Ensemble statistics accessed through proxies: Range heuristic and dependence on low-level properties in variability discrimination

Jonas Sin-Heng Lau
Department of Psychology, University of California, San Diego, La Jolla, CA, USA

Timothy F. Brady
Department of Psychology, University of California, San Diego, La Jolla, CA, USA

People can quickly and accurately compute not only the mean size of a set of items but also the size variability of the items. However, it remains unknown how these statistics are estimated. Here we show that neither parallel access to all items nor random subsampling of just a few items is sufficient to explain participants' estimations of size variability. In three experiments, we had participants compare two arrays of circles with different variability in their sizes. In the first two experiments, we manipulated the congruency of the range and variance of the arrays. The arrays with congruent range and variability information were judged more accurately, indicating the use of range as a proxy for variability. Experiments 2B and 3 showed that people also are not invariant to low- or mid-level visual information in the arrays, as comparing arrays with different low-level characteristics (filled vs. outlined circles) led to systematic biases. Together, these experiments indicate that range and low- or mid-level properties are both utilized as proxies for variability discrimination, and people are flexible in adopting these strategies. These strategies are at odds with the claim of parallel extraction of ensemble statistics per se and random subsampling strategies previously proposed in the literature.

Introduction

Suppose you work in a strawberry farm and your task is to screen strawberries before they are packaged for sale. Small berries should be removed from the conveyor belt as they pass through, because customers generally do not enjoy them. Batches of strawberries also come in different average sizes. Some batches have larger berries in general, other have smaller ones. To perform your screening task well, you need some mental representations of the overall size of the batch of berries. Yet, since berries vary in size, having only a representation of the overall size and a simple cutoff score are not sufficient. You also need a representation of how variable they are. The variability measure gives you a sense of how small a berry is abnormal, and hence should be taken away. Anomaly detection such as this berry screening task, and many other real-world judgment tasks, require quick access to statistical summary representations such as the average and variability of a set of items (Ma, Navalpakkam, Beck, Van Den Berg, & Pouget, 2011).

Since Ariely (2001), ample evidence has shown that statistical summary representations in an array can be efficiently extracted. Studies have shown that mean size (e.g., Ariely, 2001), orientation (e.g., Morgan, Chubb, & Solomon, 2008), spatial position (e.g., Alvarez & Oliva, 2008), and even emotion and gender (e.g., Haberman & Whitney, 2007, 2009) of an array can be extracted with little effort. The ability is believed to be relevant to scene perception since when a large number of objects are present in a scene, they may not each be perceived individually but instead, represented as a set or ensemble (Brady, Shafer-Skelton, & Alvarez, 2017; Haberman & Whitney, 2012; Greene, 2013).

In recent years, there has been a significant interest in understanding how this efficient statistical summary representation process is implemented cognitively. Is there a special mechanism designed to compute summary statistics efficiently? Or is this ability just the result of smart sampling strategies where we attend to and remember a few items in working memory, and use them to derive summary statistics? Ariely (2001) and Chong and Treisman (2003, 2005) argued that mean size extraction involves a parallel process, in part because calculating the average size of an array seems quick and effortless. In addition, they showed that participants do not necessarily have access to the identities of individual items even when they have

Citation: Lau, J. S.-H., & Brady, T. F. (2018). Ensemble statistics accessed through proxies: Range heuristic and dependence on low-level properties in variability discrimination. *Journal of Vision*, *18*(9):3, 1–18, https://doi.org/10.1167/18.9.3.

https://doi.org/10.1167/18.9.3 Received April 17, 2018; published September 5, 2018 ISSN 1534-7362

access to the mean size (Ariely, 2001), and that performance at judging mean size is relatively unaffected by variations in the number of items shown, the variability of those items, and the duration they are shown (Chong & Treisman, 2003). However, there are alternative strategies that could allow participants to estimate the average size of items without the need to invoke a fully parallel process focused on calculating mean size. For example, Myczek and Simons (2008) proposed several alternative strategic accounts to the global, parallel process for mean size extraction. They showed that size discrimination could be performed through subsampling the arrays. For example, simulated accuracy patterns when two or three items were sampled and averaged from an array could produce accuracy patterns close to, or exceeding that of human participants. The authors were cautious that participants might not actually carry out this exact subsampling heuristic, but noted that similar sampling strategies provided a means to perform the discrimination task. Since Myczek and Simons (2008) proposed this account, a great deal of work has focused on parsing out the actual cognitive mechanisms of extracting ensemble information about the mean of a set (e.g., Chong, Joo, Emmanouil, & Treisman, 2008; Simons & Myczek, 2008). For example, Allik, Toom, Raidvec, Averin, & Kreegipuu (2013) suggested that most of the variance in a mean discrimination task could be explained by a simple model taking internal noise and sampling into account. Others have found evidence more consistent with some "smart" subsampling strategies (e.g., Marchant, Simons, & de Fockert, 2013; Maule & Franklin, 2016). On the other hand, some findings have provided support for more parallel mechanisms—for example, outliers tend to be discounted in extracting the mean (Haberman & Whitney, 2010), inconsistent with a straightforward random subsampling account; and in pairs of arrays where only a single item changes between arrays, participants can recognize the change in the mean without knowing which item changed (Haberman & Whitney, 2011; see also Ward, Bear, & Scholl, 2016).

However, nearly all of these previous studies have focused on extracting the mean. Higher-level statistical summary representations, such as variability and kurtosis, have remained understudied. Given the importance of variability in real-world applications like outlier detection and visual search, and given the computational difficulty of extracting a set's variability from a subsampling strategy, this paper focuses on understanding the mechanisms of size variability extraction. In particular, we examine whether size variability discrimination can be carried out efficiently, and, if so, how this calculation is cognitively implemented. As has been done in the case of mean size, we contrast an account where (a) the size of all of the items is preattentively available and utilized to compute the variability in parallel; with (b) a subsampling strategy where four to five items are randomly sampled; and (c) a set of alternative, more parallel strategic accounts, where people make use of cognitive strategies to infer variability without directly carrying out the computation. In particular, we focus on the possibility that participants make use of a visual search for the largest and smallest items (a *range*-based account), and an account where the variability is computed using low- or mid-level properties of the array like spatial frequency or texture. Together, our experiments suggest that size variability of a set of items is accessed through proxies, and this view may be extended to other types of statistical summary representations.

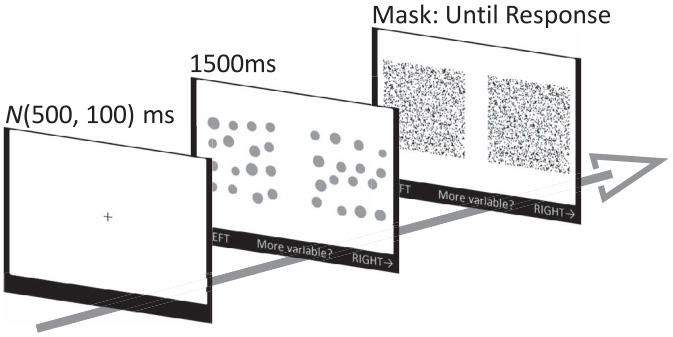

Figure 1. Sample array used in Experiment 1. The participants' task was to decide which of the two sides had a larger variability in size. A static mask was shown upon removal of the circle arrays from the screen.

The variability discrimination task and possible mechanisms

Imagine we show participants two arrays of circles simultaneously, one on each side of fixation, and ask them to judge the relative variability of these arrays (as in Solomon, Morgan, & Chubb, 2011). The two arrays have roughly equal mean sizes, but one of the arrays has more variability (see Figure 1 for an example). The participants' task is to say in which of the two arrays there is more variability in the size of the items. How do people perform this variability discrimination task? Here, we propose four possible mechanisms based on those proposed for mean size discrimination.

(1) A cognitively demanding serial processing of all items

Imagine you were asked to perform this variability task, but rather than a visual task with the size of circles, you were given numeric values from two different lists and asked to compute which had the higher variability. To do this, you would need to

mathematically calculate variance for each list and compare them. To use this strategy in the visual array, a numerical value of the size of each individual item (X) in the array is first extracted. The values within an array are then summed and divided by the number of items in the array (N). A mean of each array (μ) is then calculated by the formula:

$$\mu = \frac{\sum X}{N} \quad (1)$$

This is the general formula for calculating the arithmetic mean of a population (μ). The squared deviations between each item in the array and the mean are then summed. The sum of squared deviations is then divided by N to form a variance (σ^2).

$$\sigma^2 = \frac{\sum (X-\mu)^2}{N} \quad (2)$$

The variance of the two arrays can then be compared. This process guarantees that all individual items in the arrays are utilized.

