How can observers use perceived size? Centroid versus mean-size judgments

Laris M. Rodriguez-Cintron Department of Cognitive Sciences, University of California, Irvine, Irvine, CA, USA

Charles E. Wright Department of Cognitive Sciences, University of California, Irvine, Irvine, CA, USA

Charles Chubb Department of Cognitive Sciences, University of California, Irvine, Irvine, CA, USA

George Sperling Department of Cognitive Sciences, University of California, Irvine, Irvine, CA, USA

Summary statistical representations are aggregate properties of the environment that are presumed to be perceived automatically and preattentively. We investigated two tasks presumed to involve these representations: judgments of the centroid of a set of spatially arrayed items and judgments of the mean size of the items in the array. The question we ask is: When similar information is required for both tasks, do observers use it with equal postfilter efficiency (Sun, Chubb, Wright, & Sperling, 2016)? We find that, according to instructions, observers can either efficiently utilize item size in making centroid judgments or ignore it almost completely. Compared to centroid judgments, however, observers estimating mean size incorporate the size of individual items into the average with low efficiency.

Introduction

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 preattentively—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, and range, and the variance of features like size, motion, location, and orientation (Ariely, 2001; Chong & Treisman, 2003, 2005; Alvarez & Oliva, 2008; Alvarez, 2011; Marchant, Simons, & de Fockert, 2011; Robitaille & Harris, 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 article is to compare postfilter efficiency of size estimation in two tasks: the mean-size task and the centroid task (Drew, Chubb, & Sperling, 2010; Sun, Chubb, Wright, & Sperling, 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.

Much of the previous research on mean-size judgments has concluded that in 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 that experiment, set size was varied (four, eight, 12, or 16 items) and four distinct sizes were used within each set.

In follow-up work, Chong and Treisman (2003, 2005) varied the heterogeneity of the disk sizes, the

Citation: Rodriguez-Cintron, L. M., Wright, C. E., Chubb, C., & Sperling, G. (2019). How can observers use perceived size? Centroid versus mean-size judgments. *Journal of Vision*, *19*(3):3, 1–14, https://doi.org/10.1167/19.3.3.

https://doi.org/10.1167/19.3.3 Received July 18, 2018; published March 18, 2019 ISSN 1534-7362

presentation mode (sequential vs. simultaneous), and their numerosity and density. Across all these manipulations, observers achieved results of 75% accuracy with a difference in size between 6% and 8%. The fact that these discriminations were performed following relatively brief exposures (50–1,000 ms) and that increasing the size of the sample set did not affect performance led the researchers to conclude 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, 2013; Allik, Toom, Raidvee, Averin, & Kreegipuu, 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 & Simons, 2008; Im & Halberda, 2013; Allik et al., 2013), suggesting subsampling as a possible strategy. In a subsampling 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 the experiments of Ariely (2001) and Chong and Treisman (2003, 2005) and suggested that observers could be using subsampling as one of their strategies when making the mean-size discriminations. This interpretation assumes that all the errors in an 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 article, we present an experiment that compares performance for two summary statistical representations: centroid and mean size. We use a postfilter efficiency analysis as a common framework to compare performance across these two tasks. The procedure used to estimate postfilter 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. Most importantly, our interpretation of the postfilter efficiency analysis emphasizes the idea that failure to register stimulus items is only one source of error in these tasks.

Observers viewed sets of three or nine squares and were then asked, in different sessions, to perform one of three tasks: to estimate the centroid of the squares, ignoring variations in item size; to estimate the centroid of the squares, weighting items in proportion to their size; or to estimate the mean size of the squares. For the rest of this article, these three tasks will be referred as, respectively, the equi-weighted centroid task, the size-weighted centroid task, and 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).

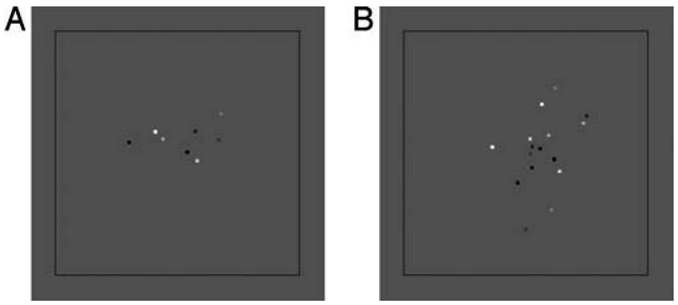


Figure 1. Displays presented in the centroid task. Participants were asked to determine the centroid of either all or a subset of these items determined by their luminance. (A) A sample set of eight dots. (B) A sample set of 16 dots. Reprinted from "Precise attention filters for Weber contrast derived from centroid estimations," by S. Drew, C. Chubb, & G. Sperling, 2010, *Journal of Vision, 10*(10):20. Copyright 2010 by ARVO. Reprinted with permission.

