Not all summary statistics are made equal: Evidence from extracting summaries across time

Bjorn Hubert-Wallander Department of Psychology, University of Washington, Seattle, WA, USA

Geoffrey M. Boynton Department of Psychology, University of Washington, Seattle, WA, USA

Over the past 15 years, a number of behavioral studies have shown that the human visual system can extract the average value of a set of items along a variety of feature dimensions, often with great facility and accuracy. These efficient representations of sets of items are commonly referred to as summary representations, but very little is known about whether their computation constitutes a single unitary process or if it involves different mechanisms in different domains. Here, we asked participants to report the average value of a set of items presented serially over time in four different feature dimensions. We then measured the contribution of different parts of the information stream to the reported summaries. We found that this temporal weighting profile differs greatly across domains. Specifically, summaries of mean object location (Experiment 1) were influenced approximately 2.5 times more by earlier items than by later items. Summaries of mean object size (Experiment 1), mean facial expression (Experiment 2), and mean motion direction (Experiment 3), however, were more influenced by later items. These primacy and recency effects show that summary representations computed across time do not incorporate all items equally. Furthermore, our results support the hypothesis that summary representations operate differently in different feature domains, and may be subserved by distinct mechanisms.

Introduction

The human visual system is constantly confronted with a large, complex, and dynamic stream of information that far outstrips its processing capacity. One way the visual system is thought to deal with this problem is through the use of summary representations (also referred to as ensemble representations, summary statistics, or set representations), wherein a central tendency is extracted from a set of stimuli that vary along one or more feature dimensions. Recent behavioral studies show that participants can accurately report summary representations for mean size (Ariely, 2001; Chong & Treisman, 2003, 2005a, 2005b), mean orientation (Dakin, 2001; Parkes, Lund, Angelucci, Solomon, & Morgan, 2001; Robitaille & Harris, 2011), mean position (Alvarez & Oliva, 2008; Greenwood, Bex, & Dakin, 2009; Spencer, 1961, 1963), mean color of a group of objects (de Gardelle & Summerfield, 2011), and even the mean expression or identity contained in a set of faces (de Fockert & Wolfenstein, 2009; Haberman, Harp, & Whitney, 2009; Haberman & Whitney, 2007, 2009). Given the accuracy, efficiency, and automaticity with which they appear to be computed, some have speculated that summary representations are an important contributor to our subjective impression of a complete visual world, since they could potentially provide a rough sketch of areas or objects we're not currently focusing on (Whitney, Haberman, & Sweeny, 2013).

Extracting summaries across time

Most studies of summary representations have used static arrays, where all the samples that participants are expected to summarize are presented concurrently. However, real world visual input is inherently dynamic, and the properties of objects that we wish to know about often change across time. For example, the expression on a friend's face evolves as he or she listens to what we say. If the physical object is unchanging, the retinal input may still change, as the visual angle of the retinal image changes with depth and the color and brightness of objects shift with lighting or viewpoint changes. Finally, even for a static visual stimulus,

Citation: Hubert-Wallander, B., & Boynton, G. M. (2015). Not all summary statistics are made equal: Evidence from extracting summaries across time. *Journal of Vision*, *15*(4):5, 1–12, doi:10.1167/15.4.5.

doi: 10.1167/15.4.5 Received November 13, 2014; published June 8, 2015 ISSN 1534-7362

information is effectively sampled serially in time through shifts in visual attention and eye position.

While several studies have shown that it is *possible* to extract a summary representation from stimuli presented across time, (Albrecht & Scholl, 2010; Albrecht, Scholl, & Chun, 2012; Corbett & Oriet, 2011; Haberman et al., 2009; Piazza, Sweeny, Wessel, Silver, & Whitney, 2013), very little is known about *how* the summary is built in this case. How are the individual stimuli incorporated into a summary? Work by Juni and colleagues (Juni, Gureckis, & Maloney, 2012) has demonstrated that summary-like judgments are sensitive to manipulations of how informative early and late items are, with more reliable items contributing more to subjects' judgments than less reliable ones. But it still remains to be seen how a summary is constructed from a set presented over time in the default case where all items carry equal information. Do all items or parts of the information stream contribute equally to the perceived mean, or do early items or late items contribute more heavily?

Unequal weighting of information over time, here referred to as primacy and recency, might be consequences of underlying neuronal processes or mechanisms, or might reflect optimal behavior for a particular task. For example, if making a decision rapidly is important and the first few items provide sufficient information for the task, one might expect primacy. The costs of time and cognitive resources to incorporate later items might outweigh any additional accuracy they might contribute. On the other hand, one might expect recency if an accurate representation of the most recent state of the world is desired, or if early information is lost due to memory or attentional limitations.

Extracting summaries across feature domains

Summary representation has been invoked to describe behavior across a wide variety of visual features, but surprisingly little is known about how summarization mechanisms might or might not differ across those domains. Is summary extraction a general cognitive mechanism that operates similarly across stimulus domains, or does it depend on the feature domain of interest?

Some studies have attempted to address this question by comparing various properties of summaries across different feature domains, but reports on these properties are generally few, unclear, or conflicted. For example, some researchers have compared the accuracy of summary representations in two or more domains (Albrecht et al., 2012; Emmanouil & Treisman, 2008), but using accuracy measures may not be optimal since, as Albrecht et al. note, it is likely highly dependent on the statistical properties of the particular stimuli used. One can also compare domains by examining how summary accuracy varies with the number of items present, but even within a domain accuracy sometimes increases with set size (Ariely, 2001; Chong & Treisman, 2003, 2005b; Parkes et al., 2001) and sometimes does not (Robitaille & Harris, 2011; Solomon, Morgan, & Chubb, 2011). Finally, one could compare domains by examining what portion of items in a set are incorporated into a summary. However, not enough is known about this property to understand whether it differs across domains. Most researchers are only able to conclude that more than two but fewer than all items present are used to summarize (Dakin, 2001; Morgan & Glennerster, 1991; Piazza et al., 2013; Solomon, 2010; Solomon et al., 2011; Watamaniuk & Duchon, 1992). Comparing how summaries are computed across time in different feature domains may help to address this unresolved issue.

