the probe display's variance type. All error bars indicate the 95% confidence interval.

Discussion

In Experiment 4, we tested if mean size estimation bias was influenced by variance congruency between two sets. Results showed that having congruent variance between the two sets made the mean size comparison easier, evidenced by overall smaller estimation error and error variance in ‘congruent’ than ‘incongruent’ condition. But when variance was incongruent, bias increased/decreased in opposite directions depending on the standard display type. When high variance was adjusted to match the low variance standard display, error remained low. However, when low variance was adjusted to match the high variance standard display, there was drastic increase resulting in overestimation. This means that mean estimation for low variance sets are relatively accurate regardless of the probe type. In contrast, mean estimation for high variance sets are more difficult, also evidenced by higher error variance for high variance standard displays. In fact, it has been found in the previous studies that high variance increases mean size discrimination threshold (Chong & Treisman, 2005; Corbett et al., 2012; Dakin, 2001; Fouriezos et al., 2008; Haberman et al., 2015). However, the important point to note is that even if standard display has high variance, using the same high variance set to report the mean facilitates mean estimation performance (decreasing error and error variance). If the ensemble’s size representation was indeed represented in a form of a single average size, there is no reason why variance disparity should yield larger errors or response variance. Because regardless of size variance, both ensembles would have similar average.

Therefore, the results suggest that variance information is included in the mean representation so that having variance congruency helps mean estimation easier.

GENERAL DISCUSSION

This study examined whether mean size is represented as a single size. For this purpose, we conducted a mean size estimation task and examined how estimation error is influenced by the disparity between the standard and probe display, by varying set size and variance. In Experiment 1, we compared mean size estimation performance between using a single probe versus a set probe. Replicating the macro trend across previous studies, estimation error was greater in the single probe condition than the set probe condition. Error variance was also smaller when set probe was used. In Experiment 2, we further divided probe's set size into four levels. Results showed that mean estimation error and response variance reduced as set size disparity decreased between the standard and probe displays. Experiment 3 was conducted to check whether this observed difference between set size disparity may have been due to differences in sensory memory overlap. We examined if results are different when the probe display is presented in a different location from the standard display versus the same location. Results showed that there was no significant difference between the same location and different location conditions, suggesting that set size disparity effect was indeed due to representation differences. Experiment 4 examined how variance congruency between the standard and probe displays influenced mean size estimation. There was a large congruency effect for mean estimation error when variance of the standard display was high. Variance of error was always smaller when variance was congruent than when variance was incongruent.

The current study, thus, suggests that the size ensemble is not represented as a single size, but includes ensemble of statistical properties, such as variance and numerosity. If the size ensemble was indeed reduced to a single average size, there is no reason why such numerosity or variance disparity should make mean estimation more difficult (larger errors and response variance). When viewing multiple items, the items' various sizes are summarized into a representation that embraces other statistical properties of the feature distribution so that two sets have different representation even if they have identical means. Accordingly, when the task requires the mean information to be reduced down to a single unit or compared with another representation, the original representation should go through a conversion process that translates it into a comparable unit, which in the course yields estimation errors and loss of precision. This comparison, on the other hand, is easier when using a similar set because no information has to be reduced and converted.

~~The fact that ensemble representation is summed to a single mean implies that~~ the pooling account of ensemble averaging (e.g., Allik et al., 2013) needs revised understanding (Comparison in Figure 10). The pooling account explains that individual measurements are averaged to a higher level of abstraction, during which the random noise of individual measurements cancel out. The resulting ensemble representation is an average which is more precisely represented than the individuals in the set. This benefit of averaging is bigger when noise in individual measurements is less correlated, and more individual measurements are averaged (Alvarez, 2011). However, the results of this study suggest that these individual measurements do not accrue to one dominant peak, but more information about the distribution of composing individual representation is preserved such as how variable the sizes are and how numerous the ensemble is. In fact, we believe that the ensemble size representation could be understood as a population summary where individual measurements are summed into a distribution of magnitude, which includes mean, variance, and numerosity. This concept of population coding could be found in Georgopoulos, Schwartz & Kettner (1986) where the final direction of motion was determined by population activation of individual neurons tuned for different directions. Similarly, the individual measurements need not be reduced to a single average in an additional level of abstraction. The various individual measurements could simply be integrated as a population vector, which simply summarizes the shape of the distribution including the range, variability of size, and even different frequencies on the size spectrum (numerosity). In other words, mean size is represented as a pooled activation of population of individual representations tuned for different sizes each making weighted contributions.

