UNIVERSITY OF CALIFORNIA,
IRVINE

Mean-Feature and Centroid Estimations:
A study on how the human visual system processes two classes of summary statistical representations

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Cognitive Neurosciences

by

Laris M. Rodriguez-Cintron

Dissertation Committee:
Charles E. (Ted) Wright, Ph.D. - Chair
Charlie Chubb, Ph.D.
Emily Grossman, Ph.D.

2019

DEDICATION

To
Gabriel and Isaac
My inspiration, my everything!

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express the deepest appreciation to:

Antonio, thank you for always giving me your unconditional love and support. I am the luckiest to have you as my husband.

My family in California and in Puerto Rico: Nylsa and Tomas, I am very fortunate to have you as my parents. I owe everything to you. Omar: you are the best brother and an excellent writer!

My dear friend Carol, is simple… without you this would have been impossible.

The Chubb Wright Lab members – thank you for all your help running the experiments and all your support.

Charles (Ted) Wright, Charlie Chubb and George Sperling for you advice and support and the contributions to the published paper.

Jeffrey Kirkham and Emily Grossman, for your advice and support along the way.

Finally, but definitively not least… the best advisor and mentor there is. Ted, this is all because of you. You made the impossible… possible. THANK YOU!

CURRICULUM VITAE

Laris M. Rodriguez Cintron

Department of Cognitive Sciences

Irvine, CA

Email: larisr@uci.edu

larismrodriguez@gmail.com

EDUCATION

2000: B.S. Mechanical Engineering - University of Puerto Rico, Mayaguez, PR

2008: MFA Film and TV production - University of Southern California, Los Angeles, CA

2019: Ph.D. Cognitive Neurosciences, University of California, Irvine, CA

WORK EXPERIENCE

2000-2003: Powertrain Engineer - Ford Motor Company, Dearborn, Michigan

2005-2008: Freelance Video Editor

2012-2013: Supply Engineer - Boston Scientific, San Clemente, CA

2017-2018: Community College teaching internship - OCC, Costa Mesa, CA

TEACHING EXPERIENCE

2013-2018: Teaching Assistant, University of California, Irvine, CA

PUBLICATIONS

2019: How can we use perceived size? Centroid versus Mean-size Judgments - Journal of Vision (Chapter 3 in this thesis)

ABSTRACT OF THE DISSERTATION

Mean-Feature and Centroid Estimations:
A study on how the human visual system processes these two summary statistical representations

By

Laris M. Rodriguez Cintron

Doctor of Philosophy in Cognitive Neurosciences

University of California, Irvine, 2019

Professor Charles E. Wright

When presented with a complex scene our visual system relies on perceptions that occur pre-attentively, or in just a fraction of a second. This ability, known as summary statistical representations, allows us to get the gist of a group of items. The studies presented in this thesis are focused on two of these summary statistical representations, the centroid and the mean-feature estimations, and are designed to understand mean-feature estimations in the centroid paradigm framework. First, we compared observers' performances between mean-size and centroid estimations. There were two different centroid tasks: in one, observers were asked to give more weight to larger items and in the other they were asked to ignore size. In a third task, observers judged the mean size of the same stimuli used in the centroid tasks. Results from this study showed that observers were able to either ignore or use size to make centroid responses efficiently, but that mean-size judgments were inefficient. To asses the generality of these results, in a second study we asked observers to estimate the mean luminance of a group of items and compared their performance when asked to estimate the centroid giving more weight to lighter items and less weight to darker items or to estimate the centroid ignoring the luminance of the items. Again, we found high Efficiency in both centroid

tasks, but low Efficiency in the mean-luminance task. Finally, we explored whether the distinction between the ventral-dorsal streams of visual processing as described by Goodale & Milner (1992) might provide a plausible explanation for the difference in performances between the mean-size and centroid task. We found that when the relative mean size of one of three groups was used to select and guide movements (dorsal processing) to that group, observers were again able to use size information efficiently.

Chapter 1

Literature Review

As ensemble is defined as a visual representation that minimizes individual features to provide a single description of a group of items (Ariely, 2001). When representing multiple objects as an ensemble, our visual cognition is enhanced to process the gist of the scene and we use selective attention to process only the most relevant information (Alvarez, 2011; Chong & Treisman 2005). While this ensemble representation provides us with an accurate description of the gist of the group, it restricts the processing of the details of any individual features. For example, when we are scanning a supermarket stand to buy the apple, it would be nearly impossible to attend to, and remember the individual sizes of each apple. However, these restrictions are a small price to pay since summary statistical representations provide us with a helpful and efficient way to process large amounts of information for multiple items.

The consensus of most of the previous research is that people are very good at the discrimination of the average size (or mean size) of a group. Ariely (2001) and Chong & Treisman (2003, 2005) found that observers were able to judge better the mean size of a set independently of its set size, heterogeneity, numerosity and density than determining if a single item was a member of that set. Based on these results they concluded that our visual system is able to judge the mean size of a set through a parallel process using a global strategy (Ariely 2001, Chong & Treisman 2003, 2005). This means that when judging the mean size of a group of items, our visual system gathers information from the whole display, processing all the items presented in the stimuli simultaneously.

Ariely (2001) claimed that observers were not only very good at calculating the mean size of a group, but that increasing the number of items in the display did not impair the

accuracy of the estimation. In his experiment, observers were presented a display for 500ms. Each display had circles of four different sizes, and each display could consist of 4, 8, 12 or 16 circles. A second display followed, in which a single circle was displayed (Figure 1.1). At that point observers were asked to perform two different tasks: to determine if the circle presented in the second display was a member of the first group of items or if the circle was smaller or larger than the mean size of the group of circles presented in the first display. The results showed that observer's performance was above chance, for a threshold between 6-12%, when they were estimating the mean size of the set. This performance was substantially better than when asked if the second circle was a member of a set.

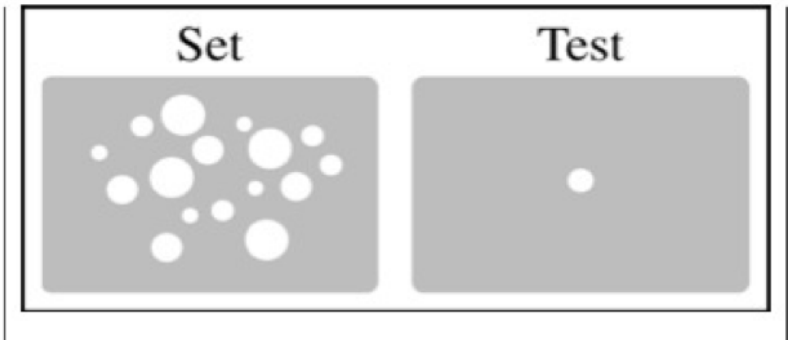


Figure 1.1: Schematic representation of Ariely (2001) experiment. On the left there is an example of a display with 16 circles with groups of 4 different sizes. To the right is the second display, which prompted the observers' response.

Building on the results presented by Ariely (2001), Chong and Treisman (2003) conducted several experiments in which they tested if any of a series of manipulations that added complexity to the task would impair observers' performance. In their first experiment, observers were presented a divided screen (Figure 1.2) and were asked to discriminate the mean size between the two sides of the displays. In each trial, the divided displays presented different conditions for comparison: heterogeneous sets, homogenous sets or single items (clearly no averaging calculation is needed in the

single item condition and, arguably, this is true also for the homogenous condition). The sessions consisted of mixed blocks of two modes of stimuli presentation: simultaneous or successive (Figure 1.2). In the successive presentation, the first display was located on one side of the visual field and after 100ms or 2 seconds a second display followed in the opposite side of the visual field. In the simultaneous condition, the divided screen presented the two displays at the same time. The simultaneous version of this experiment, along with Attarha, Moore & Vecera (2014) which will be described later, were used as an inspiration for the display we used in the experiment described in Chapter 5.

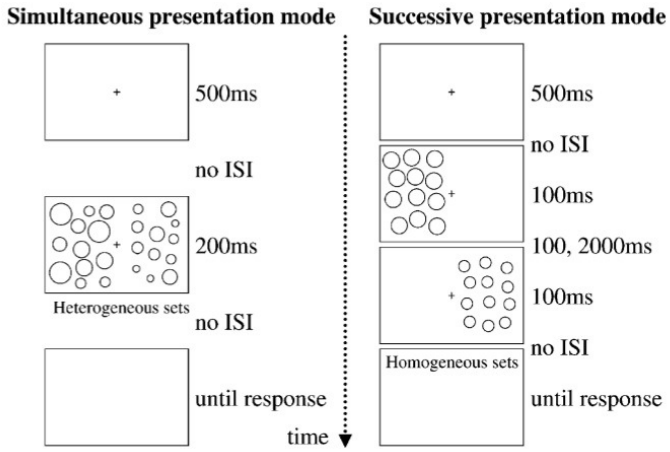


Figure 1.2: Schematic Representation of Chong and Treisman (2003) experiment. On the left an example of simultaneous presentation for a heterogeneous set and on the right an example of a successive pretention for a homogenous set.

Observers achieved 75% accuracy, when the difference in mean diameter was between 6-8% for a simultaneous presentation, and between 8– 10% for the successive presentation. Presentation time had little effect on the thresholds for the homogeneous arrays of circles (Figure 1.3). The thresholds for the heterogeneous arrays (mean) and the single circles did increase with delay, but they argued that this increase was minimal

(2%). They concluded that observers' performance under brief exposures was unaffected, due to a global, parallel process and suggested that it is unlikely that, with such brief exposure, observers have the time to use any other strategy (Chong and Treisman, 2003).

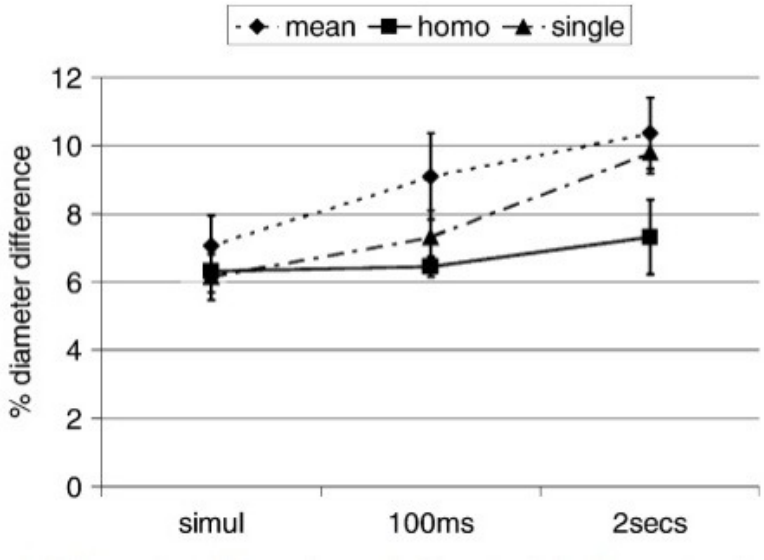


Figure 1. 3: Results Experiment 1 Chong and Treisman (2003). X-axis shows the duration of the stimulus exposure, the y-axis shows the percent diameter difference between the two displays on any given trial.

The second experiment was very similar to the simultaneous displays used in the first experiment, except that they varied the display presentation time within blocks. In the first experiment the exposure duration ranged from 100 ms to 2 seconds, but in this experiment it ranged from 50 ms to 1 second. Performance was not impaired by a brief successive exposure of 50 ms, for all conditions - single size items, homogeneous and heterogeneous groups.

The third experiment was conducted to test if observers were really averaging sizes. They presented observers with groups of circles that were chosen from four different distributions – uniform, two peaks, normal, and homogenous (Figure 1.4). They argued that these distributions present a challenge to the observer that eliminates the possibility of using any strategy other than averaging the whole display. How? They posit that when presented with different distributions, observers were forced to compare the means of the whole group (Chong and Treisman, 2003). They do not see a subsampling strategy as plausible strategy being used because, they argue, that if observers were basing their judgments on a fixed-size subsample, then an increase in set size should lead to a decrease in performance, which they did not find. Observers did obtained thresholds 2% higher when discriminating displays from the same distribution. Their claim again is that a 2% is just a marginal increase of threshold.

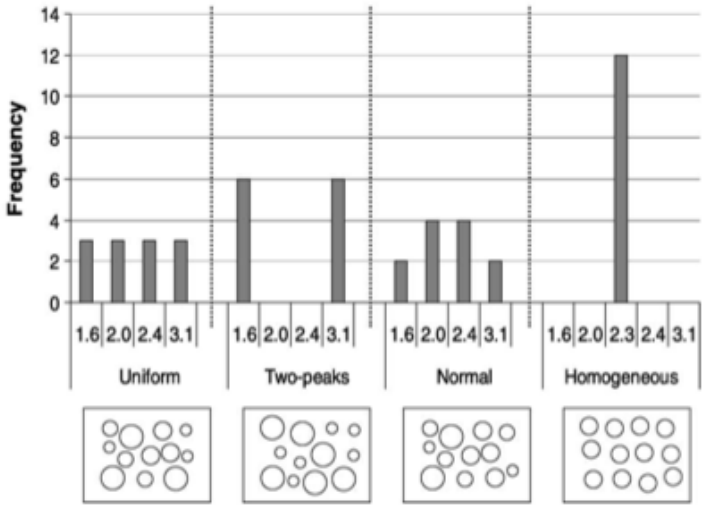


Figure 1. 4: In Experiment 3, Chong and Treisman varied the distribution of the display with that aim of presenting more difficult challenge for the mean size estimation. Since performance was minimally affected they argue that observes must be using a global strategy.

In another study conducted by Chong and Treisman (2005) they tested observers' ability to estimate the mean size of a group in three different experiments where: (1) numerosity and density of the display was manipulated (Figure 1.5), (2) mean size had to be determined in sets with two different colors (blue and green), and (3) when the colors of the members of the set were cued or non-cued.

The aim of the first experiment in this study was to understand the weight observers gave to numerosity and density when judging the mean size. The display was split in two arrays - one array was presented on the left visual field and the other on the right visual field (Figure 1.5). Each trial presented observers with blocks in which the density and numerosity of one of the visual fields was either matched or mismatched to the other visual field. The task was to determine which visual field had dots with the larger the mean size. They found that the manipulation of numerosity or density in their displays had little effect in the observers' performance.

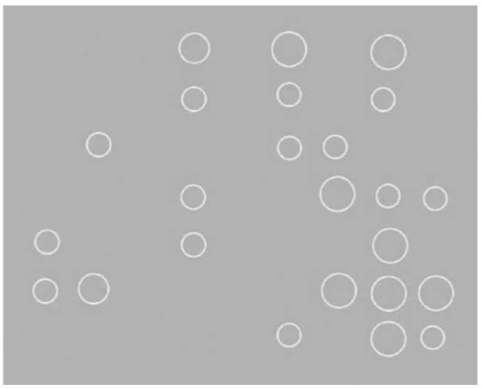


Figure 1. 5: Schematic representation Chong and Treisman (2005) Experiment 1. In this example the array on the left was less dense (they called it sparse - S) that the one on the right (dense - D).

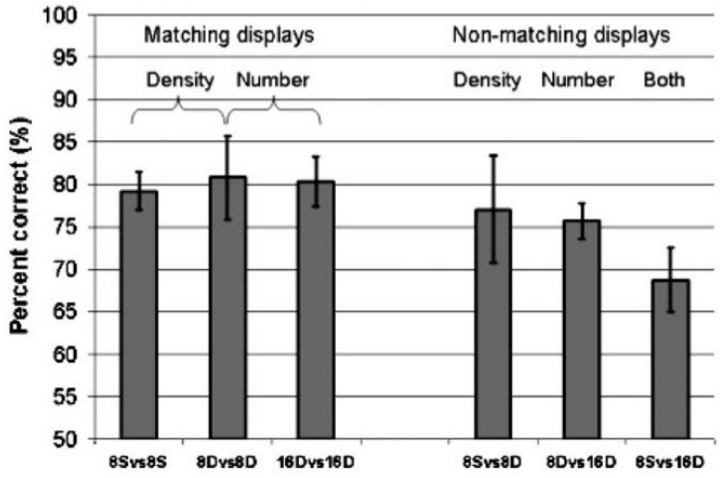


Figure 1. 6: Results from numerosity and density experiment (Chong and Treisman, 2005). S stands for the ***sparse*** condition and D for the ***dense*** condition. In the matching displays, performance was not impaired by the size of the set or how sparse or dense they were. The performance in the non-matching displays performance was slightly worse than performance in the matching displays, but it was still 15 % over chance level.

In the second and third experiments color was introduced as a variable. The display presented either 1 green and 1 blue circle or a set of 12 circles of each color. The task was to discriminate which was the subset with the larger mean size between the two colors by selecting either blue or green. The thresholds obtained were consistently low; with observer performing slightly better when judging the group or 24 items, than the comparing only two circles. These threshold were also better that those obtained in the previous experiment (Chong & Treisman, 2003). Note that this task required that the observer use some filtering between the two colors, an idea further explored in the next experiment.

In the third experiment a filtering task was added. A predetermined relevant color (blue of green) was cued or non-cued. The non-cued condition included trials in which the stimulus and the same conditions used in experiment 2 – 12 blue and 12 green circles. Other trials presented 12 circles of different size in a single color. Two test circles of the "relevant color for that task" followed these displays. The task was to determine which

of the two test circles matched the mean size of the color subset. For the cued condition the observers were presented with a more sophisticated filtering challenge. To cue a color, the fixation cross that preceded the first display included two lines of the color that was relevant for that trial. They found no difference in performance between the cue and non-cued conditions.

For all three experiments the results showed that observers were able to discriminate the mean size independently from the numerosity or density of the display (Figure 1.6). In the matching displays, performance was not impaired by the size of the set or how sparse or dense they were. Even though performance with the non-matching displays was not as good as in the matching displays, performance was 15 % over chance level. Also, they found that even when color was a variable, observers were able to determine which color subset had the larger mean size accurately (around 80%). Additionally, the fact that the relevant color was cued or non-cued, did not affect performance either.

The experiments described above, led Ariely (2001) and Chong and Treisman (2003 & 2005) to conclude that observers were using information from the whole display to make mean size judgments. They base this claim on the fact the performance of observers was minimally affected by the many factors they tested – set size, color (cued or non cued), density, numerosity, size distribution of the set and display exposure. Their main arguments are the good performance after brief exposures - as low as 50 ms - suggest a "parallel, or very rapid serial, extraction of the mean" (Chong and Treisman, 2005), and that a global strategy explains the fact that performance was not impaired by an increase of set size (Ariely, 2001). If an observer is truly using a sub-sampling strategy, he or she will be missing more relevant data when a set with a large number of items presented. Therefore, performance should be increasingly impaired as set size is increased.

In addition to the global and parallel claims mentioned above, Ariely (2001) proposed that the visual system reduces and combines the details of a group of items to create an ensemble representation and that the details of its individual members are then discarded. Alvarez & Olivia (2008) conducted three multiple-object tracking experiments to study how the visual system processes withdrawing of attention from targets. All three experiments required observers to identify and follow targets. These targets were a set of four flashing items, among a group of four non-flashing distractors (Figure 1.7). After an uncertain amount of time, between 6-10 seconds, the items stopped moving and a number appeared in the center of screen indicating if either one or more items were missing from the display. Observers then were asked to click where: (1) one item was missing or (2) if there were more items missing to click the centroid of that group of items. In two experiments the targets were the same color as the distractors (black), while in a third experiment the target was white and the distractors were black, making the identification process easier.

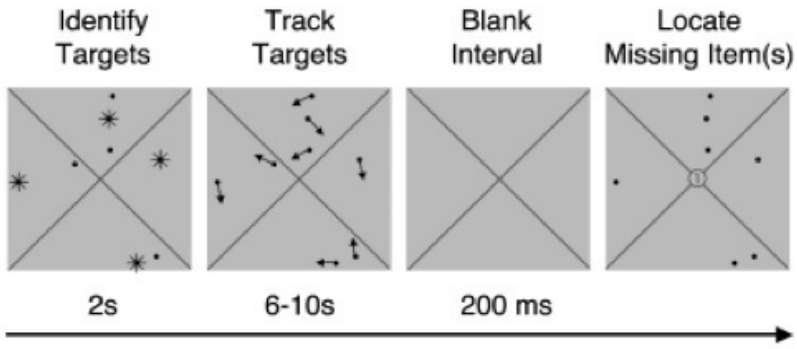


Figure 1. 7: Schematic Representation of Alvarez and Olivia (2008) experiment. Observers were asked to identify and track targets. After a 200ms they were asked to locate single or the centroid of missing items. Observers performed better when they were asked to locate the centroid of the missing group of distractors.

In all three experiments Alvarez and Olivia (2008) found that observers' performance was worse when just a single item was missing; which confirms the claim that our ability

to process features of single items is worse than when we generate summary statistical representations to an ensemble (Ariely 2001). But they also found that performance was better than chance at selecting the location of the centroid of the distractors when a group of items was missing. These results suggest that at some level, observers are processing both targets and distractors in the display, even though attention has been withdrawn from the distractors (Sun, Chubb, Wright, Sperling, 2018).

Myczek and Simons (2008) challenged the claim that our visual system relies on a parallel system that uses information from the whole display to calculate the mean size (Ariely 2001, Chong Treisman 2003 & 2005). The alternative proposed by Myczek and Simons (2008) did not eliminate the possibility of a parallel mechanism that uses a global strategy as plausible explanation, but they claimed that observers could be using other strategies, such as sub-sampling, to calculate mean size of a sample set with the same accuracy. The results of previous experiments showed that observers were able to calculate quickly and accurately, the mean size of a sample set and that set size and heterogeneity did not impede this ability. Myczek and Simons replicated the results of previous experiments, specifically the ones conducted by Ariely (2001) and Chong and Treisman (2003 & 2005), and put them through a series of computer simulations.

In these simulations an (otherwise) ideal observer processed only a random sub-sample of the stimuli: i.e., the computer would choose one, two, or three of the stimulus objects at random and calculate the mean size of a sample. Figure 1.8 shows the results of one of the simulations in which sub-sampling as few as 2 random items, out of a set if 8, could achieve similar results to the data of the real observers.

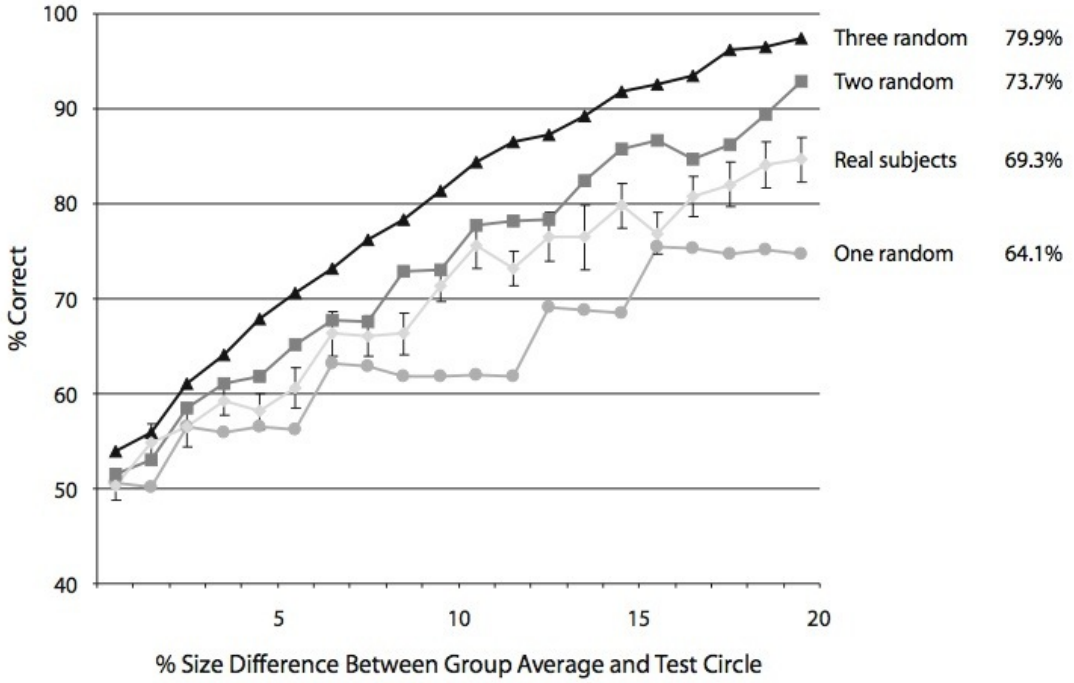


Figure 1. 8: Results from one of the Myczek and Simons (2008) simulations. Computer simulations were able to achieve similar performance to previous experiments by randomly selecting 1, 2 or 3 items.

