



Grouping by proximity and the visual impression of approximate number in random dot arrays



Hee Yeon Im^{a,c,*}, Sheng-hua Zhong^{b,c}, Justin Halberda^c

^a Department of Radiology, Harvard Medical School/Massachusetts General Hospital, United States

^b College of Computer Science & Software Engineering, Shen Zhen University, China

^c Department of Psychological and Brain Sciences, Johns Hopkins University, United States

ARTICLE INFO

Article history:

Received 31 January 2015

Received in revised form 21 August 2015

Accepted 21 August 2015

Available online 26 September 2015

Keywords:

Perceptual grouping

Numerosity estimation

ABSTRACT

We address the challenges of how to model human perceptual grouping in random dot arrays and how perceptual grouping affects human number estimation in these arrays. We introduce a modeling approach relying on a modified *k*-means clustering algorithm to formally describe human observers' grouping behavior. We found that a default grouping window size of approximately 4° of visual angle describes human grouping judgments across a range of random dot arrays (i.e., items within 4° are grouped together). This window size was highly consistent across observers and images, and was also stable across stimulus durations, suggesting that the *k*-means model captured a robust signature of perceptual grouping. Further, the *k*-means model outperformed other models (e.g., CODE) at describing human grouping behavior. Next, we found that the more the dots in a display are clustered together, the more human observers tend to underestimate the numerosity of the dots. We demonstrate that this effect is independent of density, and the modified *k*-means model can predict human observers' numerosity judgments and underestimation. Finally, we explored the robustness of the relationship between clustering and dot number underestimation and found that the effects of clustering remain, but are greatly reduced, when participants receive feedback on every trial. Together, this work suggests some promising avenues for formal models of human grouping behavior, and it highlights the importance of a 4° window of perceptual grouping. Lastly, it reveals a robust, somewhat plastic, relationship between perceptual grouping and number estimation.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

We possess the remarkable ability to nonverbally extract the numerosity of collections of multiple items through a near-instantaneous impression of approximate number. This ability can be useful in real world contexts, which often contain structures formed by groups of similar objects clustered together (e.g., trees in a forest or buildings and cars on the street, etc.). In such cases, it is often impractical to directly count items one-by-one: the number of items may be too large, separating already-counted items from not-yet-counted ones may be very difficult, the viewing time may be limited, and so on.

The situations in which we most naturally extract the approximate number of visual elements are also situations that naturally invite perceptual grouping of items into clusters. In the lab,

perceived numerosity has often been explored by presenting human observers with simplified images of multiple items that are distributed over space and asking observers to estimate or discriminate numerosity (e.g., Gilmore et al., 2013; Halberda, Sires, & Feigenson, 2006; Izard & Dehaene, 2008; Jevons, 1871; Smets, Gebuis, Defever, & Reynvoet, 2014; Whalen, Gallistel, & Gelman, 1999). Previous findings consistently show that observers can apprehend the approximate number of items from a very brief exposure (e.g., 100 ms without a mask: Izard & Dehaene, 2008; 500 ms of presentation, followed by a mask: Halberda et al., 2006; even from 66 ms of presentation, followed by a mask: Im & Halberda, unpublished data). The rapidity of numerosity estimation seems somewhat surprising given that counting takes about 300 ms per item (Simon & Vaishnavi, 1996). The mechanisms that allow us to quickly and easily perceive the numerosity of up to 100 items within a 100 ms display remain a mystery (e.g., Izard & Dehaene, 2008). The fact that counting requires 300 ms per item also motivates the suggestion that perceived numerosity of a large number of elements may be achieved relying on a distinct

* Corresponding author at: Department of Radiology, Harvard Medical School/Massachusetts General Hospital, United States.

E-mail address: him3@mgh.harvard.edu (H.Y. Im).

mechanism from that operating for serial counting of individual elements (for review, see Feigenson, Dehaene, & Spelke, 2004). Furthermore, the ability to extract the approximate number of items in visual collections is present from human infancy (Xu & Spelke, 2000), and is also shared by other animal species (Hauser, Carey, & Hauser, 2000; Meck & Church, 1983). This further suggests that there is a very basic visual mechanism for approximating the number of items in a visual display – a mechanism that does not require schoolroom teaching.