In fact, several studies have found evidence that the visual system represents the shape of distribution. In Utochkin and Yurevich (2016)'s study, the segmentability of the orientation feature distributions affected search duration. More distractor distributions made search slower despite the same distance between the target and the nearest distractors. Moreover, study in ensemble statistical learning have found that response times depend on the shape of the preceding distractor distribution, suggesting observers can implicitly learn properties of distractor distributions (Chetverikov, Campana, & Kristjánsson, 2016, 2017). Another related evidence that suggests shape of distribution coding is that mean estimation becomes less accurate when variance is high (Chong & Treisman, 2005; Corbett et al., 2012; Dakin, 2001; Fouriezos et al., 2008; Haberman et al., 2015; and more). This makes sense if we are representing the distribution of the ensembles rather than one mean. Mean is easier to identify when distribution is homogenous. But when variance is high or is bipolar, finding one representative spot in the distribution is not so easy.

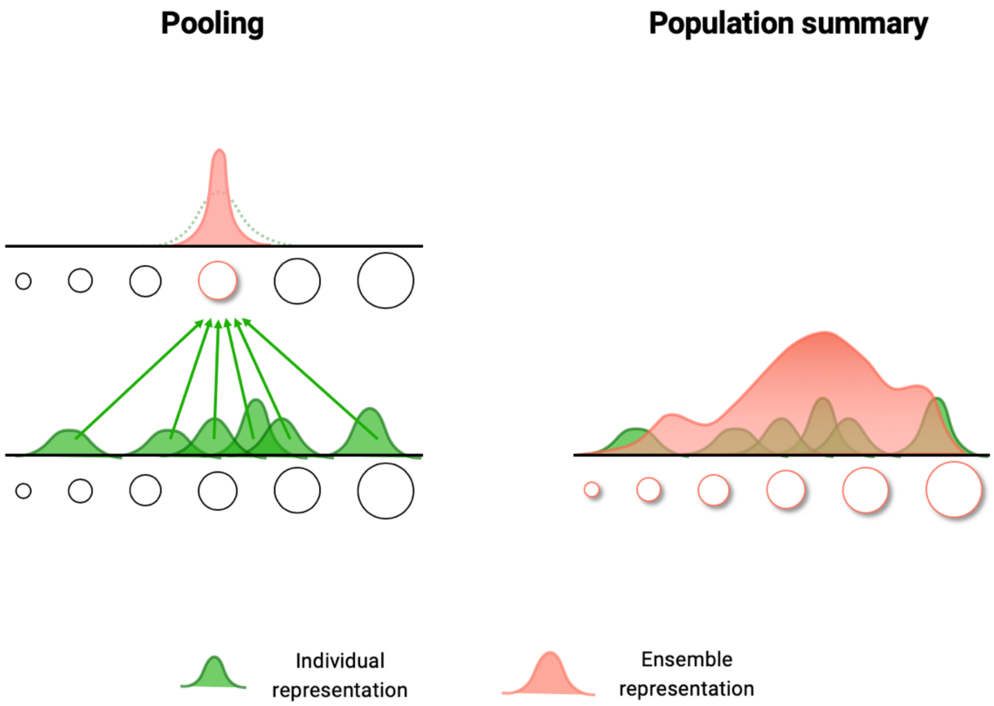

Figure 10. Simple pooling model of ensemble averaging versus population summary. In pooling account, individual representations are pooled together to a single average in a higher level of abstraction. During the course, the random noises of individual measurements will cancel out creating an ensemble representation which is more precise than the individual measurements. In

contrast, population summary does not require an additional level of abstraction, as individual measurements are simply summed as a population activation. The population vector summarizes the shape of the distribution, which portrays the prominent activation (mean) as well as the range, variability, and numerosity information.

Understanding mean size representation as a population summary does not contradict with the previous findings that seemingly show that ensemble is being averaged into a single mean. For example, studies have showed evidence for the benefit of averaging such that observers are able to judge the average size of a set more accurately than the size of individual items (Ariely, 2001), which suggests that averaging individual measurements to an ensemble mean can create more precise representation (Alvarez, 2011). In fact, noise cancellation is still relevant in population summary as it involves the same process of pooling individual measurements to a simplified form. The difference is that it is not reduced to a single peak, but trims down to a distributional shape as many activations accrue to portray most prominent peak(s). In addition, Brady & Alvarez (2011) have found that items are represented through multiple levels of abstraction. In their study, the remembered size of each individual item in a display was biased towards the mean size of the set of items of the same color as well as the mean of all items in the display. This could well be understood as a population summary, where items are abstracted as a population distribution through multiple levels instead of a single mean for each level. As such, understanding ensemble representation as a population summary shifts the current focus in understanding the mechanism of ensemble representation. Perhaps the key is not how individual items are selected or canceled out in the computation of an average percept, but how the probability distribution of the visual input is translated into the internal probability distribution.