Three corollaries logically follow if participants used a serial, cognitively demanding strategy like they would be forced to use with symbolic numbers as inputs: First, if the processes for the mean and variance computation are noise-free, the participants should always obtain the correct answer. Thus, to explain human performance, noise must be added either in the representations of item sizes, the decision stage, or both (see Allik et al, 2013; Solomon et al., 2011). Second, since individual items sizes are accessed and represented, participants may have memories of many or all of the items (e.g., when asked to recall whether a particular item is in the array after a delay, it is likely that participants would perform well or, at minimum, above chance). Third, each item is processed serially. As the number of items in the array increases, response time should also increase. In many ways, this "cognitive" account serves as a null hypothesis for how participants might perform the task. If participants simply compute variance by making noisy estimates of each item's size, this provides a possible account of the visual version of the task, though it predicts both (a) an intensely cognitively demanding task, and (b) a serial and slow task, as would be the case if given symbolic numbers rather than visual circles.

(2) Parallel extraction of some statistical summary properties

(2A) Method (1) above posits that each individual item is accessed and used for variability computation. In the ensemble statistics literature, a number of researchers have argued against such a strong position. Using a recognition test, Ariely (2001) concluded that his participants did not have representations of individual items during mean discrimination. Chong and Treisman (2003) supported that idea. They found that increasing the set size of the arrays did not affect size discrimination performance. This was taken to argue for a preattentive, parallel processing for size discrimination. Thus, one possibility is that participants have preattentive access to the size of each item without performing any cognitively demanding task to ascertain size. They can directly feed this preattentive estimate of the size of each circle not only into an "ensemble average" mechanism but also into an "ensemble variability" mechanism. This account predicts that participants make use of almost all information from the display but do not do so in a cognitively demanding way that requires processing of each item.

(2B) Alternative parallel accounts of variability estimation are also possible. For example, low-level visual information, such as luminance, spatial frequency, and texture can be used as a global, parallel mechanism for estimating statistical properties (Im & Halberda, 2012). More homogeneous arrays usually denote a smooth texture and fewer spatial frequencies, more heterogeneous arrays have rougher textures and more spatial frequencies are represented. Detection of different textures can be very efficient (e.g., Morgan et al., 2008). Thus, one possible range of accounts of variability discrimination is that participants make use of low-level or mid-level proxies for variability without directly accessing the size of items or relying on preattentive size representations.

(3) Random subsampling

Myczek and Simons (2008) proposed a computationally viable alternative to the parallel processing account (2A) for mean size discrimination. In their formulation, a subset of items from each array is randomly sampled into working memory and then an estimator (the mean) is calculated based on the sample. The estimator may be calculated using an analytic method, such as Method (1) above. It may also be extracted with some parallel process; such as Method (2) above. Once an estimator for each array is generated, an inference of variability can be made by comparing the two estimators. This model has the cognitive simplicity of assuming people use known mechanisms for selecting and holding onto a subset of items, and, in the case of estimating the mean, only a few samples are needed for accurate performance, meaning this proposal can explain average performance quite well. However, Myczek and Simons (2008) also suggested that observers might not be implementing a truly random sampling algorithm in the case of mean discrimination. As we explain below, this random sampling method is even less unlike to apply to variability discrimination, as the implementation of a

random subsampling account for a variability discrimination task requires very large numbers of samples to be stored in the working memory, which seem infeasible given limited working memory capacity.

(4) Smart subsampling

Another alternative mechanism for estimating the variability of a display is a form of smart subsampling. In particular, people may use efficient mechanisms to select a set of items that is particularly informative about variability, and rely only on that subset of items in making their judgment. Smart subsampling for variability discrimination differs from random subsampling. To make use of the strategy, people must be strategically selecting items in the array. In the case of variability discrimination, smart subsampling can be implemented by instead trying to estimate the range of the array, the largest item, the smallest item or some other proxies of variability available in the arrays (Fouriezos, Rubenfeld & Capstick, 2008). Range often works very well as a proxy for variability because a set of objects with large variability tends to have a large range as well. Using only the largest or the smallest item as proxy might work well for the same reason.

To carry out the smart subsampling, such as a range heuristic, participants have to first search for the largest and smallest items in each of the two arrays. They then compare two pairs of items and see which array has the larger range. The one with a larger range is likely to have a larger variability. Other, more local range-based strategies are also possible. For example, participants could examine only a subset of the items, like those near fixation, and find the largest and smallest among those items. This would result in a noisier estimate of the range of the display but would also be consistent with the use of the range as a proxy.

In general, any one of these strategies—or some combination of them—could explain performance in the described variability discrimination task. To try to understand the relative role of these strategies, in the current study we use a mix of both simulations and experiments. We conclude that people rely on both parallel mechanisms involving low-level or mid-level features (2B) and on a spatially constrained range heuristic (4) in computing variability, rather than either a fully serial strategy (1), a fully parallel processing of the items in the display (2A), or a straightforward subsampling algorithm (3).

Existing evidence about variability discrimination

Previous research has shown that computing the variability of a set of items is both possible and done relatively accurately (e.g., Morgan et al., 2008). Broadly, this research has suggested that people make use of a large fraction of the items when performing the discrimination task.

In particular, Morgan et al. (2008) had participants look at two Gabor arrays, arranged side-by-side. Each array had the Gabor stimuli arranged in an 11×11 imaginary grid. The stimuli varied in orientation. One of the arrays was the *standard* array, which had a certain orientation variability among the Gabor stimuli, and it was compared to the *test* array. The test array has a larger variability compared to the standard array. The arrays were shown for 200 ms. The participants' task was to determine which of the two arrays had a larger variability (i.e., identify the test array). Participants were quite good at this task. However, the cognitive mechanism—how participants performed this judgment—was relatively unexplored. In a later study by the same group, Solomon et al. (2011, experiment 3) examined whether the just noticeable difference (JND) to detect variability differences increased as the variability of the standard array increased. This time, the authors focused on size variability instead of orientation. They systematically manipulated the variability differences between the standard and test arrays displayed in succession. They found that as the variability of the standard array increased, a larger difference between the two arrays was needed for participants to detect the difference. This is consistent with Weber's law and thus with most other domains of quantity judgment, which provides some evidence that participants may have been directly estimating variance from the size of the items in the arrays. However, the results of Solomon et al. (2011) also suggested that a simple parallel processing account cannot explain their data. They concluded that participants utilized only a subset of items in each array, and then use them to compute the summary statistics. In particular, their efficiency analysis led to an estimate that participants used five to eight items' worth of information from each array (relative to all eight items in each array) for the variability discrimination task. Interestingly, participants' reports of variability were also more accurate than one would expect based on their reports of mean size in a previous experiment. Solomon et al. (2011) suggested that this is accounted for by some form of late decision noise in the computation of mean size, but an alternative suggestion is that participants simply use different cognitive strategies in computing the two (e.g., Yang, Tokita, & Ishiguchi, 2018).

More direct evidence about the extent to which participants engage in direct computation of size variability is provided by Tokita, Ueda, and Ishiguchi (2016, experiment 3). They had participants perform a variability discrimination task and attempted to clarify

whether all or a random subset of items were used during variability discrimination. Two levels of variability differences between the standard and test arrays, as well as four levels of set sizes were used. Employing a statistical efficiency analysis, the authors rejected the possibility of a random subsampling model, as participants were too efficient at variability discrimination at higher set sizes to be relying on a purely random subset of items. Using a perceptual adaptation method, Norman, Heywood, and Kentridge (2015) asked if a mental representation of variability exists. They showed that participants were less sensitive in variability discrimination after being adapted to high variability stimuli. However, these studies did not directly address other cognitive strategies that might play a role without involving sampling all items (e.g., range-based accounts; low- or mid-level proxy-based accounts).

Taken together, the existing work provides strong evidence that participants can successfully perform size variability tasks and provides initial evidence that a random subsampling account is insufficient to explain performance in these tasks. However, the particular strategies used by participants have remained relatively unexplored.

Random subsampling is not a feasible strategy for variability discrimination

The idea of subsampling has been explored mostly in the context of mean size. Interestingly, however, both simulations and best-case estimation (below) show that mean size is much more easily approximated using a subsampling strategy of three to four items (i.e., the typical working memory limits) than is variability estimation. To perform a mean size task using a subsampling strategy, participants have to only sample a few items in the arrays, and generate a single representation of the mean. Sampling just a few items provides a relatively accurate estimate of the mean (as pointed out by Myczek & Simons, 2008). This is compatible with idea that only three to four items can be held in working memory.

In addition, it is the case that variability is more difficult to estimate than mean size, even with the same number of samples per array. In particular, the variance of an estimate of the mean of a set of items is σ^2/n, where σ^2 is the variance across items and n is the number of items sampled. By contrast, the variance in an estimate of the *variance* of a set of items is $2\sigma^4/n$. This means you have much more uncertainty about the variance of a set of items from a given number of samples, n, than about the mean from that same number of samples.