When deciding on stimuli to use in this experiment, we were concerned that observers, when presented with filled squares, could use mean luminance in estimating mean size. However, we were also concerned that outline squares might not 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 (Drew et al., 2010; Sun et al., 2016), observers could judge the centroid of a group of dots when asked to attend to all dots or while selecting stimuli with a specific feature, such as attending darker dots versus lighter dots (Figure 1). Drew et al. (2010) 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 and when they were asked to attend 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 estimating mean size may be a less efficient task. One of the simulations presented in that experiment showed that an ideal observer, attending to only two items out of a group of eight when estimating mean size, could still perform as well as the observers in

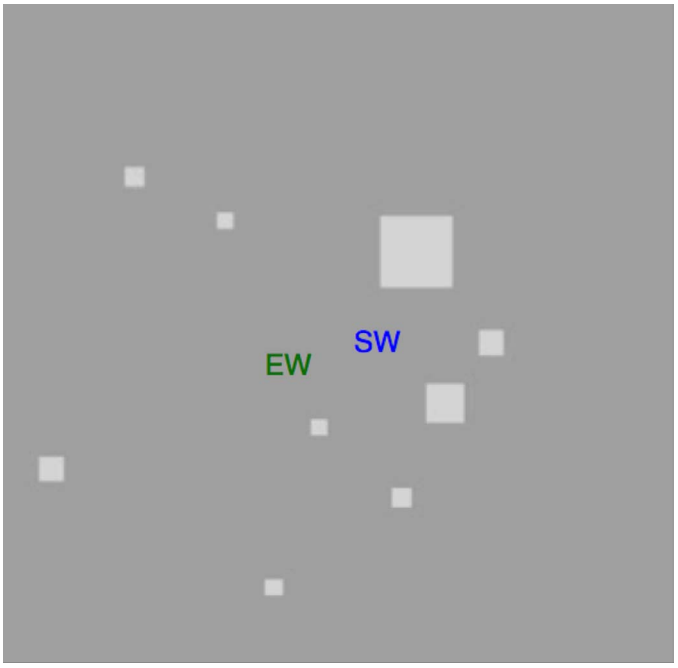


Figure 2. Example showing how the centroid response changes when participants are asked to give equal weight to all the items (equi-weighted task, EW) and to give more weight to larger items (size-weighted task, SW) in the centroid task for a set size of nine squares.

experiments reported by Ariely (2001) and Chong and 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 them 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 estimating 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 in the centroid task, we presented observers with a probe square whose initial size 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 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 sizes of the stimuli accurately and that they combine both types of information accurately when estimating the centroid. Figure 2 shows how the location of the centroid for a stimulus differs across these two tasks.

Methods

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 female and four were male, between the ages of 17 and 40 years. All observers reported having normal or corrected-to-normal vision. The 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 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; Kleiner et al., 2007) for MATLAB (Version 7.1).

Screenshots illustrating the two types of stimuli used in this experiment are shown in Figure 3. The size of the stimulus area was 500 × 500 pixels and the viewing angle was approximately 15°. The outlined squares (Figure 3A) were constructed using white (116 cd/m^2) lines 2 pixels wide; the interior of each square matched the gray background luminance (46 cd/m^2). The other stimuli (Figure 3B) were filled white squares (116 cd/m^2) on a gray (46 cd/m^2) background. The display was

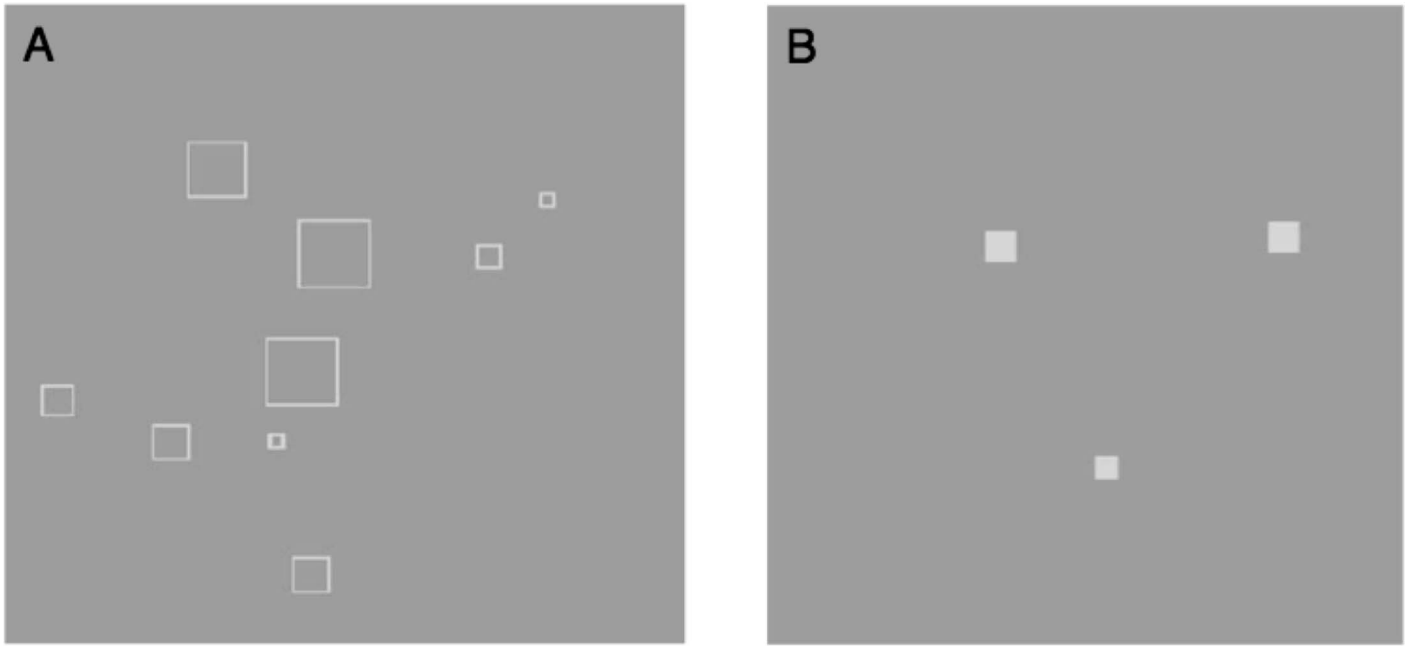


Figure 3. Two screenshots of the displays used in the experiment. (A) A set size of nine squares using outlined squares. (B) A set of three squares using filled squares.