One may argue that all estimation error results shown in the current study can well be explained by the pooling account. The bias can still be accommodated with single average representation if we think of it as visual system pitting against two biased means: ensemble is reduced to a single average size but positive biases occur because the mean representation itself is biased due to some characteristic of a set (like a perceptual distortion), rather than a result of representational conversion. This perceptual distortion shall be reflected when this biased mean is to be matched up with a single probe, but it will not be reflected when an identical set is used because it simply compares two biased sets of representations. However, the fact that precision of response (response variance) is also larger in the single probe condition suggests that representing the ensemble size in terms of a single size entails more uncertainty compared to the set probe condition. Thus, representing the size of multiple items in terms of a single size unit is more difficult, or leastways, not the most natural way it is represented in the visual system.

One thing to note is that the current study only provides an indirect support that observers are encoding the shape of distribution. Experiment 4 did not compare different shapes of distributions but only controlled for variance. In other words, depending on the way items were sampled, it could have produced different distribution shapes even for the same variance condition. It is worth designing another experiment that compares more complex distribution shapes (e.g., uniform, normal, polar, and skewed) and examine whether congruency of these specific shapes is more relevant to the mean representation than simple feature variance or range.

Not all set disparities yielded mean estimation errors. We believe the boundary condition where errors occur could be related to numerosity discrimination threshold. In Experiment 2, bias in 16:9, 16:16 conditions had no significant difference. In Experiment 3, there was no difference between 9:4, 9:9, 9:16 conditions when presented on the same side. These numerosity differences may not be big enough difference to produce errors because the visual system does not perceptually distinguish between these set-size difference as different numerosities. Therefore, not much conversion was necessary between those two conditions as they are considered as more or less the same numerosity. In fact, studies on infant's number representation show that infants require the ratio of at least 2:3 1:2 to discriminate between number of elements (Xu, Spelke, & Goddard, 2005). From this, we can infer that with a standard display consisting of 16 items, it would require the probe display to have 11 to 8 items or less for the sets to seem as perceptually different numerosities. In the same sense, 9:9 & 9:16 conditions of Experiment 3 should rather have been 9:9 & 9:18 to be considered different numerosities. The fact that set size effect correlates to numerosity discrimination threshold is another strong evidence that numerosity property is included in the mean representation.

Why is the direction of bias most often 'positive'? In other words, why does conversion result in overestimation and not underestimation? This can be explained by saliency-weighted averaging where large items are salient and therefore biases the mean size to be estimated larger than the arithmetic mean (Kanaya et al., 2018). The saliency of large items as opposed to smaller items was shown through several studies. For example, detecting a large target among small distractors is easier and faster than vice versa, showing a search asymmetry for a larger target (Treisman & Gormican, 1988). The set with the largest mean size attracted selective attention in the same way as larger individual objects (Im, Park, & Chong, 2015; Proulx, 2010). Infant studies demonstrated that this perceptive preference for larger objects is observed from early infancy (Newman, Atkinson, & Braddick, 2001). From a developmental perspective, it has been speculated that visual aspects of magnitude such as size, height or quantity may fall on a mental scale with "taller, larger or more" at the positive end and "shorter, smaller or less" at the negative end, which is why children prefer and master the "positive" terms (i.e., taller, larger or more) earlier than the "negative" terms (i.e., shorter, smaller or less; Barner & Snedeker, 2008; Donaldson & Balfour, 1968; Klatzky, Clark, & Macken, 1973). Due to this large item saliency, the set that include larger items, such as those with larger set size or higher variance are perceptually viewed as having larger sizes on average. Accordingly, when it is compared against a single probe, mean size is overestimated. In other words, having larger apples makes it seem like the apple in that bundle is larger on average. But we have to note that this does not mean people calculate one size to represent the bundle of apples. It simply means that larger items have more weight in the mean representation.