The present study

Here, we explored both how summary statistics are constructed when stimuli are presented serially across time and whether such summary computation differs across feature domains. We found that not all items contribute equally to summary representations built across time, with different temporal weighting profiles appearing in different feature domains. In particular, judgments of mean object position appeared special, apparently operating differently from those of mean object size, mean facial expression, and mean motion direction. These differences in how information is used over time to compute a summary representation imply that different mechanisms are associated with different feature domains.

Experiment 1: Averaging position and size across time

To understand how summary representations are constructed when the items are presented serially across time, we presented participants with a sequence of small white dots, asked them to report either the mean size or the mean position of the group. We then quantified the influence of the temporal position of each dot on the participants' estimates across many trials.

Method

Participants

Twenty-five students at the University of Washington were recruited from the Department of Psychology's

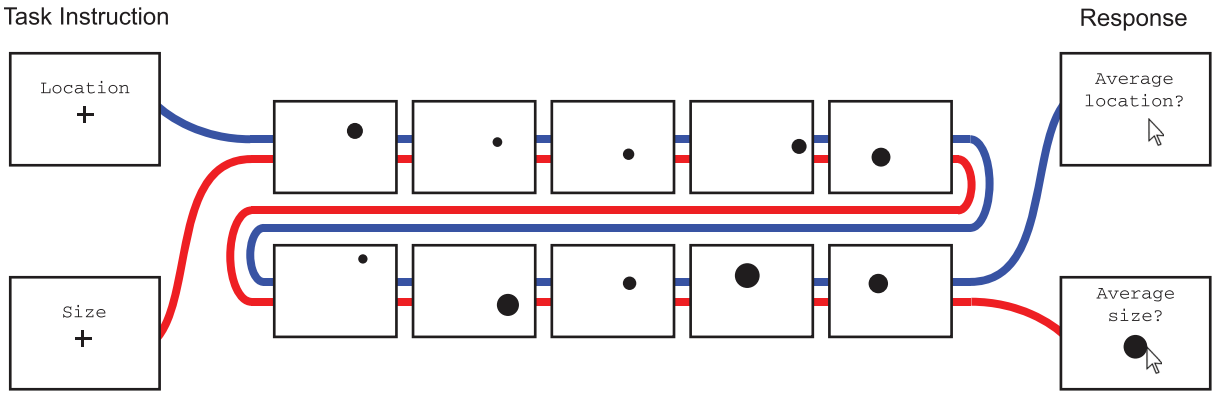

Figure 1. Trial schematic from Experiment 1. After task instruction period (1000–2000 ms), ten dots were presented over 2000 ms, with a 50 ms blank after each dot, followed by the response period, which lasted until participant response. On location trials, participants reported the "average position" or "center" of the dots shown by clicking on a point on the screen. On size trials, participants reported the "average size" of the dots shown by adjusting a test dot. Dots varied in both position and size on all trials. Trial type was blocked.

undergraduate participant pool, where students may volunteer as participants for studies in exchange for course credit. This number of participants to recruit was decided upon after initial simulations conducted based on pilot testing showed that approximately 20 participants would result in relatively stable estimations of group-level weights (see Results section) in both tasks. All participants had normal or corrected-to-normal vision. Recruitment and study procedures in all experiments presented here were conducted in accordance with the ethical policies set forth by the University of Washington's Human Subjects Division, and those in the Declaration of Helsinki.

Apparatus

All study procedures took place in a dimly lit room, with the participants seated 50 cm from a CRT monitor subtending 40.4° × 30.8° of visual angle. A chinrest was used to ensure constant viewing distance to the monitor, on which all stimuli were shown against a black background. A faint grey grid pattern was always present as part of the background in order to provide spatial reference. All stimuli were generated by custom software written using the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997) for MATLAB (The Mathworks, Natick, MA).

Stimuli and procedure

As shown in Figure 1, participants were precued at the beginning of each trial with the word "Location" or "Size" presented slightly above a white central cross approximately 0.8° in width and height. Despite the presence of the central cross, no instructions about fixation were given and participants were free to look wherever they wished over the course of the experiment. This word precue was present for a random period of time between one and two seconds and indicated what the participant would be asked to report at the end of the trial.

Each trial consisted of a series of ten white dots that varied in their position and size. Each dot was shown for 150 ms and was followed by a 50 ms blank inter-dot interval, resulting in a dot presentation rate of 5 Hz (Figure 1). On a given trial, dot locations were chosen by sampling ten times from a bivariate Gaussian probability distribution with a standard deviation of 2.3°. The center of the bivariate Gaussian distribution was drawn randomly on each trial from a square-shaped uniform probability distribution centered on the middle of the screen and subtending 21.1° × 21.1°, corresponding to 70% of the vertical height of the screen. If a dot's sampled location was outside the borders of the screen, it was moved to the point on the screen nearest to its originally sampled location. Dot radii were similarly sampled from a Gaussian distribution, the center of which was sampled on each trial from a uniform distribution ranging from 0.3° to 1.5°. The standard deviation of the Gaussian distribution was always exactly 0.3 times the center of the same distribution in order to minimize the possibility of sampling dot radii below zero. If a dot's radius was sampled to be below zero, it was resampled until it was above zero. The series of ten dots was followed by a blank period of 300 ms, followed by a response period.

During the response period, participants were reminded what to report by text appearing near the middle of the screen. On location trials participants reported the "average location" or "center" of the dots seen on that trial by moving the mouse cursor and clicking on their perceived center. The mouse cursor

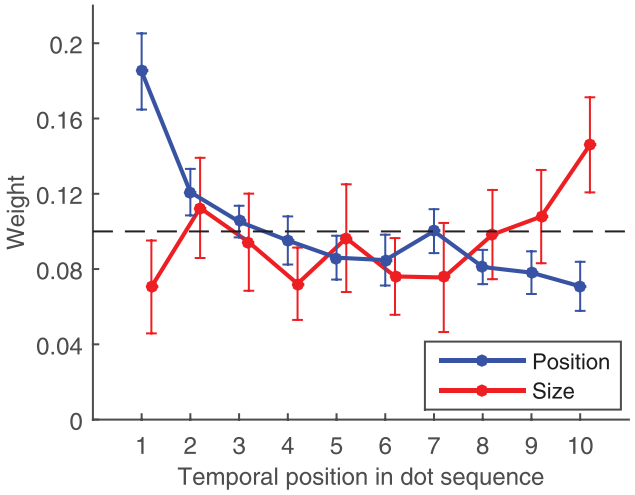

Figure 2. Results from Experiment 1. Mean weights as a function of temporal position across all participants. X-axis indicates the temporal order of the dots shown, where 1 refers to the first dot presented on each trial and 10 refers to the last dot presented on each trial. Y-axis indicates the relative influence of each dot on participants' responses in each task. These weights were obtained by fitting a weighted average model to the data (see Results text for details). Dashed line indicates the weights expected if all dots contributed equally to responses and no other source of noise or bias were present. Weights from the size trials have been displaced rightward for readability. Error bars indicate 95% confidence intervals across participants.