Similar simulations were conducted for all the experiments published by Ariely (2001) and Chong and Treisman (2003 & 2005). In one of this simulations they found that for a set stimuli with small difference in sizes subsampling 1 item out of a 4, 8, 12 or 16 set size, was enough to achieve the 75% accuracy as defined by Ariely, (2001). For some of the Chong and Treisman (2003 & 2005) experiments, Myczek and Simons (2008) found that a subsampling four out of eight items produces simulated performance comparable to that of actual observers.

The subsampling simulations of Myczek and Simons (2008) are similar to the Efficiency calculations performed for the centroid paradigm presented by Sun, Chubb, Wright, & Sperling (2016). However, Sun et al. point out the almost certainty that an observer's response is influenced by other sources of error that should not be overlooked analysis. Because of this, Sun et al. view Efficiency as a "lower bound on the proportion of display

items included by the observer in his/her centroid computation". As we shall see in Chapter 3, this insight from the Efficiency analysis could help reconcile the results obtained in Myczek and Simons (2008) simulations with the claims that mean size estimates result from a process that is global.

Sun, et al. (2016) introduced the Efficiency parameter, and it has been used in the analysis of several experiments in which the centroid paradigm is tested. The centroid task consists of estimating the center of mass (or the average location of a group of items. In this particular experiment, observers were asked to judge the centroid of (1) all dots, (2) lighter-than-background dots and (3) darker-than-background dots. The Efficiency values among observers were between 75% and 90%. The interpretation of these results is that observers are processing most of the relevant data presented in the stimuli (at least 75%), in both conditions – filtering and no filtering. As will described in Chapter 2, this Efficiency analysis provides a mathematical framework that allows us to compare the results performance between two different tasks such as the centroid and the mean size task.

Summary statistical representations such as the centroid, mean size or mean luminance are subject to several sources of errors. In their centroid research, Sun et al. name two main categories of errors: (1) corruption of responses by random error, and (2) corruption of responses by nonrandom error. Missing some of the items in the stimuli is an example of a random error. Some of the errors an observer could also be making are: mislocalizing of the item(s), inaccurate calculation of the centroid (or mean size of the set) and movement errors when clicking on the screen to indicate the location of the centroid.

Both the sub-sampling simulations (Myczek and Simons, 2008) and the Efficiency calculations (Sun et al., 2016) use missed stimuli to account for the error produced by observers. The difference between the two analyses is that the Efficiency approach views missed stimuli as only one of a number of sources of error and attempts to identify and quantify as many of the other sources of error as possible. Any error due to a process other than missed stimuli that is not accounted for will produce an overestimate of the number of stimuli missed and thus an underestimate of Efficiency. The Myczek & Simons (2008) simulations assumes that all the error is due to the proportion of stimuli missed, so it is reasonable to assume that they must be overestimates. A high efficiency value (ranges between 0-1) means that the observer is making a sensible computation given that all the other errors are accounted for by an unbiased estimate of the standard deviation of his/her response.

Chong and colleagues (2008) argued that the sub-sampling strategy proposed by Myczek and Simons (2008) is hard for humans to implement. To test their claim, they conducted a series of experiments to confirm that a parallel mechanism that uses a global strategy is still the most plausible explanation for this type of calculation.

In their first experiment observers were asked to discriminate the larger mean size in two sets presented side to side. One of the aims of this experiment was to test if observers were switching between focused attention strategies (sub-sampling, calculating the mean size for the largest and smallest circles, or choosing the largest circle on each side and compare them). They presented observers three conditions: a heterogeneous display, a homogeneous display; and finally, they mixed the conditions of each display (a set with a uniform distribution, homogenous distribution or varied frequency distribution). Observer's accuracy was around 80% across all conditions. They argued that these results suggested that the use of attention is distributed across the whole display rather

than to just a few items. In a different experiment they found that observers performed poorly when presented with a smaller sample than when a larger sample was presented. They hypothesized that if observers used a sub-sampling strategy, a small set size should not have impaired their performance since they should be using the same number of sampled of items when calculating average on a larger set.

Robitaille and Harris (2011) also tested the claim that the extraction of summary statistics from a complex scene employs a sub-sampling strategy instead of a global mechanism. They conducted two experiments. The first one was similar to the experiment conducted by Ariely (2001) where observers were asked to estimate the average size of the circles. In the second experiment, they used orientation bars instead of circles and introduced a visual search task. A new measurement was added to both experiments - reaction time (RT). They hypothesized that: (1) if a visual search was used to form summary statistical representations, reaction time should increase with a larger sample set and that (2) if subsampling was used the reaction time would remain the same of the number of stimuli presented, because the estimation would used the same ‘‘sampled’’ stimuli.

In the first experiment observers were presented displays of 2, 4, 6, 8, 10 or 12 items controlled by a staircase procedure that was used to adjust the difficulty of the task. Similar to Ariely (2001), the task was to discriminate if a target circle was smaller or larger than the sample set. The target, which was always the same size for all trials, was presented first followed by a blank screen. Then the stimuli set was presented for 94ms (brief exposure) in some trials or until response (large exposure) in other trials (Figure 1.9).

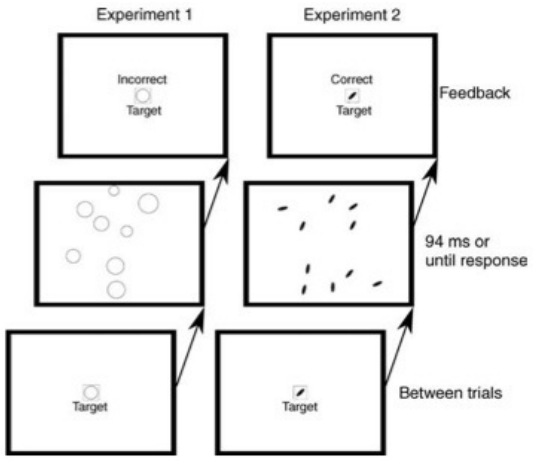


Figure 1.9: Schematic Representation of stimuli and trial sequences from the two experiments conducted by Robitaille and Harris (2011).

Surprisingly they found that both accuracy and reaction time benefited from a larger set size (Figure 1.10). Their hypotheses were that if RT slowed down, a serial search was suggested and if RT remained equal a sub-sampling strategy could be a viable alternative, because observers would be sub-sampling the same number of items in each display. However, the data showed that RT decreased with a larger set size. Robitaille and Harris (2011) reasoned that these results are better explained by the use of a global strategy, but they do not know why larger sets caused a reduction in RT. They offer two possible explanations to these behaviors. First, they reasoned that with a larger sample set, the calculation of a summary statistical representation could be benefiting form a redundancy gain effect. Alternatively, they proposed that with a larger set size the mean size of the set gets closer to the mean size of the population, making observers ability to discount items that are not essential to the mean size calculation.

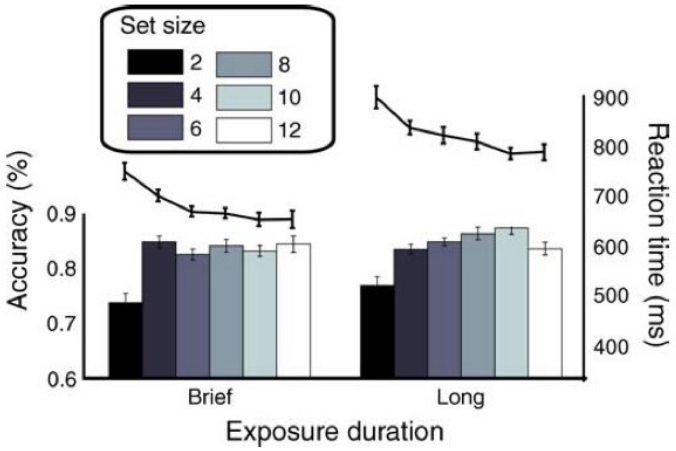


Figure 1. 10: Results from Experiment 1- Robitaille and Harris (2011). Bars represent accuracy, and a solid line represents reaction time. Reaction time benefited from set size.

The aim of their second experiment was to understand how reaction times (RT) and accuracy are affected by the task differences of a visual search task and a summary statistical task. The set consisted of orientation bars – sets of 2, 4, 6, 8, and 10. This time observers were presented with two different tasks: to determine if the target bar was more horizontal or more vertical than the stimulus set (summary statistics) or if the target was "present" or "absent" from the set (visual search). For the summary statistic task, set size had no effect in accuracy, but the RT decreased with a larger set - replicating the results of experiment 1. The results were different for the visual search task – RT slowed down with an increase in set size and accuracy was decreased by larger set size (Figure 1.10). They concluded that, in the summary statistic tasks, observers are using a rapid strategy that required attention to be distributed across the whole display to make their estimates (parallel mechanism), but that a slow serial search is used to process individual items.

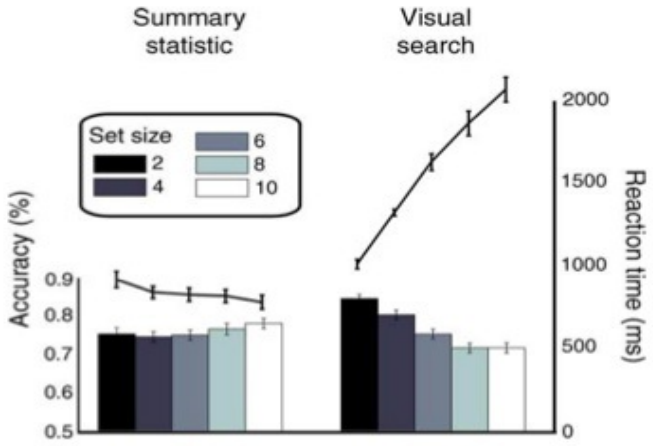


Figure 1. 11: Results from Experiment 2- Robitaille and Harris (2011). Bars represent accuracy, and a solid line represents reaction time. A visual search in larger sets impaired both accuracy and reaction time.

From the results of these two experiments Robitaille and Harris (2011) concluded that a parallel mechanism and global strategy is responsible, to explain observers' accuracy performance. However, these results do not eliminate the possibility of the mean size judgment being a low-Efficiency task. The fact that RT benefited from a larger set, does not negate the possibility that the observer are using a sub-sampling strategy, making errors such as miscalculating the mean size of the group, misregistration of the types of different items, or motor response execution (Sun et al., 2016).

Marchant, Simons and de Fockert (2011) also challenged the claim that a mean size calculation is result of a global strategy. In their experiments, they set out to study the capacity and limitations of a parallel mechanism when calculating the average size of a set. They saw these limitations as possible explanation the results of Myczek and Simons (2008) simulations. They noted that studies that claimed that the calculation of average size was a parallel process used limited heterogeneity (two sets of different circle sizes) in their samples (Chong & Treisman, 2003 & 2005). With the aim of adding more complexity to the displays, they added more sizes among sample sizes (4 or 8

sizes) to increase the heterogeneity of the sample and they called these *Irregular Sets*. In addition and to re-create previous results, Marchant and colleagues presented their observers displays with circles of only two sizes (which they called '*Regular Sets*'). The task was the same in both experiments. Observers viewed a set of circles and were asked to estimate the mean size by adjusting the size of a single circle presented in a second display, by pressing keys on their keyboard (Marchant, Simons & de Fockert 2011).

Marchant and colleagues hypothesized that if the judgment of average size is a parallel process that uses a global strategy, set size and its heterogeneity shouldn't matter. Alternatively, if increasing heterogeneity of item sizes impairs observer's performance, a parallel mechanism, which uses a global strategy, could not support the estimation of mean size. Their results showed that observers' performance was indeed poorer in the *Irregular Set* trials, meaning that increasing set size and heterogeneity suggested a limited capacity. They argue that the parallel model that uses a global strategy works only for displays of limited heterogeneity and that for increased heterogeneity a sub-sampling strategy was used. The heterogeneity of the *Irregular Sets* adds more complexity to the mean size calculation. This could also mean that the observers are more prone to make mistakes and possibly making this task a low Efficiency one.

A recent study by Attarha and colleagues (2014) showed that there is a fixed processing capacity of summary statistics representations for multiple ensembles, but unlimited capacity for single ensembles. They hypothesized that unlimited capacity models predict no difference between simultaneous and sequential conditions and that limited capacity will present an advantage for sequential conditions over simultaneous. They set up three experiments in which observers were presented a display with four clusters with circles of different sizes. One of the clusters was the target and the other three were

distractors. In one experiment the task was to report whether the mean size of the target cluster was relatively smaller or larger than the mean of the distractor clusters (Figure 1.12). The results (Figure 1.13) showed a better performance in the sequential condition suggesting that the calculation of mean size engaged a fixed-capacity process when averaging is required for multiple clusters (or ensembles). A control experiment was conducted to eliminate the possibility that other factors were affecting performance in the first experiment described above. The task was the same, but now the calculation of mean size was removed from the task by showing clusters of circles of the same size. The results of this experiment showed no advantage to of sequential over simultaneous presentations, confirming that the results from the first experiment were due to the average size calculation. In the third experiment, their goal was to analyze the capacity of a single ensemble. To do this they combined the four clusters presented in the first experiment into one single cluster. All aspects of the experiment were kept the same, but in this case observers were asked to report whether the average of the single set was smaller or larger than the probe circle that had been presented throughout the practice block. In this experiment observers' performance did not change under sequential and simultaneous conditions suggesting an unlimited capacity process (Figure 1.14). They argued that an ensemble representation of the environment bypasses the limitations of our perceptual and cognitive systems and serves to guide later visual processes (Attarha, Moore & Vecera 2014).

In the next chapters we will study the performances between mean size and centroid estimations. Each experiment will present observers with different tasks with the aim to understand how this means perceptions compared with centroid in the framework of an Efficiency analysis.

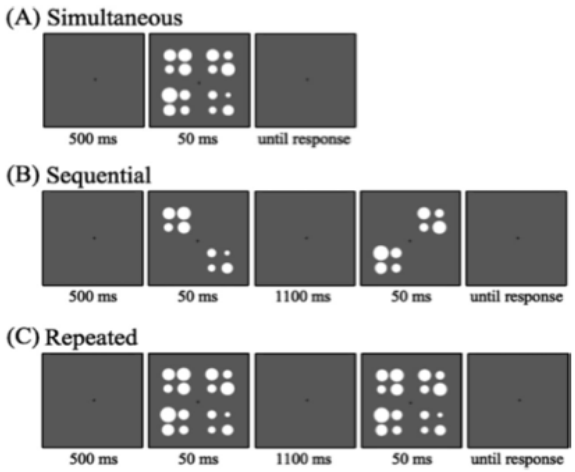


Figure 1. 12: Illustration of display of the experiment conducted by Attarha & Colleagues (2014). In their experiment observers had to report is the target cluster was smaller or larger than the distractors.

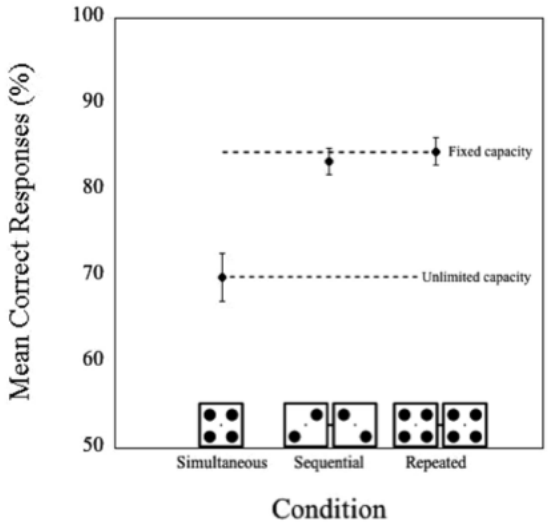


Figure 1. 13: Results of the one of the experiments conducted by Attarha & Colleagues (2014). When the task was to report whether the mean size of the target cluster was relatively smaller or larger than the mean of the distractor clusters, performance was better in a sequential condition; suggesting a fixed capacity.

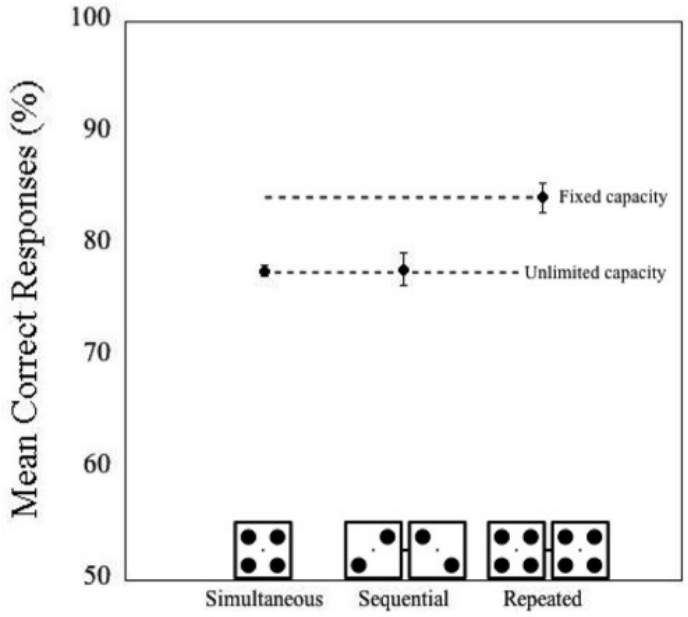


Figure 1. 14: Results of the one of the experiments conducted by Attarha & Colleagues (2014). When observers were asked to asked to report whether the average of the single set was smaller or larger than the probe circle that had been presented throughout the practice block, performance was the same for both simultaneous and sequential displays; suggesting an unlimited capacity.

CHAPTER 2:

Overall Research Approach

The three studies presented in this thesis are designed to understand mean-feature estimations in the centroid paradigm framework. We were able to compare performances across all the tasks by modifying the Efficiency analysis as described by Sun et.al (2016) and using Efficiency as a common measure by which to compare results across tasks. Before we explain how the Efficiency value was obtained and used, a brief description of the three studies follows.

Overview of the three studies:

Chapter 3: How can we use perceived size? Centroid versus Mean-size Judgments

In Chapter 3 we compared performances between two variants of the centroid task and the mean-size task. Observers were presented with displays containing 1, 3, or 9 items that varied in size with the sizes drawn from a triangular distribution. In different sessions that used similarly generated and presented stimuli they were asked to: (1) estimate the mean size by adjusting the size of a probe; (2) estimate the centroid ignoring the individual sizes of the stimulus items and (3) estimate the centroid giving more weight to larger items and less weight to smaller items.

Chapter 4: How do Mean-Luminance judgments compare to Centroid estimations?

For this study, observers were presented with displays containing 1, 3, or 9 squares of the same size, but different luminance (brightness). Each set was created with luminance levels that were randomly selected without replacement from a discrete triangular distribution. This discrete distribution was constrained to nine equally spaced

levels, linearized in luminance space - with the middle value solely used for the background. These values were specifically adjusted for the monitor used in the experiment and room illumination.

In different sessions that used similarly generated and presented stimuli observers were asked to: (1) estimate the mean luminance by adjusting the brightness of a probe; (2) estimate the centroid ignoring the luminance level of the items, and (3) estimate the centroid giving more weight to lighter items and less weight to darker items.

Chapter 5: The availability and use of size in mean size estimations

In this experiment, the display consisted of three clusters of stimuli, each containing four items varying in size, randomly positioned around points equidistant from fixation and evenly spaced around the circle. In the simplest form of this experiment, the size-choice condition, observers were asked to choose, from among the three clusters presented, the cluster with the largest and smallest mean sizes. The observer made a correct response by moving the mouse to click anywhere within a pie-shaped portion the subsumed one-third of the circle that contained the chosen cluster; a response in some ways analogous to the one used in the centroid estimations. Some sessions also required the observer to select the correct clusters by clicking as close as possible to the centroid of that cluster. In some of these sessions, the centroids were to be formed ignoring the size of the of the stimulus items in the cluster; in other sessions, the centroid were to be formed giving more weight to large items and less weight to small items. Although their accuracy was not emphasized in the size-choice condition, the location of the centroid responses was measured in all sessions. In the two conditions emphasizing centroid accuracy, this was treated as a secondary task, with the choice of the correct large and small clusters constituting the primary task.

B. Overall Analysis

In the experiments reported here, the stimulus types are: chapter 3, squares of different sizes; chapter 4, squares of the same sizes, but different luminance, and chapter 5, the mean size of clusters of four items that vary in size. The tasks in chapter 3, and chapter 4 are based on two target filters. In the equi-weighted centroid task, the target filter gives equal weight to all the squares presented: for chapter 3 of all 9 sizes and for chapter 4 of all 9 levels of luminance. In the size-weighted centroid task and for the mean size task of chapter 3, the target filter gives a weight to each stimulus proportional to the size (more weight to larger items). A similar weighting scheme also should apply in chapter 5. In the luminance-weighted centroid task and for the mean luminance task of chapter 4, the target filter gives a weight to each stimulus proportional to the luminance (more weight to lighter items).

Influence Functions and Efficiency

The data from all three tasks were analyzed using procedures similar to those described by Sun, Chubb, Wright & Sperling (2016), with minor modifications for the data from the mean-estimation task. The first step in these analyses generates estimates of the observer's attention filter, f_φ. An observer's attention filter is the vector of weights (one for each of the 8 square-widths used in our stimuli) used by the observer when performing a task with a particular target filter φ. The tasks in these experiments are based on two target filters. In the equi-weighted centroid task, the target filter φ gives equal weight to the squares of all 8 widths *w*: i.e., $\varphi(w_i) = 1/8$, for all *i* from 1 to 8. In the weighted centroid task and the mean-estimation task, the target filter φ gives weight to each square equal to its guiding feature: i.e., $\varphi(w_i) = w_i/\sum_i w_i$.

In the centroid task with target filter $\varphi(w)$, the correct response, *T,* on a given trial has *x*- and *y*-coordinates

$$T_x = \frac{\sum_i \varphi(w_i) x_i}{\sum_i \varphi(w_i)} \quad \text{and} \quad T_y = \frac{\sum_i \varphi(w_i) y_i}{\sum_i \varphi(w_i)} \qquad (1)$$

where the sum is over all squares *i* in the display, w_i is the width of the square *i*, and x_i and y_i are the *x*- and *y*-coordinates of its location. Typically, however, the response of the observer deviates from this target location.

We assume that the *x*- and *y*-coordinates of the observer's response on trial *t* are given by

$$R_{t,x} = \mu_{t,x} + Q_{t,x} \text{ and } R_{t,y} = \mu_{t,y} + Q_{t,y} \qquad (2)$$

where $Q_{t,x}$ and $Q_{t,y}$ are independent, normally distributed random variables with mean 0 and some standard deviation σ, and for some function $f_\varphi(w)$,

$$\mu_{t,x} = \frac{\sum_i f_\varphi(w_{t,i}) x_{t,i}}{\sum_i f_\varphi(w_{t,i})} \quad \text{and} \quad \mu_{t,y} = \frac{\sum_i f_\varphi(w_{t,i}) y_{t,i}}{\sum_i f_\varphi(w_{t,i})}. \qquad (3)$$

In Eq. (3) $w_{t,i}$, $x_{t,i}$, and $y_{t,i}$ are the width and *x*- and *y*-coordinates of the i^{th} square in the stimulus on trial *t*, and $f_\varphi(w)$ is the attention filter that the observer uses to perform the task.