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, respectively, 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 eight 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°) to each other. In addition, because the standard deviation of the distribution of the centroids would normally be reduced by $\sqrt{3}$ when going from three to nine stimuli, 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 4 shows the timeline of events for both the centroid and the mean-size tasks (using filled squares). The mask stimulus constructed for each trial consisted of a 10×10 jittered grid that filled the display area with a random sample of squares of sizes drawn from the triangular distribution used to generate the stimuli.

Procedure

The present study consisted of three tasks: 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; the size-weighted centroid task, in which they 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 the mean-size task, in which they 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 squares 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) and 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 in the following. In all tasks, the observer terminated the feedback screen by pressing any key.

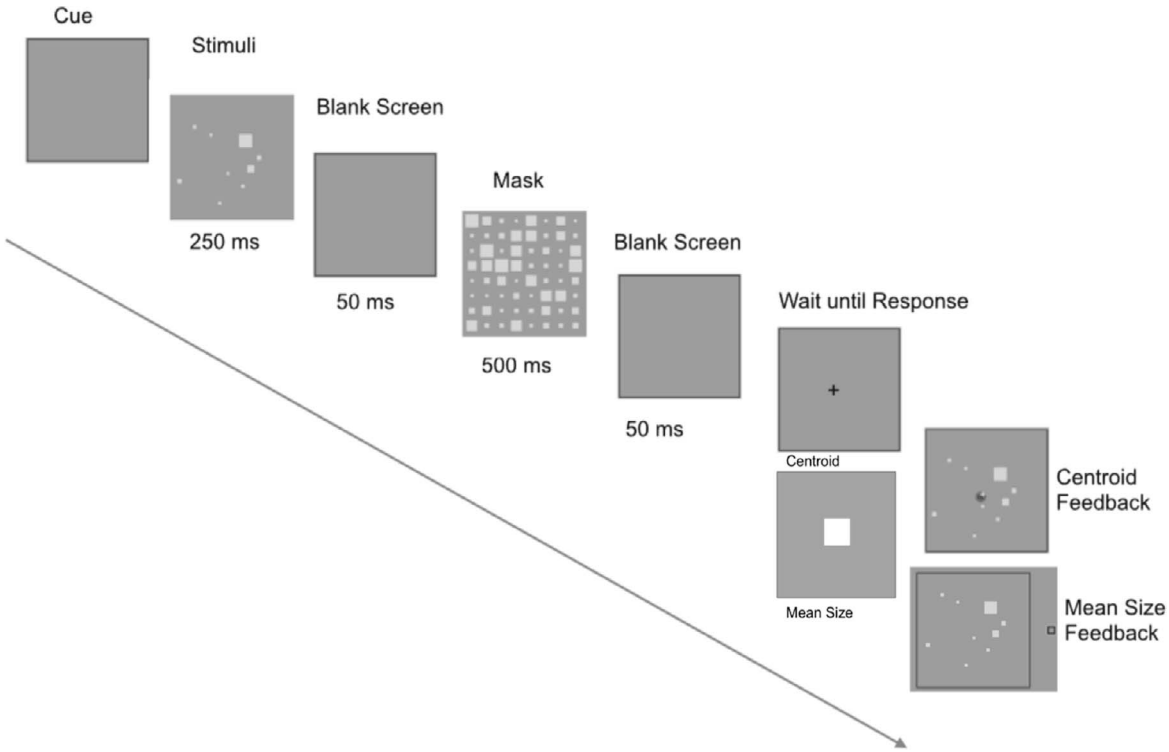


Figure 4. The timeline of a trial (from a nine-item condition using an example based on filled squares). The two final frames show two possible (1) response screens and (2) feedback screens, one for the mean-size task and one for both the equi-weighted and size-weighted centroid tasks.

Feedback and the response screens in the centroid tasks

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, and it also had a white cross that showed the location the observer chose as the centroid, as well as a black bull's-eye centered at the correct centroid location (Figure 5) depending on the weighting function.