was visible to the participants only during the response period of location trials. On size trials participants reported the "average size" of the dots seen on that trial by adjusting the size of a centrally-presented test dot, the radius of which varied with the horizontal location of the mouse. Participants clicked to submit their response when the perceived mean size was obtained. The test dot's initial radius was chosen randomly on each trial from 0.1° to 3.2°, which also served as the limits of possible responses. The response period ended when the participant submitted his or her response, and was followed by a 1500 ms inter-trial interval.

Each participant received full instructions from an experimenter and then completed approximately ten practice trials in view of the experimenter before beginning the experimental trials. Each participant completed 320 experimental trials in blocks of 40 trials. Blocks alternated between all location trials and all size trials, with the first block type seen counterbalanced across participants. The participants were free to take breaks after each block, but could also do so at any point during the experiment by simply waiting to submit their response for a given trial. A full experimental session lasted about an hour.

Results

Weights that quantified the relative influence of each dot number (one through ten) on participants' responses were obtained by fitting a weighted average model (as used in Juni et al., 2012) to each participant's data separately for size and location trials. The model took the form

$$R_j = \sum_{i=1}^{10} w_i x_{ij} \quad (1)$$

where R_j is the participant's response for trial j, x_{ij} is the position or radius of the dot at temporal position i and trial j, and w_i is the weight for temporal position i. Linear regression was used to obtain the least-squares best fitting set of weights for each participant for each task. The model fit the data well in all participants in both location (model R^2: mean = 0.98, $SD = 0.02$; all $p < 0.001$) and size (model R^2: mean = 0.74, $SD = 0.10$; all $p < 0.001$) trials. Mean weights across participants for both trial types in Experiment 1 are shown in Figure 2.

The dashed line in Figure 2 shows the weight that would be obtained for each dot if all dots contributed equally to participants' responses and no other source of noise or bias were present. It is clear that participants did not weigh each of the ten dots evenly in either feature domain. Instead, we found *primacy* for participants' mean position judgments, with dots that appeared early in the sequence contributing more to the perceived mean position than later dots. The pattern was reversed for judgments of mean dot size, where participants showed clear *recency*; later dots contributed more to the perceived mean size. The size of the effect was considerable for both feature domains, where the most-weighted dots contributed approximately two to three times as much as the least-weighted dots. Additionally, the effect was relatively smooth in both cases, with weights gradually increasing or decreasing as the dot number increased, though pooling across many trials may have averaged out more discrete effects. A two-way ANOVA on the weight data showed a significant dot number x trial type interaction, $F(9,216) = 10.40$, $p < 0.001$, $\eta_p^2 = .30$, and one-way ANOVAs performed separately on location, $F(9,216) = 23.15$, $p < 0.001$, $\eta_p^2 = .49$, and size trials, $F(9,216) = 3.10$, $p = 0.002$, $\eta_p^2 = .11$, showed significance in both cases.

A set of five additional experiments were conducted separately in the two feature domains to test for the replicability of these findings. Under a variety of experimental manipulations, we consistently found primacy for mean position judgments and recency for mean size judgments (see Supplemental Material for methods and results). Together, these results indicate that summary representations of mean size and mean

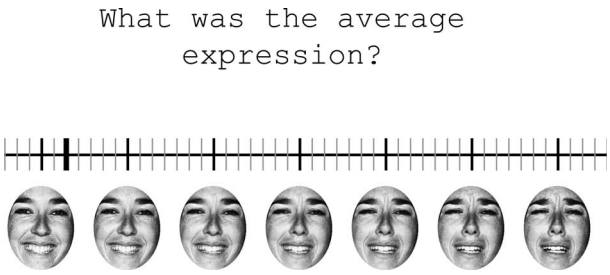

Figure 3. Response screen used in Experiment 2. Participants used this screen to indicate the "average expression" of the faces shown on that trial. Participants chose one of fifty vertical ticks by moving the indicator (the bold tick shown here at face number six) with the mouse and clicking. Seven "landmark" faces were always visible under their corresponding ticks (the black ticks at 4, 11, 18, 25, 32, 39, and 46) for participants to use as reference. Landmark faces did not change from trial to trial. Participants were encouraged use the non-landmark ticks if they thought the average lay between two landmark faces.

position extracted from items presented across time do **not** incorporate all presented items equally. Instead, we find that perception of mean size and position across time favor different portions of the information stream, with mean size favoring more recently presented items and mean position favoring earlier presented items.

Experiment 2: Averaging facial expression across time

Results from Experiment 1 suggest that how summary representations are computed across time may differ across feature domains. Experiments 2 and 3 explored this possibility by extending the method to two new feature domains. In Experiment 2, we used emotive face stimuli created by Haberman and Whitney (2007, 2009) to explore how summary representations of mean facial expression are generated across time.

Method

Participants, apparatus, stimuli, and procedure in general mirrored those of Experiment 1, except where otherwise noted.

Participants and apparatus

A new group of twenty participants was recruited from the University of Washington undergraduate student body using different participants from the same Department of Psychology participant pool as before. The decision to use this smaller number of participants than used in Experiment 1 was made ahead of data collection since in Experiment 2 all trials (rather than half) would contribute to a single set of weights per participant. The apparatus as in Experiment 1 was used, except that stimuli were presented on a medium gray background instead of black, and no grid was present.

Stimuli and procedure

Each trial began with the presentation of a black central cross approximately 0.4° in width and height, though participants were free to look wherever they wished over the course of the trial. After 500 ms, the cross changed color to red as a trial start warning and was present for a random period of time between one and two seconds. A series of eight faces was then presented.