Similarly, in the mean-estimation task with target function φ, we assume that the observer's response on trial *t* is

$$R_t = \mu_t + Q_t$$

where Q_t is a normally distributed random variable with mean 0 and some standard deviation σ, and

$$\mu_t = \frac{1}{N}\sum_i f_\varphi(w_{t,i}), \qquad (3)$$

where N is the number of squares in the display, either 3 or 9 depending on the condition, and f_φ is the attention filter achieved by the observer in this task.

A Bayesian procedure was used to derive parameter estimates. This method used a Markov chain Monte Carlo simulation to extract a sample of vectors from the joint posterior density characterizing the model parameters (Gelman et al., 2014). Each iteration of this process required evaluation of the likelihood function (or more properly, of the log of the likelihood function). The likelihood function for the centroid task model given in Eqs. (2) and (3) is

$$\Lambda(f_\varphi, \sigma) = \prod_t \frac{1}{2\pi\sigma^2}\exp\left[\frac{-(R_{t,x} - \mu_{t,x})^2 - (R_{t,y} - \mu_{t,y})^2}{2\sigma^2}\right], \qquad (4)$$

where the product is over all trials *t*. And similarly, the likelihood function for the mean-estimation task is

$$\Lambda(f_\varphi, \sigma) = \prod_t \frac{1}{\sqrt{2\pi}\sigma}\exp\left[\frac{-(R_t - \mu_t)^2}{2\sigma^2}\right]. \qquad (5)$$

For simplicity, we use uniform prior distributions on all parameters whose bounds are well-outside what might reasonably be expected.

In any MCMC process, one starts with some arbitrary guess at the parameter vector *V* (which will eventually be thrown away) and sets $S_1 = V$. (In the current application, the vector *V* contains guesses at the 8 values of the function f_φ as well as a guess at σ.) Then, one iterates the following steps some large number, N_{iter}, of times:

Pick a candidate parameter vector, C, in the neighborhood of the last sample, S_{n-1}. Then for

$$P = \frac{\Lambda(C)}{\Lambda(S_{n-1})},$$

if P > 1, set $S_n = C$; otherwise, set

$$S_n = \begin{cases} C & \text{with probability } P \\ S_{n-1} & \text{with probability } 1 - P \end{cases}$$

Provided that the procedure for choosing candidates satisfies certain conditions, as N_{Iter} goes to infinity, this process produces a sample from the posterior joint density characterizing the model parameter vectors (Hastings, 1970). For both the size and the centroid analyses, the initial values of $f_\varphi(w_i) = 1/8$ for all *i*, and the initial value of σ was 10. To ensure that the samples of this process used to generate estimates were stable, N_{iter} was 20,000 and the first 10,000 samples were discarded. To ensure that the samples used to generate estimates were independent, of the remaining 10,000 samples only every fortieth was retained.

A key measure that we have adapted from Sun et al. (2016) to characterize the results of this experiment is Efficiency. Efficiency is particularly useful because it is a measure that can be used to compare the response error observed in tasks as disparate as the centroid tasks and the mean-estimation task. Sun, et al. (2016) developed this measure

for centroid data but simply called it Efficiency. The Efficiency value was estimated as the proportion of the stimulus squares that would need to be processed by an ideal observer using the *observer's* estimated attention filter, $f\varphi$, rather than the target filter, φ. The value Efficiency ranges from 0 to 1. Because this is the estimate for an ideal observer, it is a lower bound on the proportion of squares that would have been processed by the actual observer.

Efficiency estimates were obtained using a univariate optimization function in Matlab, fminsearch. To evaluate a proposed value of Efficiency, 100 decimations of the stimulus cloud used on each trial was generated. For every decimation, each square in the cloud had a probability equal to the Efficiency value of being included in the centroid (or size) calculation. The observer's estimated attention filter was used to weight the included squares in that calculation. The difference between the estimated centroids (sizes) and the actual responses were combined across decimations and trials to guide the optimization process.

Figure 7 illustrates how the Efficiency analysis works. Panel A shows a 9-item stimulus in the equi-weighted centroid task. The bullseye indicates the target centroid. To get a sense of how the Efficiency calculation works, consider Panel B, which shows an example in which an ideal observer, processing this display with Efficiency of .89, has based the centroid estimate on a random subset of 8 from the 9 tokens, producing an estimate that is, in this case, slightly shifted from the true estimate. Because the decimation is done independently for each item in the display, the ideal observer operating with an Efficiency of 0.89 would not always process 8 tokens, this is simply the expected number of items processed since 8 = 9 × .89; however, since it is the probability that an item is decimated that is fixed, sometimes the simulated ideal observer would be expected to process 8 or even all 9 tokens, and sometimes fewer

than 8. As shown in Panel C, an ideal observer operating with Efficiency of 0.22, would be expected to produce the centroid estimate using only two tokens, although it could be more or less, and so would be expected to produce a larger error. These examples show a particular subset of the stimuli being used in the centroid calculation; however, as described above, the actual estimation was averaged over 100 subsets for the stimulus cloud used on each trial.

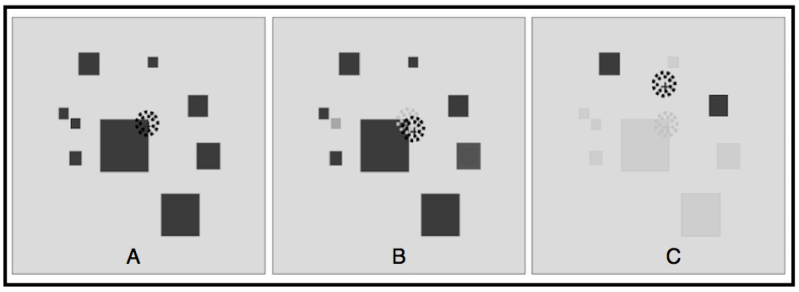


Figure 2.1: Panel A is representation of an equi-weighted centroid estimation with an Efficiency value of 1. Panel B shows a typical equi-weighted centroid estimation with an Efficiency value of .89. Note that the 0.89 Efficiency indicates that at a minimum 8 items out of the 9 are included in the estimation. The observer's response shown by the black bullseye is still close to the correct response, which is shown by the gray bullseye. Panel C, shows a typical response when the Efficiency is .22. In this example, the observer's centroid estimate (black bullseye) is far from the correct response (gray bullseye).

Sun et al. (2016) describe Efficiency as a lower bound on the number of squares processed by the observer. This is because the observer's response is likely to be corrupted by sources of error other than decimation of the stimulus. For example, the locations or sizes of squares may be registered incorrectly, or the memory of the centroid estimate may deteriorate before the response can be completed. The Efficiency statistic treats the error from all of these sources as if it resulted only from random decimation of squares from the display. However, with this caveat of interpretation, Efficiency provides a useful way to compare the response error produced in different tasks.

CHAPTER 3:

How can we use perceived size? Centroid versus Mean-size Judgments

When looking at a group of flying birds, we easily detect the general direction the birds are flying, the center of mass of the group, their approximate number, and the average size of the birds. Most of the time these perceptions occur pre-attentively - in just a fraction of a second. Visual researchers refer to this ability as the formation of a statistical summary representation. This ability allows us to get the gist of a group of items by effectively calculating the mean size of the objects in it, their centroid, numerosity, range and the variance of features like size, motion, location and orientation (Ariely 2001; Chong & Treisman 2003, 2005; Alvarez & Olivia, 2008; Alvarez, 2011; Marchant, Simons & de Fockert 2011; Robitaille, N., & Harris, I. M., 2011).

The estimation of the mean size of a group of items has provoked the interest of many visual researchers studying summary statistical representations (Ariely 2001, Chong & Treisman 2003, 2005). A recurring finding from this research is that observers can estimate the average size of the items in a group relatively well, certainly better than they can identify individual stimuli displayed (Ariely 2001). Building on these results and previous research on mean size, one of the goals of this paper is to compare Efficiency of size estimation in two tasks: the mean-size task and the centroid task (Drew, Chubb & Sperling, 2010; Sun et al. 2016). Of particular interest will be a variant of the centroid task in which observers weight stimulus items in proportion to their size, because in this weighting task observers must make use of both location and size information.

As described in the Literature Review (chapter 1), most of the previous research on mean-size judgments has concluded that when making them, the visual system relies on

a global, parallel perception mechanism. This suggests that observers incorporate most, if not all, of the displayed items into the mean-size estimate (Ariely, 2001; Chong & Treisman 2003, 2005). Initially, Ariely (2001) found that observers were able to judge the mean size of a group of disks better than they were able to determine if a single disk was a member of that set, independently of set size. In his experiment, set size was varied (4, 8, 12, or 16 items) and used 4 distinct sizes within each set.

In a follow-up paper, Chong & Treisman (2003, 2005) varied the heterogeneity of the disk sizes, the presentation mode (sequential vs. simultaneous) and their numerosity and density. Across all these manipulations Chong & Treisman (2003, 2005) achieved results of 75% accuracy with a difference in size between 6-8%. The fact that these discriminations were performed following relatively brief exposures (from 50 ms to 1000ms) and that increasing the size of the sample set did not affect performance, led them to the conclusion that the estimation of mean size was based on including most, if not all, of the items presented on the screen.

Recent research has supported the claim that the size of an individual item cannot be measured with complete accuracy in an ensemble representation such as the perception of the mean size of a group (Im & Halberda, 2012; Allik et.al, 2013). However this research has also challenged the claim that observers used most, or all of the items presented in a display in judging the mean size of a group of items (Myczek and Simons, 2008; Im & Halberda 2012 Allik et.al, 2013), suggesting subsampling as a possible strategy. In a sub-sampling strategy, an otherwise ideal observer uses only a few items from the full set displayed to make the mean-size discrimination rather than attempting to include all of the items presented in the display. Myczek and Simons (2008) simulated Ariely (2001) and Chong & Treisman (2003, 2005) experiments and suggested that observers could be using sub-sampling as one of their strategies when

making the mean-size discriminations. This interpretation assumes that all the errors in the observer's responses are due to the observer failing to include all of the display items in their estimation. However, this assumption can be misleading as a model of human performance since other sources of error are almost certainly involved.

In this paper, we present an experiment that compares performance for two summary statistical representations: centroid and mean size. We use an Efficiency analysis as a common framework to compare performance across these two tasks. The procedure used to estimate Efficiency and the differences between it and the measure originally proposed for the centroid task by Sun, et al. (2016) will be described later, but for this discussion, it can be understood as a lower bound on the proportion of information contained in the display that is incorporated into an observer's judgment. But most importantly, our interpretation of the Efficiency analysis emphasizes the idea that failure to register stimulus items is only one source of error in these tasks.

Observers viewed sets of 3 or 9 squares and were then asked, in different sessions to perform one of three tasks: (1) to estimate the centroid of the squares ignoring variations in item size, (2) to estimate the centroid of the squares weighting items in proportion to their size or (3) to estimate the mean-size of the squares. For the rest of this paper, these three tasks will be referred as: (1) the equi-weighted centroid task, (2) the size-weighted centroid task and (3) the mean-size task. In these tasks, the size of the squares was defined as the length of a side, not the area (Solomon, Morgan & Chubb, 2011).

When deciding on stimuli to use in this experiment, we were concerned that, when presented with filled squares, observers could use mean luminance when estimating mean size. However, there was also the concern that outline squares might fail to be

detected, especially those presented more peripherally. Because of these competing concerns, observers were presented with two types of stimuli in separate conditions: outlined squares and filled white squares. As we will show, performance was similar for both classes of stimuli supporting the conclusion that observers were using size and not luminance in their estimations.

In both centroid tasks used in this experiment (equi-weighted and size-weighted), we asked observers to estimate the center of mass (centroid) of a set of items. In previous research observers could judge the centroid of a group of dots when asked to attend to all dots and/or while selecting stimuli with a specific feature, such as attending darker dots versus lighter dots (Drew, Chubb & Sperling, 2010; Sun et al. 2016). They found that, with little training, observers were able to accurately determine the required centroids with efficiencies between 75% and 90%. These high efficiencies were obtained both when observers were asked to attend to all the dots or just to some targets. These results suggest that centroid estimation is a highly efficient task.

In contrast, results from Myczek and Simons's (2008) simulations, suggest that mean size may be a less efficient task. One of the simulations presented by Myczek and Simons (2008) showed that an ideal observer, attending to only 2 items out of a group of 8 when estimating mean size, could still perform as well as the observers in experiments reported by Ariely (2001) and Chong & Treisman (2003, 2005). In other words, the mean-size task yields Efficiency as low as 25% (or 2÷8), much lower than the Efficiency estimations in the centroid task (75%-90%) found by Sun et al. (2016).

In order to compare directly the efficiencies of these two tasks, we designed an experiment that minimizes the differences between the tasks other than the summary statistical representation to be estimated. For instance, the versions of the mean-size and centroid tasks that are typically studied have a procedural difference that might complicate comparing their results. In most studies of mean-size, observers submit binary responses, pressing one of two keys to indicate whether a probe disk is larger or smaller than the mean size of the stimuli (Ariely, 2001). In other variations, the observer is asked to judge which side of the screen has the larger (or smaller) mean size by pressing a key on the keyboard (Chong & Treisman 2003, 2005). This presented a major methodological difference between the typical mean-size task and the centroid task since in the centroid task observers provide their responses in a continuous fashion by moving the mouse and clicking where they estimate the center of the mass is located.

To make the observer's response in the mean-size task similar to that of the centroid task, we presented observers with a probe square - the initial size of which was randomly selected by the computer - and asked them to indicate their response by moving the mouse to adjust the size of the probe square until it matched their remembered percept. Observers clicked on the mouse when they felt they had reached the size that represented their estimation of the group mean size.

Another difference between the mean-size and the centroid tasks is that they require the observer to process different aspects of the stimuli: sizes or locations. To explore this difference, we presented observers with a variation of the centroid task that we called the size-weighted centroid task. This task requires a judgment based on two aspects of the stimuli: observers estimate the centroid giving proportionally more weight to the larger squares. Good performance - i.e., high Efficiency - in this task requires two things: that the observers register both the locations and the sizes of the stimuli accurately and

that they combine both types of information accurately when estimating the centroid. Figure 3.1 shows how the location of the centroid for a stimulus differs across these two tasks.

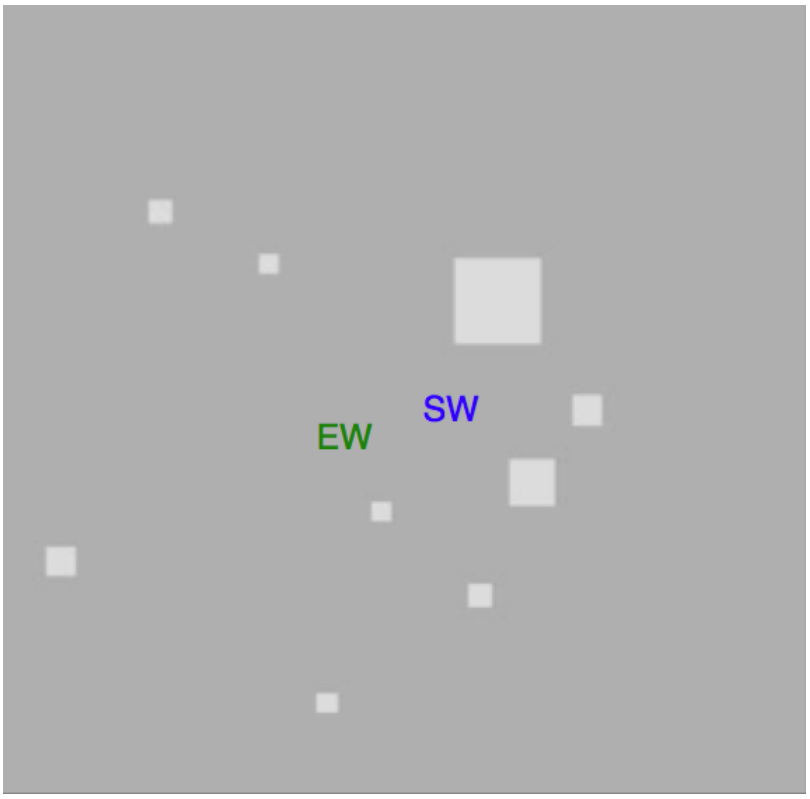


Figure 3. 1: Example showing how the centroid response changes when asked to give equal weight to all the items (equi-weighted –EW- task) and when asked to give more weight to larger items (size weighted - SW- task) version of the centroid task for a set size of 9 squares.

Method

Observers

Eight observers, including the first author, participated in the experiment. Four were novice observers, and the other four were experienced with the centroid task. All were students at the University of California, Irvine. Four were females and four were males between the ages of 17 and 40. All observers reported having normal or corrected-to-normal vision. The present study was conducted in accordance with the regulations of the Institutional Review Board of the University of California, Irvine.

Apparatus and Stimuli

The observer sat in an adjustable height chair in a dark room and viewed the stimuli presented on an iMac (Mac OS X) with a 54 cm screen controlled by an ATI Radeon HD 4670 graphics card from a distance of about 84 cm. The stimuli were generated using the Psychophysics Toolbox Version 3.0.8 (Brainard, 1997; Pelli, 1997; Kleiner et al, 2007) for MATLAB (Version 7.1).

Screen shots illustrating the two types of stimuli used in this experiment are shown in Figure3.2. The size of the stimulus area was 500 x 500 pixels and the viewing angle was approximately 15 degrees. The outlined squares (Figure 3.2a) were constructed using white, 2-pixel-wide lines (115.80 cd/m^2); the interior of each square matched the gray background luminance (46.00 cd/m^2). The other stimuli (Figure 3.2b) were filled white squares (115.80 cd/m^2) on a gray (46.00 cd/m^2) background. The display was constructed using squares of eight fixed sizes (0.23°, 0.27°, 0.34°, 0.45°, 0.52°, 0.67°, 0.81°, 0.99°). Each set was created with sizes that were randomly selected without replacement from a discrete triangular distribution. The probability assigned to each of

the eight possible sizes to appear was: 5.63%, 10.25%, 14.75%, 19.38%, 19.38%, 14.75%, 10.25%, or 5.63%. This discrete distribution was constrained to have only 8 levels, because we wanted to be able to estimate the influence of each level on the size and centroid judgments. Given this constraint, this seemed a reasonable approximation of the Gaussian distribution used to determine item location. The dispersion of the location of the squares was determined by a Gaussian distribution with a standard deviation of 110 pixels (1.98°) centered in the middle of the screen. The sampling from this distribution was constrained so that the edges of two squares were never closer than 6 pixels (0.11°) from each other. In addition, because, when going from 3 to 9 stimuli, the standard deviation of the distribution of the centroids would normally be reduced by, after the stimulus clouds were generated, their centroids were then translated to a location separately chosen from a Gaussian distribution centered in the middle of the screen with a standard deviation of 63.5 pixels (1.9°).

Figure 3.3 shows the timeline of events for both the centroid task and the mean-size task (using filled stimuli). The mask stimulus constructed for each trial consisted of a 10x10 jittered grid that filled the display area with a random sample of squares with sizes drawn from the triangular distribution used to generate the stimuli.

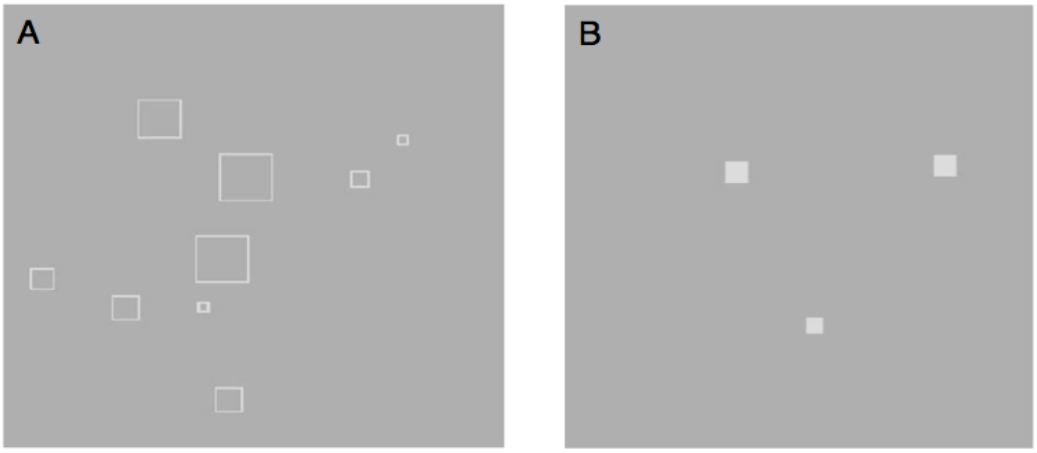


Figure 3. 2: Two screen shots of the displays used in the experiment. Figure 3.2A shows an example of a set size of 9 squares using the "outlined stimuli," and Figure 3.2 B shows an example of a set of 3 squares using the "filled stimuli."

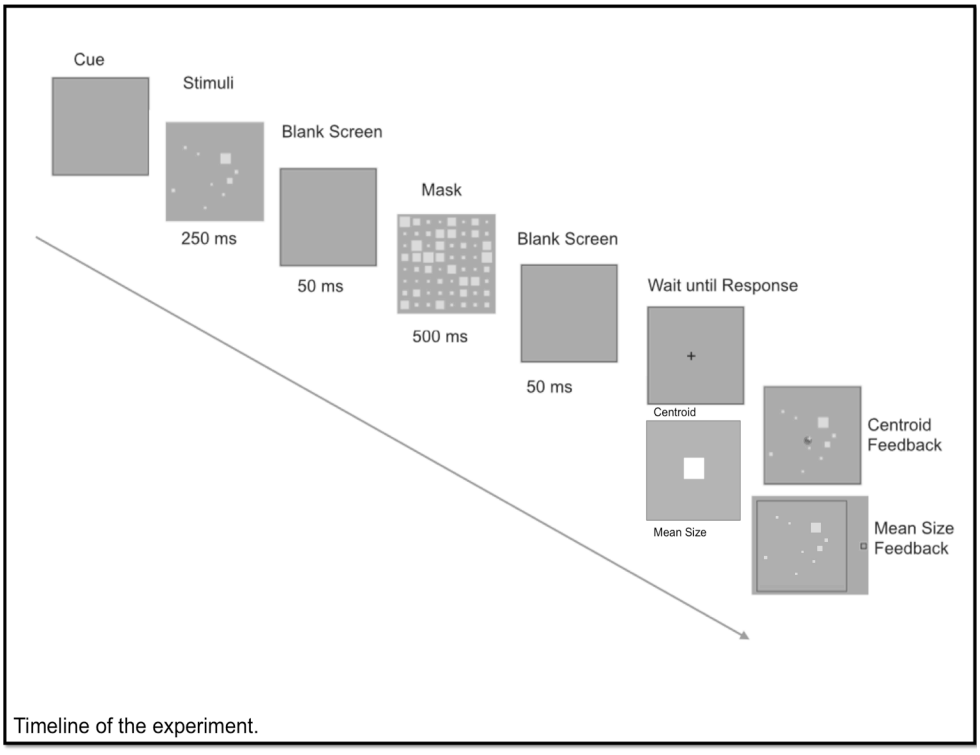


Figure 3. 3: The timeline of a trial (from a 9-item condition using an example based on filled squares). The two final frames show two possible (1) response screens and (2) feedback screens, one for the mean-size task and one for both the equi-weighted and size-weighted centroid tasks.