To see the exact effect this has on a variability judgment task, we simulated performance of variability discrimination as a function of the number of random subsamples taken. In our simulation, we assumed that a participant (in this case, the computer) was given a task similar to that of Figure 1. In each trial, two arrays were shown, each with N items. In our simulation, we used an arbitrary N, and set it to $N = 50$, though the results are invariant to N as long as N is greater than the number of sampled items. The average sizes of the two arrays were always the same, with 10 units. One of the arrays was a *standard* array, which had a standard deviation of 1.0 unit. The other was a *test* array, which had a standard deviation of 1.0, 1.2, 1.4, 1.6, or 1.8 units. These translated into a 0% to 80% difference in standard deviation between the standard and test arrays. The observer's task was to identify the array with larger variability in size (i.e., the test array).

In each trial, k items were randomly sampled from each of the two arrays, so a total of $2k$ items were stored in working memory. The standard deviations of the two samples were then separately calculated and compared. The one with the larger sample standard deviation was chosen as the response. The same procedure was repeated 20,000 times at each level of $2k$ and each standard deviation difference. The responses were aggregated at each level.

Figure 2a shows the simulation results. When there were no differences between the standard and test arrays, increasing the number of items stored in working memory ($2k$) would not increase the rate of selecting the array with larger variability (i.e., the test array). When there was a real difference between the two arrays, accuracy (the rate of selecting the test array) grows approximately logarithmically with the number of items stored in working memory ($2k$).

Thus, the nature of variability discrimination tasks makes it such that very large numbers of samples are needed to perform well. In particular, our simulations show that size variability discrimination can be theoretically achieved by subsampling only with unreasonably large samples. For example, according to Figure 2a, for a moderate difference in standard deviation between the standard and test arrays (difference = 0.4), around eight items have to be sampled from each array to achieve an accuracy of 80%. That is, to achieve an 80% accuracy, participants have to maintain 16 independent representations in working memory. This working memory capacity requirement is well above estimated capacities for individual items (e.g., Luck & Vogel, 2013). For this reason, a pure random subsampling heuristic with analytical variability extraction is not feasible for a variability discrimination task.

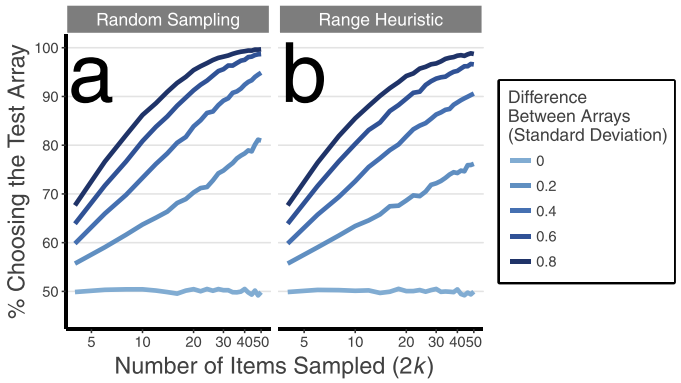

Figure 2. (a) Simulation of random subsampling strategy on a variability discrimination task. The simulation assumed that two arrays were shown on each trial, where the two arrays had the same average size, but one (test array) had a larger variability than the other (standard array). The participant's task was to identify the one with a larger variability (i.e., the test array). Each of the lines indicates a particular difference in standard deviation between the standard and test arrays. In general, the test array was increasingly likely to be chosen as more items ($2k$) were utilized and stored in working memory, provided that there was a true difference between the arrays. (b) Efficiency simulation of the range heuristic in a variability discrimination task. In the simulation, k items from each array (hence, a total of $2k$) were sampled. The largest and smallest items among the sample were extracted, and used to calculate the ranges of the arrays. The array with a larger range was judged to be more variable (and selected as the *test* array). At each level, 20,000 pairs of arrays were generated and the rate of picking the actual test array was recorded. All simulations in this paper were done in R.

The current experiments

Using simulations, we have shown that neither directly processing and remembering the size of each item nor a random subsample of items is likely to be the method implemented by human observers when performing variability discrimination tasks. The limiting factor is working memory capacity. To obtain reasonable performance with such a strategy on variability discrimination tasks, a large number of items have to be stored in working memory, which is beyond the capacity limit of working memory. To perform the task, observers are likely to implement other strategies.

Thus, in a series of experiments, we explore possible strategies that they might use. In Experiments 1 and 2, we test whether they might use a heuristic based on the range of the arrays. In Experiments 2B and 3, we test whether participants' performance is invariant to differences in low- and mid-level visual information. We found evidence for a range heuristic and for the usage of low- or mid-level visual features, suggesting multiple heuristics are at play in estimating variability of a group of items. These strategies provide partial explanation to the relatively efficient performance when variability discrimination is needed in daily life.

Experiment 1

Given that it is unlikely that people perform straightforward random sampling strategies, how do participants perform variability discrimination tasks? Experiment 1 aimed at testing if participants were using a particular form of smart sampling strategy: the range heuristic, relying primarily on the size of the smallest and largest items on each trial; or whether they instead had some direct access to the variability of the array (akin to claims of direct access to mean size). The range heuristic is a feasible strategy because, on average, range is correlated with variance.

Simulation of range heuristic

To support the idea that range heuristic is a viable alternative of the random subsampling strategy, we illustrate the efficiency of the range heuristic with a simulation. The simulation was similar to what we did for the random sampling strategy reported in the Introduction. The stimuli used were identical to the previous simulation. Two arrays of items, each with N items, were shown in each trial. The task was to identify the array with a larger variability.

Instead of storing k items from each array, we simulated a case where k items were *searched* in each array, and the ranges of the arrays were estimated based on the size of the largest and smallest items in the set of k items that was searched. The array with a larger estimated range was deemed to be the test array. Because, in this case, the k items do not need to be held in mind, but only searched, this strategy is far less cognitively demanding than the random subsampling account proposed above. Is such a strategy effective?

To assess how effective this range heuristic would be, 20,000 samples were simulated at each level, and the percentage of correctly selecting the test array reflects the efficiency of the range-heuristic. As can be seen in Figure 2b, when there were no differences between the standard and test arrays, the rate of selecting the test array did not increase with number of items sampled (as expected). When there was a moderate difference between the two arrays (difference = 0.4 *SD*), searching a set of only five items in each array and estimating the range of these items (total items searched = $2k = 10$) yielded an accuracy of 72.6%. When the difference between the two arrays is large (difference = 0.8 *SD*), employing the range heuristic on 10 searched items

yielded an impressive 85.5% accuracy. The rates of improvement are slower than utilizing $2k$ items in working memory for random subsampling (see Figure 2a), but in this case, only a simple visual search is needed. The entire set of display need not be held in working memory. Instead, only the largest and smallest items of each array have to be stored in working memory. That is, regardless of number of items searched, only a total of four items from both arrays need to be stored in working memory. Hence, working memory requirements for the range heuristic are very low compared to the random subsampling method, and are hence within the limits of working memory capacity.

Thus, if participants could search 5–10 items per side of the display for the largest and smallest items, they could accurately perform variability discrimination using what we could consider a smart subsampling strategy rather than a purely parallel mechanism.

Behavioral evidence of range heuristic

This kind of visual search has the potential to be a cognitively feasible strategy. Previously, studies have shown that people are quite efficient at visual search for the smallest or largest items in an array. Response times in the search are hardly affected by an increase in the number of distractors in the array when participants need to search for the largest or smallest item, and thus large numbers of items being searched is quite feasible. This is especially true when the target size is made known to the participants, the distractors are largely homogeneous, and the target is linearly separable from the distractors in the feature space (Hodsoll & Humphreys, 2001; Hodsoll, Humphreys, & Braithwaite, 2006). However, in cases when the distractors are heterogeneous, or when the smallest and largest target items are similar to the distractors in the feature space, the predictions are less clear (Duncan & Humphreys, 1989). In our task, participants could make use of a relatively efficient relational strategy (e.g., "find the largest"; Becker, 2010), suggesting this is likely to be a quite efficient search—and in any case, searching a set of k items is significantly easier than computing and remembering each of their sizes.

To test whether participants actually utilize the range heuristic in estimating variability, on each trial, two 16-circle arrays were shown, one on the left, the other on the right. The log-transformed mean diameters of the two arrays within a trial were always the same. One of the two arrays was the *standard* array, in which the circles had a 0.1-log-pixel standard deviation in log-transformed diameter. The other array was the *test* array, in which the circles were more variable in size. Participants were instructed to select the array that was more variable (i.e., the array in which the circles were more different from each other).

We manipulated the *range-variance congruency* across trials. Recall that the *test* array was always more variable compared to the *standard* array. In the congruent trials, the test array also had a larger range compared to the standard array. In the incongruent trials, the test array had a *smaller* range compared to the standard array. The spatial positions (left/right) of the test and standard array were randomized, so the participant's task was always a two-alternative forced choice in both range congruent and range incongruent conditions.