Face stimuli consisted of eight human faces, presented one after another in the center of the screen. Each face was present for 252 ms and was followed by an 82 ms blank inter-face interval, resulting in a presentation rate of 3 Hz. The set of faces used was a subset of stimuli used by Haberman and Whitney (2007, 2009) in a series of experiments showing that humans can accurately and efficiently extract the mean emotional expression contained in a set of human faces, and consisted of fifty faces from the same person. The extreme two faces were actual photos, one showing a happy expression and one showing a sad one. The remaining 48 faces were regularly-spaced interpolations between the two emotionally extreme ones, created with image morphing software. See Haberman and Whitney (2009) for further details on the generation and properties of the faces, and see Figure 3 for example faces. For the purposes of stimulus sampling and modelling, each face in the stimulus set was assigned a number from 1 to 50, such that the distance between sequential faces is defined as one "eu," or *emotional unit*. When presented, the face stimuli subtended 6.0° vertically and 4.5° horizontally. The exact faces shown on each trial were chosen by sampling eight times from a Gaussian distribution (rounding to the nearest eu) with a standard deviation of 6 eu and a center that was itself drawn on each trial from a uniform distribution over face space between faces 11 and 39, inclusive. Any face that was sampled outside of the range 1 to 50 was resampled until it fell within the showable range.

The series of faces was followed by a blank period of 500 ms, followed by a response screen (see Figure 3) containing a horizontal line with fifty evenly spaced vertical tick marks in it, corresponding to the fifty possible response faces. Seven of the tick marks (numbers 4, 11, 18, 25, 32, 39, and 46) were larger than the others and had images of their corresponding face

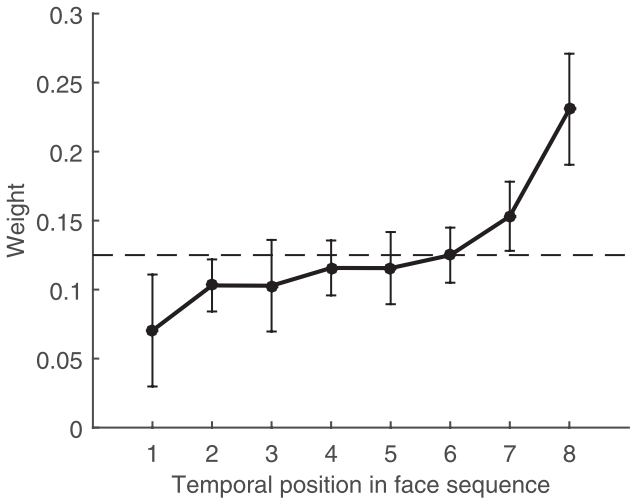

Figure 4. Results from Experiment 2. Mean weights as a function of temporal position across all participants. X-axis indicates the temporal order of the faces shown, where 1 refers to the first face presented on each trial and 8 refers to the last face presented on each trial. Y-axis indicates the weight of each temporal position on participants' responses. Weights were obtained using linear regression to fit a weighted average model to the data (see Results for details). Dashed line indicates the weights expected if all faces contributed equally to responses and no other source of noise or bias were present. Error bars indicate 95% confidence intervals across participants.

displayed below them. The purpose of these ticks and faces was to act as landmarks that participants could use to base their responses on. Once the response screen appeared, participants were asked to report the "average expression" in the faces shown on that trial by clicking the mouse cursor on the tick mark that most closely matched their estimate. Participants were explicitly instructed to not limit themselves to the landmark ticks or faces and to select one of the ticks in between if they thought the correct answer was between two landmark faces. The initial cursor position was chosen randomly at the start of each response period. The response period ended when the participant clicked on a tick mark, and was immediately followed by the start of the next trial.

Each participant completed 320 trials in blocks of 40 trials. The break, instruction, and practice procedures used in Experiment 1 were also used here. A full experimental session lasted about 50 minutes.

Results

The same weighted average model used in Experiment 1 was fitted to the present data and a set of eight weights was obtained for each participant that quantified the average contribution of the eight temporal positions to the participants' responses. The model once again described a significant portion of the variance, though the amount explained was in general lower and more variable than seen in Experiment 1 (model R^2: mean $= 0.63$, $SD = 0.17$; all $p < 0.001$). There are a number of possible reasons for this. Both the stimulus and response space resolution were lower in Experiment 2, where only fifty unique values were possible compared to the virtually unconstrained response spaces of Experiment 1. Another potential source of noise is the fact that participants were presented only a subset of all possible face images while making their selection. Indeed, inspecting the frequency of subject responses across the fifty possible faces reveals that subjects tended to choose ticks at the seven sample faces more often than ticks in between. While this tendency may add to the overall variability of the results, it should not affect the profile of weights over time. Finally, it is possible that the weighted average model simply describes the cognitive operation underlying facial expression averaging less well than for location or size.

Mean weights as a function of temporal position across all participants are shown in Figure 4. Just as with estimates of mean size and mean location, not all faces contributed equally to the participants' responses. Instead, as in the size domain, estimates of mean facial expression exhibited clear recency, where the later faces influenced participants' reports of the mean more than the earlier ones. Again, the size of the effect was relatively large, with each of the last two faces contributing, on average, between 1.5 and 4 times more than each of the first two faces. The effect was once more relatively smooth, with the average weight increasing gradually from face number one to eight. A one-way ANOVA confirmed the statistical significance of the effect, $F(7,133) = 9.12$, $p < 0.001$, $\eta_p^2 = .32$.

Experiment 3: Averaging motion direction across time

Results from Experiment 2 suggest that computation of a summary representation for facial expression uses a temporal weighting profile similar to that used in computation of mean dot size, and distinct from that used in computation of mean dot location. Experiment 3 extended the method from Experiment 2 into a fourth feature domain: motion direction.