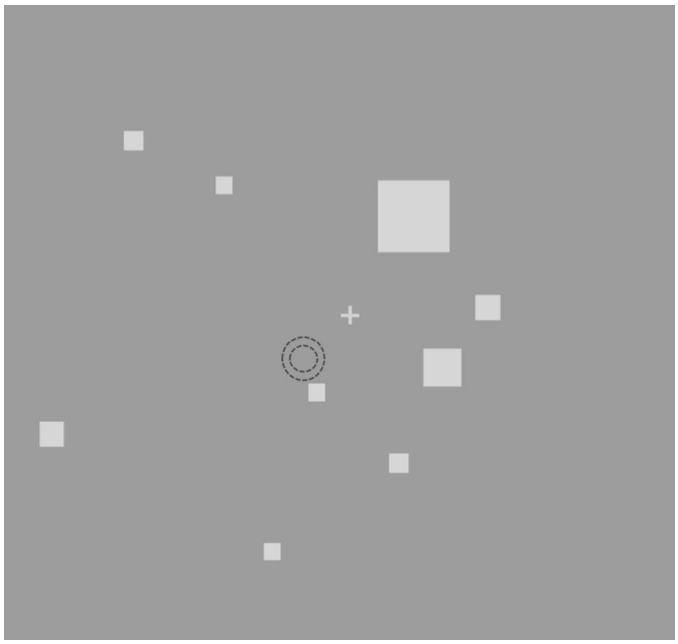


Figure 5. Schematic representation of the feedback screen for the equi-weighted centroid task, for a set size of nine (filled) squares. The dark gray bull's-eye represents the correct centroid and the cross shows the observer's response.

Feedback and the response screens in the mean-size task

The initial response screen in the mean-size task consisted of a probe square with a size randomly selected from 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

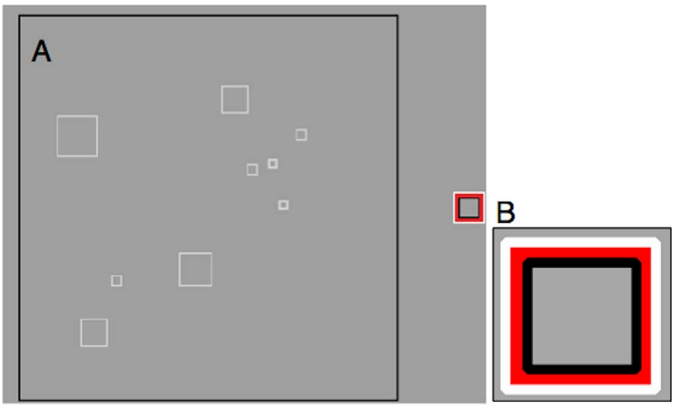


Figure 6. (A) Feedback screen for the mean-size task with a set of nine outlined squares. The black outline is the correct mean size; the red outline around the single square on the side represents the observer's error; the white outline is the observer's response. (B) A zoomed-in schematic representation of the feedback square in the mean-size task.

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 6. 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 6B).

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 the set size (1, 3, or 9 squares). A session consisted of two blocks—one per stimulus type—of the same task. Across sessions, the task was varied using a 3×3 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 nine sessions (three sessions per task). The order of filled versus outlined stimuli within a session was 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 in the following. A block consisted of 105 trials, five of which 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., response motor error). On the remaining trials, groups of three or nine squares were presented 50 times each. The order of the numerosity condition within a block was randomly determined.

Analysis

The data from all three tasks were analyzed using procedures similar to those described by Sun et al. (2016), with minor modifications for the data from the mean-size 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 eight square widths used in our stimuli) used by the observer when performing a task with a particular target filter φ. The three tasks in this experiment are based on two target filters. In the equi-weighted centroid task, the target filter φ gives equal weight to the squares of all eight widths w—i.e., $\varphi(w_i) = 1/8$, for all i from 1 to 8. In the size-weighted centroid task and the mean-size task, the target filter φ gives weight to each square equal to its size:

$$\varphi(w_i) = w_i \Big/ \sum_i w_i.$$

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

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

where the sum is over all squares i in the display, w_i is the width of 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

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

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 $f_\varphi(w)$ we have

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

In Equation 3, $w_{t,i}$, $x_{t,i}$, and $y_{t,i}$ are the width and x- and y-coordinates of the ith 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-size 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

$$\mu_t = \frac{1}{N}\sum_i f_\varphi(w_{t,i}), \quad (4)$$

where N is the number of squares in the display (either three or nine, 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 Equations 2 and 3 is

$$\Lambda(f_\varphi, \sigma) = \prod_t \frac{1}{2\pi\sigma^2} \times \exp\left[\frac{-(R_{t,x} - \mu_{t,x})^2 - (R_{t,y} - \mu_{t,y})^2}{2\sigma^2}\right], \quad (5)$$

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

$$\Lambda(f_\varphi, \sigma) = \prod_t \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-(R_t - \mu_t)^2}{2\sigma^2}\right]. \quad (6)$$

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

In any Markov-chain Monte Carlo 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 eight 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 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 40th was retained.

A key measure that we have adapted from Sun et al. (2016) to characterize the results of this experiment is postfilter efficiency. Postfilter 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 and mean-size tasks. Sun et al. developed this measure for centroid data but simply called it *efficiency*. Here we use *postfilter efficiency* to emphasize that this 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_φ rather than the target filter φ. The value of postfilter 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.

Postfilter efficiency estimates were obtained using the fminsearch univariate optimization function in MATLAB. To evaluate a proposed value of postfilter efficiency, 100 decimations of the stimulus cloud used on each trial were generated. For every decimation, each square in the cloud had a probability equal to the postfilter 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 was combined across decimations and trials to guide the optimization process.