The first experiment aimed to show that (a) participants can perform variability discrimination more accurately than would be expected under the random subsampling account, which predicts very poor performance, and (b) participants make use of the range heuristic, a form of smart subsampling, to perform the discrimination task. A substantial drop in accuracy is expected when there is a range-variance incongruency.

Methods

Design

Experiment 1 was a complete within-subjects design. Range-variance congruency was manipulated within subjects, such that half of the trials were range-variance congruent, the other half were incongruent. Trials also varied in terms of the standard deviation difference between the standard and test arrays. For a particular trial, the difference in log-transformed standard deviations between the two arrays could be any of the set $\{0.01, 0.02, 0.04, 0.06, 0.08\}$. To prevent participants from anchoring their judgments based on a fixed stable mean context (Tong, Ji, Chen, & Fu, 2015), we also varied the mean size of the circle arrays across trials. Across trials, the mean size of the circles could be small (around 40 pixels in diameter), or large (around 50 pixels in diameter).

Participants

Participants were recruited from University of California, San Diego's Psychology Subject Pool. Thirty-four participants (22 women, 12 men) were recruited, with a mean age of 20.9. Participants gave informed consent before participating in the experiment. All participants participated for partial course credit.

For the major range-variance congruency effect that we were testing, our pilot study indicated a large effect size (estimated Cohen's d around 0.9). A power analysis indicated that we could achieve a power of 80% with a

sample size $n = 12$. To ensure normal distribution of the sampling distributions, we decided to include at least 30 participants in each of our experiments.

Apparatus and stimuli

The circle arrays were shown on a Dell E173FPc 17-in. LCD monitor with 4:3 aspect ratio. At a viewing distance of approximately 60 cm, the monitor's visible area was 31.2° wide and 25.4° tall in visual angle. The resolution of the screen was set to 1,024 × 768.

Twenty pairs of circle arrays were generated to serve as the stimuli (see Supplementary Appendix). The same set of stimuli was shown to all participants. It included combinations of two levels of range-variance congruency and five levels of standard deviation difference between the standard and test arrays.

Two circle arrays were presented simultaneously, side-by-side (Figure 1). Each array contained 16 circles, positioned on a 4 × 4 imaginary grid. Individual circles had diameters between 29 to 67 pixels. At a viewing distance of approximately 60 cm, the diameters of the circles translated to 0.91° to 2.09° in visual angle. Adjacent circles within the same array had a mean center-to-center distance of 100 pixels (3.1°). The exact position of each circle was jittered such that the circles did not appear to be on a static grid, with the constraint that the circles did not overlap. The locations of the circles within each array were randomized. Center-to-center distance between the two arrays was maintained at 512 pixels (15.9°).

Procedure

Participants were shown the instructions of the experiment, followed by a short quiz to make sure they understood the instructions. Upon passing the quiz, participants were instructed to sit at an arm's length from the computer monitor. The experiment started with a practice block with 10 trials. Each trial began with a fixation cross at the center of the screen for approximately 500 ms. The exact duration was drawn from a Gaussian distribution with a mean of 500 ms and a standard deviation of 100 ms. Two circle arrays then showed up for 1500 ms. Only arrays with the two largest standard deviation difference levels (i.e., 0.06 and 0.08) were used during the practice block. The two arrays were replaced by a static mask once the time had elapsed. Participants then pressed one of the two buttons on a standard keyboard to indicate which of the two arrays had a larger variability. A sound was emitted to indicate whether the decision was correct. A new trial then began. Participants had to achieve at least 70% accuracy in the practice block to proceed, otherwise they were asked to repeat the practice block.

Each experimental block contained 40 trials, with each of the 20 stimulus pairs being presented twice. The order of these arrays was randomized. Feedback was given throughout the experiment.

Participants began by doing an experimental block as another form of practice, the results of which were not analyzed. Then, to provide incentives for good performance, participants were told that the length of the experiment depended on how well they performed. Specifically, if participants achieved an accuracy below 65% in a particular block, an extra block of trials would be appended to the end of the experiment. All participants went through at least 10 experimental blocks of trials, with a cap of 15 blocks within each experimental session.

After the variability discrimination task, participants were given a size estimation task similar to that of Brady and Alvarez (2011). In each trial, participants were shown 32 circles of various sizes for 500 ms. They were instructed to estimate the mean size of the circles. Upon disappearance of the circle array, a sliding scale showed up. Participants adjusted the sliding scale to create a circle on the screen that matched the mean size of the circle array that was shown previously. Across trials, circles on each display had a log-transformed standard deviation of 0.11, 0.12, 0.14, 0.16, or 0.18 pixels.

Upon completion of both the variability discrimination and size estimation task, participants were given an exit survey. The survey asked them if they used any strategies for the two tasks. Participants were then debriefed about the purpose of the study.

Results

Data from the practice block and the first experimental block were excluded from analysis. Participants' performance at the predefined within-subject levels was aggregated. The results are shown in Figure 3. To look at the effect of range-variance congruency as well as standard deviation difference, the data were submitted to a 2 × 5 ANOVA, with range-variance congruency and differences between arrays as the factors. Importantly, accuracies on the range-variance congruent trials were significantly higher ($M = 73\%$, 6.4%) compared to those of incongruent trials ($M = 66\%$, $SD = 5.3\%$), $F(1, 33) = 28.5$, $p < 0.001$, Cohen's $d = 0.92$. The results showed that participants relied on the range of circle sizes for variability discrimination. In addition, accuracy improved as standard deviation differences increased, $F(4, 132) = 118.1$, $p < 0.001$. The interaction between the two factors is also significant, $F(4, 132) = 4.7$, $p = 0.001$, indicating differential reliance on the range heuristic across levels of standard deviation difference, likely due to the failure

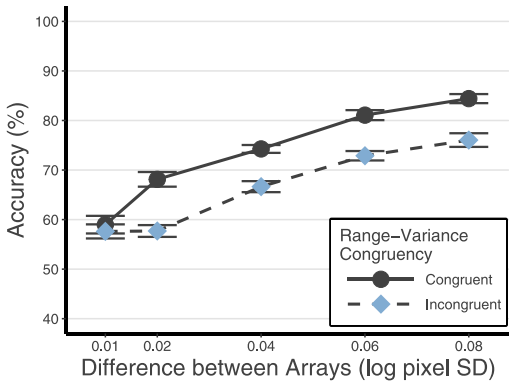

Figure 3. When range conveyed the same information as the variance, participants were more accurate on variability discrimination. The points are jittered slightly along the *x*-axis to provide clarity. All error bars in this paper denote within-subject standard errors of the means.

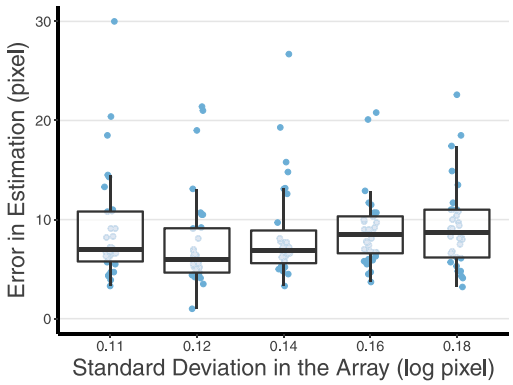

Figure 4. Performance was high in the mean size estimation task. An increase in size variation of the circle array did not hamper mean size estimation.

to discriminate when the two arrays had very similar variabilities.

Thirty participants completed the mean size estimation task (Figure 4). Performance was high. The mean absolute deviations in their estimates ranged from 7.83 to 8.99 pixels. A one-way repeated-measures ANOVA did not reveal any differences between the standard deviation conditions, $F(4, 116) = 1.558$, $p = 0.19$. The result is consistent with previous research (e.g., Allik et al., 2013; Chong & Treisman, 2003).

Discussion

Experiment 1 showed that participants had access to size variability in an array, as performance on the variability discrimination task was well above chance level (50%). The experiment also showed that arrays with a larger range were more likely to be judged as more variable especially when variability and range conveyed congruent information. This clearly indicated that the range of circle sizes was used as a proxy for variability, providing evidence in favor of a heuristic account of how participants extract size variability. Furthermore, because feedback was given throughout the experiment and yet participants continued to use the range as a proxy even when it was misleading, this reliance on range appeared to be obligatory or at least difficult to overcome. Nevertheless, even in range-incongruent arrays, participants maintained some ability to discriminate variability, suggesting that participants did not entirely rely on the range of the items in estimating variability.

Experiment 2

Experiment 1 led us to believe that some heuristics were used to indirectly perform variability discrimination. However, it is also possible that participants were affected by the range of the array despite using information from all circles (e.g., they could rely on the largest and smallest items by weighting them higher, rather than using such a visual search strategy).