Procedure

The present study consisted of three tasks: (1) the equi-weighted centroid task, in which observers strove to estimate the centroid of the stimulus array giving equal weight to all squares regardless of size, (2) the size-weighted centroid task, in which observers strove to estimate the centroid of the stimulus array weighting items in proportion to their size, with size being defined as the length of the square, and (3) the mean-size task, in which observers were asked to determine the mean size of the squares in the stimulus, ignoring their locations, by adjusting the size of a single square. An initial screen displayed the instructions for each session: whether to assess the size-weighted centroid, the equi-weighted centroid, or the mean size of the target stimuli. The initial screen also displayed examples of each of the stimulus sizes using the type of stimuli to be judged - outlined or filled. At the start of each trial, which began 500 ms after the initial block screen or the feedback from the previous trial ended, the observer was cued with a screen containing just the cue square, a white line that outlined the stimulus region (500 ms) that was followed by the stimulus (250 ms); then came a blank screen (50 ms), the mask (500 ms), another blank screen (50 ms), and then the display that the observer used to respond; finally, the feedback display was presented. The feedback and response displays used for the different tasks are described below. In all tasks, the observer terminated the feedback screen by pressing any key.

Feedback and the Response Screens in Centroid Task

On the response screen for the centroid task, a white-cross appeared at the center of the display area. It functioned as a cursor, tracking the movements of the mouse. The appearance of this cursor prompted the observer to move the mouse and click on the

location of the estimated centroid. After the location was selected, a feedback screen followed. The feedback screen redisplayed the stimulus used in that trial, but it also had a white cross that showed the location the observer chose as the centroid, and a black bullseye centered at the correct centroid location (Figure 3.4) depending on the weighting function.

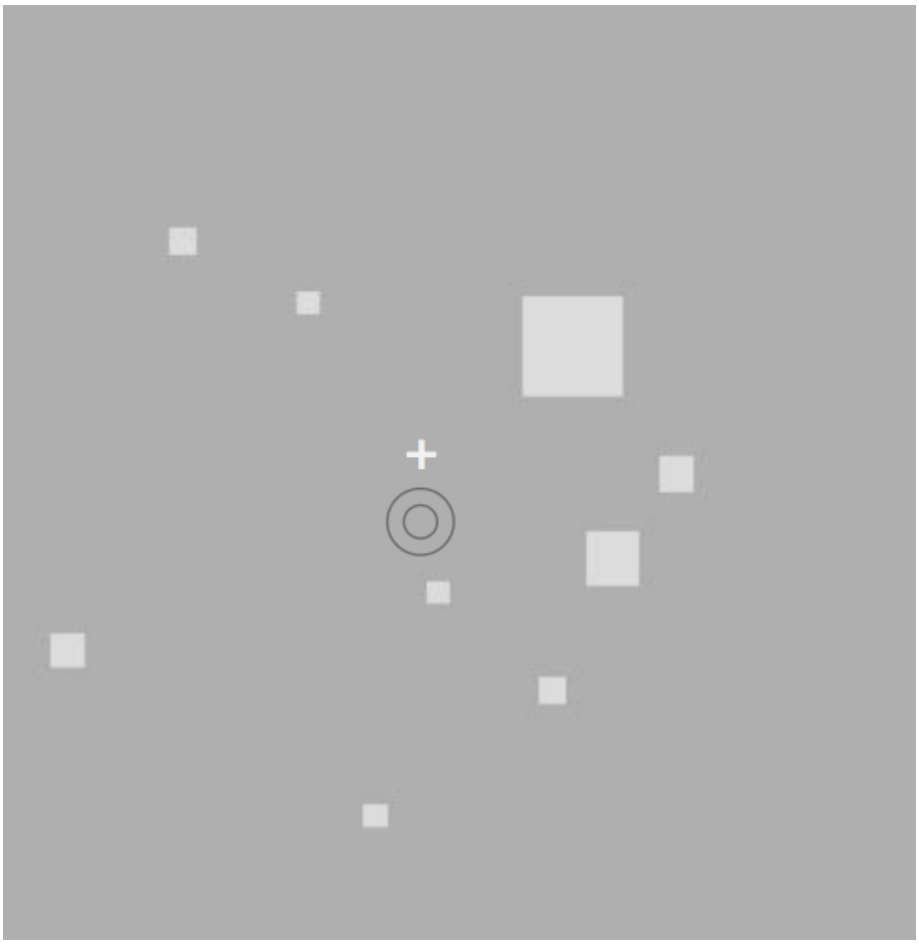


Figure 3. 4: Schematic representation of the feedback screen for the equi-weighted centroid task, for a set size of 9 (filled) squares. The dark gray bullseye represents the correct centroid and the cross shows the observer's response.

Feedback and the Response Screens in Mean-Size Task

The initial response screen in the mean-size task consisted of a probe square with a size randomly selected in the range of the stimulus sizes. By moving the mouse horizontally, the observer changed the size of the probe square until its size matched the size of the estimated mean of the stimuli. Moving the mouse to the right made the probe square larger; moving the mouse to the left made it smaller; vertical movement was ignored. The probe square was either outlined or filled, to match the squares used in the current condition. The observer terminated the response process with a mouse click. An example of the feedback screen is shown in Figure 3.5. The screen showed the stimulus used in that trial and the probe square with the response. A white outlined square showed the observer's response. A black outlined square showed the correct response. Between these two outlined squares, the region in red indicated the response error (Figure 3.5b).

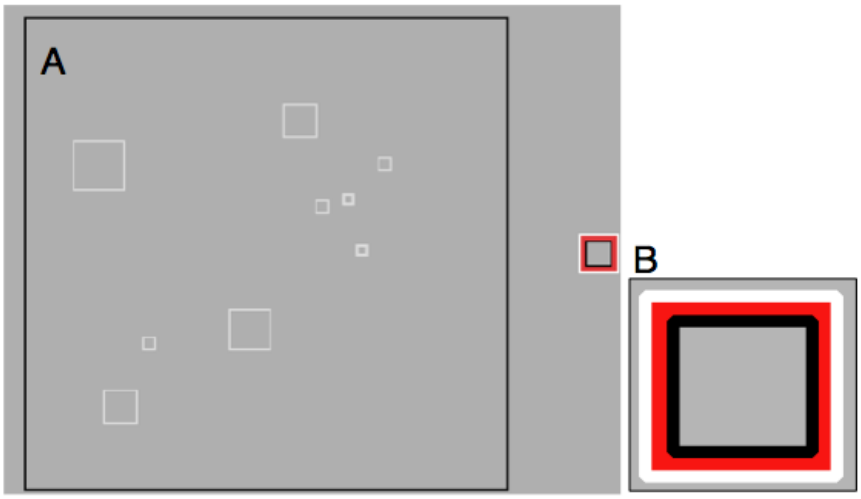


Figure 3. 5: Feedback screen for the mean-size task with a set of 9 outlined squares. The black outline is the correct mean size; the red outline around the single square on the side represents the observer's error; the white outline is the observer's response. Figure 3.4B: A zoomed in schematic representation of the mean-size task feedback square.

Design

The conditions in this experiment were constructed from the factorial combination of three factors: the task (equi-weighted Centroid, size-weighted Centroid, or Mean size), the type of stimuli (Outlined or Filled) and set size (1, 3, or 9 stimuli). A session consisted of two blocks - one per stimulus type - of the same task. Across sessions, the task was varied using a 3x3 Latin square, with the conditions for each observer taken from a different row. The conditions specified by the Latin square were mirrored twice resulting in the sequence A-B-C-C-B-A-A-B-C, so that each observer ran 9 sessions (3 sessions per task). The order of filled versus outlined stimuli within a session were switched across the mirrored repetitions. We monitored each observer's mean squared error to ensure that large improvements associated with learning did not occur after the first three sessions, which were dropped from the analyses reported below. A block consisted of 105 trials of which five were singleton trials, on which only a single square was presented. Singleton trials were included to estimate the error due to processes that were not associated with estimating the mean size or the centroid (e.g., sensory mislocalization of items and response motor error). On the remaining trials, groups of 3 or 9 squares were presented 50 times each. The order of the numerosity condition within a block was randomly determined.

Results

All observers ran 9 sessions, 3 sessions per task. We measured the root mean square error (RMSE) of the responses in each session and compared them. The RMSE was stable for the last 6 sessions, and for most observers the error was stable and constant for all 9 sessions. For all observers, only the data from the last 6 sessions are reported here.

We expected experts to be better than naïve observers at least in the centroid task. The actual difference was small; the observed efficiencies were .88 and .85, respectively (Δ = 0.03, SD = 0.09, t (6) = 0.523, p = 0.62, Bayes Factor[1] BF=0.764). The main effect of stimulus type was negligible. Since there are also no reliable interactions involving stimulus type or level of expertise, the reported results are collapsed across these factors. Also, to simplify the summary, we will consider the data from the singleton trials separately, so that, for most of the summaries, only results for trials with three and nine items are reported. Finally, we will focus on two pre-planned contrasts for the task factor: one comparing the results in the equi-weighted and size-weighted centroid tasks and one comparing the size-weighted centroid tasks results with the mean-size task results.

Efficiency

Observers achieved higher, and almost identical, Efficiencies[2] in the two centroid tasks, and lower efficiencies in the mean-size task (Figure 3.6). The preplanned contrast comparing both centroid tasks suggests that efficiencies for the size-weighted centroid task are essentially identical to those from the equi-weighted centroid task (Δ = 0.01, SD = 0.02, t (7) = 1.460, p = 0.188, BF = .74). The preplanned contrast comparing the

[1] Bayes Factor computed using the calculator at http://pcl.missouri.edu/bayesfactor (Rouder, Speckman, Sun, Morey, & Iverson, 2009).
[2] Procedure explained in Chapter 2

Efficiency for the size-weighted centroid task with that in the mean-size task very strongly suggests that observers were able to use size more effectively when estimating the centroid of a group of squares than when estimating the mean size of the same group (Δ = 0.35, SD = 0.15, t (7) = 6.485, $p < 0.001$, BF = 45.9).

Figure 3.6 also shows that observers achieved higher efficiencies when presented with three items than when presented with nine items. A t-test provided evidence for a reduction of Efficiency with increased numerosity (Figure 3.6) for all three tasks (Δ = -0.14, SD = 0.08, t (7) = -4.810, p = 0.002, BF = 24.06).

No interactions were found between: (1) stimulus type and numerosity, (2) stimulus type and task, (3) numerosity and task and (4) stimulus type, task and numerosity. The biggest t-value associated with the interactions of any of these factors was equal to 1.42 with a p-value of 0.198 and BF = 0.71.

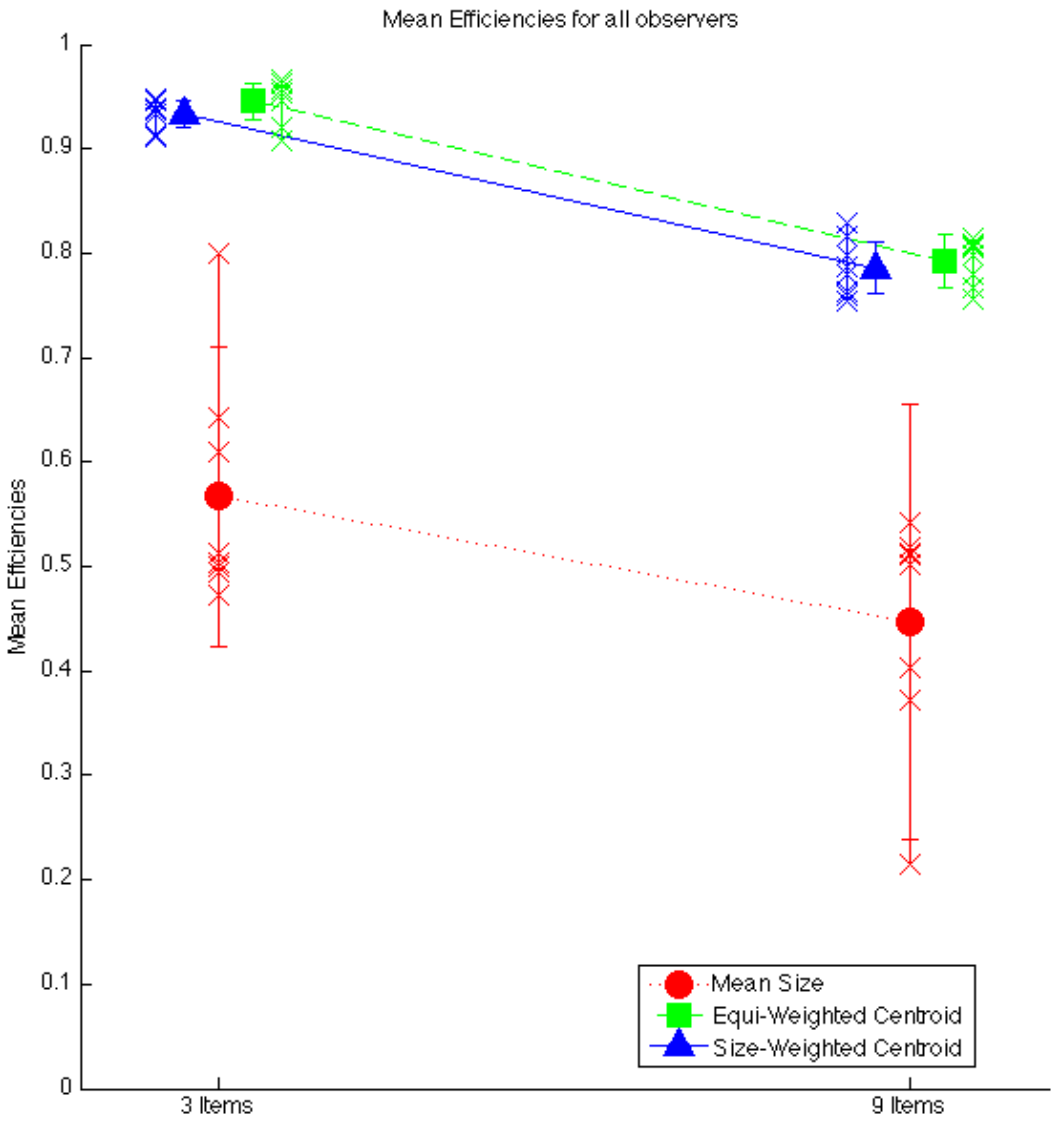


Figure 3. 6: Mean Efficiencies for all observers as a function of set size for equi- and size-weighted centroid judgments and for mean-size judgments. The filled plotting symbols represent the mean across observers in each condition; the x's are the Efficiencies for each observer. The error bars display the 95% confidence intervals for the averages.

Influence Functions

Figure 3.7 shows the influence functions for both centroid tasks, averaged across observers and collapsed across level of expertise and stimulus type. In the size-weighted centroid task, the slope of the ideal influence function is one. The average data follow this ideal closely. With 9 stimuli, observers tended to overweight the larger stimuli and underweight the smaller stimuli relative to this ideal, but with 3 stimuli, observers produced the opposite pattern. In the case of the equi-weighted centroid task, observers were asked to give equal weight to all squares independently of their sizes, so the ideal influence should have a slope of zero. Although the resulting influence functions are flatter than those for the size-weighted centroid task, the observers substantially underweighted the smaller squares and overweighted the larger ones. We have omitted a figure showing the mean influence functions for the mean-size task because the influence functions estimated for each observer are not well constrained by the data; this makes sense with Efficiency values of 0.5. The wide confidence intervals obtained for this task makes these data hard to interpret.

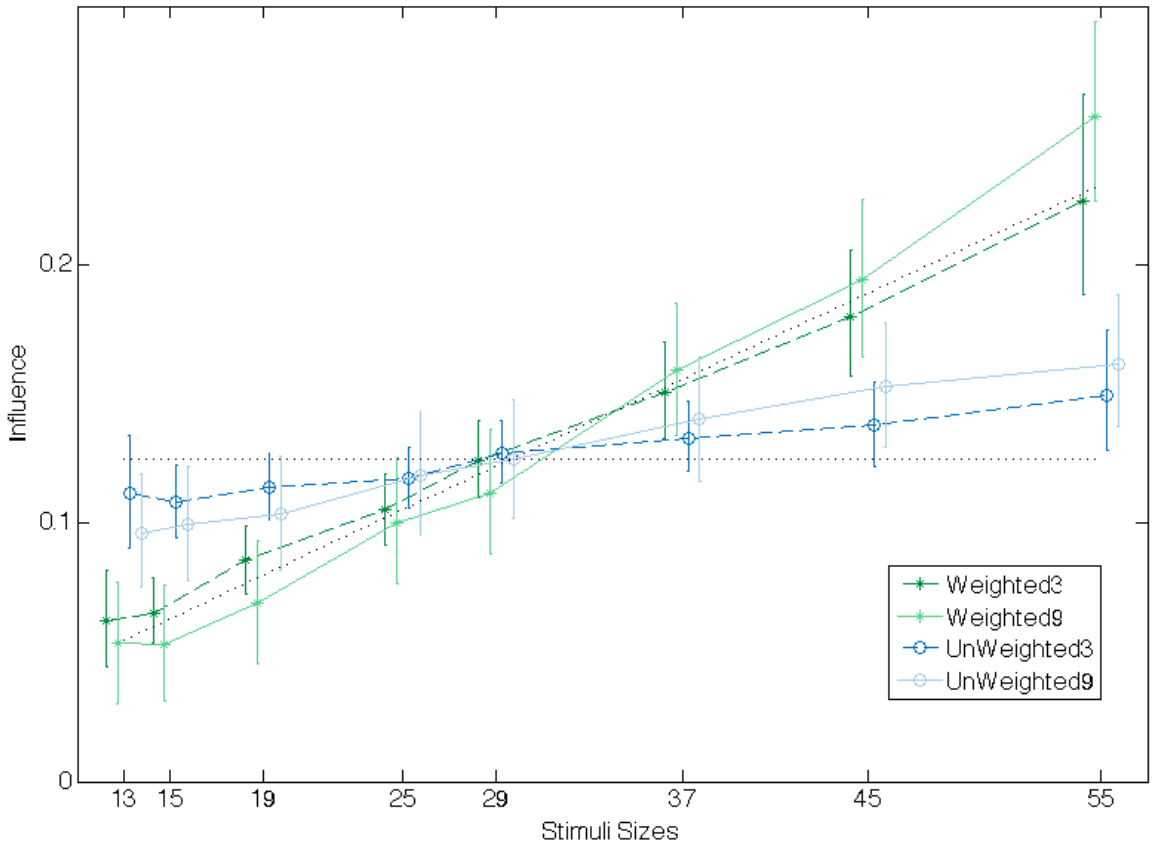
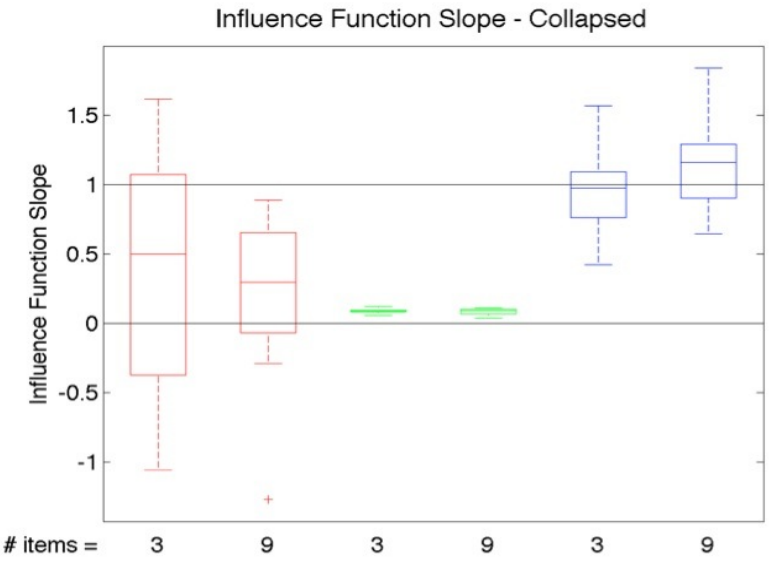


Figure 3. 7: Influence as a function of stimulus size for the two centroid tasks and two types of stimuli averaged across observers and stimulus types. The intervals are 95%confidence intervals based on the variation across observers. The black dashed lines represent the ideal influence: slope = 0 for the equi-weighted task and slope =1 for the size-weighted task.

To precisely characterize the difference between the influence functions across task and numerosity and to create a summary that could be applied to the mean-size task, linear regression was used to estimate the slope of the influence function in each condition. Figure 3.8 provides a summary of the slope estimates from this analysis. The means shown by the bars in the figure confirm the general impressions provided by Figure 3.7 for the equi-weighted and size-weighted centroid tasks. As shown by the summary at the bottom of Figure 3.8, the slope estimates for the equi-weighted centroid task are close to zero and do not differ with numerosity. Because there is little variability across observers, the mean of these slopes collapsed across numerosity is clearly different from the ideal of zero, but the confidence intervals show how close to zero it is (slope= 0.085 [0.068 .10], t(7) = 11.98, p = .0000, BF = 141.169). There is substantially more variability across observers looking at the slope estimates for the size-weighted centroid task. Despite this variability, there is a reliable numerosity effect. However, as shown Figure 3.8, for neither of numerosities 3 or 9 is the slope distinguishable from the expected slope of one, and this result still holds if the estimates for numerosities 3 and 9 are averaged (slope = 1.055 [.76 1.33], t(7) = 0.43, p = .68, BF = 0.485).

As noted previously, there was substantial variability in the influence function estimates for the mean-size task both within and across observers. Despite this variability, Figure 3.8 includes a summary of the slope estimates for the mean-size task. Not surprisingly, these slopes vary more across observers than in the size-weighted centroid task, with slope estimates ranging from less than -1 to over 1.5. With this caveat, we note several things based on these slope estimates. First, there is no evidence for an effect of numerosity on these slopes. Second, averaged across numerosity, the slope in the mean-size task differs reliably from one (slope = 0.270 [-.35 .89], t(7) = -2.790, p =

0.030, BF = 3.980), the expected value for the slope in this task. Further, averaging across numerosity in both cases, the slope in the mean-size task differs reliably from that in the size-weighted centroid task (Δ = -0.78, SD = 0.71, t (7) = -3.113, p = 0.017, BF = 4.32).



	Ideal Slope =1		Ideal Slope =0		Ideal Slope =1	
Average	0.370	0.170	0.090	0.080	0.960	1.150
CI at 95%	[-0.41 1.14]	[-0.41 0.75]	[0.07 0.10]	[0.06 0.10]	[0.67 1.23]	[0.85 1.44]
SD	0.930	0.691	0.018	0.026	0.336	0.359
	H0=1		H0=1		H0=1	
t(7)	-1.920	-3.392	-138.382	-99.206	-0.375	1.163
p-value	0.096	0.012	0.027*10^-11	.279 *10^-11	0.718	0.283
BF	1.630	5.830	277	276	0.477	0.779
	H0=0		H0=0		H0=0	
t(7)	1.150	0.699	13.329	8.966	8.051	9.030
p-value	0.302	0.507	0.031 *10^-4	0.437 *10^-4	0.875 *10^-4	0.417 *10^-4
BF	0.770	0.554	159	91.9	74.9	93.1

Effect of Numerosity			
	H0 = 0		
Ave. Difference	0.200	0.010	0.190
CI at 95%	[-.39 .78]	[-0.01 0.02]	[-0.30 -0.08]
SD	0.700	0.021	0.134
t(7)	0.790	0.671	-4.049
p-value	0.455	0.523	0.005
BF	0.586	0.545	11.998

Figure 3. 8: The influence function slope for all three tasks, summarized separately for three and nine items. The summary at the bottom of the figure displays results from two sets of t-tests. The upper sets of results are for data collapsed across stimulus type, but separated by numerosity; those in the lower set examine the effect of numerosity. Note that, for the mean-size and size-weighted centroid tasks, the ideal slopes should be equal to 1, so the null hypothesis for these tests are highlighted in gray. In the equi-weighted centroid task, the ideal slopes should be equal to 0, so it is the tests for this null hypothesis that are highlighted in gray.