Figure 7 illustrates how the efficiency analysis works. Figure 7A shows a nine-item stimulus in the equi-weighted centroid task. The bull's-eye indicates the target centroid. To get a sense of how the efficiency calculation works, consider Figure 7B, which shows an example in which an ideal observer, processing this display with an efficiency of 0.89, has based the centroid estimate on a random subset of eight from the nine 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 eight tokens; this is simply the expected number of items processed, since $8 = 9 \times 0.89$. However, since it is the probability that an item is decimated that is fixed, sometimes the simulated ideal observer would be expected to process eight or

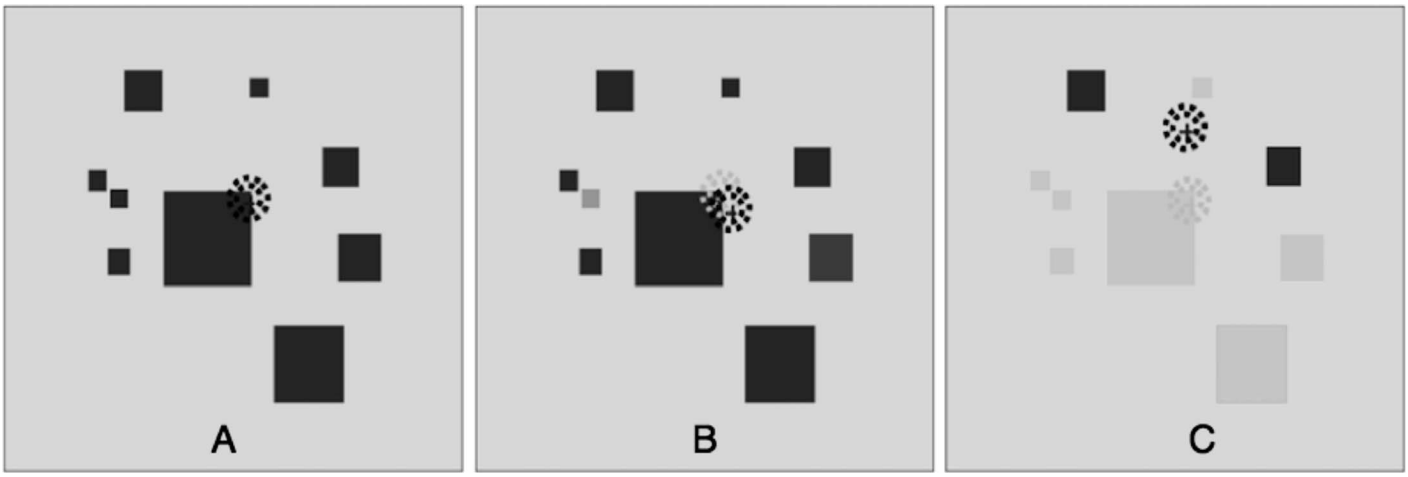


Figure 7. (A) An equi-weighted centroid estimation with an efficiency value of 1. (B) A typical equi-weighted centroid estimation with an efficiency value of 0.89. Note that the 0.89 efficiency indicates that at a minimum eight items out of the nine are included in the estimation. The observer's response shown by the black bull's-eye is still close to the correct response, which is shown by the gray bull's-eye. (C) A typical response when the efficiency is 0.22. In this example, the observer's centroid estimate (black bull's-eye) is far from the correct response (gray bull's-eye).

even all nine tokens, and sometimes fewer than eight. As shown in Figure 7C, an ideal observer operating with an 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 already described, the actual estimation was averaged over 100 subsets for the stimulus cloud used on each trial.

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.

Results

All observers ran nine sessions—three 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 six sessions, and for most observers it was stable and constant for all nine sessions. For all observers, only the data from the last six sessions are reported here.

We expected experts to be better than unpracticed observers at least in the centroid task. The actual difference was small; the observed efficiencies were 0.88 and 0.85, respectively, $\Delta = 0.03$, $SD = 0.09$, $t(6) = 0.523$, $p = 0.62$, Bayes factor BF = 0.764.[1] The main effect of stimulus type was negligible. Because 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 preplanned contrasts for the task factor: one comparing the results in the equi-weighted and size-weighted centroid tasks, and one comparing the results of the size-weighted centroid task and the mean-size task.

Postfilter efficiency

Observers achieved higher, and almost identical, postfilter efficiencies in the two centroid tasks, and lower efficiencies in the mean-size task (Figure 8). 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, $\Delta = 0.01$, $SD = 0.02$, $t(7) = 1.460$, $p = 0.188$, BF = 0.74. The preplanned contrast comparing the postfilter efficiency for the size-weighted centroid task with that for the mean-size task very strongly suggests that observers were able to use size

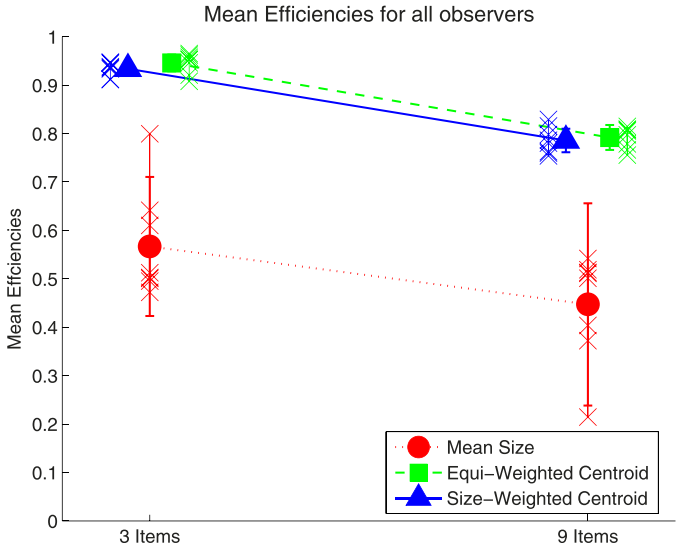


Figure 8. Mean efficiencies for all eight observers as a function of set size for equi- and size-weighted centroid judgments and for mean-size judgments. The filled plotting symbols represent the mean across observers in each condition; the Xs are the efficiencies for each observer. The error bars display the 95% confidence intervals for the averages.