If our proposed multistep approach to variability discrimination is implemented, the first step would resemble a classic visual search paradigm. From the visual search literature, we know that searching for the largest and smallest circles among heterogeneous distractors is efficient but not parallel (Becker, 2010; Duncan & Humphreys, 1989; Hodsoll & Humphreys, 2001; Hodsoll et al., 2006).

Thus, to provide additional evidence for this account of why range mattered to participants, in Experiment 2A, we manipulated the spatial arrangements of the circles in an attempt to influence this visual search process. In half of the trials, circles with the greatest range information (e.g., the greatest distinction in size) were placed close to the fixation. With this spatial arrangement, the range of the full circle array was readily accessible, as the biggest and smallest circles could be found without eye movements (and participants tend to focus most on items near the center of array when performing search; Tseng, Carmi, Cameron, Munoz, & Itti, 2009). In the other half of the trials, circles with the least range information were placed close to the fixation. The range of the sizes in the array was therefore less accessible. With a fixed presentation time for both types of arrays, we made two hypotheses for Experiment 2A. First, as in Experiment 1, when the arrays showed a range-variance incongruency, variability discrimination performance would drop. Second, regardless of the spatial arrangements of individual circles, participants would tend to utilize

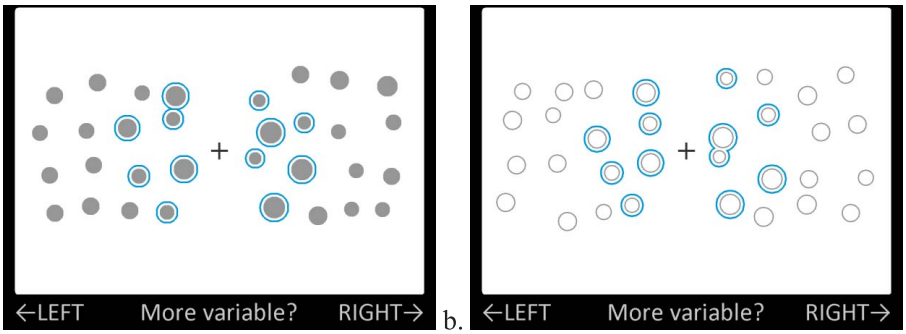

Figure 5. (a) A sample close-range trial used in Experiment 2A. The three largest and three smallest circles in each array were placed close to the fixation. The items with largest size contrast were denoted by the blue circles. The blue circles are for illustration only—they did not appear in the actual experiment. (b) Stimuli used in Experiment 2B were exactly the same as in Experiment 2A, except that the circles were outlined instead of filled. The blue circles indicate the six locations closest to the fixation, they were not shown in the actual experiment.

only the ones close to the fixation in computing the range of the display (as these would be most easily found in a visual search).

In Experiment 2B, we also examined the influence of manipulating low- and mid-level factors on size variability judgment in preparation for Experiment 3. The designs of Experiments 2A and 2B are identical, except that Experiment 2A utilized filled circles (Figure 5a), and Experiment 2B utilized outlined circles (Figure 5b).

Methods

Design

As in Experiment 1, circle arrays were used in the variability discrimination task. The design of Experiments 2A and 2B was identical to that of Experiment 1, with an additional factor. We manipulated the availability of different sized circles near the fixation. In the *close-range* condition, the three largest and three smallest circles within an array were placed closest to the fixation (marked with red circles in Figure 5a). Other circles were randomly placed in the rest of the imaginary grid. In the *far-range* condition, the largest and smallest circles were placed far from fixation, with the remaining items placed closest to the fixation. Hence, most range information was located far away from the fixation, and in the incongruent far-range trials, the items near fixation were not incongruent (e.g., the range of the items near fixation was congruent with the variability, since the incongruent items were far from fixation). Close- and far-range trials were interleaved within each block.

Simulations and predictions

Because the predicted results of this task are not entirely straightforward, we ran a simulation where we know that the computer is (a) using a range heuristic, and (b) likely to primarily sample items near fixation, in order to demonstrate the predicted pattern from this strategy. The computer was shown two arrays at a time, similar to the task given to the human participants. It was to sample one, two, or three pairs of circles that were closest to the fixation. The largest and smallest circles of the sample from each array were extracted, and the range was calculated. The array with the larger sampled range was judged to be the one with a larger variability.

The two arrays shown on each trial had either congruent or incongruent range information, and the circles with the greatest amount of range information was either placed near or far away from the fixation. Each condition was run 2,000 times. The rate of correctly picking the array with a larger variability was recorded. Figure 6 shows the aggregated results for each condition.

In the far-range condition, the largest and smallest items were far from fixation. The range heuristic was applied to a subset of circles that were near fixation, and thus the sampled range was different from the actual range of the full array. Thus, accuracy was high regardless of the number of items being sampled (left panel of Figure 6), and even when range and variance contained contradicting information in the full array, since the misleading range information was unlikely to be sampled by the simulation and thus played a minor role in performance. This shows, in general, that this more locally restricted range heuristic is a useful strategy in the incongruent condition when the range information of the full array is place far away from the fixation.

In the close-range condition, the largest and smallest items are near fixation and thus likely to be sampled. Accuracy was high in the case only when range and variance of the full array convey congruent information. When the range provided misleading information, however, accuracy dropped as more circles were

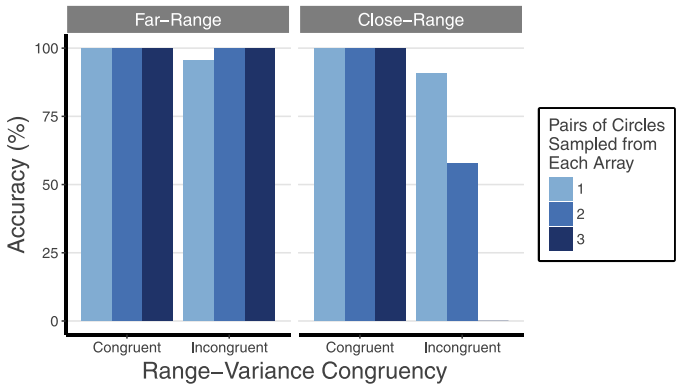

Figure 6. Simulated data examining the effects of spatial arrangement of circles and range-variability congruency on variability discrimination. The task was to select the circle array with a larger variability. In the simulation, the difference in standard deviation was 0.08 log pixel. Performance dropped only in the case when range and variance of the array convey incongruent information. In that situation, performance was negatively correlated with the number of circles sampled.

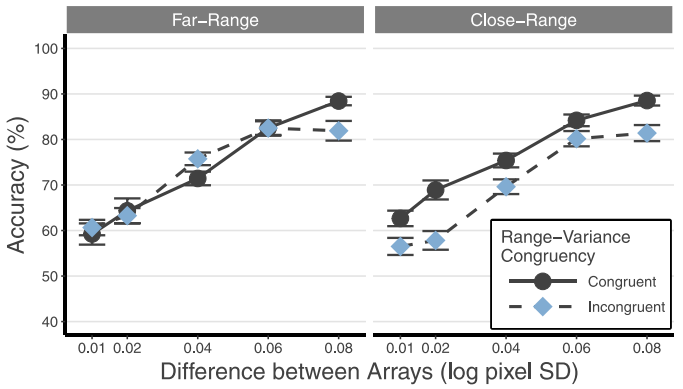

Figure 7. Accuracy between range-variance congruent and incongruent arrays did not differ when circles were homogeneous at the fixation. Range-variance congruency affected accuracy only when the largest and smallest circles were placed close to the fixation.

sampled (the bars on the far right, Figure 6). This is perhaps not surprising, as the ranges of the sampled circles were a misleading metric of the overall variability of the full arrays.

Behavioral experiment—Participants

Participants were recruited from the same subject pool described in Experiment 1. Forty-three undergraduate students (33 women, 10 men) volunteered in Experiment 2A, and 32 undergraduate students volunteered in Experiment 2B (25 women, 7 men) in exchange for partial course credit. The mean age of the participants was 20.5.

Procedure, apparatus, and stimuli

Apparatus and stimuli used in Experiment 2 were identical to those in Experiment 1. The 40 arrays used in Experiment 1 were rearranged into two possible spatial arrangements, close-range or far-range (see Supplementary Appendix). Hence, 40 pairs of circle arrays were generated. Each pair of array appeared once in the 40-trial block. As in Experiment 1, participants performed 10 to 15 blocks of trials.

In Experiment 2A, the circles in the arrays were filled. In Experiment 2B, outlined circled were used. The outline was 3-pixels thick.

Results

We performed a $2 \times 2 \times 5$ repeated-measures analysis of variance (ANOVA) with the accuracy data, with range-variance congruency, spatial arrangement, and standard deviation difference as factors.