Discussion

The central result here is that efficiencies were high in both centroid tasks, but substantially lower in the mean-size task. Based on previous literature, these results were expected for the mean-size task and the equi-weighted centroid task. The simulations reported by Myczek & Simons (2008) suggested that the estimate of the mean size of a group of items is obtained with low Efficiency. Also, previous research from our lab (Drew, Chubb & Sperling, 2010; Sun et al. 2016) found that equi-weighted centroids could be estimated with high Efficiency. The surprising result is that locating the centroid, while weighting items in proportion to their size, can also be done with high Efficiency. This is surprising because one might expect Efficiency in size-weighted centroid task to be no better than lesser of that obtained in the mean-size task or the equi-weighted centroid task. Our results show that observers achieved almost identical Efficiency in the two centroid tasks and that Efficiency was much lower in the mean-size task. However, this implies, counter-intuitively, that a summary statistical representation based on a combination of two distinct kinds of information, location and size, appears to be substantially easier for observers than a summary statistical representation based on only one of these components, size.

These results suggest that the estimation of mean size is different and perhaps more difficult for observers than a centroid task that also involves size information. First, the high efficiencies achieved in the size-weighted centroid task show that both location and size information are accurately registered for most, if not all of the squares. Second, the influence function analysis suggests that, although observers can achieve a weighting rule that accurately gauges the sizes of display squares in the size-weighted centroid task, they are unable to achieve such a weighting rule in the mean-size task.

Figure 3.9 shows the root mean squared error (RMSE) for the mean-size task in degrees of visual angle, broken out by number of items on the abscissa. Each of the colored lines connecting the x plotting symbols, reflect the data from one observer. The black circles show the mean error for each numerosity averaged across observers. The black solid line is the best linear fit. The data for the 3- and 9-item conditions are a "raw" version of the data used as the basis for the Efficiency analysis; this is a raw summary because it does not depend on the influence function analysis. The singleton data were not included in the Efficiency analysis. Given that the stimulus items ranged in size from 0.22° to 0.99°, the standard deviation of 0.17° [0.15° 0.19°] of the singletons suggests that observers were able to perceive and then recall a single size fairly accurately with the adjustment procedure used here.

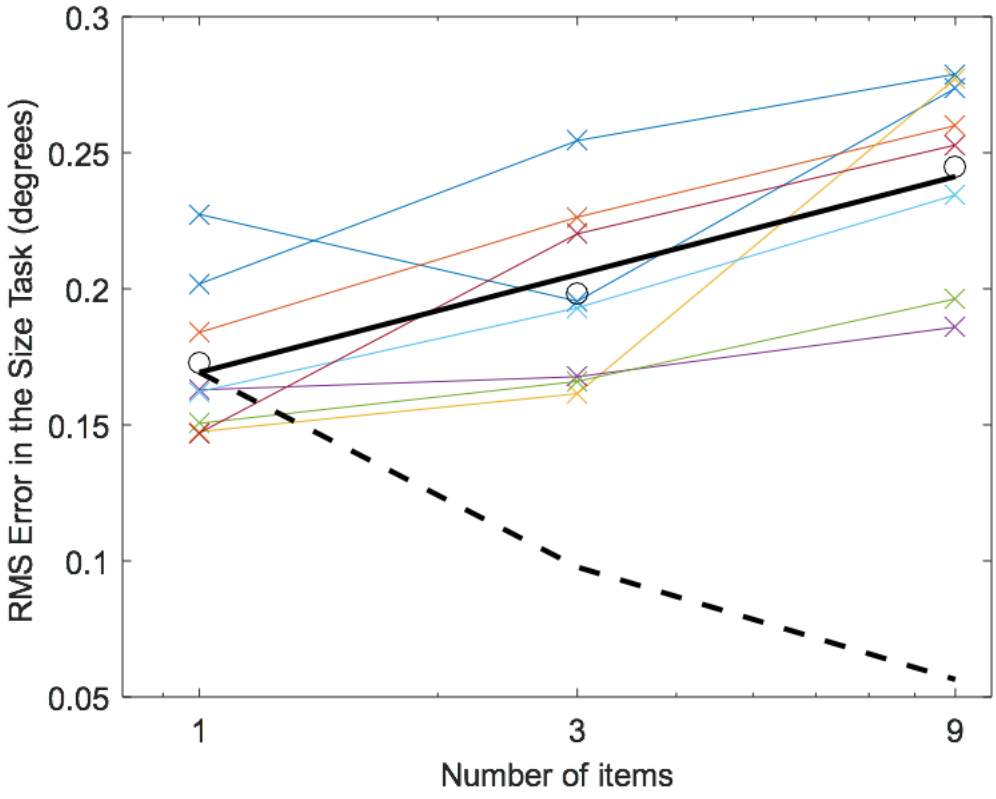


Figure 3. 9: (Previous page) RMSE in the mean-size task as a function of number of items (on a log scale). The × plotting symbols connected by colored lines are the data from each of the eight observers. The black circles are the means over observers. The solid black line is the best fitting linear function. The dashed line represents the predictions of a model that assumes that all of the error in this task is due to misperception of item size so that mean-size error is, where *n* is the number of items and is the error in the singleton task (*n* = 1).

Because the spacing on the abscissa of Figure 3.9 is logarithmic, it appears that the mean-size error increases linearly with the logarithm of the numbers of items (slope = 0.033 [0.020 0.046], t(7) = 5.927, p = .001, BF = 63.97). What makes this observation striking is that it suggests that something other than the misperception of the sizes of the items must be contributing to the error observed in the 3- and 9-item conditions. We reach this conclusion because the mean-size error due to misperception of the items sizes would be expected to decrease as one over the square root of the number of observations (items). Under the extreme assumption that all the singleton error is due to size misperception, the dashed black line shows the predicted RMSE. Another possibility is that the error in the mean-size task, rather than being due to size misperception, arises from "late" sources: i.e., error depending on processes that come after the mean-size estimate has been created. Two examples of late sources of error are memory errors that result from having to keep a perceived mean size in memory while making the response and reproduction errors that arise because of problems correctly reproducing the correctly-remembered mean size. One characteristic of late error is that it should not depend on the number of items included in the mean. Thus, an alternative but equally extreme model based on the assumption that all size error arises from late sources predicts that the dashed line in Figure 3.9 should be flat. However, neither size misperception errors, late errors, nor some combination of the two predict the increase in the RMSE with an increasing number of items that was observed. This argument suggests that there is some other component of error in the mean-size task that produces the observed increase in RMSE with *n*.

One clue that at least some of the error in the mean-size task results from the misperception of size is that, not surprisingly, the variability of the error increased with the size of the item being reproduced. To quantify this, a Markov chain Monte Carlo simulation was used to fit a three-parameter model to the singleton data from the mean-size task. The three parameters were bias, (the amount that an observer systematically over- or under-estimated the size of the item) and the two parameters of a linear model for the standard deviation of the size response error (an additive term and a slope). This analysis showed there might have been a slight bias, in this case, a tendency for observers to underestimate the true item size (-0.042° [-0.092° 0.008°], $t(7) = -1.979$, $p = .088$, $BF = 0.80$), but the evidence for this is weak. There was evidence for an additive component of the standard deviation of the error (0.062° [0.031° 0.094°], $t(7) = 4.691$, $p = .002$, $BF = 21.53$) and even stronger evidence that the standard deviation of the size error also increased as the size of the item being estimated increased (0.106 [0.072 0.140], $t(7) = 7.300$, $p = .00016$, $BF = 182.9$). One way to get a sense of the relative importance of the additive and multiplicative contributions to the standard deviation is to compare the contribution of the multiplicative component for an average size item (0.485°) with that of the additive component: $0.485° \times 0.106 / 0.062° = 1.19$. This suggests that the additive and multiplicative components contribute about equally to the standard deviation of the size estimation error for the singletons, with the multiplicative component possibly being slightly stronger.

In the singleton task, size responses were strongly correlated with item size ($r = .86$). That correlation along with the comparison above showing that the multiplicative component made a substantial contribution to the overall error in size judgments for singletons, gives us confidence that observers were able to perceive the size differences of the stimuli used and report sizes using the response method employed in this

experiment. Another window on the accuracy with which the item sizes could be perceived in the stimulus displays is provided by a comparison of the results in the size-weighted and equi-weighted centroid tasks. This comparison was done by extending the Efficiency analysis (Chapter 2) to allow for the perturbation of the item sizes. For each observer, the analysis of the data from the size-weighted centroid task used the estimated Efficiency in the equi-weighted centroid task as a fixed value determining what proportion of the items in a stimulus cloud would be retained after the simulated decimation process. In addition, in this expanded analysis, the size of each stimulus item was randomly perturbed prior to computing the simulated centroid judgment. The size perturbations were drawn from a Gaussian distribution with mean zero and a standard deviation that depended on item size. The Matlab optimization function fmincon() was used to estimate the slope and intercept of a linear function relating the standard deviation of item perturbation to item size so that the centroid response error produced in the simulation matched that produced by the observer in the size-weighted centroid task.

Starting with the approximation that the centroid response error in the equi-weighted task does not reflect the size variation of the stimulus items, if one also accepts the assumption that additional centroid response error observed in the size-weighted task is only due to incorporating size information into the centroid judgments (and not, for example, the recruitment of some completely different centroid judgment process), then the size error estimated by this expanded analysis provides an upper bound on the variability in the misperception of size for these stimuli. This is an upper bound because all of the additional centroid response error in the size-weighted task is ascribed to size misperception; however, it seems plausible that the process of forming a size-weighted centroid introduces some of the additional error.

For the size-weighted centroid of 3 items, this elaboration of our Efficiency analysis estimated the additive component of the size misperception error to be 0.053° ([0.033° 0.074°], $t(7) = 6.085$, $p = .0005$, BF = 72.78); for 9 items this was 0.044° ([0.033° 0.055°], $t(7) = 9.366$, $p = .0000$, BF = 692.5). Because there is only weak evidence for a difference between these estimates (Δ=0.010° [-0.006° 0.026°], $t(7) = 1.490$, $p = .180$, BF = 1.09), we will consider their average, 0.049° ([0.034° 0.063°], $t(7) = 7.837$, $p = .0001$, BF = 265.4). The slope relating the size misperception error to item size for the 3-item task was 0.050 ([0.002 0.099], $t(7) = 2.445$, $p = .044$, BF = 2.073); for 9 items the slope was 0.038 ([-0.010 0.085], $t(7) = 1.859$, $p = .105$, BF = 0.65). Because there is only weak evidence for a difference between these estimates (Δ=0.013 [-0.069 0.094], $t(7) = 0.369$, $p = .723$, BF = 0.356), we will consider their average, 0.044 ([0.018 0.070], $t(7) = 4.045$, $p = .005$, BF = 11.473). What is striking here is that the estimate of the additive component of the size misperception error computed in this way is similar to that estimated above for the singleton trials in the mean-size task (0.049° versus 0.062°; Δ = 0.014° [-0.025° 0.053°], $t(7) = 0.850$, $p = .423$, BF = 0.451), but the slope of the multiplicative component is substantially smaller (0.049 versus 0.106; Δ = 0.057 [0.022 0.093], $t(7) = 3.804$, $p = .007$, BF = 8.975). We interpret this as evidence that the information about this size of the stimulus items in the size-weighted centroid task is more accurate than that incorporated into the mean-size judgments.

If, as these analyses of the of the size-weighted centroid task suggest, the sizes (and locations) of up to 9 items can be perceived accurately and incorporated effectively into a centroid judgment, why are the mean-size judgments so inefficient? The analysis above suggests that, at least in part, this reflects degradation in the quality of the size information available to the mean-size calculation. However, the data summarized in Figure 3.9 suggests that the problem goes further than this. One possibility is that the

calculation of the mean size itself is a substantial source of error. The fact that size information can be used effectively in the size-weighted centroid task suggests that the brain has processes that can accurately perceive and calculate with this information, but, apparently, the mean-size responses do not tap these processes. Ours is not the only demonstration that comparing the mean size of a set of items with the size of a single item could be problematic; Chong & Treisman (2003) found reduced thresholds when observers were asked to compare the mean size of two stimulus arrays, even when they were presented sequentially. One speculation about the source of this difference between the centroid and mean-size tasks is that the centroid judgments may be produced by a mechanism in the dorsal visual pathway whose purpose is to guide movements (Goodale & Milner, 1992). In this interpretation, mean-size judgments result from a ventral mechanism that either has poor access to size information or that combines that information inefficiently.

An issue that presents a potential complication for the interpretation of these results is that, depending on the task, observers may be registering size in different ways. Because it is a reproduction task, the mean-size task requires observers to register and then produce their judgment using absolute sizes. By contrast, for the size-weighted centroid task, observers could be using relative sizes; it is possible to perform this task perfectly well with size information that only preserves the proportional sizes of the stimuli. We should point out, however, both that there is nothing in our results that suggests that observers were, in fact, using relative size estimates in the size-weighted centroid task and that we unaware of any literature that shows that using such relative sizes would be easier than actually using absolute sizes. Also, as discussed above in the analysis of the singleton data from the mean-size task, there is evidence that

suggests that, at least in this case, observers were able to perceive and report absolute size with good accuracy.

A secondary result is that there was no effect on performance due to the two types of squares used in this experiment. Both influence functions and efficiencies were very similar for both outlined and filled squares. These findings suggest that the observer is actually using the sizes of the squares to make their judgments and is not being influenced by the luminance of the screen (e.g. using mean luminance to make their estimation).

With the aim of exploring if there are systematic, individual differences across tasks, we conducted a correlation analysis of the efficiencies for all four variants of the three tasks, i.e., the variants due to stimulus type and by set size. These correlations, averaged over stimulus type and set size, are summarized in Table 1. There was a strong, positive correlation of the Efficiency estimates both within (i.e., across the variants) and across the two centroid tasks, suggesting that the differences in Efficiency across observers in these tasks reflects a common mechanism. In contrast, there was little or no correlation among the variants of the mean-size tasks or between them and the centroid tasks. Given that there are large Efficiency differences across observers and the variants of the mean-size task (ranging from .2 to almost .9), these correlations close to zero suggest two separate conclusions. First, the Efficiency variations across observers in the mean-size task derive from a different source than the Efficiency variations in the size-weighted centroid task. Even more troubling for those interested in the mean-size task is the lack of correlation across its variants, which suggests that any variation across observers in their ability to make mean-size judgments is swamped by other, unrelated sources of error. Of course, since these correlations are being computed based on only

eight observers, these estimates are not precise; however, the differences are large enough to suggest that there is an effect here worth considering.

Efficiency Correlations (Average)

Mean Size to Mean Size	0.06
Mean Size to Equi-Weighted Centroid	-0.19
Mean Size to Size-Weighted Centroid	-0.03
Equi-Weighted Centroid to Equi-Weighted Centroid	0.85
Size-Weighted Centroid to Equi-Weighted Centroid	0.80
Size-Weighted Centroid to Size-Weighted Centroid	0.79

Table 3.1: A summary of the correlations of the efficiencies between the three tasks.

Interim Conclusions

The primary result reported here is that size information can be used substantially more efficiently in a size-weighted centroid judgment than in a mean-size judgment. Other research has shown that human gaze tends to prefer the centroid of items and that saccades land closer to the center of mass, suggesting why performance in both centroid tasks was better than in the mean-size task (Fehd & Seiffert, 2008 & Melcher and Kowler, 1999). A paper by Christie, Hinchey and Klein (2013) suggests that inhibition of return (IOR) is primarily driven by the center of gravity of the attended stimuli. Specifically, they found that when observers were presented with multiple cues, both manual and saccade-detection responses were considerably affected by the center of gravity and there was a stronger IOR for the center of gravity than for the actual

stimuli. They suggest that the calculation of the centroid of a set of stimuli is an important, exogenous cue used to guide attention and the planning of future movements. Our findings elaborate these claims and suggest that reported judgments of mean size may not accurately reflect the information about the sizes of individual items available to later processes from a briefly perceived group of items.

Chapter 4

How do Mean-Luminance judgments compare to Centroid estimations?

The general finding from the experiment described in the previous chapter, was that observers were able to use or ignore size when asked to estimate the centroid, but that they were less efficient when asked to estimate the mean size of the same group of items. To assess the generality these results, in this study we used a new feature - luminance - in the context of the same two tasks. Here, observers were asked to estimate the mean luminance of a group of three or nine items, and we compared their performance with centroid estimations in which their response was a result of one of two instructions: to locate the centroid by (1) ignoring the luminance of the items or (2) giving more weight to lighter items.

The luminance, or brightness, of an item has been previously used to study perception of summary statistics, such as the centroid (Drew et al., 2010 and Sun et al., 2016). Results from Drew et al. (2010) show that observers can successfully estimate the centroid of a group of items when instructed to attend (1) to items that are lighter than the background, (2) to items that are darker than the background and (3) equally all items. In their study, they used a display that consisted of 8 levels of luminance which were separated by increments or decrements of ¼ in a Weber Contrast scale from -1 to 1, with '0' being defined as the background; and depending on the condition, observers were instructed to click on the screen at location that they thought was the centroid of the targets – items that were lighter or darker than background; or the location that they thought was the centroid of the whole set. Figure 4.1 shows an example of the displays

used in Drew et al. (2010) experiment and figure 4.2 shows the three possible responses, which varied depending on the instructions for a block of trials.

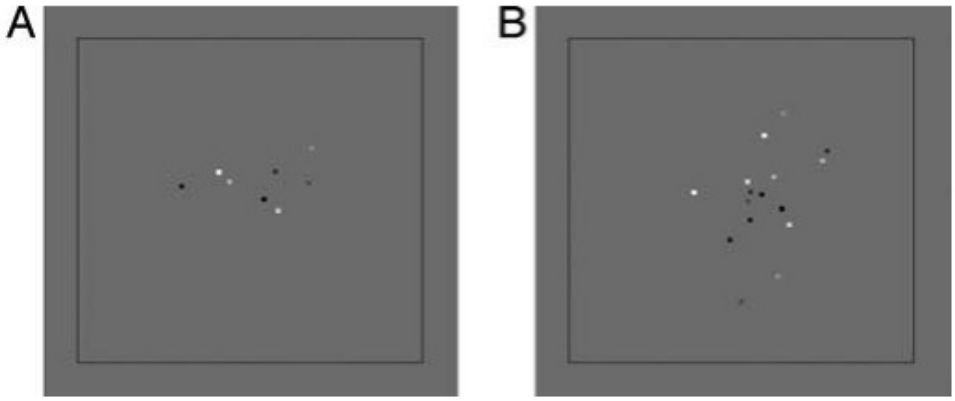


Figure 4. 1: Displays presented in the centroid task (Drew, Chubb & Sperling 2010). Observers were asked to determine the centroid of a group set. Figure A has a sample set of 8-items, Figure B has a sample set of 16-items.

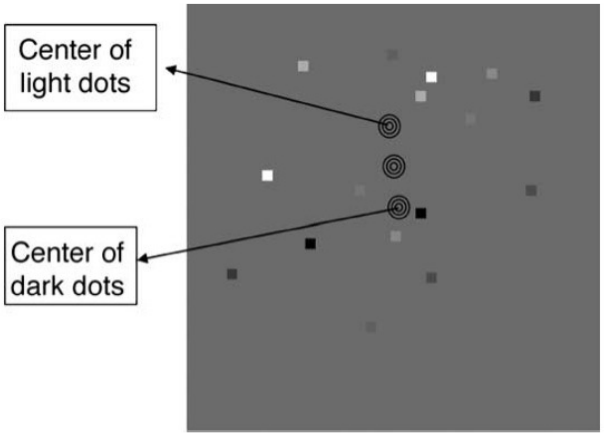


Figure 4. 2: Feedback screen in the presented in the centroid task (Drew, Chubb & Sperling 2010). This shows observers where the correct centroid estimations were located in the display. The upper bulls-eye shows the correct centroid when the instruction given to observers was to attend to lighter items; the middle bulls-eye shows the correct centroid when the instruction given to observers was to attend all items; the lower bulls-eye shows the correct centroid when the instruction given to observers was to attend to darker items.

For each display, the levels of luminance were drawn, with replacement, from eight levels of luminance, four that were lighter and four that were darker than the background. In our experiment we used two different instructions for the centroid task: (1) to weight items according to their brightness when judging the location of the centroid and (2) to give equal weight to all items, in other words to estimate the centroid of the whole set. This type of display and methodology was used in a different study conducted by Sun et al. (2016). We will use a similar approach to the one Sun et al. (2016) used to create the stimulus displays in our experiment, and it will be described later in the Methods section.

In their study Sun et al. (2016) - using stimuli that consisted of 8 possible types of Gabor patterns with contrasts that varied by 1/8 from each other - asked observers to locate the centroid by giving (1) equal weight to all the Gabor patterns (Figure 4.3, top row), (2) by giving more weight to higher contrast Gabor patterns (Figure 4.3, middle panel) or (3) by giving more weight to lower contrast Gabor patterns (inverse-weighting condition, Figure 4.3, bottom panel). Conditions (1) and (2) of this experiment, using grating contrast, are analogous to the conditions in our experiment in which the weight was based on luminance.

Sun et al. (2016) found that observers, depending of the instruction, can generally produce differently weighted centroids. The influence functions shown the top panel in have slopes close to 1, meaning that the observers were able weight items according to their contrast. They also found that can equally weight all the items, by ignoring their contrast, with the exception of the lowest contrast items.

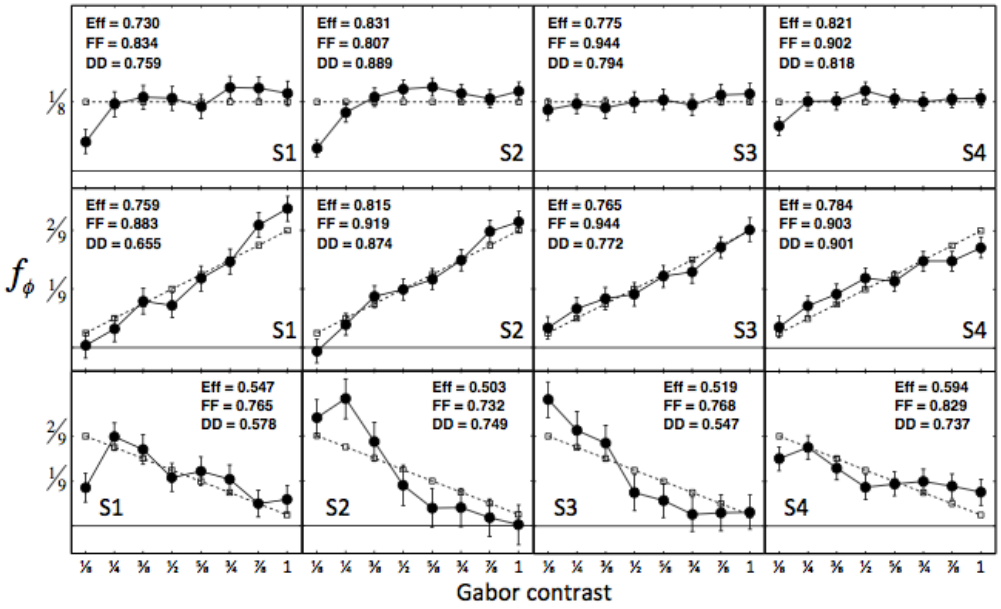


Figure 4. 3: Results from the centroid task experiment (Sun et al., 2016). (Top Panel): attending all items, (B) grading condition, (C) for inverse-grading condition.