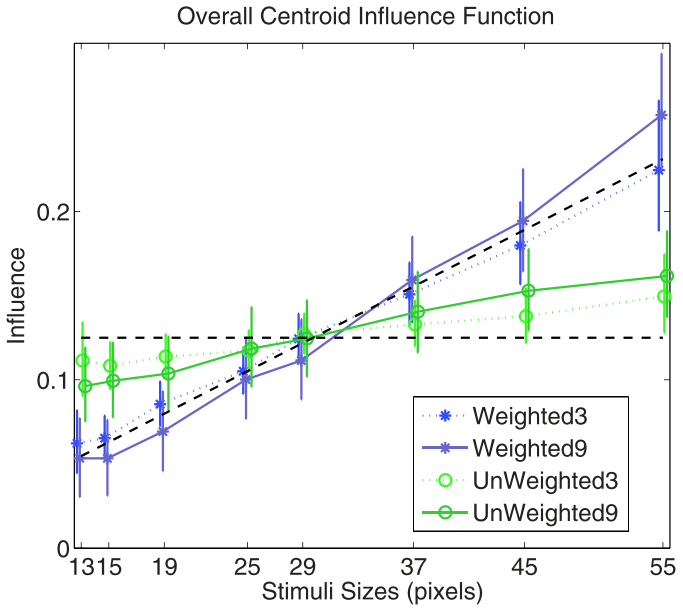


Figure 9. Influence as a function of stimulus size for the two centroid tasks. These plots are averaged across observers and stimulus types. The intervals are 95% confidence intervals based on the variation across observers. The black dashed lines represent the ideal influence: slope = 0 for the equi-weighted task and 1 for the size-weighted task.

more effectively when estimating the centroid of a group of squares than when estimating the mean size of the same group, $\Delta = 0.35$, $SD = 0.15$, $t(7) = 6.485$, $p < 0.001$, BF = 45.9.

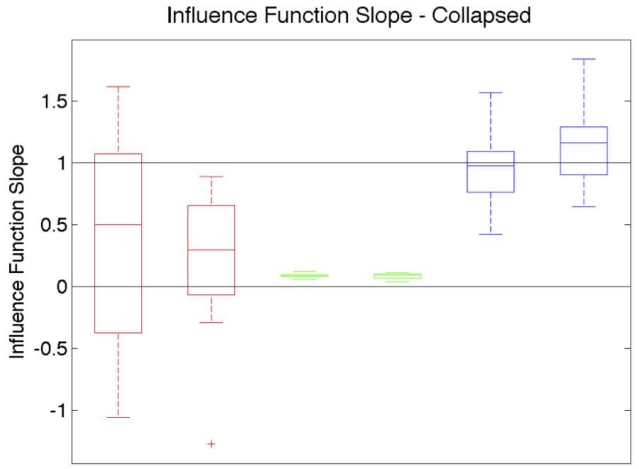
Figure 8 also shows that observers achieved higher efficiencies when presented with three items than when presented with nine items. A t test provided evidence for a reduction of postfilter efficiency with increased numerosity (Figure 8) for all three tasks, $\Delta = -0.14$, $SD = 0.08$, $t(7) = -4.810$, $p = 0.002$, BF = 24.06.

No interactions were found between stimulus type and numerosity, stimulus type and task, numerosity and task, or stimulus type, task, and numerosity. The biggest t value associated with any of these interactions was 1.42, with a p value of 0.198 and a BF of 0.71.

Influence functions

Figure 9 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 1. The average data follow this ideal closely. With nine squares, observers tended to overweight the larger squares and underweight the smaller squares relative to this ideal, but with three squares they produced the opposite pattern. In 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 0. Although the resulting influence functions are flatter than those for the size-weighted centroid task, 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 postfilter efficiency values of 0.5. The wide confidence intervals obtained for this task make these data hard to interpret.

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, we used linear regression to estimate the slope of the influence function in each condition. Figure 10 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 9 for the equi-weighted and size-weighted centroid tasks. As shown by the summary at the bottom of Figure 10, the slope estimates for the equi-weighted centroid task are close to 0 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 0, but the confidence intervals show how close to 0 it is—slope = 0.085, 95% confidence interval (CI) [0.068, 0.10], $t(7) = 11.98$, $p = 0.0000$, BF = 141.169. There is substantially more variability across observers in the slope estimates for



	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

Collapsed by Numerosity

	H0=1	H0=0	H0=1
Difference	0.200	0.010	0.190
CI at 95%	[-.35 .89]	[.07 .1]	[.76 1.3]
SD	0.740	0.020	0.340
t(7)	-2.790	-128.400	0.430
p-value	0.030	0.000	0.680
BF	3.980	277.400	0.485

Figure 10. 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 (filled/outlined squares) 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 hypotheses 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.

the size-weighted centroid task. Despite this variability, there is a reliable numerosity effect. However, as shown in Figure 10, the slope is not distinguishable from the expected slope of 1 for numerosity 3 or 9, and this result still holds if the estimates for numerosities 3 and 9 are averaged—slope = 1.055, 95% CI [0.76, 1.33], $t(7) = 0.43$, $p = 0.68$, 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 10 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 more than 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 them. Second, averaged across numerosity, the slope in the mean-size task differs reliably from 1, the expected value in this task—slope = 0.270, 95% CI [−0.35, 0.89], $t(7) = -2.790$, $p = 0.030$, BF = 3.980. Further, averaging across numerosity in both cases, the slope in the mean-size task differs reliably from that in the size-weighted centroid task, $\Delta = -0.78$, $SD = 0.71$, $t(7) = -3.113$, $p = 0.017$, BF = 4.32.