There are several features of numerosity estimation that may help to determine the underlying mechanism. Previous studies of numerosity estimation consistently find that observers underestimate the actual numerosity (e.g., Indow & Ida, 1977; Izard & Dehaene, 2008; Krueger, 1982, 1984). While underestimation is present from the very first trial (Krueger, 1982), Izard and Dehaene (2008) have shown that observers' numerosity estimations can also be calibrated such that observers adjust their estimation either to increase or decrease the amount of underestimation when they are provided with explicit feedback. The source of this underestimation remains to be described, and one possibility is that this underestimation emerges from the heuristic, or algorithm, for extracting approximate number from the visual display.

Human observers' numerosity judgments also display an inherent variability or noise that increases linearly with the signal – scalar variability (discussed as the coefficient of variation, CV: Cordes, Gelman, & Gallistel, 2001; Crollen, Castronovo, & Seron, 2011; Frank, Everett, Fedorenko, & Gibson, 2008; Le Corre & Carey, 2007; or also as the Weber fraction, w : e.g., Dehaene, 2003; Dehaene & Changeux, 1993; Meck & Church, 1983; Stoianov & Zorzi, 2012). CV reflects the normalized standard deviation of assumed Gaussian distributions for internal representations, which is inversely related to the precision of the internal representation. Therefore, the precision of numerosity estimation can be quantified by CV, with lower CV indicating more precise number estimation.

Another feature of numerosity estimation that may inform proposed mechanisms is the lack of a demonstrated upper bound for number estimation. Unlike serial counting of individual objects for small, precise number (e.g., subitizing; Trick & Pylyshyn, 1993, 1994), extracting approximate number is not constrained by the limited capacity of object-based attention. For example, observer's error rate and response time do not increase with the absolute numerosity, suggesting that extracting approximate number does not rely on the serial, limited process of object-based attention (Barth, Kanwisher, & Spelke, 2003). For these reasons, several researchers have suggested that a global process might support the estimation of approximate number, and there are many such global processes that could be relevant. For example, it has been suggested that textural information about the whole scene such as the density of elements within a given area can support the rapid extraction of large, approximate numerosity of elements in a visual array (Dakin, Tibber, Greenwood, Kingdom, & Morgan, 2011; Tibber, Greenwood, & Dakin, 2012). Such models are consistent with suggestions that numerosity is not perceived directly, that is, as an independent visual property, but rather is calculated indirectly via texture density (Durgin, 2008). Indeed, one would expect that density and numerosity would be highly inter-related in the environment (e.g., more items goes with more density).

Relatedly, it has been found that how dots are spatially organized can modulate perceived numerosity – e.g., a uniform layout of items throughout the display area results in a scene that appears more numerous than the same number of items clustered into multiple sub-groups (Frith & Frith, 1972); and dots occupying a more extended region of the display area results in a scene that appears to be more numerous than the same number of items

clustered into a smaller display region (Bevan, Maier, & Helson, 1963; Binet, 1890; Ponzo, 1928). From results such as these, it seems likely that the mechanisms that support the extraction of approximate number will involve early, rapid, global processing – perhaps with some additional later algorithms that may be attention-dependent.

The models on visual density and texture perception (e.g., Dakin et al., 2011; Tibber et al., 2012) have been popular not only because they are computationally simple and biologically plausible but also because they can very precisely predict human observer's response bias in numerosity estimation. However, the conclusions from these models may mislead one to overlook the fact that human observers are also able to perceive the visual dots in different levels of hierarchy, from individual objects (e.g., how many dots) to configuration of higher-level groups (e.g., how many clusters). Other work suggests that number judgments rely on interactions across multiple levels (e.g., groups and items).

A fourth feature of number estimation is the effect of visual grouping on number judgments. Approximate number estimation is modulated by how elements are grouped and bound together into higher-order objects. The same number of items will look more numerous when regularly arranged than when randomly distributed (Ginsburg, 1976; Taves, 1941), and random patterns look more numerous than clustered patterns (Ginsburg & Goldstein, 1987). The grouping of elements in a display also affects number estimation latencies such that several groups of dots spread out in the periphery of the display are enumerated faster than the same number of dots clustered into one group in the center of the display (van Oeffelen & Vos, 1982), suggesting that parsing of elements into subgroups may occur before enumeration of the elements. Extending these grouping effects into more advanced visual processing, it has also been shown that when higher-order objects are presented (e.g., 3D-like objects consisting of two squares and a connecting line between the squares; Franconeri, Bemis, & Alvarez, 2009; He, Zhang, Zhou, & Chen, 2009), observers tend to more drastically underestimate the number of squares than when the same number of squares are presented as disconnected “lollipops”. Note that in such cases the number of elements (e.g., squares and connecting lines), the size of elements, lower-level visual texture cues, and the overall display area, were held constant – suggesting that it is the higher-order grouping cues that drive the effect. These findings together provide evidence that visual grouping cues affect estimation of approximate number, but they do not provide a computational account for grouping and its effects on approximate number.