Method

Observers

Eight observers, including the first author, participated in this experiment. Seven observers were experienced with the centroid task. All were students at the University of California, Irvine. All observers reported having normal or corrected-to-normal vision. The present study was conducted in accordance with the regulations of the Institutional Review Board of the University of California, Irvine.

Apparatus and Stimuli

The observer sat in an adjustable height chair in a dark room and viewed the stimuli presented on an iMac (Mac OS X) with a 54 cm screen controlled by an ATI Radeon HD

4670 graphics card from a distance of about 84 cm. The stimuli were generated using the Psychophysics Toolbox Version 3.0.8 (Brainard, 1997; Pelli, 1997; Kleiner et al., 2007) for MATLAB (Version 7.1).

The size of the stimulus area was 600 x 600 pixels and the viewing angle was approximately 15 degrees. The display was constructed using squares of a fixed size 2-pixel wide (0.36^{o}) and eight fixed luminance levels ($0cd/m^2$, $2.15cd/m^2$, 8.47cd/, cd/m^2, $18.67cd/m^2$, 31.49 cd/m^2, 48.88 cd/m^2, 69.26 cd/m^2, 92.19 cd/m^2, $118.1cd/m^2$) (Figure 4.4).

Observers were presented with displays containing of 1, 3, or 9 squares of the same size, but different luminance (brightness). We adjusted the luminance values to reflect the specific characteristics of the monitor by following a psychophysical adjustment described by Chubb et al. (2007) and Silva & Chubb (2014). The general idea of this procedure was to adjust the mid gray for the monitor, by a comparing with a low luminance level (black) and a high luminance level (white). After the texture grid – with the three luminances appeared on the screen - an observer, who was sitting at a predetermined distant from the monitor, adjusted the level of the mid gray until it disappeared. This procedure ensured that the luminance levels were specifically adjusted for the monitor used in the experiment and room illumination. After the adjustment was completed a vector of 9 values from linearized luminance space: (0 .0000, 0.1250, 0.2500, 0.3750, 0.5000, 0.6250, 0.7500, 0.8750, 1.0000) was generated with the middle value (.5000) used for the background luminance. The luminance levels displayed in each trial were randomly selected, with replacement, from the remaining 8-luminance levels. RGB values associated with these levels were used to generate the display and the perceived response. To get the expected mean luminance for a

particular stimulus cloud, the luminance was averaged (i.e., the "0-1" values of the stimuli).

Luminance levels

Darker than background				Background	Lighter than Background			
0 cd/m^2	2.15 cd/m2	8.47 cd/m2	18.67cd/m2	31.49 cd/m2	48.88 cd/m2	69.26 cd/m2	92.19 cd/m2	118.1 cd/m2

Figure 4. 4: Schematic representation of the 8-luminance levels used in the experiment, including the background luminance level.

Procedures

An initial screen displayed the instructions for each session, which indicated to the observers whether to assess the luminance-weighted centroid, the equi-weighted centroid, or the mean luminance of the target stimuli. The initial screen also showed examples of each of the stimulus luminance levels. At the start of each trial, which began 500 ms after the instruction screen or the feedback from the previous trail ended, the observer was cued with a screen containing just the cue square, a white line that outlined the stimulus region (500 ms), that was followed by the stimulus (250 ms), then came a blank screen (50 ms), the mask (500 ms), another blank screen (50 ms), and then the display that the observer used to respond; finally, the feedback display was presented. The feedback and response displays used for the different tasks are described below. In all tasks, the observer terminated the feedback screen by pressing any key.

Response and Feedback Screens in Mean Luminance Task:

We used a procedure similar to the one used for the mean-size task, described in chapter 3. The observers were asked provide their mean- luminance estimations in continuous manner, similar to what it is obtained in the centroid task. The initial

response screen in the mean luminance task consisted of a probe square with a luminance randomly selected from a uniform distribution from the range of possible luminance levels. By moving the mouse horizontally, the observer changed the

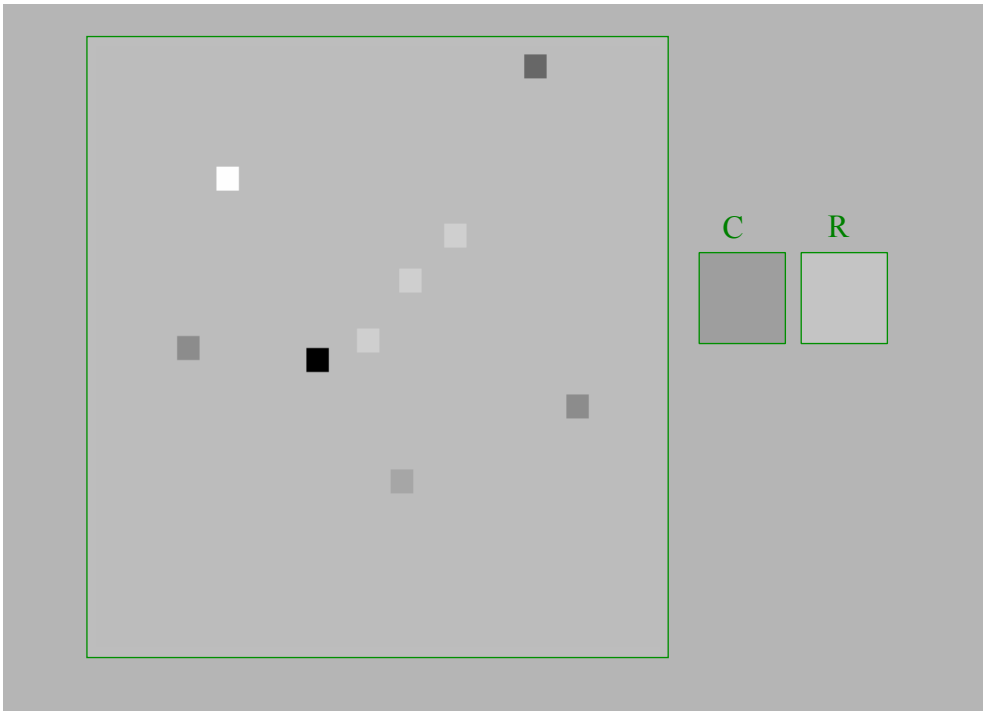


Figure 4. 5: Schematic representation of the feedback screen for the Mean Luminance task for a set luminance of 9 squares. The square with the C on top shows the correct luminance; the square with the R on top shows the observers response. In this example the observer perceived the mean luminance of the group to be lighter than the correct answer.

luminance of the probe square until its luminance matched the luminance of the estimated mean of the stimuli. Moving the mouse to the left made the probe square darker; moving the mouse to the right made it lighter. The observer terminated the response process with a mouse click. After that feedback was provided (Figure 4.5), in which we presented the observer with the stimulus used in that trial and two square with the correct response (labeled 'C') and a square with the observer's response (labeled 'R').

Response and Feedback Screens in Centroid Task:

As described in Drew et al., (2010), Sun et al., (2016) and in the experiment described in the previous chapter, observers moved the mouse and clicked in the location of the estimated centroid. Following this response, a feedback screen presented the original stimulus display along with a bullseye to indicate the location of the correct centroid and a white cross marking the observer's response (Figure 4.6).

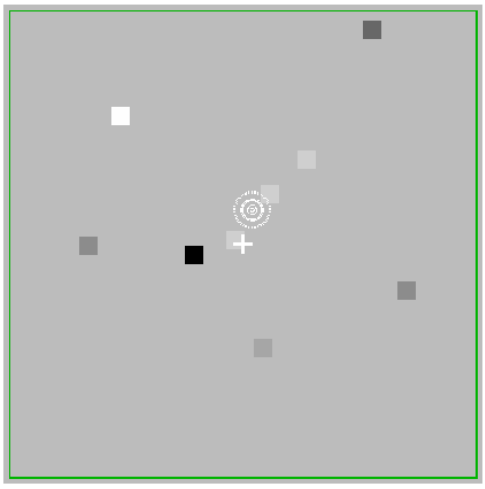


Figure 4. 6: Screen shot of the display used in one of the trials use in the luminance-weighted centroid task. The white cross represents the observers' response and the white bulls eye represents the correct response.

Design

The present study consisted of three tasks: (1) the equi-weighted centroid task, in which observers were instructed to estimate the centroid of the stimulus array by giving equal weight to all squares regardless of luminance, (2) the luminance -weighted centroid task, in which observers strove to estimate the centroid of the stimulus array weighting items in proportion to their luminance, giving more weight to lighter items, and (3) the mean

luminance task, in which observers were asked to determine the mean luminance of the squares in the stimulus, ignoring their locations, by adjusting the luminance of a single square.

The conditions in this experiment were constructed from the factorial combination of two factors: the task (equi-weighted Centroid, luminance-weighted Centroid, or Mean Luminance) and the number of items displayed (1, 3, or 9 items). A session consisted of 6 blocks with each task appearing twice in A-B-C-C-B-A order. Across sessions the task order was varied using a 3x3 Latin square, with the conditions for each observer taken from a different row. After at least one training session, the data from which are not reported here, each observer ran 2 test sessions (a total of 4 blocks per task). Experimental session blocks consisted of 105 trials of which 5 were singleton trials. Singleton trials were included to estimate the error due to processes that were not associated with estimating the mean luminance or the centroid (e.g., sensory mislocalization of items and response motor error). On the remaining trials, trials with of 3 or 9 items were presented 50 times each. The order of the numerosity condition within a block was randomly determined.

Results

We computed the root mean square error (RMSE) of the responses across the across the four blocks for each task and compared them. The RMSE was stable for seven out of the eight observers. Due to the unstable performance for the remaining observer, the data from that observer were eliminated from the analysis.

To simplify the summary, we will consider the data from the singleton trials separately, so that, for most of the summaries, only results for trials with three and nine items are reported. We will focus on two pre-planned contrasts for the task factor: one comparing the results in the equi-weighted and luminance-weighted centroid tasks and one comparing the luminance-weighted centroid tasks results with the mean- luminance task results.

Efficiency

Observers achieved higher, and almost identical, Efficiency in the two centroid tasks, and lower efficiencies in the mean-luminance task (Figure 4.7). The preplanned contrast comparing both centroid tasks suggests that efficiencies for the luminance-weighted centroid task are essentially identical to those from the equi-weighted centroid task (**Δ** = 0.01, SD = 0.03 t(6) = 0.625, p = 0.552, BF = 0.56). The preplanned contrast comparing the Efficiency for the luminance-weighted centroid task and the mean-luminance task very strongly suggests that observers were able to use luminance more effectively when estimating the centroid of a group of squares than when estimating the mean-luminance of the same group (**Δ** = 0.23, SD = 0.05, t(6) = 12.011, p = 0.00, BF = 57.6).

Figure 4.7 also shows that observers achieved higher efficiencies when presented with three items displays than when presented with nine items displays for all three tasks. A t-test provided evidence for a reduction of Efficiency with increased numerosity (Figure 4.7) for all three tasks (Δ=-0.19 SD = 0.06, t(6) = -8.117, p <0.00, BF = 37.14), but that there was no interaction between set size and task (Δ=-0.01 SD = 0.04, t(6) = .695, p = 0.509, BF = .59).

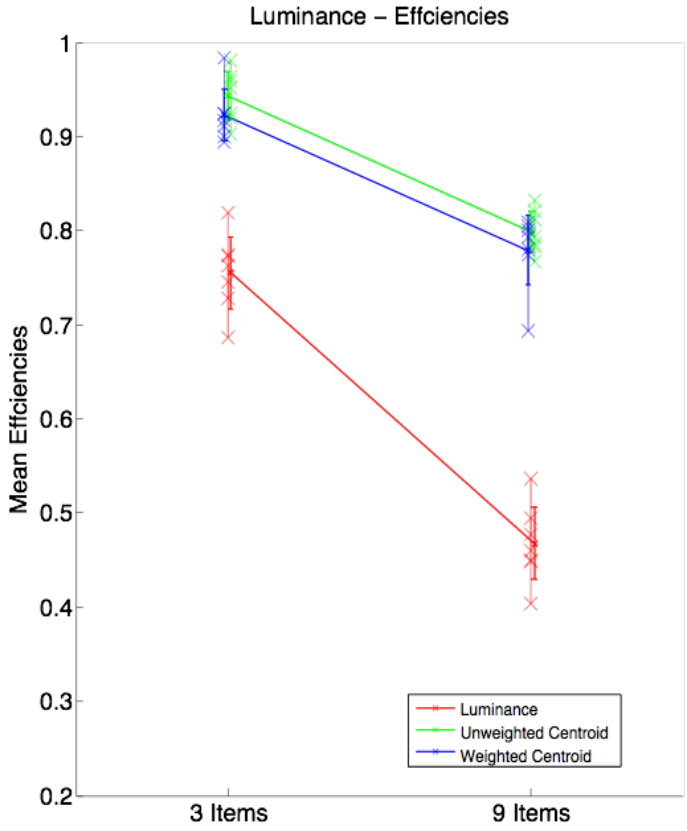


Figure 4.7: The efficiencies from all observers for all three tasks. Individual data is presented by the X marks. Note that the Mean Luminance tasks yields lower efficiencies than the two-centroid tasks.

Influence Functions

Figure 4.8 shows the averaged influence functions for all observers for all three tasks: (1) equi-weighted centroid, in black (2) the luminance-weighted centroid, in red and, (3) the mean-luminance task, in blue. In the equi-weighted centroid task, the expected slope of the ideal influence function is zero. The average data follow this ideal closely- observers tended to give equal weight to all the items, regardless of their luminance level. In the luminance-weighted centroid task we found that in general observers are giving weight according to a square's luminance level, as instructed but instead of the expected linear influence function, it appears that observers tend to assign the weights following a pattern similar to a S-shaped function. This inference based on the average data may be misleading, however. Looking at the data from the individual observers, only two clearly exhibit a sigmoid pattern.

The variability and the larger confidence intervals obtained from the mean-luminance data, makes it difficult to make concrete comments about an overall tendency. The expectation was that observers will generate a linear influence function, similar to the function expected in luminance-weighted centroid task, which gives weight to the squares according to their luminance level. We see that, in general they are able to do this for the squares darker than background items, but that this response flattens out for the squares brighter than the background.

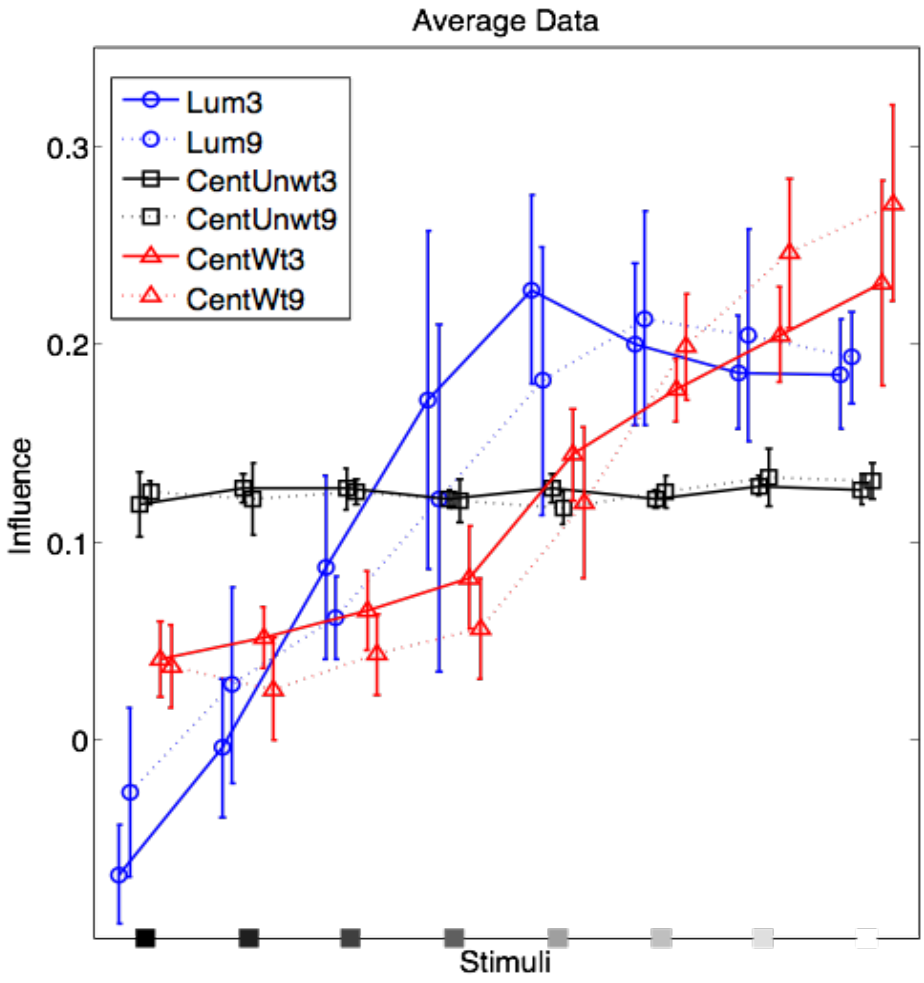


Figure 4. 8: Overall influence function for all observers in the three tasks for three and nine items. In the equi-weighted centroid task (black) observers are able to give equal weights to all the items in the display; in the luminance-weighted centroid task (red) observers are able to give equal weights more weight to lighter items: and in the mean-luminance task (black) observers are able to give more weights to all the items in the display.

Slopes Analysis

A slope analysis was conducted for the three tasks. Different models had to be fitted to the influence functions obtained in each task, therefore the results of the slopes analysis will be reported separately.

Equi- Weighted Centroid Slopes

As shown in figure 4.8 (in black) observers were able to equally weight all the items in spite of their luminance level. A t-test analysis (Table 4.1) show that the influence function obtained from the equi-weighted centroid task closely resembles the expected flat function with slope = 0. (Mean = 0.02, SD = 0.13, t (6) = 0.448, p = 0.670, BF = .51). No effects of set size were found (Δ=0.01, SD = 0.02, t (6) = 0.740, p = 0.487, BF = .59).

Luminance-Weighted Centroid Slopes

At first glance, the influence function obtained from the Luminance-Weighted Centroid looks like its linear slope only approximates a sigmoid function. We tested to see whether the sigmoid function actually fits better, but we found that the sigmoid model seems appropriate for the data averaged across observers, but for most of the individual data a linear fit is more appropriate. Figure 4.9 shows an example of the data of one observer for which the sigmoid function was not the best fit. A linear regression of the average data shows that there is a mild effect of numerosity for this task (Δ=0.12, SD = 0.09, t (6) = 3.566, p = 0.012, BF = 6.5), but the slopes for both functions are close to one (Table 4.1).

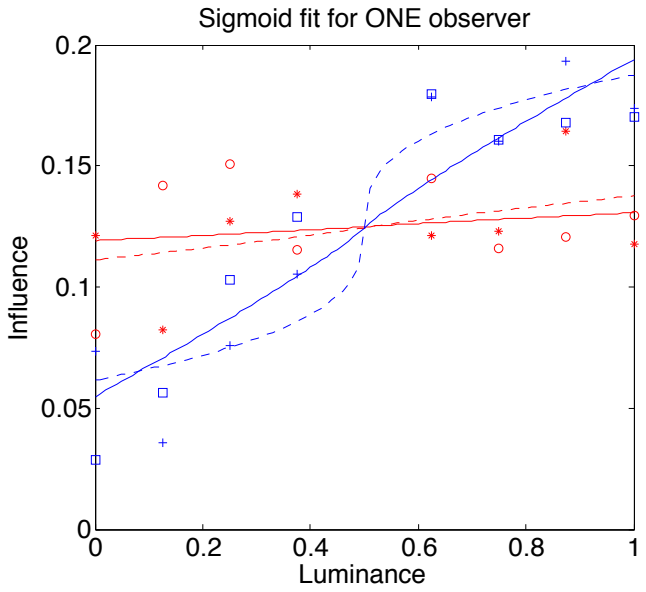


Figure 4. 9: The slopes of each observer were fit to a sigmoid function. For most of them, as shown in this figure, this was not an appropriate fit. Therefore linear fit was used for the slopes analysis.

	Equi-Weighted Centroid		Luminance - Weighted Centroid	
	3	9	3	9
	Ideal Slope =0		Ideal Slope =1	
Average	0.02	0.03	0.80	1.05
CI at 95%	[-0.10 0.13]	[-0.10 0.16]	[0.69 0.92]	[.88 1.23]
SD	0.12	0.14	0.12	0.19
t(6)	0.35	0.52	17.14	14.91
p-value	0.74	0.62	0.00	0.00
BF	0.50	0.53	71.85	66.90
Effect of Numerosity				
Difference	0.01		0.12	
CI at 95%	0.02		0.09	
SD	[-0.01 0.02]		[0.04 0.21]	
t(6)	0.74		3.57	
p-value	0.49		0.01	
BF	0.60		6.56	

Table 4. 1: Slope analysis for both centroid tasks. The equi-weighted centroid task generates slopes close to zero and the luminance-weighted centroid task generates slopes close to one, as expected.

Mean Luminance Slopes

The influence function for this task (Figure 4.8 in blue) shows a linear function from the 1st level of luminance (darkest item) and around the 5th level of luminance the function flattens out. We fitted the data to a two-segment linear model with different slopes around a transition point (joint), which was estimated from the fit. We found that, in general to the left of the 5th luminance level, the influence function has a positive slope, while function to the right of the 5th luminance level flattens out. There was a mild effect of numerosity for the items to the left of the 5th luminance level. Details are summarized below in table 4.2.

Mean Luminance -Task				
	3-items		9-items	
	Ideal Slope =1			
	Left Slope	Right Slope	Left Slope	Right Slope
Average	0.67	-0.15	0.43	-0.07
CI at 95%	[0.39 0.98]	[0.00 -3.1]	[0.683 0.184]	[0.20 -0.35]
SD	0.30	0.17	0.27	0.30
t(6)	5.95	-2.44	4.25	-0.66
p-value	0.00	0.05	0.01	0.54
BF	21.70	2.60	10.13	0.57
	Effect of Numerosity			
Difference	-0.24		0.08	
CI at 95%	[-0.46 -0.02]		[-0.147 0.308]	
SD	0.24		0.25	
t(6)	-2.36		0.86	
p-value	0.04		0.42	
BF	2.43		0.65	

Table 4. 2: Analysis of the slopes for the Mean Luminance task. Note that a two-segment linear model was fitted to the different slopes on either side of a joint point, which was estimated from the fit.

Discussion

We found than observers are able to use the luminance of a group of items more efficiently when asked to estimate the centroid of a group of items than when asked to estimate the mean-luminance. Efficiencies for both centroid tasks are higher- and almost identical - suggesting that the luminance level of an item could be either ignored or used in the context of the centroid task, but not as efficiently in the mean-luminance task. These results confirm what was found in the size experiment (Chapter 3), that a summary statistical representation based on a combination of two distinct kinds of information, in this case location and luminance, is easier for observers to produce in the centroid context, than a summary statistical representation based on only one of these components, size when the task is estimating the average of that feature. The high efficiencies achieved in the luminance-weighted centroid task show that observers perceived both location and luminance levels and that information is accurately registered for most, if not all of the squares.

One of the purposes of including singleton trials was to understand how the feature being studied is perceived free of any model-based interpretation. We conducted a root mean squared error analysis (RMSE) of the raw data from the mean luminance task for all three numerosities. Each of the colored lines connecting the x plotting symbols, reflect the data from one observer. The black circles show the mean error for each set size, averaged across observers. The black solid line is the best linear fit. The error for the singleton trials (in cd/m^2) yields a standard deviation of 2.91. The luminance levels of the squares used as stimuli ranged from 2 to 118.18 (in cd/m^2). These results suggest that observers were able to accurately perceive and remember the luminance of the single square by using the procedure presented in this experiment (i.e. by moving the mouse to adjust its luminance level).