It makes sense that mean representation should include various statistical properties if we consider the functional role of perceptual averaging in the natural world. Averaging a pool of similar items into an ensemble makes it easier to efficiently reduce redundancy. But if mean is indeed represented only in terms of a single size, then we would be facing many confusions distinguishing between things that have similar averages, but yet have different variance or numerosity. For example, when we view a panorama of autumn foliage where all trees are dyed in diverse yet similar mixture of red, brown, yellow hues, we do not have trouble distinguishing one tree from another. We are able to grasp the sense of subtle differences such as, which tree still has more green leaves remaining, which is more homogenous in color etc. Because we are representing the shape of distribution along the feature spectrum (whether it be color or size), it enables us to distinguish sets' different distributional characteristics based on other properties even if the two ensembles are similar in average.

In conclusion, this study offers strong evidence that mean representation is not form of a single average, but includes ensemble of statistical properties, such as numerosity and variance. Error and response variance of mean estimation were contingent on the similarity of probe displays' variance and set size, suggesting that various properties are bound to the mean representation. Contrast to the simple pooling model, ensemble summary may be better understood as a population activation representing the shape of feature distribution. Such representation helps us to distinguish the subtle differences in a complex visual scene where the average is similar. Previous studies showed that various statistical properties are correlated to the mean. The current study extends these findings to show that properties are not only correlated, but a part of one composite mean representation.

REFERENCES

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

Allik, J., Toom, M., Raidvee, A., Averin, K., & Kreegipuu, K. (2013). An almost general theory of mean size perception. *Vision Research, 83*, 25–39.

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends Cognitive Sciences, 15*(3), 122–131.

Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science, 19*(4), 392–98.

Ariely, D. (2001). Seeing sets: representation by statistical properties. *Psychological Science, 12*(2), 157–62.

Banno, H., & Saiki, J. (2012). Calculation of the mean circle size does not circumvent the bottleneck of crowding. *Journal of Vision, 12*(11), 13.

Barner, D., & Snedeker, J. (2008). Compositionality and statistics in adjective acquisition: 4-year-olds interpret tall and short based on the size distributions of novel noun referents. *Child Development, 79*(3), 594–608.

Bauer, B. (2009). Does Stevens's power law for brightness extend to perceptual brightness averaging?. *Psychological Record, 59*, 171–86.

Bauer, B. (2017). Perceptual averaging of line length: Effects of concurrent digit memory load. *Attention, Perception, & Psychophysics, 79*, 2510–2522.

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

Chetverikov, A., Campana, G., & Kristjansson, A. (2016). Building ensemble representations: how the shape of preceding distractor distributions affects visual search. *Cognition, 153*, 196–210.

Chetverikov, A., Campana, G., & Kristjansson, A. (2017). Representing Color Ensembles. *Psychological Science, 28*(10), 1510–1517.

Chong, S. C., Joo, S. J., Emmanouil, T.A., & Treisman, A. (2008). Statistical processing: not so implausible after all. *Perception & Psychophysics, 70*(7), 1327–34.

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

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

Corbett, J. E., Wurnitsch, N., Schwartz, A., & Whitney, D. (2012). An aftereffect of adaptation to mean size. *Visual Cognition, 20*(2), 211–231.

Dakin, S. C. (2001). Information limit on the spatial integration of local orientation signals. *Journal of the Optical Society of America A, 18*(5), 1016–26.

de Fockert, J. W., & Marchant, A. P. (2008). Attention modulates set representation by statistical properties. *Perception & Psychophysics, 70*(5), 789–94.

Demeyere, N., Rzeskiewicz, A., Humphreys, K. A., & Humphreys, G. W. (2008). Automatic statistical processing of visual properties in simultanagnosia. *Neuropsychologia, 46*(11), 2861–64.

Donaldson, M., & Balfour, G. (1968). Less is more: a study of language comprehension in children. *British Journal of Psychology, 59*(4), 461-471.

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A, 4*(12), 2379–2394.

Fischer, J., & Whitney, D. (2011). Object-level visual information gets through the bottleneck of crowding. *Journal of Neurophysiology, 106*(3), 1389–98.

Fitts, P. M., & Deininger, R. L. (1954). S-R compatibility: Correspondence among paired elements within stimulus and response codes. *Journal of Experimental Psychology, 48*, 483–492.

Florey, J., Clifford, C. W. G., Dakin, S., & Mareschal, I. (2016). Spatial limitations in averaging social cues. *Scientific Reports, 6*, 32210.

Fouriezos, G., Rubenfeld, S., & Capstick, G. (2008). Visual statistical decisions. *Perception & Psychophysics. 70*(3), 456–64.

Georgopoulos, A. P., Schwartz, A. B., & Kettner, R. E. (1986). Neuronal population coding of movement direction. *Science, 233*(4771), 1416–1419.