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 and Simons (2008) suggested that the estimate of the mean size of a group of items is obtained with low postfilter efficiency. Also, previous research from our lab (Drew et al., 2010; Sun et al., 2016) found that equi-weighted centroids could be estimated with high postfilter efficiency. The surprising result is that locating the centroid while weighting items in proportion to their size can also be done with high postfilter efficiency. This is surprising because one might expect that postfilter efficiency the in size-weighted centroid task would be no better than the lesser of the two obtained in the mean-size task and the equi-weighted centroid task. Our results show that observers achieved almost identical high efficiencies in the two centroid tasks and that the postfilter efficiency was much lower in the mean-size task. However, this implies, counterintuitively, 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 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.

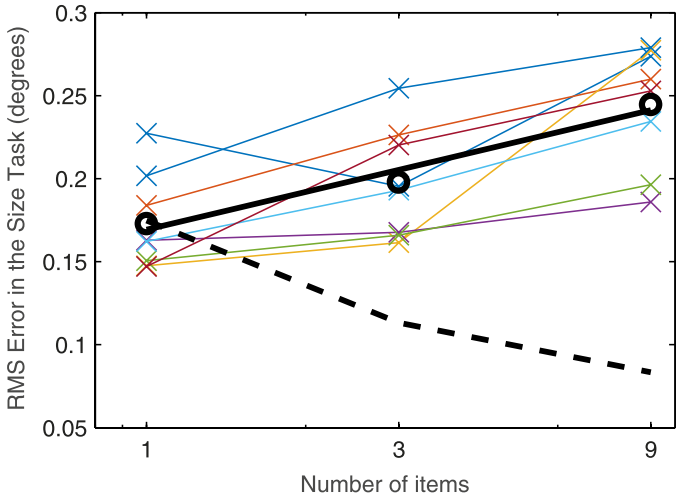


Figure 11. Root mean square error in the mean-size task as a function of number of items (on a log scale). The X plotting symbols connected by colored lines are the data from each of the eight observers. The black circles are the means over observers. The solid black line is the best-fitting linear function. The dashed line represents the results of a simulation that assumes that all of the error in the mean-size task is due to misperception of item sizes. This simulation was run separately for each observer. Based on a linear function fitted to the singleton data for that observer relating the standard deviation of the estimated size to the actual size, for each trial, 50 separate samples of the perceived size of each item were drawn and averaged to construct mean-size estimates. Root mean square error was then computed by comparing the estimate for each sample to the true mean size, pooling across samples and trials within the three- and nine-item conditions. The displayed results are averaged over observers.

Figure 11 shows the 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 reflects 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 three- and nine-item conditions are a "raw" version of the data used as the basis for the postfilter 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 postfilter efficiency analysis. Given that the stimulus items ranged in size from 0.22° to 0.99°, the standard deviation of 0.17°, 95% CI [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.

Because the spacing on the abscissa of Figure 11 is logarithmic, it appears that the mean-size error increases linearly with the logarithm of the numbers of items—slope = 0.033, 95% CI [0.020, 0.046], $t(7) = 5.927$, $p = 0.001$, 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 three- and nine-item conditions. We reach this conclusion because the mean-size error due to misperception of item sizes would be expected to decrease as 1 over the square root of the number of observations (items). Under the extreme assumption that all singleton error is due to size misperception, the dashed black line shows the predicted RMSE. Another possibility is that rather than being due to size misperception, the error in the mean-size task 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 11 should be flat. However, neither size misperception errors, late errors, nor some combination of the two predict the observed increase in the RMSE with an increasing number of items. 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 underestimated 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 that there may have been a slight bias—in this case, a tendency for observers to underestimate the true item size, –0.042°, 95% CI [–0.092°, 0.008°], $t(7) = -1.979$, $p = 0.088$, 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°, 95% CI [0.031°, 0.094°], $t(7) = 4.691$, $p = 0.002$, 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, 95% CI [0.072, 0.140], $t(7) = 7.300$, $p = 0.00016$, 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° × 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 ($r = 0.86$). That correlation, along with the previous comparison 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 postfilter efficiency analysis (see the description in the Analysis subsection under Methods) to allow for the perturbation of item sizes. For each observer, the analysis of the data from the size-weighted centroid task used the estimated postfilter efficiency from 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 0 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 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 some of the additional error is introduced by the process of forming a size-weighted centroid.