The plot for the best linear fit (Figure 4.10) shows that (see Chapter 3), the mean-luminance error increases linearly with the logarithm of the numbers of items, but not reliably (slope = 0.875 [-1.81 3.57], t(6) = .794, p = .457, BF = 0.63). The assumption is that if the errors were due to the misperception of the luminance levels, such error should decrease with set size (black dashed line in Figure 4.10, one over the square root of the number of items). These results suggest that something other than the misperception of the luminances of the items must be contributing to the error observed in the 3- and 9-item conditions.

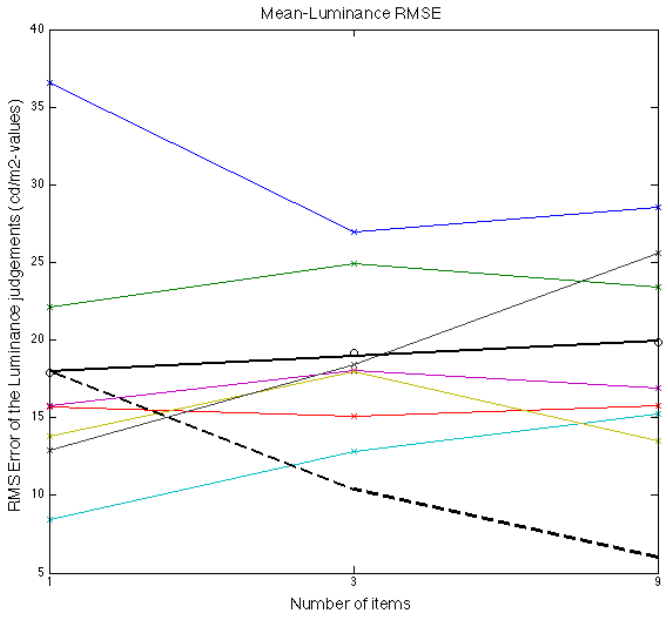


Figure 4. 10: RMSE plotted against the number of items. The colored lines represent the error for each observer. The black line represents the best-fit function for all observers and the black – dashed line represents the expected RMSE function, which is to decrease as the set size increases.

The influence functions from both centroid tasks are consistent with those found by Sun et al. (2016). In the equi-weighted centroid task observers were able to successfully give equal weight all the 8 levels of luminance used in the display. In the luminance-weighted task the expectation was that observers would apply a linear weighting rule, meaning

that they will give less weight to darker than background items and more weight to lighter than background items. The influence function for the luminance-weighted centroid task shows that in general observers give more weight to brighter items and less weight to darker items, as instructed. The high efficiencies obtained from the luminance-weighted centroid task suggest that observers have used a weighting rule that accurately gauges the luminance of display squares needed to complete this task.

But for the mean-luminance task, Figure 4.8 (in blue) shows how observers applied a different weighting rule that could be described as a positive linear weighting rule for darker items, and a flatter function for the lighter than background items. It has been documented that the discernibility of darker than background items is higher when compared to lighter than background items (Lu & Sperling, 2012). The positive slope to the left side of the inflection point - located near the background luminance level - suggests that it the observer is able to gradually weight darker than background items. This tendency reaches a maximum near the background luminance level. Then, the influence function to the right side of the inflection point flattens out, suggesting that as the items become lighter (brighter) it becomes more difficult to apply the appropriate weight according to their luminance level.

The influence functions obtained from the mean-luminance task resemble one of the selectivity functions found by Silva & Chubb (2014). In their study they identified four field channels. They described these channels as a pre-attentive mechanism that transforms grayscale images into visual system perception. In their experiment they found that how observers are able to weight more one of the four channel types which are activated depending on what type of gray scrambles (task) that was presented. One of the resulting sensitivity functions appears to be similar to the influence function that we found in the mean luminance condition, the **"**up-ramped channel.**"** This sensitivity

function has zero sensitivity for the black level of luminance, increases linearly with brightness until reaches a limit near mid-gray, and then flattens out.

Efficiency across features: Comparing size and luminance results

Table 4.3, shows the Efficiency values for this and the previous experiment in which what was varied as size and luminance, respectively. We found no significant difference when we asked observers to locate the centroid of a group of items when asked to weight size or luminance. Also, there was no significant difference when observers are asked to use the attended feature in the weighted- centroid estimation. We found a mild difference between the efficiencies obtained in both mean-feature tasks – which were slightly higher for the mean-luminance task.

	Mean Estimation	Equi-Weighted Centroid	Feature-Weighted Centroid
	Efficiencies		
Luminance	0.68	0.87	0.88
Size	0.51	0.88	0.86
Difference	0.17	-0.01	0.02
SD	0.12	0.05	0.06
t-value	2.646	-0.285	0.501
DF	13	13	13
p-value	0.02	0.78	0.625
BF	2.8	0.53	0.55

Table 4.3: Efficiencies comparison between Size- feature experiment (Chapter 3) and current study. Note that the efficiencies for the Mean Size experiment are lower that those achieved in the Mean Luminance, while the centroid efficiencies in both studies are almost identical.

Interim Conclusions

The main finding of this experiment is that the efficiencies obtained from the Mean - Luminance task are lower that those obtained in the two-centroid task. These results replicated what it was found in the experiment in which the size was the guiding feature in the observers' estimations.

The results from this experiment provide more support to the claim that mean estimations are different and perhaps more difficult than the centroid estimations. One speculation, which will be address in the following chapter, is that this difference in performance between the centroid and the mean estimations, could be explained by the distinctions between the ventral-dorsal pathways as described by Ungerleider and Mishkin (1982) and Goodale & Milner (1992). In the following chapter we explore this idea by using size as the guiding feature in both mean size and centroid judgments.

Chapter 5

The availability and use of size in mean size estimations

So far, we have seen how observers achieve higher efficiencies in two variations of the centroid task than in mean-feature estimation tasks. In addition to that, in the previous chapter we presented how participants seem to use luminance better than size when asked to estimate an average of each feature in a group of squares, but that their performance was still worse than when using these feature dimensions to weight centroid estimations. What is it that makes mean estimations more difficult than centroid estimations weighted by the same features?

The main goal of the study presented in this chapter is to explore if the difference in performance between the size-weighted centroid and the mean size task, could be explained by the distinction between the ventral-dorsal pathways described by Ungerleider and Mishkin (1982) and Goodale & Milner (1992). Ungerleider and Mishkin (1982) first described the functional and anatomical differences between the dorsal and ventral visual pathways in monkey brain. Subsequently, research conducted in patients with optic ataxia (damage in parts of the dorsal pathway) or visual agnosia (damage in parts of the ventral pathway), helped to differentiate the functions between these two visual pathways in humans. The case study of patient D.F., who had damage in a part of the ventral stream, illustrated the implications of these two processes (Goodale & Milner, 1992). When D.F. was asked to describe the orientation of a slot in a mailbox, she was unable to do so, and claimed not to be able to see the slot; but when she was asked to put the letter inside the slot, she was successful at aligning the letter to match the orientation of the slot every time. Goodale and Milner (1992) suggested that this difference in performance was due to the damage of her ventral stream processing,

which impaired her awareness of the orientation of the slot. But since there was not damage reported in her dorsal pathway her ability to use guide movements using the visual information, the “how” processing of the dorsal stream, was not impaired.

I hypothesize that the mean size estimations, as presented in the previous experiment, required ventral stream processing, while centroid responses, were processed in the dorsal stream. According to the “what/how-where” characterization proposed by Goodale & Milner (1992), processing in the ventral stream is associated with the recognition and description of visually-perceived objects, while the dorsal stream is associated with using visual information for motor planning. The crux of the hypothesis is that, in the case of the mean size task, observers were required to identify and characterize the size of each item as a separate object, possibly in working memory, and then use that information to estimate the mean size of the set (so the response is based on “what” processing). By contrast, in the centroid tasks, size (or brightness) along with the location of the items is operated on directly, bypassing object identification and working memory, to determine the endpoint of a movement (so the response is based on “how/where” processing). We speculate that performance of observers might be improved for a task in which the mean-size of a group of items is used directly to select a movement endpoint, in a similar fashion to the way that they respond in the centroid task.

Some support for this suggestion is provided by the data from an fMRI MVPA study that showed that the relevance of the task and the features of the object play a role in the dorsal pathway (Vaziri-Pashkam et.al, 2017). Results from three experiments, in which the strength of color and the shape of an object were manipulated to either be relevant or not relevant to the task, suggest that the dorsal pathway was significantly influenced when the feature of object was relevant to the task, while only a minor effect was found in the ventral system. They argued the dorsal system deals more with how the stimulus

is used in the context of the task. For example, in the centroid task the observer uses stimulus features such as the location and size to guide their response, which makes these features relevant to the task. We posit that if we present observers with a task in which the mean size of clusters of objects is used to determine an end point location, it is reasonable to think that the dorsal stream will be used to select and guide the movements of the observer's response, generating measures of performance similar to those achieved in the centroid task.

Another aim of this study is to address the possibility that in the study described in chapter 3, the centroid estimations might have been based on the relative size of the items while the mean size estimations required the use of absolute size information. In an experiment conducted by Chong and Treisman (2005)[3] they used a display that allowed observers to estimate mean size by using relative size. Their display consisted of an array of circles, which was split into to two subgroups; one array was located to the right and the other to the left visual field. In each trial, they varied numerosity and density on each visual field. Using a 2AFC, observers were asked to click the key that represented the side of the screen with the largest mean size. They found that observers achieved 75% accuracy when the size difference between the displays was between 6-8%.

The experiment presented in this chapter eliminates, in a similar way, the necessity for using absolute sizes to carry out the task – although the task can be completed with absolute sizes. It also, in contrast to the 2AFC approach used by Chong and Treisman (2005), allows observers to respond using the mouse and clicking on the group of items selected as larger or smaller.

[3] Described in detail in the Literature Review: Chapter 1

The display consisted of three clusters of items, spaced 120° apart with a randomly chosen starting orientation, on a circle surrounding fixation. In the simplest version of the experiment, the size-choice task, the responses need only be directed toward the correct cluster, which was the one with either the largest or the smallest mean size; however, the response was only required to select the cluster – no feedback was given comparing the endpoint of this movement and the centroid of the cluster. In the other two conditions, observers were asked to indicate which of the three groups contained the largest (or smallest) items by accurately locating the centroid of the group selected. Constructing an accurate centroid response now potentially becomes a secondary task that must compete for processing resources with the task of determining the appropriate group. Of special interest is the relationship between the size processing done for the cluster selection task and the equi-weighted centroid task, and whether the equi-weighted centroid response is disrupted by the cluster selection task.

Method

Observers

Seven observers, including the first author, participated in the experiment. All observers had previous experience with the centroid task and were students at the University of California, Irvine. Six were females between the ages of 17 and 40. All observers reported having normal or corrected-to-normal vision. The present study was conducted in accordance with the regulations of the Institutional Review Board of the University of California, Irvine.

Apparatus and Stimuli

The observer sat in an adjustable height chair in a dark room and viewed the stimuli presented on an iMac (Mac OS X) with a 54 cm screen controlled by an ATI Radeon HD

4670 graphics card from a distance of about 84 cm. The stimuli were generated using the Psychophysics Toolbox Version 3.0.8 (Brainard, 1997; Pelli, 1997; Kleiner et al, 2007) for MATLAB (Version 7.1).

A screen shot illustrating the display used in this experiment is shown in Figure 5.1. The size of the stimulus area was 600 x 600 pixels and the viewing angle was approximately 15 degrees.

Observers were presented with a display that consisted of three clusters of four squares (Figure 5.1). The sizes of the squares were drawn from a set of eight fixed sizes (0.23°, 0.27°, 0.34°, 0.45°, 0.52°, 0.67°, 0.81°, 0.99°) that were randomly selected and assigned to each cluster according to a discrete triangular distribution (5.63%, 10.25%, 14.75%, 19.38%, 19.38%, 14.75%, 10.25%, 5.63%). The locations of the center of each group were at a fixed radius of 2.9° from the center of the screen, spaced 120° apart with a random starting angle. The color of the squares was fixed to white (90.48 cd/m^2) on a gray (58.73 cd/m^2) background.

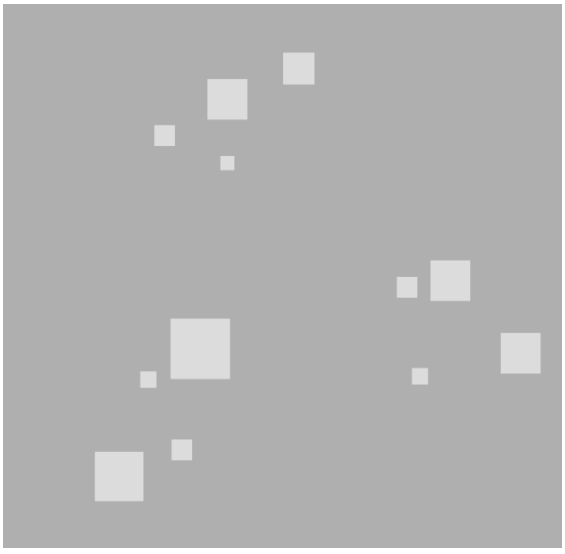


Figure 5. 1: Screen shot of the display used in one of the trials. It consisted of three clusters, which were randomly separated by 120^o. Each clusters had four items that were chosen randomly from 8 predetermined sizes.

Procedure

On each trial, observers were asked to identify both the cluster with the largest mean size and the cluster with the smallest mean size. Depending on the session they were presented with one of the following three tasks: a simple task (Size-Choice) and two combined tasks (Size-Choice-Equi-Weighted and Size-Choice-Size-Weighted). In the size choice (SC) task, after being presented with the three clusters, observers were asked to select the cluster with the largest mean size and the cluster with the smallest mean size. In Size-Choice-Equi-Weighted (SCEW), observers were asked to select the cluster with the largest (smallest) mean size by clicking on the centroid of the chosen cluster, giving equal weight to all the items in the cluster. The procedure was similar in the Size-Choice-Size-Weighted (SCSW) task, except that the centroid of the selected clusters, computed by giving more weight to larger items, was now the expected response. An initial screen displayed the instructions for each session and showed examples of each of the stimulus sizes.

The events on a trial are illustrated in Figure 5.2. At the start of each trial, the observer was cued with a blank screen that was followed with the stimulus, as described above, displayed for 250 ms. The stimulus was followed by a blank screen (50 ms), a mask made up of a jittered array of randomly sized items covering the display (500 ms), and another blank screen (50 ms) followed before the observer's response was required.

On the response screen, a red letter "L" or a blue letter "S" appeared initially at the center of the display area. The letter functioned as a cursor, tracking the movements of the mouse. The appearance of the letter "L" (or "S") prompted the observer to move the mouse and click on the location of the cluster with the largest mean size or the smallest mean size. Half of the observers were asked to identify the largest cluster first.

Immediately after that, depending on the order assigned for that specific observer, the second letter appeared at the center of the display area and it also functioned as a cursor. The appearance of the second letter “S” (or “L”) also prompted the observer to move the mouse and click on the location of the cluster with the smallest mean size or largest mean size.

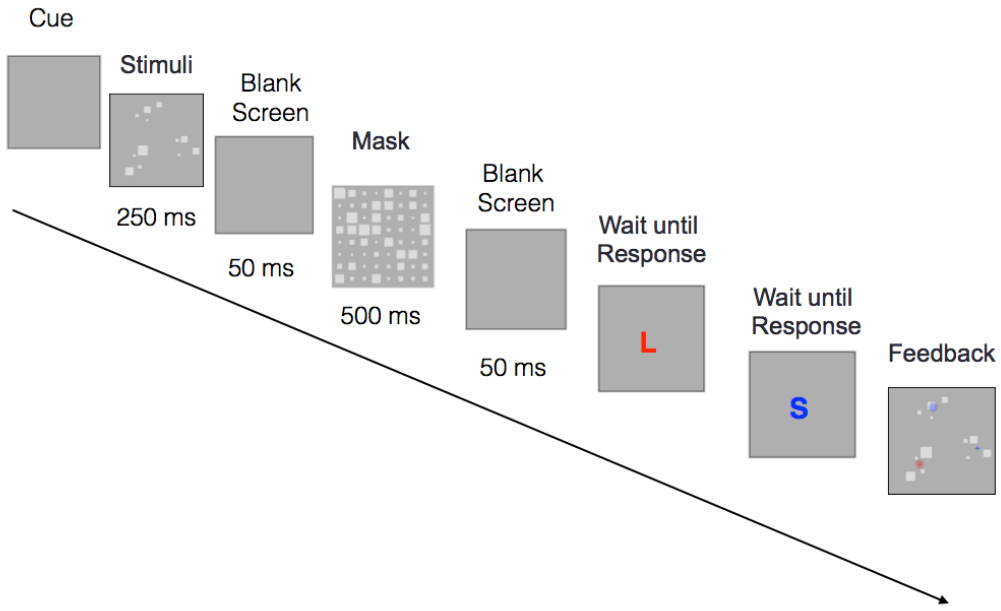


Figure 5. 2: Experiment Timeline

After both responses were made, a feedback screen followed. The feedback screen displayed the stimuli used in that trial. For the SC task, the feedback screen was divided into three areas as illustrated in Figure 5.3. Two areas were colored with a semi-transparent overlay: red (for the segment containing the large cluster) and blue (for the segment containing the small cluster). The third cluster was uncolored. Red and blue

crosses marked the location of the large and small responses, respectively. As long as the large and small response locations were in the colored area, the response was counted as correct.

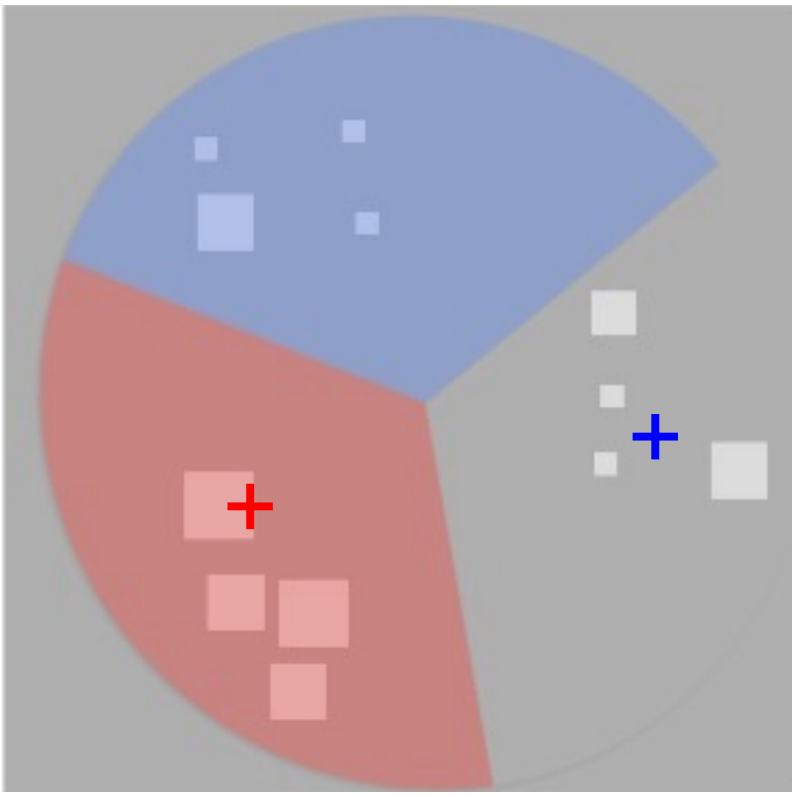


Figure 5. 3: Feedback screen used in the Size Choice condition, for which no centroid feedback was provided. In this trial the red area has the cross on it, meaning that the observer selected the correct large cluster. However, the small size estimation was incorrect.

For sessions in which responses were also judged based on how close they are to the centroid of the correct group, a feedback screen like that in Figure 5.4 was used. The screen displayed the stimuli used in that trial, a red cross showing the perceived location of the largest group and a blue cross showing the perceived location of the smallest group. Instead of colored regions marking the regions of acceptable responses, this feedback screen had a red and blue bullseye centered at the centroid of the largest (red) and the smallest (blue) group, respectively. The location of these bullseyes was based

on either the unweighted or the weighted centroids of the items in the group, depending on the condition.

Following each block, summary feedback in the form of an overall score was provided. Observers were encouraged to improve their score from block to block. The components of this score varied depending on the condition tested: the accuracy of the cluster responses and the distance error of the centroid responses. After blocks for which only the accuracy of the cluster responses was emphasized, the score in the block level feedback was simply a count of the number of errors in that block. In the blocks for which the positional accuracy of the response was also emphasized, the observer received feedback for both cluster selection error and the root mean square error of the distance between the response position and the correct centroid, only on trials where the cluster selection was correct.

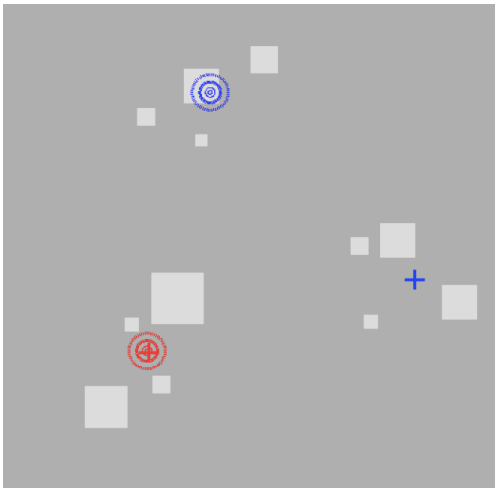


Figure 5. 4: Feedback screen used in the combined conditions. In these conditions feedback for Size-Weighted and Equi-Weighted centroid was provided in addition to the main task, which was the cluster selection task. In this trial the observer chose the correct “large” cluster, shown in red, but chose the incorrect “small” cluster. The feedback for the correct cluster selection (large) show how close this particular observer was to the correct centroid.

Design

Each observer participated in 12 sessions, divided into 4 sessions per day. On each day an observer ran four sessions, with the first and last ones being the Size-Choice (SC) task. The two sessions in the middle were counterbalanced to be either the Size-Choice-Equi-Weighted (SCEW) task or the Size-Choice-Size-Weighted (SCSW) task. Half of the observers followed the order shown in the figure 5.5 and, for the other half, the order of the SCEW and SCSW sessions was reversed. A session consisted of six blocks of 25 trials each, with block feedback provided after each block.

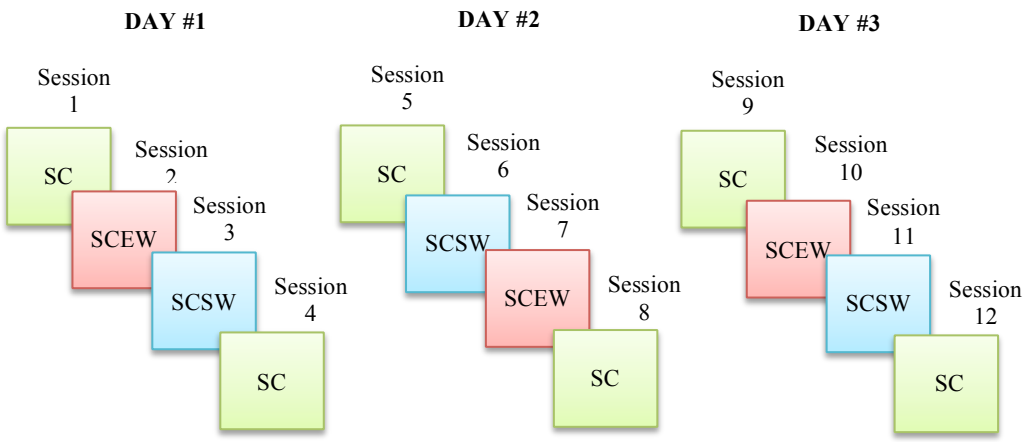


Figure 5. 5: Observers ran the experiment in three days. Each of the days the observer began and ended with the Size Choice condition. In between these two conditions they were asked to judge the centroid in addition of selecting the correct cluster.