Haberman, J., Brady, T. F., & Alvarez, G. A. (2015). Individual differences in ensemble perception reveal multiple, independent levels of ensemble representation. *Journal of Experimental Psychology: General, 144*(2), 432–46.

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

Haberman, J., & Whitney, D., (2011). Efficient summary statistical representation when change localization fails. *Psychonomic Bulletin & Review, 18*(5), 855–59.

Halberda, J., Sires, S. F., & Feigenson L. (2006). Multiple spatially overlapping sets can be enumerated in parallel. *Psychological Science*, *17*(7), 572–576.

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

Hurewitz, F., Gelman, R., & Schnitzer, B. (2006). Sometimes area counts more than number. *Proceedings of the National Academy of Sciences, 103*, 19599–19604.

Im, H. Y., & Chong, S. C. (2014). Mean size as a unit of visual working memory. *Perception, 43*(7), 663–76.

Im, H. Y., Park, W. J., & Chong, S. C. (2015). Ensemble statistics as units of selection. *Journal of Cognitive Psychology, 27*(1), 114–127.

Im, H. Y., & Halberda, J. (2013). The effects of sampling and internal noise on the representation of ensemble average size. *Attention Perception & Psychophysics, 75*(2), 278–86.

Jeffreys, H. (1961). *Theory of probability* (3rd ed., pp. 470). Oxford: Oxford University Press.

Joo, S. J., Shin, K., Chong, S. C., & Blake, R. (2009). On the nature of the stimulus information necessary for estimating mean size of visual arrays. *Journal of Vision. 9*(9), 7.

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, 285*, 20172770.

Klatzky, R. L., Clark, E. V., & Macken, M. (1973). Asymmetries in the acquisition of polar adjectives: linguistic or conceptual?. *Journal of Experimental Child Psychology, 16*, 32–46.

Kornblum, S., Hasbroucq, T., & Osman, A. (1990). Dimensional Overlap: Cognitive Basis for Stimulus-Response Compatibility--A model and taxanomy. *Psychological Review, 97*(2). 253–270.

Kruijne, W., & Meeter, M. (2016). Long-term priming of visual search prevails against the passage of time and counteracting instructions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(8), 1293–1303.

Lee, H., Baek, J., & Chong, S. C. (2016). Perceived magnitude of visual displays: Area, numerosity, and mean size. *Journal of Vision, 16*(3), 12.

Leon, M., Oden, G. C., & Anderson, N. H. (1973). Functional measurement of social values. *Journal of Personality and Social Psychology, 27*(3), 301–10.

Leon, M. I., & Shadlen, M. N. (2003). Representation of time by neurons in the posterior parietal cortex of the Macaque, *Neuron, 38*, 317–327.

Li, K. A., & Yeh, S. L. (2017). Mean size estimation yields left-side bias: role of attention on perceptual averaging. *Attention Perception & Psychophysics, 79*, 2538–2551.

Marchant, A. P., Simons, D. J., & de Fockert, J. W. (2013). Ensemble representations: effects of set size and item heterogeneity on average size perception. *Acta Psychologica, 142*(2), 245–50.

Mareschal, I., Morgan, M. J., & Solomon, J. A. (2010). Attentional modulation of crowding. *Vision Research, 50*, 805–809.

Maule, J., & Franklin, A. (2016). Accurate rapid averaging of multihue ensembles is due to a limited capacity subsampling mechanism. *Journal of the Optical Society of America A. 33*, A22–29.

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

Myczek, K., & Simons, D. J. (2008). Better than average: alternatives to statistical summary representations for rapid judgments of average size. *Perception & Psychophysics, 70*(5), 772–88.

Newman, C., Atkinson, J., & Braddick, O. (2001). The development of reaching and looking preferences in infants to objects of different sizes. *Developmental Psychology, 37*(4), 561–572.

Onoe, H., Komori, M., Onoe, K., Takechi, H., Tsukada, H., & Watanabe, Y. (2001). Cortical networks recruited for time perception: A monkey positron emission tomography (PET) study. *Neuroimage, 13*, 37–45.

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

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

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

Parkes, L., Lund, J., & Angelucci, A. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience, 4*(7), 739–44.

Pinel, P., Piazza, M., Bihan, L. D., & Dehaene, S. (2004). Distributed and overlapping cerebral representations of number, size, and luminance during comparative judgments. *Neuron, 41*, 983–993.

Proulx, M. J. (2010). Size matters: large objects capture attention in visual search. *PloS ONE, 5*(12), e15293.