Experiment 2A

As in Experiment 1, there are a main effect of range-variance congruency, $F(1, 42) = 6.76$, $p < 0.05$, Cohen's $d = 0.40$, a main effect of standard deviation difference level, $F(4, 168) = 129.8$, $p < 0.001$, and an interaction between the two factors, $F(4, 168), = 3.23$, $p < 0.05$. Importantly, there is also a significant interaction between range-variance congruency and spatial arrangement, $F(1, 42) = 8.02$, $p < 0.01$. When range-relevant items were far from fixation (far-range conditions), variability discrimination did not differ regardless of whether range and variability information were congruent (Congruent: 73%, Incongruent: 73%), $F = 0.03$. When range-relevant circles were close to the fixation (close-range conditions), participants were able to utilize the range of the full array as proxy for variability. As a result, worse performance was seen in the range-variance incongruent trials ($M = 69\%$, $SD = 17\%$), compared to the congruent trials ($M = 76\%$, $SD = 15\%$), $F(1, 42) = 19.1$, $p < 0.001$, Cohen's $d = 0.41$. The behavioral results are consistent with our simulation above. They suggest that participants were primarily (a) using the range, but (b) sampling mostly circles close to fixation. Results are summarized in Figure 7.

Experiment 2B

The pattern of results is similar to that of Experiment 2A. Consistent with Experiments 1 and 2A, there are a main effect of range-variance congruency, $F(1, 31) = 11.4$, $p = 0.002$, Cohen's $d = 0.60$, a main effect of standard deviation difference level, $F(4, 124) = 113$, $p <$

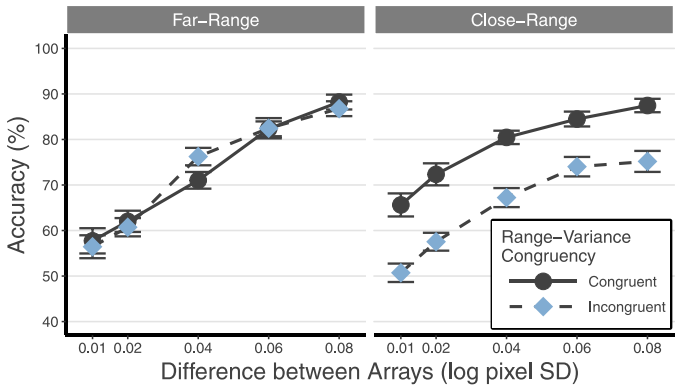

Figure 8. As in Experiment 2A, range-variance congruency did not affect accuracy when low contrast circles were placed at fixation (left panel). When high contrast circles were placed at fixation, participants depended heavily on range heuristic for variability discrimination.

0.001. Unlike in Experiment 2A, the interaction between the two factors is not significant, $F < 1$. It shows that the improvements in accuracy from increasing standard deviation difference are similar regardless of whether the range and variance were congruent.

Importantly, as in Experiment 2A, there is a significant interaction between range-variance congruency and spatial arrangement, $F(1, 31) = 33.0$, $p < 0.001$. In the far-range condition, variability discrimination did not differ regardless of whether range and variance were congruent (Congruent: 72%, Incongruent: 73%), $F = 0.01$. In the close-range condition, the range of the full array was readily available to the participants. They seemed to utilize the range information as a proxy for variability. Participants performed worse in the range-variance incongruent trials ($M = 65\%$, $SD = 16\%$), compared to the congruent trials ($M = 78\%$, $SD = 15\%$), $F(1, 31) = 35.0$, $p < 0.001$, Cohen's $d = 0.75$]. Results are summarized in Figure 8.

Comparing Experiments 2A and 2B

Experiments 2A and 2B had exactly the same design, and they only differed in whether the circles were filled. A between-subjects t test was conducted on the accuracy data, suggesting that the difficulty in variability discrimination for filled circles ($M = 72.8\%$, $SD = 5.8\%$, Experiment 2A) was comparable to that for outlined circles ($M = 72.0\%$, $SD = 5.5\%$, Experiment 2B), $t(73) = 0.61$, $p = 0.54$, Cohen's $d = 0.14$.

Discussion

Experiments 2A and 2B provided further evidence that range serves as a proxy for variability discrimination via a search process, but showed that participants do not exhaustively search for the largest and smallest item on each side but primarily focus on the items near fixation. When circles with the most distinction in size were close to the fixation (close-range), participants found these items and used them to judge the variability. When the items with the most distinction in size were far from fixation (far-range), participants may not have found these items and instead relied on the range of items close to fixation, leading to correct answers even in the incongruent case. To account for all of this data together, given that participants did not know what kind of trial they were about to see, requires that participants computed variability using a multi-stage process. First, the largest and the smaller circles that are selected, then ranges of the respective arrays are generated; participants are likely to search only items near fixation for the largest and smallest items in those spatial locations. Finally, the two ranges are compared, and the array with the larger range is chosen as the one with more variability. This process accounts for the difference between range-variance congruent and incongruent trials when circles with the most diverse sizes were placed closest to the fixation or far from fixation.

Importantly, the pattern of the data was captured by the simulations that relied entirely on range and did so primarily by sampling items near fixation. This is because the circles in the array that were sampled in the far-range conditions were not incongruent. A sampled range obtained based on the subset of the array near fixation is an accurate proxy for variability in far-range incongruent displays.

However, unlike the simulated data, human participants do not show perfect performance when range information was utilized. This suggests that participants did not sample all three pairs of circles closest to the fixation, and there may be perceptual and decision noise that was not captured by the simulations. In addition, the participants' data reveal that the range heuristic based solely on items close to fixation is not necessarily the full story. Unlike the simulation, even in the close-range incongruent condition, variability discrimination was reliably above chance performance. This shows that either some other heuristics are in place, that participants do not always find the largest or smallest item even when they are near fixation, or that variability information is available to participants through some other processes.

Experiment 3

In Experiments 2A and 2B, we show that for both filled circles and outlined circles, participants made use

of a range heuristic, but even when this range heuristic was not useful, or misleading (e.g., close-range incongruent) participants could still estimate size variability at above chance levels. Was this because they made use of a parallel size computation mechanism that could be used to compute variance? Or because other cognitive strategies were available for participants to make use of in addition to the range heuristic?

When comparing arrays of different variability, not only item-based properties like the variance of individual sizes is affected. In addition, low-level and mid-level properties like the overall texture of the array, including the density and spatial frequency of the array, are also affected (e.g., Dakin, Tibber, Greenwood, & Morgan, 2011). For example, the standard array, having mostly similar-sized circles, also had a smoother texture and fewer spatial frequencies represented compared to the test array. Thus, there may be alternative strategies for participants to estimate which array is more variable in size without explicitly representing any of the items. These strategies are in line with models that attempt to explain performance on high-level tasks by appealing to peripheral representations that summarize the array in terms of its texture (Balas, Nakano, & Rosenholtz, 2009; Chang & Rosenholtz, 2016; Ehinger & Rosenholtz, 2016; Rosenholtz, Juang, Raj, Balas, & Ilie, 2012) or sensitivity to spatial ensemble patterns (Brady et al., 2017; Alvarez & Oliva, 2009). Importantly, however, these techniques for representing an array as a set of low- or mid-level features are not truly estimating the size of the items. Thus, unlike models that propose a parallel size estimation process (e.g., Chong & Treisman, 2005), these techniques should be sensitive to seemingly irrelevant low-level properties of the array.

In other words, one possible cognitive strategy that might allow participants to estimate which array has higher variability without actually relying on estimating the items' sizes is using peripheral low- and mid-level representation mechanisms (e.g., Balas et al., 2009). The distinction between this low-level representation and an actual parallel computation of size is that the low-level mechanism predicts that differences in these low- and mid-level features are not abstracted over (as when items' sizes are computed), but are inherently part of the representation that is used to estimate variability.

Thus, in Experiment 3, we ask whether participants are impacted by being asked to compare the variability across two arrays that differ in low-level properties (filled vs. outlined). This task should be more difficult than the previous ones if they are using a representation that includes low-level or mid-level information but not if size is first extracted from all of the items (independent of their features) and then undergoes a variability estimation procedure. Thus, unlike the previous experiments in which all the circles were either filled (Experiments 1 and 2A) or outlined (Experiment 2B), each trial in Experiment 3 contained one outlined circle array, and one filled circle array. If participants utilized low-level or mid-level representations in the arrays, comparing abstract representations across these different low-level features should be difficult, even though both array types on their own allow for efficiency and equally good variability estimation (Experiments 2A and 2B).