These features of approximate number estimation can help inform proposals for mechanisms that support the extraction of approximate number. Proposed mechanisms might provide a principled explanation for the underestimation bias, they could explicate proposals that rely on rapid and global processes, and they might include a role for visual grouping effects in numerosity perception. Because perceptual grouping organizes the visual scene into units, and because it can operate rapidly across the entire image, we focus here on the possibility that perceptual groups may be crucial higher-units for the rapid extraction of approximate number in random dot arrays.

Before our empirical investigation, we first consider the literature on perceptual grouping in greater detail. Even when there is no explicit grouping cue such as connecting lines (e.g., Franconeri et al., 2009), the visual system can organize the visual scene easily and flexibly. When similar items are randomly distributed over space, observers can readily and near-instantaneously organize the global structure from the scene by grouping items together based on proximity (Pomerantz, 1981). Visual grouping has been a significant focus of perception research since it was first emphasized by Gestalt psychologists (Wertheimer, 1924). The law of

grouping by proximity states that “when the field contains a number of equal parts, those among them which are greater in proximity will be organized into a higher unit”, (Koffka, 1935, pp. 164–165). The mental computations for proximal grouping have been suggested to operate in a purely bottom-up fashion (Pomerantz, 1983) and grouping is achieved at a pre-attentive stage of visual-processing (Neisser, 1967).

Although existing reports demonstrate observers' perceptual grouping behavior when presented with different images, formal descriptions of the underlying mechanisms have been somewhat lacking. Very reasonably, much of the evidence presented in support of perceptual grouping has focused on phenomenological demonstrations. Only a few attempts have been made to propose and evaluate formal models of perceptual grouping by proximity (e.g., the CODE algorithm proposed by van Oeffelen and Vos (1982) and evaluated by Compton and Logan (1993, 1999). Existing computational models of perceptual grouping by proximity might be improved and refined such that the models would be capable of explaining human grouping behavior more efficiently, with simpler estimation procedures (e.g., with fewer free parameters). Here, we propose and test one such approach. Our new approach to modeling human grouping behavior relies on a modified version of the k -means clustering algorithm from computer vision. Similar to the standard k -means algorithm, our window-based clustering algorithm finds a solution that partitions the dots in a random dot array into k clusters. Each dot is assigned either to a multi-dot cluster, or becomes a cluster unto itself. The critical modification for our window-based clustering algorithm is that it returns the number of clusters, k , as a result of varying the size of the grouping window throughout a range, whereas the k -means algorithm is given a k value as an input for a goal state. We will show that our window-based clustering approach provides a robust, quantitative measure of human observers' grouping assignments, determined by the grouping window size (the only free parameter for the model). In Experiment 1, we focus on describing and testing this algorithm with respect to human judgments for the number of visual groups present in random dot arrays. In Experiment 2, we extend this approach to test the effects of perceptual groups on human number judgments for the total number of elements in random dot arrays. Given the pre-attentive, global nature for perceptual grouping, we reason that perceptual groups as “higher-units” are achieved before numerical estimation and may support the rapid read-out of a visual array containing multiple items to be enumerated. The relevance of the perceptual grouping problem for the study of number estimation is highlighted by the previous work, in that the stimuli used for numerosity estimation usually contain similar items that are randomly located over space (e.g., Choo & Franconeri, 2014; Franconeri et al., 2009; Halberda et al., 2006; Izard & Dehaene, 2008; Miller & Baker, 1968; Whalen et al., 1999).

Previous research, reviewed above, suggests that visual numerical approximation is likely to involve multiple levels of processing – beyond just early visual density – and suggests that perceptual grouping may affect numerical judgments. But, previous research has not provided a quantitative analysis of the relationship between grouping and number estimation. Given the constructive nature of the visual system, we reason that perceptual grouping based on proximity is likely to occur prior to enumeration, and that it may provide some of the inputs for the rapid extraction of approximate number. To test these ideas, we explored human observers' grouping behavior and how their grouping judgments might systematically affect their numerosity judgments.