Method

A new group of twenty participants was recruited just as in Experiments 1 and 2, and the apparatus,

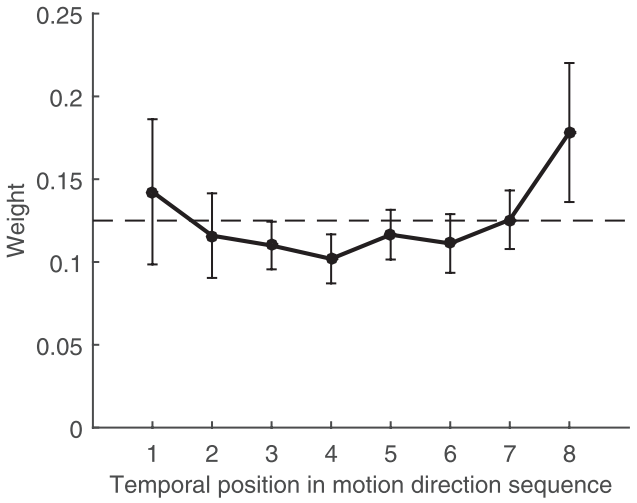

Figure 5. Results from Experiment 3. Mean weights as a function of temporal position across all participants. X-axis indicates the temporal order of the motion directions shown, where 1 refers to the first motion direction epoch shown on each trial and 8 refers to the last direction epoch on each trial. Y-axis indicates the relative influence of each motion direction epoch on participants' responses. Weights were obtained by fitting a weighted average model to the data (see Results text for details). Dashed line indicates the weights expected if all motion direction epochs contributed equally to responses and no other source of noise or bias were present. Error bars indicate 95% confidence intervals across participants.

stimuli, and procedure used closely mirrored those of Experiment 2, except where otherwise noted.

Stimuli and procedure

Each trial began with the presentation of a central red square marker approximately 0.5° in width and height, though once again participants were free to look wherever they wished over the course of the trial. After 200 ms, the square disappeared as a trial start warning. After a random period of between one and two seconds, a moving dot fields was then presented.

The moving dot field consisted of a square region (width = 8°) with a circular aperture (diameter = 8°) overlaid on it, both centered on the screen. Ten small (width = 0.12°) black dots moved within a square field, though only those inside the circular aperture were visible to the participant. The ten dots always moved with 100% coherency at 10°/s. Each dot had a maximum lifetime of 200 ms (12 frames at a 60 Hz monitor refresh rate) and was initialized with a random starting "age" between 0 and 11 monitor refresh frames. Initial dot location in the field was also random. When a dot reached the maximum age, it was destroyed and a new dot with age zero was created at a random point in the field. Whenever a dot's motion carried it outside the square field, its location was wrapped around to the opposite edge of the field.

Over the course of a single trial, the dot field moved in eight distinct directions, one after another, for 333 ms in each direction. Transitions between motion directions were abrupt, though dots persisted through the transition if their age permitted it. The set of eight motion directions were chosen on each trial by sampling from a Gaussian distribution with a standard deviation of 30° and a center that was itself sampled on each trial from a uniform distribution across all possible directions.

After the sequence of eight dot fields (lasting 2667 ms in total) there was a blank period of 300 ms, followed by a response screen. The response screen contained a central red square marker with a red line of length 2° radiating from it in a random initial direction. Once the response screen appeared, participants were asked to report the "average" or "overall" direction of motion present during that trial by using the mouse to adjust the radial direction of the red line until it pointed in the direction of the perceived average motion. The response period ended when the participant submitted their response with a mouse click and was followed by a 1000 ms inter-trial interval.

Each participant completed 320 trials in blocks of 40 trials. The break, instruction, and practice procedures used in Experiments 1 and 2 were also used here. A full experimental session lasted about 40 minutes.

Results

Though the dot motion parameters were chosen to maximize, for any given motion direction, the likelihood of perceiving coherent motion in that direction, some participants found the task very difficult, apparently due to the motion reversal illusion. Of the 20 initial participants, data from 3 were discarded due to clearly outlying mean absolute response errors (MAEs). The discarded participants' MAEs were 24.5°, 33.5°, and 39.0°, compared to an average MAE of 14.4° ($SD = 3.5°$) in the rest of the participants. Within the remaining 17 participants, high error trials were removed from the data, where high error was defined as a response error greater than three standard deviations from the mean response error for that participant. This resulted in discarding, on average, 1.2% of the data (~3.7 trials) from each participant, with 2.2% of the data (7 trials) being removed in the most affected participant.

Best-fitting weights describing the average influence that each of the eight motion direction epochs had on participants' responses were obtained by fitting the same model described in Experiments 1 and 2 to the remaining data. Just as in Experiment 1, the data were

described very well by the model (model R^2: mean = 0.98, $SD = 0.01$; all $p < 0.001$).

Mean weights from the 17 included participants are shown in Figure 5. As was seen in Experiments 1 and 2, weights deviated from even weighting across temporal position. A one-way ANOVA confirmed this, $F(7,112) = 2.90$, $p = 0.008$, $\eta_p^2 = .15$. In particular, the last motion direction appeared to contribute somewhat more (~1.5 times) than the rest of the individual motion directions, in an apparent recency effect. Interestingly, and in contrast to our findings in other domains, the average weight given to the first motion direction seen on each trial also appeared to be relatively large, suggesting some degree of primacy. However, pairwise t-tests between weights for the first motion direction and motion directions two through seven showed that no pairs were statistically different, with all six ps > 0.13 before multiple comparisons correction. These results suggest that summary representations for average motion direction over time (at least when the motion directions are discrete and separated in time) are computed more like those of size than location.

General discussion

Our primary finding is that the influence an item has on a summary statistic depends on its temporal position in the information stream. Specifically, summary representations of mean position were more strongly influenced by earlier items (primacy) and summary representations of mean size, mean facial expression, and mean motion direction were more strongly influenced by later items (recency). Across the experiments, the effect was reasonably large, with the most influential items in the stream contributing on average about 1.5 to 3 times more than the least influential items. Below, we consider a variety of explanations for the particular pattern of results we observed.

Strategies for summary computation

One argument for why primacy was observed in the mean location task is that it can be a highly functional strategy. Kiani and colleagues (Kiani, Hanks, & Shadlen, 2008) found a form of primacy in a task where nonhuman primates had to integrate motion information across time. The dot motion stimulus they used moved with 0% coherence on average, but over the course of each individual trial there were fluctuations in the moment-to-moment motion direction. The investigators found that motion information in the early portion of the stimulus influenced the monkey's eventual decision more than that in the later portion. They and others went on to show that monkey and human behavior in this task is consistent with a model of information collection where the cost of sampling additional information increases as more and more evidence is accrued (Drugowitsch, Moreno-Bote, Churchland, Shadlen, & Pouget, 2012). In other words, if an observer is optimizing for energy or time expenditure in addition to accuracy, primacy might be an optimal strategy. Later information in a stream might not be worth integrating into the summary if the benefit to the estimate is outweighed by the cost of collecting it. This framework might explain why we found primacy for judgments of mean location, but it does not explain why we found recency in the other feature domains.