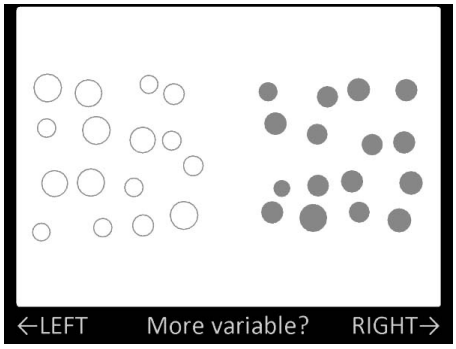

Figure 9. In Experiment 3, one of the two arrays contained outlined circles, the other contained filled circles. As in previous experiments, the participants' task was to identify the array that had a higher variability in circle sizes (i.e., the test array).

Methods

Design

The design of the experiment was identical to Experiment 1, except for an additional factor. The factor controls whether the circle array with larger variance (i.e., the test array) was outlined or filled. All factors were manipulated within-subjects. Unlike Experiments 2A and 2B, we did not manipulate the spatial arrangements of the circles. All 16 circles in each array were equally likely to be placed close to or far away from the fixation.

Participants

Thirty participants (24 women, 6 men) volunteered in the experiment in exchange for partial course credits. Participants were recruited from the same subject pool described above. The mean age of the participants was 20.5.

Procedure, apparatus, and stimuli

The same procedure and apparatus from previous experiments were used. In all the trials, an outlined circle array was shown against a filled circle array, side by side (Figure 9). In half of the trials, the filled circle array was

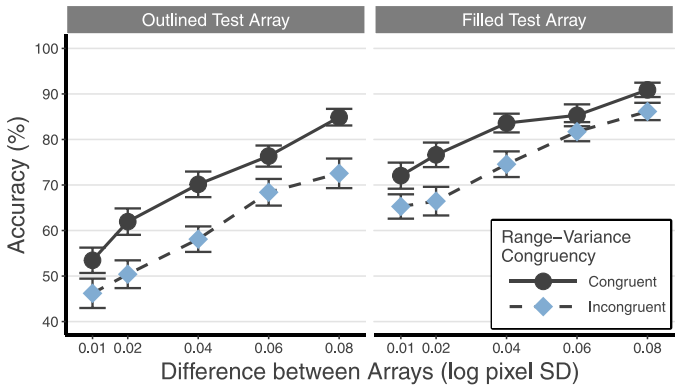

Figure 10. Outlined circle arrays have less low-level perceptual information. Participants depended more heavily on the range heuristic in that case (left panel). Filled circle arrays have more low-level perceptual information, such as luminance, spatial frequencies, and texture, so range-variance congruency effect was less prominent (right panel).

the test array, with larger variability between the two. In the other half of the trials, the outlined circle array was the test array. Filled and outlined circle arrays were equally likely to be on either side of the screen.

Results

A $2 \times 2 \times 5$ repeated measures ANOVA was used on the accuracy data, with range-variance congruency, fill (filled vs. outlined test array), and standard deviation difference as factors.

Figure 10 clearly shows that accuracy was higher when the test array was filled ($M = 78\%$, $SD = 11\%$), compared to when it was outlined ($M = 64\%$, $SD = 11\%$, comparing left and right panels), suggesting that participants could not completely abstract beyond the low-level features in variability discrimination, $F(1, 29) = 14.8$, $p < 0.001$, Cohen's $d = 0.70$. Other comparisons in the ANOVA were consistent with previous experiments. There is a main effect of standard deviation difference, $F(4, 116) = 98.8$, $p < 0.001$, indicating an increased performance as differences in variability between the standard and test arrays increased. A highly significant range-variance congruency, $F(1, 29) = 37.9$, $p < 0.001$, Cohen's $d = 1.12$, suggests that participants were more accurate when range and variance conveyed consistent information ($M = 75\%$, $SD = 6.3\%$), compared to the incongruent condition ($M = 67\%$, $SD = 5.5\%$).

An interaction between standard deviation difference level and range-variance congruency, $F(4, 116) = 3.0$, $p < 0.05$, suggests that the range-variance congruency advantage was different as the difference between standard and test arrays increased. Lastly, a marginally significant interaction between fill and range-variance congruency, $F(1, 29) = 3.4$, $p = 0.08$, suggested that the range-variance congruency advantage for outlined circles (10.2%) trended toward larger than the one for filled circle arrays (6.9%).

Comparing Experiment 3 with previous experiments

Experiment 3 was most similar to Experiment 1, in the sense that there were no manipulations in spatial arrangements of the circles. Comparing overall performance of the two experiments, we found no systematic differences (Experiment 1: $M = 69.8\%$, $SD = 4.3\%$; Experiment 3: $M = 71.3\%$, $SD = 4.6\%$), $t(60) = 1.32$, $p = 0.19$. This shows that the bias toward selecting filled test arrays in Experiment 3 was compensated by the same bias against outlined test arrays.

Difference in performance between Experiments 2A, 2B, and 3 was also found to be minimal (within 1.5%). In particular, while mean accuracy in Experiment 3 was 71.3% ($SD = 4.6\%$), that of Experiment 2A was 72.8% ($SD = 5.8\%$), $t(70) = 1.24$, $p = 0.22$, and Experiment 2B was 72.0% ($SD = 5.5\%$), $t(59) = 0.54$, $p = 0.59$. This shows that participants were flexible in adopting different strategies, using range, low-level visual information, or a combination of these strategies to attain a reasonably high-level performance.

Discussion

Experiment 3 shows that participants chose the filled circle arrays as being more variable more often, compared to the outlined circle arrays, when the two contained the same variability information. The failure for participants to abstract over low-level features when making their comparisons suggests that some low- or mid-level features in the array were used when making variability discriminations. This is consistent with the idea of using peripheral texture representations or spatial ensemble representations, among a number of other possible uses of low- and mid-level information, but not consistent with a parallel computation of items' sizes.

Consistent with previous experiments, participants exploited the fact that arrays with larger variance usually have larger ranges, and this information could be used for variability discrimination.

General discussion

In three experiments, we showed that participants could perform size variability discrimination at an above-chance level under a variety of situations. In all three experiments, we manipulated range-variance

congruency of the arrays. In half of the trials, the array with larger range in size also had a larger variability. In the other half, the array with larger range had a smaller variability. Participants were less accurate in the range-variance incongruent condition, indicating the use of range information as a proxy for variability discrimination. It is important to note that while we refer the strategy as *range* heuristic, participants might have employed other forms of smart subsampling heuristics other than using the range information.

With evidence of the range heuristic, Experiment 2A examined if the heuristic took into account all of the circles (e.g., found the true largest and smallest items), or whether spatial arrangement was important, as you might expect if participants perform a visual search task to find the largest and smallest items. In addition to the range-variance congruency, we manipulated the arrangements of circles such that sampling nearby items and applying range heuristic would result in accurate or inaccurate performance on different trials. Specifically, in half of the trials, we placed the three largest and three smallest circles in each array close to the fixation. This condition allowed participants to find the smallest and largest circles more easily, which facilitated the application of the range heuristic. In the other condition, the most similar six circles were placed close to the fixation, with the largest and smallest circles far from fixation, so that if participants relied primarily on the range of items close to fixation, they would accurately respond even when the overall range was "incongruent" with the variability. The behavioral results were consistent with our simulation, in that participants only utilized the range of items close to fixation. In particular, accuracy of the far-range condition was comparable to the close-range condition when range-variance congruent.

Experiments 2B and 3 looked at the contribution of low-level factors to variability estimation. Outlined circle arrays were used in Experiment 2B. The results were qualitatively similar to those in Experiment 2A, indicative of variability discrimination and comparable performance with both outlined and filled circles. Experiment 3 directly tested whether low-level information affects variability discrimination. We found that participants were unable to abstract over the visual features when performing variability estimation, and were biased to believe filled circle arrays were more variable as a result.

In short, we showed that participants were fairly flexible in applying heuristics to estimate the variability of a set of arrays. Participants seemed to rely on the range of items close to fixation across a variety of conditions. Low-level visual information was also used and affected variability judgments when low- and mid-level information differed between sides of the display. It is also likely that participants employed other unknown heuristics to perform the task.

Random subsampling heuristic

In the Introduction, we showed that variability discrimination cannot be performed by randomly subsampling only a few items from the arrays, but only if observers use many items from each array. We argued that subsampling cannot work the same way it does as in the case of mean size discrimination (Myczek & Simon, 2008). In the case of mean size discrimination, a single representation of the sample array can be generated. The representation can then be compared with a test array with one item. In variability discrimination, the array cannot be reduced to a single representation. Multiple items in an array have to be retained in working memory for the task. A random subsampling heuristic that implements analytical methods would simply overload working memory. Therefore, a pure random subsampling heuristic is unlikely.

Experiments 2A and 2B showed that participants were not always utilizing all the information in the array for their decisions, as they were seriously affected by the spatial positions of the smallest and largest items. The general accuracy pattern mirrors that of our simulation. This indicates that only partial information in the array was utilized, or that information near the fixation was weighted differentially. Together, these pieces of evidence support a "smart" subsampling view on variability discrimination, in which people are using a set of efficient but not perfectly parallel heuristics to perform the task. These smart subsampling heuristics provide information about the variability of the set without directly computing the variance.