For the size-weighted centroid of three items, this elaboration of our postfilter efficiency analysis estimated the additive component of the size misperception error to be 0.053°, 95% CI [0.033°, 0.074°], $t(7) = 6.085$, $p = 0.0005$, BF = 72.78; for nine items it was 0.044°, 95% CI [0.033°, 0.055°], $t(7) = 9.366$, $p = 0.0000$, BF = 692.5. Because there is only weak evidence for a difference between these estimates, $\Delta = 0.010°$, 95% CI [–0.006°, 0.026°], $t(7) = 1.490$, $p = 0.180$, BF = 1.314, we will consider their average, 0.049°, 95% CI [0.034°, 0.063°], $t(7) = 7.837$, $p = 0.0001$, BF = 265.4. The slope relating the size misperception error to item size for the three-item task was 0.050, 95% CI [0.002, 0.099], $t(7) = 2.445$, $p = 0.044$, BF = 2.073; for nine items it was 0.038, 95% CI [–0.010, 0.085], $t(7) = 1.859$, $p = 0.105$, BF = 1.102. Because there is only weak evidence for a difference between these estimates, $\Delta = 0.013$, 95% CI [–0.069, 0.094], $t(7) = 0.369$, $p = 0.723$, BF = 0.356, we will consider their average, 0.044, 95% CI [0.018, 0.070], $t(7) = 4.045$, $p = 0.005$, 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 previously for the singleton trials in the mean-size task—0.049° versus 0.062°, $\Delta = 0.014°$, 95% CI [–0.025°, 0.053°], $t(7) = 0.850$, $p = 0.423$, BF = 0.451—but the slope of the multiplicative component is substantially smaller: 0.049 versus 0.106, $\Delta = 0.057$, 95% CI [0.022, 0.093], $t(7) = 3.804$, $p = 0.007$, BF = 8.975. We interpret this as evidence that the information about the 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 nine items can be perceived accurately and incorporated effectively into a centroid judgment, why are the mean-size judgments so inefficient? The foregoing analysis 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 11 suggest 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 and Treisman (2003) have found reduced thresholds when asking observers 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 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, in the size-weighted centroid task they could be using relative sizes; it is possible to perform this task perfectly well with size information that preserves only 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 previously 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 observers are actually using the sizes of the squares to make their judgments and are 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 set size. These correlations, averaged over stimulus type and set size, are summarized in Table 1. There was a strong, positive correlation of the postfilter efficiency estimates both within (i.e., across the variants) and across the two centroid tasks, suggesting that the differences in postfilter efficiency across observers in these tasks reflect 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 postfilter efficiency differences across observers and the variants of the mean-size task (ranging from 0.2 to almost 0.9), these correlations close to 0 suggest two separate conclusions. First, the postfilter efficiency variations across observers in the mean-size task derive from a different source than those in the size-weighted centroid task. Even more troubling for the use of the mean-size task to estimate of the amount of size information available to an observer 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.

Postfilter efficiency correlations (average for all observers)	
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 1. A summary of the correlations of the efficiencies between the three tasks. These efficiencies are averaged for all eight observers and collapsed by numerosity and stimulus type.

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 the 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 (Melcher & Kowler, 1999; Fehd & Seiffert, 2008). Christie, Hinchey, and Klein (2013) suggest that inhibition of return 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 inhibition of return for the center of gravity than for the actual stimuli. The researchers 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.

Keywords: feature-based attention, centroid task, mean-size judgments, summary statistical representations, visual attention, efficiency

Acknowledgments

We would like to thank Sang Chul Chong, Joshua Solomon, and an anonymous reviewer for their useful comments on an earlier draft of this article. We would also like to thank the members of the Chubb–Wright Lab at the University of California, Irvine, for their participation in this experiment and for their comments during the analysis process.

Commercial relationships: none.
Corresponding author: Charles E. Wright.
Email: cewright@uci.edu.
Address: Department of Cognitive Sciences, University of California, Irvine, Irvine, CA, USA.

Footnote

[1] Bayes factor computed using the calculator at http://pcl.missouri.edu/bayesfactor (Rouder, Speckman, Sun, Morey, & Iverson, 2009).

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*, *19*(4), 392–398.

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

Brainard, D. H. (1977). The Psychophysics Toolbox. *Spatial Vision*, *10*, 433–436.

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., & Klein, R. M. (2013). Inhibition of return is at the midpoint of simultaneous cues. *Attention, Perception, & Psychophysics*, *75*, 1610–1618.

Drew, S., Chubb, C., & Sperling, G. (2010). Precise attention filters for Weber contrast derived from centroid estimations. *Journal of Vision*, *10*(10):20, 1–16, https://doi.org/10.1167/10.10.20. [PubMed] [Article]

Fehd, H., & Seiffert, A. E. (2008). Eye movement during multiple object tracking: Where do participants look? *Cognition*, *108*, 201–209.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: CRC Press.

Goodale, M. A., & Milner A. D. (1992). Separate visual pathways for perception and action. Trends Neuroscience, *15*(1): 20–25.

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

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.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in psychtoolbox-3. *Perception*, *36*(14), 1–16.

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

Robitaille, N., & Harris, I. M. (2011). When more is less: Extraction of summary statistics benefits from larger sets. *Journal of Vision*, *11*(12):18, 1–16, https://doi.org/10.1167/11.12.18. [PubMed] [Article]

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.

Solomon, J. A., Morgan, M., & Chubb, C. (2011). Efficiencies for the statistics of size discrimination, *Journal of Vision*, *11*(12): 13, 1–11, https://doi.org/10.1167/11.12.13. [PubMed] [Article]

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*, *78*, 474–515, https://doi.org/10.3758/s13414-015-0978-2.