However, recency can also be understood as a functional behavior. Recency might reflect the way the brain makes predictions about what the next item in a series will be based on previous items. For example, if previously shown facial expressions predict the facial expression that is likely to come next, then it makes sense to be most sensitive to the predicted incoming expression and less sensitive to unpredicted expressions. It has been shown that this type of adaptive gain control leads to recency when observers are asked to integrate samples over time (Cheadle et al., 2014). In this account recency is produced because, by the time the last sample arrives, observers have a strong prediction about its value and are thus highly sensitive to even small deviations from that prediction. Again however, such an account cannot explain the totality of our results, since the model Cheadle et al. describe cannot naturally produce the primacy we observed for mean location.

It should also be noted that recency becomes an attractive strategy if the underlying process generating the samples is non-stationary. Many things in the real world that humans might want to summarize, such as a conversation partner's face or a moving car, change in their properties over time. In this case, a summary representation that discards old information and reflects the most up-to-date state of the world is obviously valuable. However, two problems exist for this as a satisfying explanation of our recency findings. First, the underlying processes generating the stimuli in our experiments were stationary over the course of a trial in all cases. Second, it is unclear why the presumption of non-stationarity would apply to size, facial expression, and motion direction but not to location, where primacy was found.

Finally, related recent work on serial dependence has shown that the perceived orientation of a stimulus is systematically biased toward the orientation of recently seen stimuli (Fischer & Whitney, 2014). This

result has been interpreted as evidence of a "continuity field" that promotes visual stability over time by effectively acting as a low-pass temporal filter that biases the current perception of the world towards recent events (Liberman, Fischer, & Whitney, 2014). On its face, the serial dependence effect is reminiscent of our recency effect, but it seems instead to predict primacy for judgments of the mean in our tasks, since the perception of the later samples would in theory be pulled towards those of the early samples. This would presumably result in earlier samples being represented more than later ones in the overall mean judgment. So it is possible that this newly-documented perceptual bias from recent events affects or supports the perception of the mean in our tasks, but more work would be needed to determine the exact nature of its role.

Visual memory

All of the present experimental tasks involve maintaining visual memory representations, either of specific items presented in the sequence or a running belief of the average, over the course of a trial. Given this, it is important to consider explanations for primacy and recency that involve characteristics of or limitations in visual working memory.

Though not unopposed, the traditional view of visual working memory is that it is composed of at least partially separated subsystems that are involved in storing different types of visual information. Specifically, considerable behavioral (Hyun & Luck, 2007; Logie & Marchetti, 1991; Woodman & Luck, 2004; Woodman, Vogel, & Luck, 2001), electrophysiological (Goldman-Rakic, 1996), and neuroimaging work (Courtney, Ungerleider, Keil, & Haxby, 1996, 1997; Smith & Jonides, 1997; Smith et al., 1995) supports the existence of subsystems for spatial- and object-based working memory representations. If the spatial working memory subsystem is primarily recruited in our mean location task and the object-based working memory subsystem is recruited in our mean size and mean facial expression tasks, then it is possible that primacy and recency are the result of differential characteristics of those systems. Motion direction as a visual feature, however, is inherently spatial and thus poses a problem for this account, since we found recency in that domain. However, since our motion stimulus was 100% coherent and changed abruptly from one direction to the next, the motion sequences in our task may have been encoded and summarized as a series of orientations, a feature more strongly associated with object-based processing systems than spatial-based systems. While there is some alignment between working memory subsystems and our findings across domains, this alone does not constitute a satisfying explanation for our effects. The relative lack of information about the differential properties and functioning of spatial- and object-based working memory subsystems makes it difficult to explain why one would lead specifically to primacy and the other to recency in a summarization task, for example. This problem is exacerbated by the fact that the most popular visual working memory tasks measure memory for objects that are defined by a binding of location and either color or orientation (Luck & Vogel, 1997; Zhang & Luck, 2008), confounding spatial- and object-based memory.

Could primacy and recency in our summarization task be driven by serial position effects in short-term memory? As in, do our results reflect more about memory quality for items presented in sequences than summarization processes themselves? There are indications of serial position effects in short term memory in at least two of the domains that we investigated here: location (Farrand & Jones, 1996; Farrand, Parmentier, & Jones, 2001; Guérard & Tremblay, 2008; Jones, Farrand, Stuart, & Morris, 1995) and faces (Hay, Smyth, Hitch, & Horton, 2007; Ward, Avons, & Melling, 2005). In these tasks, subjects were shown a series of to-be-memorized objects (usually 5–12 items at 0.5 Hz) and are asked to reconstruct the sequence after some retention interval (usually 0–30 s). However, little evidence of domain differences in recall as a function of position in the sequence is noted. Instead, both primacy and recency of recall are seen in nearly all face and location experiments, with several of the researchers noting that the strength of primacy or recency seems to depend more on the specific testing or recall method used than on the feature domain (Farrand et al., 2001; Jones et al., 1995; Ward et al., 2005). While these findings are clearly related to the present results, this lack of domain differences makes it difficult to conclude that serial position effects in memory are solely responsible for our summarization findings, in particular the differences we observed across domains.

Finally, a related explanation for our findings of primacy and recency in summary computation is that one or both of them are due to capacity limitations in visual working memory. In this hypothesis, primacy is produced when the limited capacity of visual working memory is filled with memory representations from the early items and little to no resources are left to store the later items. Alternatively, recency might be produced by the same limited capacity if later items push representations of earlier items out of memory. The effect of memory capacity limitations on summary computation could in theory be tested by adding a sequence length manipulation to the experiments we report on here. If either or both of primacy and recency significantly diminish or disappear with shorter se-

quences, this would constitute evidence that capacity limits play a strong role in our findings. But if either primacy or recency are still observed with sequence lengths below the traditional visual working memory capacity (about 3–4 items according to Luck & Vogel, 1997), then the contribution of memory capacity to either effect is likely limited. It should be noted, though, that even if capacity limitations are involved in our findings, it is not immediately clear how this explanation alone would produce primacy in one feature domain and recency in the others.