Parallel or serial variability extraction

Instead of a pure parallel process, as Ariely (2001) and Chong and Treisman (2003, 2005) advocated for size discrimination, or a pure random subsampling process (Myczek & Simon, 2008), our experiments suggest that variability discrimination involves a variety of different strategies. The use of low-level or mid-level visual properties for variability discrimination can be carried out quickly, and hence be considered a parallel process. However, the lack of invariance to low-level properties suggests that low-level or mid-level information is not directly used to estimate variability. Instead, this low-level visual information affects the representation of the displays in

such a way that it biases variability perception. In addition, the range heuristic requires participants to search for the smallest and the largest circles in the array. This process is not likely to be fully parallel and depends on the number of items in the array, which seems to result in participants primarily utilizing the range of items near fixation. Thus, multiple strategies seem to be at play in variability perception, and participants' choice of strategies depends on the information that is available.

Thus, we suggest that in the case of variability estimation, while people can perform the task fairly accurately and efficiently and use variance estimates to perform real-life tasks, the cognitive strategies they use to estimate variability are indirect. While useful for estimating variability in the world, they do not support the view of a dedicated, parallel process for working with the size of all of the items in the array. To put that in the context of the strawberry-screening task we described in the Introduction, our experiments suggest multiple cognitively viable strategies are utilized. They suggest that factory workers have some representations of variability, but these representations are highly unlikely to be resulted from any analytic, serial computations or random subsampling strategies. Instead, the range of the largest and smallest berries, or some low-level or mid-level representation of multiple berries on the belt, give rise to variability judgments, as used in anomaly detection. The results are at odds with suggestions that some statistical properties, such as the mean size, are always extracted efficiently. Future research may examine whether proxies are used in other types of ensemble representations, such as mean estimation, and look into other strategies that human observers employ to perform variability discrimination tasks similar to the one we reported here.

Keywords: variability discrimination, statistical summary representations, range heuristic

Acknowledgments

The authors would like to thank Sang Chul Chong, Isabel Gauthier, and Yaffa Yeshurun for their helpful comments on previous versions of the manuscript. This work was supported by a NSF CAREER Grant (BCS-1653457) to TFB.

Commercial relationships: none.
Corresponding author: Timothy F. Brady.
Email: timbrady@ucsd.edu.
Address: Department of Psychology, University of California, San Diego, La Jolla, CA, USA.

References

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

Alvarez, G. A., & 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*(2), 157–162.

Balas, B., Nakano, L., & Rosenholtz, R. (2009). A summary-statistic representation in peripheral vision explains visual crowding. *Journal of Vision*, *9*(12):13, 1–18, https://doi.org/10.1167/9.12.13. [PubMed] [Article]

Becker, S. I. (2010). The role of target–distractor relationships in guiding attention and the eyes in visual search. *Journal of Experimental Psychology: General*, *139*(2), 247–265.

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

Brady, T. F., Shafer-Skelton, A., & Alvarez, G. A. (2017). Global ensemble texture representations are critical to rapid scene perception. *Journal of Experimental Psychology: Human Perception and Performance*, *43*, 1160–1176.

Chang, H., & Rosenholtz, R. (2016). Search performance is better predicted by tileability than presence of a unique basic feature. *Journal of Vision*, *16*(10):13, 1–18, https://doi.org/10.1167/16.10.13. [PubMed] [Article]

Chong, S. C., Joo, S. J., Emmanouil, T. A., & Treisman, A. (2008). Statistical processing: Not so implausible after all. *Attention, Perception, & Psychophysics*, *70*(7), 1327–1334.

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

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

Dakin, S. C., Tibber, M. S., Greenwood, J. A., & Morgan, M. J. (2011). A common visual metric for approximate number and density. *Proceedings of the National Academy of Sciences, USA*, *108*(49), 19552–19557.

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review*, *96*(3), 433–458.

Ehinger, K. A., & Rosenholtz, R. (2016). A general account of peripheral encoding also predicts scene perception performance. *Journal of Vision*, *16*(2): 13, 1–19, https://doi.org/10.1167/16.2.13. [PubMed] [Article]

Fouriezos, G., Rubenfeld, S., & Capstick, G. (2008). Visual statistical decisions. *Perception & Psychophysics*, *70*(3), 456–464.

Greene, M. R. (2013). Statistics of high-level scene context. *Frontiers in Psychology*, *4*:777.

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology*, *17*(17), R751–R753.

Haberman, J., & Whitney, D. (2009). Seeing the mean: Ensemble coding for sets of faces. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(3), 718–734.

Haberman, J., & Whitney, D. (2010). The visual system discounts emotional deviants when extracting average expression. *Attention, Perception, & Psychophysics*, *72*(7), 1825–1838.

Haberman, J., & Whitney, D. (2011). Efficient summary statistical representation when change localization fails. *Psychonomic Bulletin & Review*, *18*(5), 855–859.

Haberman, J., & Whitney, D. (2012). Ensemble perception: Summarizing the scene and broadening the limits of visual processing. In J. Wolfe & L. Robertson (Eds.), Oxford series in visual cognition. *From perception to consciousness: Searching with Anne Treisman* (pp. 339–349). New York: Oxford University Press.

Hodsoll, J., & Humphreys, G. W. (2001). Driving attention with the top down: The relative contribution of target templates to the linear separability effect in the size dimension. *Attention, Perception, & Psychophysics*, *63*(5), 918–926.

Hodsoll, J. P., Humphreys, G. W., & Braithwaite, J. J. (2006). Dissociating the effects of similarity, salience, and top-down processes in search for linearly separable size targets. *Attention, Perception, & Psychophysics*, *68*(4), 558–570.

Im, H. Y., & Halberda, J. (2013). The effects of sampling and internal noise on the representation of ensemble average size. *Attention, Perception, & Psychophysics*, *75*(2), 278–286.

Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: From psychophysics and neurobiology to individual differences. *Trends in Cognitive Sciences*, *17*(8), 391–400.

Ma, W. J., Navalpakkam, V., Beck, J. M., Van Den Berg, R., & Pouget, A. (2011). Behavior and neural basis of near-optimal visual search. *Nature Neuroscience*, *14*(6), 783–790.

Marchant, A. P., Simons, D. J., & de Fockert, J. W. (2013). Ensemble representations: Effects of set size and item heterogeneity on average size perception. *Acta Psychologica*, *142*(2), 245–250.

Maule, J., & Franklin, A. (2016). Accurate rapid averaging of multihue ensembles is due to a limited capacity subsampling mechanism. *Journal of the Optical Society of America A*, *33*(3), A22–A29.

Morgan, M., Chubb, C., & Solomon, J. A. (2008). A "dipper" function for texture discrimination based on orientation variance. *Journal of Vision*, *8*(11):9, 1–8, https://doi.org/10.1167/8.11.9. [PubMed] [Article]

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

Norman, L. J., Heywood, C. A., & Kentridge, R. W. (2015). Direct encoding of orientation variance in the visual system. *Journal of Vision*, *15*(4):3, 1–14, https://doi.org/10.1167/15.4.3. [PubMed] [Article]

Rosenholtz, R., Huang, J., Raj, A., Balas, B. J., & Ilie, L. (2012). A summary statistic representation in peripheral vision explains visual search. *Journal of Vision*, *12*(4):14, 1–17, https://doi.org/10.1167/12.4.14. [PubMed] [Article]

Simons, D. J., & Myczek, K. (2008). Average size perception and the allure of a new mechanism. *Attention, Perception, & Psychophysics*, *70*(7), 1335–1336.

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]

Tokita, M., Ueda, S., & Ishiguchi, A. (2016). Evidence for a global sampling process in extraction of summary statistics of item sizes in a set. *Frontiers in Psychology*, *7*:711.

Tong, K., Ji, L., Chen, W., & Fu, X. (2015). Unstable mean context causes sensitivity loss and biased estimation of variability. *Journal of Vision*, *15*(4): 15, 1–12, https://doi.org/10.1167/15.4.15. [PubMed] [Article]

Tseng, P. H., Carmi, R., Cameron, I. G., Munoz, D. P., & Itti, L. (2009). Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision*, *9*(7):4, 1–16, https://doi.org/10.1167/9.7.4. [PubMed] [Article]

Ward, E. J., Bear, A., & Scholl, B. J. (2016). Can you perceive ensembles without perceiving individuals?: The role of statistical perception in determining whether awareness overflows access. *Cognition*, *152*, 78–86.

Yang, Y., Tokita, M., & Ishiguchi, A. (2018). Is there a common summary statistical process for representing the mean and variance? A study using illustrations of familiar items. *i-Perception*, *9*(1): 2041669517747297.