Sawamura, H., Shima, K., & Tanji, J. (2002). Numerical representation for action in the parietal cortex of the monkey. *Nature, 415*, 918–922.

Setic, M. Svegar, D., & Domijan, D. (2007). Modelling the statistical processing of visual information. *Neurocomputing, 70*, 1808–1812.

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

Stein, J. (1989). The representation of egocentric space in the posterior parietal cortex. *Experimental Physiology, 74,* 583–606.

Sweeny, T. D., Wurnitsch, N., Gopnik, A., & Whitney, D. (2015). Ensemble perception of size in 4–5-year-old children. *Developmental Science, 18*(4), 556–68.

Teghtsoonian, M. (1965). The judgment of size. *American Journal of Psychology, 78*, 392–402.

Tong, K., Ji, L., Chen, W., & Fu, X. (2015). Unstable mean context causes sensitivity loss and biased estimation of variability. *Journal of Vision. 15*(4), 15.

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

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*(7), 995–1007.

Walsh, V. (2003). A theory of magnitude: common cortical metrics of time, space and quantity. *Trends in Cognitive Sciences, 7*(11), 483–488.

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.

Whitney, D., & Yamanashi Leib, A. (2018). Ensemble Perception. *Annual Review of Psychology, 69*, 105–29.

Xu, F., Spelke, E. S., & Goddard, S (2003). Number sense in human infants. *Developmental Science*. 8(1), 88–101.

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

Yamanashi Leib A., Puri, A. M., Fischer, J., Bentin, S., Whitney, D., & Robertson, L. (2012b). Crowd perception in prosopagnosia. *Neuropsychologia, 50*(7), 1698–707.

국 문 요 약

평균 표상의 형태

인간은 유사한 사물들에서 평균과 같은 통계 정보를 추출하여 복잡한 시각 정보를 효율적으로 표상할 수 있다. 평균 정보의 형태에 대해선 알려진 바가 적지만 대체로 평균 정보가 단일 크기와 같은 하나의 대표값으로 표상된다고 가정하는 듯하다. 하지만 단일 원을 이용하여 평균을 보고하는 대부분의 연구에서 상대적으로 큰 오차가 나타나는 점과 평균을 포함한 다양한 통계 정보들이 서로 상관되어 있다는 점을 미루어 볼 때 평균 표상이 단일 크기 보다는 더 복잡한 형태라는 것을 유추해 볼 수 있다. 본 연구는 평균 표상이 비교하는 두 자극의 특성 차이에 따라 어떻게 달라지는 지 연구하였다. 이를 위하여 자극 화면과 검사 화면 사이의 자극 개수 차이와 자극 크기들의 변산 차이를 조작하였다. 참가자들은 자극 화면에서 짧게 제시된 다양한 크기의 원들을 본 후 검사 화면에서 나타난 원(들)의 크기를 조절하여 자극 화면에 제시된 원들의 평균 크기를 추정하였다. 실험 1에서는 응답 화면의 원이 하나인 경우와 여러 개(세트)인 경우를 비교하였다. 실험 결과 단일 원 조건이 세트 조건보다 평균 추정 오차와 오차의 변산이 큰 것으로 나타났다. 실험 2 에서는 검사 화면의 원의 개수를 네 단계로 세분화 하였다. 실험 결과, 자극 화면과 검사 화면의 자극 개수

차이가 커질수록 오차가 감소하는 것으로 나타났다. 실험 3에서는 앞서 나타난 결과가 위치 중첩으로 인해 나타난 단순 감각 기억의 유사성 이었는지 확인해보기 위해 응답 화면 자극이 나타나는 위치가 같을 때와 다를 때를 비교하였다. 실험 결과, 두 조건에 차이가 없는 것으로 나타났다. 마지막으로 실험 4에서는 자극 화면과 응답 화면 간에 변산이 일치하거나 다를 때 평균 추정이 어떻게 변하는지를 알아보았다. 실험 결과, 변산이 일치할 때 오차와 오차의 변산이 작은 것으로 나타났다. 결론적으로 평균 추정 오차와 오차의 변산은 자극 화면과 응답 화면의 유사성에 따라 달라졌다. 이는 평균 정보가 하나의 대표값으로 표상되는 것이 아니고 변산과 개수와 같은 통계 정보들이 평균 표상에 유기적으로 포함되어 있다는 것을 시사한다.

핵심어: 평균 표상, 평균 크기, 통계 정보, 개수, 크기 변산