Estimation of the influence function and efficiency for cluster selection responses

The nature of the Size Choice task allowed us to treat each of the two responses required in each trial as a separate piece of data. For this combined analysis, the responses choosing the smallest cluster were recoded so that all of the responses involved a comparison between a larger mean size, target cluster and a smaller mean size cluster. A Markov Chain Monte Carlo simulation was used to obtain weights associated with each of the item sizes in a Probit model. If the response on trial *i*, R_i, is coded as 1, when it is correct, or 0, when it is incorrect, then the predicted response under the Probit model is given by

$$\hat{R}_i = \begin{cases} 1, & \text{if } \Delta = \sum_{j=i}^{4} W(l_j) - \sum_{j=1}^{4} W(s_j) + \epsilon_i > 0 \\ 0, & \text{otherwise} \end{cases}$$

Where l_j represents the size of one of the four stimuli in the cluster with larger mean size, s_j represents the size of one of the four stimuli in the cluster with smaller mean size, $W()$ maps stimulus size onto the weight (relative subjective size) of each stimulus type in the cluster selection response, and ϵ_i is a random variable drawn from a normal distribution with mean 0 and standard deviation 1. To convert the weights estimated from this model into influence function values, they were normalized to sum to 1.

The calculation of the efficiency for the cluster selection responses was also done differently than that for the centroid responses. Because there were only 4 stimuli in each cluster, it is practical to enumerate the 16 possible ways that the items in a cluster might decimated. For a comparison involving two clusters, there are 256 possible ways that the two clusters might be decimated. Given the probability that an item will be retained – i.e., the efficiency – it is straightforward to calculate the exact probability of

occurrence for each these 256 outcomes. Using the item type weights estimated above, it is similarly straightforward to compute the probability that $\Delta > 0$ for each decimation combination should it occur. Weighting these probabilities of a correct response for each decimation combination by the probability that that combination will occur provides an exact estimate of the average proportion of correct responses across a set of trials for a given level of efficiency. Based on this, it was straightforward to use the Matlab function fminbnd() to estimate the efficiency for a set of stimuli and the associated responses.

Influence functions and efficiency for the centroid response were estimated as in the previous chapters except that we only included the data in this analysis for which the correct cluster was selected.

Results

All observers ran 12 sessions, 6 sessions total for the Size Choice (SC) task, 3 sessions for the Size-Choice-Equi-Weighted (SCEW) task and 3 sessions on the Size-Choice-Size-Weighted (SCSW) task. We will focus on two pre-planned contrasts for the task factor: one comparing the SC tasks results with the average of the two centroid tasks and one comparing the results in the SCEW and the SCSW centroid tasks.

We calculated efficiencies for the cluster selection portion in all three tasks and measured the effects of the response order ("large response first versus small response first"). We found that there are large main effects of response order for the cluster selection component: selecting the "Large" cluster first improved efficiency in all three tasks (SC: $\Delta = 0.11$, SD = 0.06, $t(6) = 3.907$, $p = 0.011$, Bayes Factor **(BF)** = 4.25; SCSW: $\Delta = 0.16$, SD = 0.04, $t(6) = 5.71$, $p = 0.002$, BF = 19.9; SCEW: $\Delta = 0.13$, SD = 0.04, $t(6) = 4.583$, $p = 0.006$, BF = 12.1). To test if the size of this effect depend on the task, we compared the response order effect between the size choice size weighted and size choice equi weighted task, ($\Delta = 0.01$, SD = 0.02, $t(6) = 0.767$, $p = 0.477$, BF = 1.91) and also we compared the size choice task with the average of the two centroid task ($\Delta = 0.02$, SD = 0.01, $t(6) = 2.153$, $p = 0.084$, BF =2). We found that the size of this effect doesn't depend on the task.

Size Choice Results (Cluster Selection)

The main task in this experiment was to select the clusters with largest and the smallest mean size. We measured performance on this selection task in two ways (Table 5.1), by tabulating the percentage of correct responses when selecting the correct cluster and with an efficiency analysis, calculated as described in the Methods section. There was a mild decrement in cluster selection accuracy when the feedback emphasized the

accuracy of the centroid response (Δ = -0.97%, SD = 1.22, t(6) = -2.260, p = 0.058, BF = 3.02). Similarly, we found a decrement in the cluster selection efficiency, when the feedback emphasized the accuracy of the centroid response (Δ = -0.01, SD = 0.01, t(6) = -6.46, p = 0.001, BF = 24.42).

	Cluster Selection			
	Efficiency		%Correct	
Response Order	Large First	Small First	Large First	Small First
Size Choice Only	0.82	0.7	82	76
SC Equi-Weighted	0.81	0.66	80.23	73
SC Size-Weighted	0.79	0.66	80.2	73

Table 5. 1: Values for efficiency and percentage of correct cluster selection separated for response order and averaged for all observers.

Centroid - (Secondary Task)

The secondary task of this experiment was to locate the centroid of each cluster while selecting the clusters with the largest or the smallest mean size. For the trials on which the cluster was correctly selected, we measured observers' performance in two ways, using efficiency and the root mean square error (RMSE) between the location of the responses and the correct location (Table 5.2). We took the RMSE of the response end point location when centroid feedback was provided and compared it with the error when such feedback was not provided. Similar efficiencies were found between the two tasks for which centroid feedback was provided (Δ = 0.90, SD = 6.91, t(6) = 0.344, p = 0.743, BF = 0.50). Observers produced higher RMSE in the Size Choice when no feedback was provided (Δ = 13.93, SD = 9.43, t(6)= 3.909, p = 0.008, BF = 8.2).

End Point location (centroid)		
Task	*Efficiency*	*RMSE*
Size Choice Only	n/a	43.2825
SC Equi-Weighted	0.665	30.485
SC Size-Weighted	0.6775	30.075

Table 5. 2: Values for efficiency and RMSE of the end point location (secondary task) averaged for all observers

Influence Functions

Figure 5.6 shows the influence functions for all observers for all three tasks: (1) SC, in green (2) SCEW in blue and, (3) SCSW, in red. The expectation is that the slopes of the influence values regressed against the normalized sizes should be close to one for the cluster selection responses, in all three tasks. Table 5.3 show that the data do not accord with this expectation. Instead, for all three tasks, the slope is quite close to 1.4 and is significantly greater than 1. There is also no difference in the slope across tasks. A slope analysis found that there is no significant difference between the two centroid (SCSW and SCEW) tasks (Δ = 0.01, SD = 0.034, t(6)= 0.756, p = 0.478, BF =0.61) and also between the SC and the average of the two centroid tasks (Δ = 0.003, SD = 0.026, t(6)= 0.299, p = 0.775, BF =0.49).

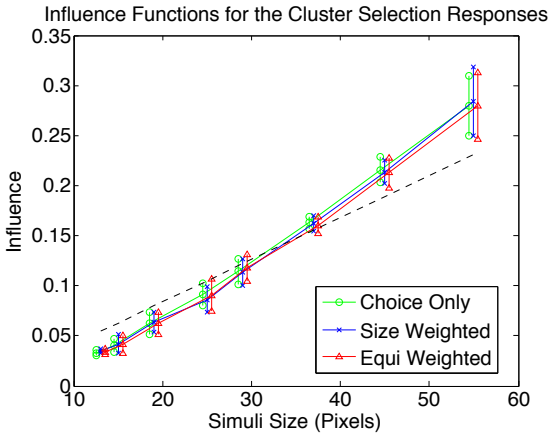


Figure 5. 6:Overall Influence Function for the cluster selection portion in all three tasks. In all three tasks observers weigh the squares according to their size.

Slope Analysis for Cluster Selection Influence Function

H0 = 1			
	Size Choice Only	SC Equi-Weighted	SC Size-Weighted
Average	1.41	1.40	1.39
CI at 95%	[1.37 1.43]	[1.37 1.44]	[1.35 1.42]
SD	0.04	0.04	0.04
t(6)	30.37	29.94	25.14
p-value	0.00	0.00	0.00
BF	84.20	84.01	81.30

Table 5. 3 Summary of the Slope Analysis conducted for the Size Choice influence functions

At first glance a similar pattern is shown in Figure 5.7. This figure shows the influence function for the centroid portion of the task. The slopes obtained from a linear regression of the influence values against the normalized sizes (Table 5.4) shows that the observers are in general weighting all the items in the display according to their size. The expected slope for the SCSW task is one, but as the table shows the observed values are closer to 0.7. The expected slope for the SCEW task is zero, but the table shows that, although smaller, these slopes were systematically larger than zero. However, the slope in the SCEW task, averaging across the large and small cluster responses, was reliably flatter than the slope obtained from the SCSW ($\Delta = 0.255$, SD = 0.175, $t(6) = 3.847$, $p=0.008$, BF = 8.44). Although there appears to be a systematic effect of cluster size on these slopes, averaging across tasks, this apparent difference was not reliable ($\Delta = 0.087$, SD = 0.141, $t(6) = 1.637$, $p=0.153$, BF = 1.25).

Slope Analysis for Centroid tasks
Influence Function

	H0 = 1		H0 = 0	
	SC Size-Weighted		SC Equi-Weighted	
	Large	Small	Large	Small
Average	0.74	0.65	0.52	0.41
CI at 95%	[0.61 0.86]	[0.48 0.86]	[0.40 0.60]	[0.28 0.50]
SD	0.13	0.21	0.11	0.12
t(6)	-5.22	-4.18	11.98	8.91
p-value	0.00	0.00	0.00	0.00
BF	16.39	9.74	57.50	42.2

Table 5. 4: Summary of the Slope Analysis conducted for both centroid influence functions

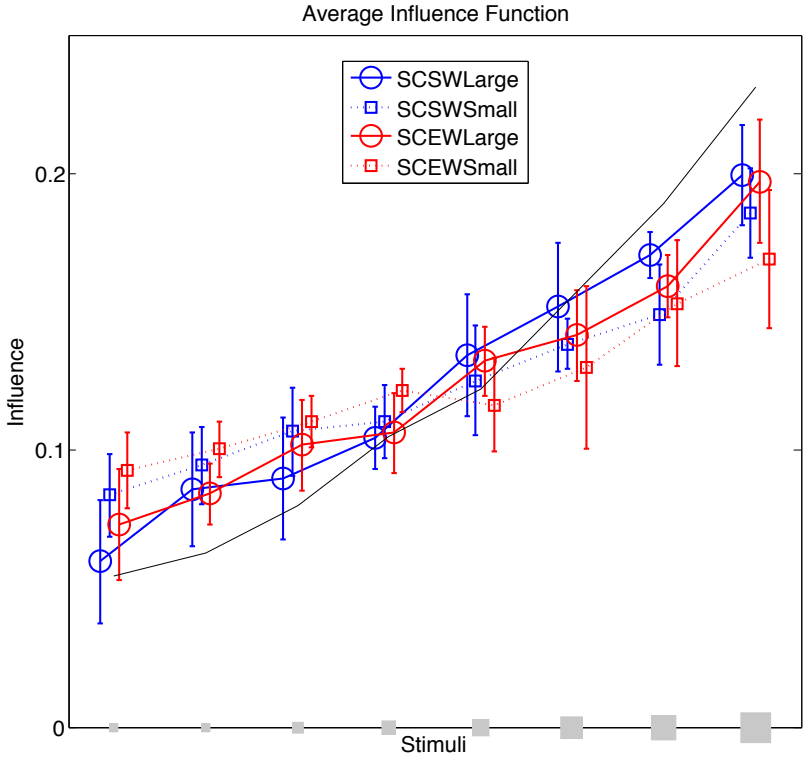


Figure 5. 7: Influence Function for the centroid portion of the experiment. The equi-weighted centroid task is plotted in red, and the size-weighted centroid task is plotted in blue. Circles identify the large clusters and squares identify the small clusters. The black line, which has a slope equal to one, is plotted as a reference. Note that regardless of the instructions, observers weigh items accordingly to their size.

Discussion

Using the efficiency values obtained in this experiment we can draw a comparison with the efficiency values obtained in the experiment described in chapter 3 (Table 5.5). The three efficiency values from the cluster selection response for all three tasks in this experiment were 0.72 or better. This is substantially larger than the efficiency of 0.55 with which observers were able to estimate the mean size of a group of items in the previous experiment. This increased efficiency was observed even though, in this experiment, observers had to estimate the size of three clusters and make two responses on each trial.

Mean Estimations (averaged for all obsservers and response order)	
Task	*Efficiency*
Size Choice Only	0.755
Size Choice with Equi-Weighted feedbak	0.725
Size Choice with Size-Weighted feedback	0.72
Mean Estimation (from Chapter 3)	0.55

Table 5.5: Efficiency comparison across the two experiments that use size as the main feature.

A possible explanation for the improvement in performance could be the nature of the response: the higher efficiency observed in this experiment could reflect a response that allows size information to be processed in the dorsal stream of visual processing, becoming more of a 'how/ where' task rather than a 'what task' as described by Ungerleider and Mishkin (1982) and Goodale & Milner (1992).

In the trials when centroid feedback was provided, observers not only had to identify and then select the largest/smallest clusters, but they also had to determine and respond with the centroid of those two clusters. Taking into account that this was the secondary

task and that two centroid locations had to be computed and then recalled on each trial, one could argue that the efficiencies on the centroid estimation portion of the experiment were relatively high, although lower than the ones obtained for the previous experiments (on average: .68 vs. .86). In addition, the centroid efficiencies obtained in this experiment are, however, still higher than the mean-size estimations obtained in Chapter 3 (.55). We suspect that observers can still reliably locate the centroid of a chosen cluster, because it has been documented that human gaze tends to prefer the centroid of items and that saccades land closer to the center of mass (Fehd & Seiffert, 2008 & Melcher and Kowler, 1999).

Another interesting finding from this experiment stems from the influence functions. Although they show that for the both the Size Choice and the Centroid portion of the experiment all observers roughly linearly weight the squares according to their sizes, there are some important differences between them. First, the influence functions for the cluster selection task are similar to one another (as expected) but with higher slopes than those for the centroid responses. In addition to that, for the centroid responses, although the SCEW slope is smaller than in the SCSW task, observers are unable to equally weight all the squares, as found previously in equi-weighted centroid tasks. It looks like the nature of the size choice task, which requires using size information to select the correct cluster, makes it difficult for observers to weight the items equally when locating the centroid of the cluster. It seems reasonable to ask whether the average data mask a mixture of observers who are able to achieve equal weight and those who produce centroids in the SCEW task using the same weighting as in the SCSW task. However, looking at the slopes of the influence functions for individual observers suggests that this is not the case: i.e., that the average data provides a reasonable summary for all observers.

As we described at the beginning of this chapter, the display that Chong and Treisman (2003) used in their experiment resembled the large cluster selection portion of our experiment. Chong and Treisman (2003) found that a 6-8% difference in size between the two clusters was needed to yield the 75% accuracy. To achieve 75% accuracy in the Size Choice only task (when no centroid feedback was provided) our observers needed a larger difference to achieve the same accuracy criterion – for large clusters 14.0% and for small clusters 13.1%. Although a larger size difference is required in our experiment, this might reflect differences in task difficulty across the two experiments. First, our display was divided in three areas, (3 clusters). Second, on each trial we asked observers to make two choices (observers here needed to select two clusters). And finally, the centroid feedback, added more to the complexity to the task, which resulted in a slightly higher difference in size required (Table 5.6).

% Difference in size (for 75% Accuracy)		
Task	*Large Cluster*	*Small Cluster*
Size Choice Only	14%	13%
SC Equi-Weighted	14%	15%
SC Size-Weighted	16%	15%
Chong & Treisman (2003) Exp.1	6-8%	n/a

Table 5.6: Difference in size (ratio) needed in all three variants of the experiment, compared with Chong and Treisman (2003) experiment.

A complication of the results obtained for the mean-size estimations described in chapter 3, was the possibility that the lower efficiency values observed were due to the fact that the task required observers to estimate and report absolute sizes. This is a complication because weighted centroid responses could have been based on judgments of relative size, although absolute sizes could also have been used. However, we were skeptical of this interpretation due to the fact that the singleton analysis and the results from the

size-weighted centroid task, suggest that the observers are registering size accurately and that the only difference leading to the low efficiency values, was the mean estimation requirement of the task.

To see if relative size was the reason of the performance improvement in the mean size estimations, we compared our results to Chong and Treisman's (2003) using efficiency. To do this we decided to take a look at the simulations conducted by Myczek and Simons (2008) (for more details see chapter 1 and 3). One of the main findings from Myczek and Simons (2008) simulations was that observers could be using a sub-sampling strategy when estimating the mean size of a group of items. More specifically, when they put Chong and Treisman (2003) experiment through this simulations, they found that a computer could sample 4 out 12 items, to achieve 75% accuracy - which could be interpreted as an efficiency = 0.33[4]. The overall efficiency value obtained in our experiment for the size choice only task, averaged for all observers, is .75. The higher efficiency obtained in this experiment could only be a result of the nature of the response used in this experiment, which is the major methodology difference between our experiment and Chong and Treisman's (2003) experiment. We argue the efficiency values obtained in our task are not due to the use of relative size because of the similarities the displays used in both experiments share. Our efficiency is higher because our mean estimation response was made to be similar to the centroid task, which we argue allows to dorsal processing, and as a result better performance.

Interim Conclusions

We found that observers achieve higher efficiency when asked to estimate the mean size of a group of items and their responses are guided by movement, suggesting dorsal processing. Using size information in the cluster selection task is slightly more accurate

[4] We interpret this ratio (4÷12) as a value similar to efficiency (as explained in chapter 2).

and efficient when there is no secondary task requiring the observer to report the centroid of each identified cluster, suggesting that observers treated the cluster selection as the primary task. In addition to that we found that, in centroid estimations, observers could not help but to weight items according to their size, regardless of the instruction. We think this is a by-product of the demands of the main task, which is cluster (size) selection, because the demands of this task require the use of size information, and this information becomes hard to ignore when estimating on the centroid of the cluster.

The use of relative size could be seen as a possible alternative explanation for the higher efficiency values, but the Myczek and Simons (2008) simulations provide some evidence against this argument. We argue that, the methodology used in our experiment, in contrast with Chong and Treisman (2003), allowed dorsal processing, resulting in a higher efficiencies for mean size estimations.

Summary and Conclusions

The aim of the three studies described in this thesis is to take a small step along the journey of understanding how our preattentive visual perception works. The first two studies give compelling evidence that observers are more efficient at making centroid estimations than mean estimations of size or luminance even when the centroid estimations require the processing of both location and either size or luminance. Efficiency values remained high (above .86), and almost identical when observers are asked to weight all the items equally or to weight items according to their size or luminance. In contrast, the efficiency of mean estimations was lower and influence functions for the mean judgments are vary considerably from the ideal for both features. Our results show how observers can efficiently use or ignore the guiding feature in the centroid context, but that they are not as efficient when estimating the mean of that feature.

In the third study we explored the distinctions between the ventral-dorsal streams as described by Goodale & Milner (1992), as a plausible explanation for the difference in performances between the mean-size and centroid task. We found that by making observers use size information to guide their movement for the response, they were able to obtain higher efficiencies. In addition to that, we found that observers are able to judge the location of the centroid in a secondary task.

The study of summary statistical representations has helped researchers understand how the human visual system uses and incorporates information that is outside of our awareness and that the summary statistical representation that underlie explicit reports – like mean estimates – may not reflect the full power and generality of summary statistical representations. A better understanding of how our visual system builds such

representations and when they are used, could increase our understanding of how the processing of a scene is guided by the task at hand, and how some features in that scene could be easily accessible in one context and not in other context. This knowledge could help us to understand the challenges facing patients with visual impairments and/or visual attention issues, but also could have applications in the visual systems for brain inspired machine designs, and marketing tools. Future research should combine this type of psychophysical research and fMRI data to further understand the implications of these findings.

References

Allik, J., Toom M., Raidvee, A., Averin, A., 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 in Cognitive Sciences, 15: 122–131.

Alvarez, G. A. & Oliva, A. (2008), The Representation of Simple Ensemble Visual Features Outside the Focus of Attention. Psychological Science. Vol 19 – Number 4.

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

Attarha, Moore & Vecera (2014). Summary Statistics of Size: Fixed Processing Capaicty for Multiple Ensembles but Unlimited Processing Capacity for Single Ensembles. Journal of Experimental Psychology

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

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

Christie, J., Hinchey, M.D. and Klein, R.M. (2013). Inhibition of return is at the midpoint of simultaneous cues, Psychonomic Society Inc.

Drew, S., Chubb, C. & Sperling, G. (2010). Precise attention filters for Weber contrast derived from centroid estimations. Journal of Vision, 1–16.

Fehd, H., Seiffert, A.E., (2008). Eye movement during multiple object tracking: Where do participants look? NIH Public Access, 108: 201-209.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B., (2014). Bayesian Data Analysis, Third Edition. Chapman & Hall.

Goodale, M.A, and Milner A.D., (1992). Separate Visual Pathways for perception and action. Trends Neuroscience.

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.

Lee, H., Beak, J., Chong, S.C. (2016). Perceived magnitude of visual displays: Area, numerosity, and mean size. Journal of Vision 16(3), 1-11

(Lu. Z.L. & Sperling, G. (2012) Black–white asymmetry in visual perception. Journal of Vision 2(10):8, 1–21

Marchant, A. P., Simons, D. J., & de Fockert, J. W., (2011). Ensemble representations: Effects of set size and item heterogeneity on average size perception. Acta Psychologica, 142: 245–250.

Melcher, D., Kowler, E., (1999). Shapes, surfaces and saccades. Vision Research 19: 2929-2946.

Myczek, K. and Simons, D. (2008). Better than average: Alternatives to statistical summary representations for rapid judgments of average size, Perception & Psychophysics, 70 (5), 772-788

Psychophysics Toolbox Version 3.0.8 (Brainard, 1997; Pelli, 1997; Kleiner et al, 2007) for MATLAB (Version 7.1)

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

Rouder, J.N., Speckman P.L., Sun D., Morey R.D., & Iverson G. (2009). Bayesian t-Tests for accepting and rejecting Null Hypothesis, Psychonomic Bulletin & Review, 16, 225-237.

Silva A. E. & Chubb, C. (2014). The 3-dimensional, the 4-channel, model of human visual sensitivity to grayscale scrambles. Vision Research 94-107

Solomon J.A., Morgan M., Chubb C. (2011). Efficiencies for the statistics of size discrimination. Journal of Vision 11(12):13, 1-11

Sun, P., Chubb, C., Wright, C.E, Sperling, G. (2016). The centroid paradigm: Quantifying feature-based attention in terms of attention filters. Attention Perception Psychophysics

Sun, P., Chubb, C., Wright, C.E, Sperling, G. (2018). High Capacity preconscious processing in concurrent grouping of colored dots. Proceedings of the National Academics of Science of the USA

Ungerleider, L.G., Mishkin, M., 1982. Two cortical visual systems, Analysis of Visual Behavior. MIT press

Vaziri-Pashkam, M. and Xu, Y. (2017), Goal-Directed Visual Processing Differentially Impacts Human Ventral and Dorsal Visual Representations. The Journal of Neuroscience.

W. K. Hastings (1970). “Monte Carlo sampling methods using Markov chains and their applications.” Biometrika 57(1), 97–109.