Implications for summary representation across domains

Since it was reported in 2001 that human observers appeared to extract the mean size of an array of discs with surprising speed, precision, and automaticity, the concept of summary statistics has been invoked to describe reports of perceptual averaging across a wide variety of stimulus domains. Size, orientation, position, brightness, color, motion direction, speed, facial expression, facial identity, biological motion, and frequency of tones have all been discussed as features across which summary statistics might be computed. However, a priori, it seems unlikely that averaging in all of these domains shows the same characteristics that made summary perception of mean size so intriguing when it was first reported, especially considering that the physiology underpinning representation of these various features and objects in the brain is very different, and in some cases still not well understood. Despite this, the extent to which averaging across these domains reflects the same computation has not been well-studied. In fact, to the best of our knowledge, the experiments reported here are the first to directly compare how summary representations are computed across a set of different feature domains.

Here we provide evidence that summary representation *behaves* differently in different feature domains, but do our results constitute evidence for distinct *mechanisms* for summary computation across time in different feature domains? Summaries that look different do not necessarily come from distinct mechanisms. A single mechanism that combines perceptual or memory representations into a summary could in theory produce different-looking results given different input. This is implied in our discussion of the role of visual memory above; perhaps the quality of early versus late item representations that are fed into a summarization mechanism simply varies across domains. This possibility combined with the lack of previous comparative work in summary representation domains makes it difficult to argue conclusively that distinct mechanisms are involved in summarizing in location and non-location domains, even if the summary computed is meaningfully different across domains.

In conclusion, it perhaps should not be surprising that the way in which summary representations are computed varies across feature domains. Just as other perceptual judgments fall along a continuum from low to higher-level processing, summary computation may do the same. Some feature domains are summarized at the sensory level. For example, a photoreceptor computes a weighted average of the spectrum of incoming light over a fixed period of time and space. Other feature domains are likely to require different, higher level cognitive processes, as in judging someone's overall moral character by his or her deeds. By understanding what similarities and differences exist between different types of summary representation, we will be better equipped to search for their underlying mechanisms, which is a major goal for this promising area of research. But until then, we conclude that not all summary statistics are created equal.

Keywords: summary statistics, ensemble coding, integration across time, size perception, face perception, motion perception

Acknowledgments

We gratefully acknowledge Jason Haberman for generously providing his face stimuli, Wendy Coard for assistance in data collection, two reviewers for helpful suggestions and feedback, and Scott Murray for valuable perspective, guidance, and advice. B. Hubert-Wallander and G. M. Boynton developed the study concept. B. Hubert-Wallander designed, implemented, and ran the experiments. B. Hubert-Wallander and G. M. Boynton analyzed the data and wrote the manuscript. This work was supported by a National Science Foundation Graduate Research Fellowship to B. Hubert-Wallander [DGE-0718124] and a grant from the National Institutes of Health to G. M. Boynton [EY12925].

Commercial relationships: none.
Corresponding author: Bjorn Hubert-Wallander.
Email: bjornhw@uw.edu.
Address: Department of Psychology, University of Washington, Seattle, WA, USA.

References

Albrecht, A. R., & Scholl, B. J. (2010). Perceptually averaging in a continuous visual world: Extracting

statistical summary representations over time. *Psychological Science*, *21*(4), 560–567, doi:10.1177/0956797610363543.

Albrecht, A. R., Scholl, B. J., & Chun, M. M. (2012). Perceptual averaging by eye and ear: Computing summary statistics from multimodal stimuli. *Attention, Perception, & Psychophysics*, *74*(5), 810–815, doi:10.3758/s13414-012-0293-0.

Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science*, *19*(4), 392–398, doi:10.1111/j.1467-9280.2008.02098.x.

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

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436.

Cheadle, S., Wyart, V., Tsetsos, K., Myers, N., de Gardelle, V., Herce Castañón, S., & Summerfield, C. (2014). Adaptive gain control during human perceptual choice. *Neuron*, *81*(6), 1429–1441, doi:10.1016/j.neuron.2014.01.020.

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research*, *43*(4), 393–404, doi:10.1016/S0042-6989(02)00596-5.

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

Chong, S. C., & Treisman, A. (2005b). Statistical processing: Computing the average size in perceptual groups. *Vision Research*, *45*(7), 891–900, doi:10.1016/j.visres.2004.10.004.

Corbett, J. E., & Oriet, C. (2011). The whole is indeed more than the sum of its parts: Perceptual averaging in the absence of individual item representation. *Acta Psychologica*, *138*(2), 289–301, doi:10.1016/j.actpsy.2011.08.002.

Courtney, S. M., Ungerleider, L. G., Keil, K., & Haxby, J. V. (1996). Object and spatial visual working memory activate separate neural systems in human cortex. *Cerebral Cortex*, *6*(1), 39–49.

Courtney, S. M., Ungerleider, L. G., Keil, K., & Haxby, J. V. (1997). Transient and sustained activity in a distributed neural system for human working memory. *Nature*, *386*(6625), 608–611, doi:10.1038/386608a0.

Dakin, S. C. (2001). Information limit on the spatial integration of local orientation signals. *Journal of the Optical Society of America, A: Optics, Image Science, & Vision*, *18*(5), 1016–1026.

de Fockert, J., & Wolfenstein, C. (2009). Rapid extraction of mean identity from sets of faces. *Quarterly Journal of Experimental Psychology*, *62*(9), 1716–1722, doi:10.1080/17470210902811249.

de Gardelle, V., & Summerfield, C. (2011). Robust averaging during perceptual judgment. *Proceedings of the National Academy of Sciences, USA*, *108*(32), 13341–13346, doi:10.1073/pnas.1104517108.

Drugowitsch, J., Moreno-Bote, R., Churchland, A. K., Shadlen, M. N., & Pouget, A. (2012). The cost of accumulating evidence in perceptual decision making. *Journal of Neuroscience*, *32*(11), 3612–3628, doi:10.1523/JNEUROSCI.4010-11.2012.

Emmanouil, T. A., & Treisman, A. (2008). Dividing attention across feature dimensions in statistical processing of perceptual groups. *Perception & Psychophysics*, *70*(6), 946–954.

Farrand, P., & Jones, D. (1996). Direction of report in spatial and verbal serial short-term memory. *Quarterly Journal of Experimental Psychology A*, *49*(1), 140–158.

Farrand, P., Parmentier, F. B. R., & Jones, D. M. (2001). Temporal-spatial memory: Retrieval of spatial information does not reduce recency. *Acta Psychologica*, *106*, 285–301.

Fischer, J., & Whitney, D. (2014). Serial dependence in visual perception. *Nature Neuroscience*, *17*(5), 738–743, doi:10.1038/nn.3689.

Goldman-Rakic, P. S. (1996). Regional and cellular fractionation of working memory. *Proceedings of the National Academy of Sciences, USA*, *93*(24), 13473–13480.

Greenwood, J. A., Bex, P. J., & Dakin, S. C. (2009). Positional averaging explains crowding with letter-like stimuli. *Proceedings of the National Academy of Sciences, USA*, *106*(31), 13130–13135, doi:10.1073/pnas.0901352106.

Guérard, K., & Tremblay, S. (2008). Revisiting evidence for modularity and functional equivalence across verbal and spatial domains in memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *34*(3), 556–569, doi:10.1037/0278-7393.34.3.556.

Haberman, J., Harp, T., & Whitney, D. (2009). Averaging facial expression over time. *Journal of Vision*, *9*(11):1, 1–13, http://www.journalofvision.org/content/9/11/1, doi:10.1167/9.11.1. [PubMed] [Article]

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

Haberman, J., & Whitney, D. (2009). Seeing the mean: Ensemble coding for sets of faces. *Journal of Experimental Psychology: Human Perception &*

Performance, *35*(3), 718–734, doi:10.1037/a0013899.

Hay, D. C., Smyth, M. M., Hitch, G. J., & Horton, N. J. (2007). Serial position effects in short-term visual memory: A SIMPLE explanation? *Memory & Cognition*, *35*(1), 176–190.

Hyun, J. S., & Luck, S. J. (2007). Visual working memory as the substrate for mental rotation. *Psychonomic Bulletin & Review*, *14*(1), 154–158.

Jones, D., Farrand, P., Stuart, G., & Morris, N. (1995). Functional equivalence of verbal and spatial information in serial short-term memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *21*(4), 1008–1018.

Juni, M. Z., Gureckis, T. M., & Maloney, L. T. (2012). Effective integration of serially presented stochastic cues. *Journal of Vision*, *12*(8):12, 1–16, http://www.journalofvision.org/content/12/8/12, doi:10.1167/12.8.12. [PubMed] [Article]

Kiani, R., Hanks, T. D., & Shadlen, M. N. (2008). Bounded integration in parietal cortex underlies decisions even when viewing duration is dictated by the environment. *Journal of Neuroscience*, *28*(12), 3017–3029, doi:10.1523/JNEUROSCI.4761-07.2008.

Liberman, A., Fischer, J., & Whitney, D. (2014). Serial dependence in the perception of faces. *Current Biology*, *24*(21), 2569–2574, doi:10.1016/j.cub.2014.09.025.

Logie, R. H., & Marchetti, C. (1991). Visuo-spatial working memory: Visual, spatial or central executive? *Advances in Psychology*, *80*, 105–115.

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

Morgan, M. J., & Glennerster, A. (1991). Efficiency of locating centres of dot-clusters by human observers. *Vision Research*, *31*(12), 2075–2083, doi:10.1016/0042-6989(91)90165-2.

Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience*, *4*(7), 739–744, doi:10.1038/89532.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442.

Piazza, E. A., Sweeny, T. D., Wessel, D., Silver, M. A., & Whitney, D. (2013). Humans use summary statistics to perceive auditory sequences. *Psychological Science*, *24*(8), 1389–1397, doi:10.1177/0956797612473759.

Robitaille, N., & Harris, I. M. (2011). When more is less: Extraction of summary statistics benefits from larger sets. *Journal of Vision*, *11*(12):18, 1–8, http://www.journalofvision.org/content/11/12/18, doi:10.1167/11.12.18. [PubMed] [Article]

Smith, E. E., & Jonides, J. (1997). Working memory: A view from neuroimaging. *Cognitive Psychology*, *33*(1), 5–42.

Smith, E. E., Jonides, J., Koeppe, R. A., Awh, E., Schumacher, E. H., & Minoshima, S. (1995). Spatial versus object working memory: PET investigations. *Journal of Cognitive Neuroscience*, *7*(3), 357–375.

Solomon, J. A. (2010). Visual discrimination of orientation statistics in crowded and uncrowded arrays. *Journal of Vision*, *10*(14):19, 1–16, http://www.journalofvision.org/content/10/14/19, doi:10.1167/10.14.19. [PubMed] [Article]

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

Spencer, J. (1961). Estimating averages. *Ergonomics, 4*(4), 317–328, doi:10.1080/00140136108930533.

Spencer, J. (1963). A further study of estimating averages. *Ergonomics, 6*(3), 255–265, doi:10.1080/00140136308930705.

Ward, G., Avons, S. E., & Melling, L. (2005). Serial position curves in short-term memory: Functional equivalence across modalities. *Memory*, *13*(3–4), 308–317.

Watamaniuk, S. N., & Duchon, A. (1992). The human visual system averages speed information. *Vision Research*, *32*(5), 931–941, doi:10.1016/0042-6989(92)90036-I.

Whitney, D., Haberman, J., & Sweeny, T. (2014). From textures to crowds: multiple levels of summary statistical perception. In J. S. Werner & L. M. Chalupa (Eds.),*The new visuall neurosciences,* (pp. 685–709). Cambridge, MA: MIT Press.

Woodman, G. F., & Luck, S. J. (2004). Visual search is slowed when visuospatial working memory is occupied. *Psychonomic Bulleting & Review*, *11*(2), 269–274.

Woodman, G. F., Vogel, E. K., & Luck, S. J. (2001). Visual search remains efficient when visual working memory is full. *Psychological Science*, *12*(3), 219–224.

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(7192), 233–235, doi:10.1038/nature06860.