To summarize our investigations, we first show that our new clustering approach provides an accurate fit to human observers' judgments of the number of clusters within random dot arrays (Experiment 1). We next investigate the relationship between

grouping and number by varying the clustering index of stimulus images (as measured by our clustering algorithm) and measuring observers' numerical estimation biases and precision (Experiment 2). We then investigate the robust nature of grouping effects on numerical cognition by providing explicit feedback for numerical judgments and looking for the effects (and non-effects) of learning on numerical judgments (Experiment 3). We also compare our numerical estimation algorithm to existing models for both perceptual grouping by proximity (e.g., CODE, van Oeffelen & Vos, 1982) and perceived numerosity (e.g., the Occupancy Model, Allik & Tuulmets, 1991) and demonstrate that the current modeling approach is accurate and generalizable in explaining how human observers group multiple items and how they enumerate. Based on our results, we propose that rapid extraction of approximate number can be achieved relying on a fast, global mechanism that is affected by a clustering pattern for perceptual groups.

2. Experiment 1

In Experiment 1 we measured observers' estimates of perceptual groups in arrays of randomly positioned dots and modeled these responses.

2.1. Method

2.1.1. Subjects

10 naive undergraduate students from Johns Hopkins University participated in the experiment for course credit. All of the subjects had normal or corrected-to-normal vision. Informed consent was obtained for experiment from the participants in accordance with the Declaration of Helsinki. The experimental protocol was approved by the Institutional Review Board of Johns Hopkins University.

2.1.2. Apparatus and stimuli

The stimuli were generated using MATLAB software, together with the Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997), and were displayed on a 17-in. LCD monitor driven by a Macintosh iMac computer. The subjects were seated approximately 50 cm from the screen and viewed the display binocularly. At this viewing distance, each pixel was approximately 0.0213° of visual angle. The stimuli were presented on a gray background and consisted of multiple blue dots (5–35 dots) each of which subtended 0.96° of visual angle. Locations of the dots were randomly chosen for each of the 180 visual images, displayed within the virtual gray area, subtending $16^\circ \times 20^\circ$ of visual angle. The overall range of dot density in the stimulus images was 0.02–0.11 dots/deg², which falls within the range in which the mechanisms for numerosity (not density) governs the numerosity judgments and Weber's law holds for numerosity discrimination (Anobile, Cicchini, & Burr, 2014). The recent study from Anobile et al., 2014 has suggested that there exists two separate mechanisms (i.e., one for numerosity and one for density) for processing random dot arrays, and that which is more dominant will vary according to the density of display. Anobile et al. (2014) showed that the dominant mechanism switches from numerosity to density at the key density of 0.2–0.3 dots/deg². That is, when a stimulus contains more than 0.3 dots/deg², visual processing of the stimulus relies more on the mechanism for density, whereas when the dots are more sparsely distributed than this, visual processing of the stimulus relies more on the mechanism for numerosity. Based on this previous finding, the dots in our visual displays seem to be very sparsely distributed, which is suitable for numerosity task.

The mask images were created to be effective by ensuring that the mask stimuli contained the same color feature as the stimulus

images (e.g., blue outlined circles: Fig. 1): this has been suggested to increase the chance to interfere with or prevent processing of a stimulus image (Haber, 1970). In order to avoid any confusion caused by adjacency in its perceived appearance (Haber, 1970), we also ensured that the mask images had up to 120 outlined circles that were randomly located and overlapping. This approach, of appropriately choosing a visual mask image, can help us interpret the data from the varying viewing times used in Experiment 1 – i.e., controlled by the onset of a visual mask image, the availability of stimulus information will “stop” at the point of the mask (e.g., from 50 to 320 ms post stimulus onset), and processing of the perceptual groups will be based solely on the information available up to that point (Schultz & Eriksen, 1977; Sperling, 1963).

2.1.3. Procedure

Fig. 1 illustrates a sample trial of the experiment. All 10 participants were presented with the same 180 stimulus images containing dots randomly located. The images were generated in advance and presented to each participant in a different sequence. After a ready signal, the stimulus array containing multiple dots was presented for varying durations (from 50 ms to 320 ms), followed by a mask array (Fig. 1). Although all the participants were shown the same 180 stimulus images, the display duration of stimulus images was randomly varied between 50 ms and 320 ms across the participants. After the mask array, a response screen was presented until a participant made a response. The response screen contained a continuous, linear scale with tick marks at every discrete value between 1 and 40 (Fig. 1). Participants were instructed to click anywhere on this response scale to make a response, using a mouse cursor. That is, participants were free to click any fractional value between 1 and 40 on the continuous response scale.

The task instructions given to the participants for Experiment 1 were minimal. Written instructions were read out loud to each participant as follows:

“On each trial, you will see many dots on the screen. After the stimulus display disappears, your task is to guess how many groups of dots were presented in the display. Please make sure that you don’t count any dots serially. Instead, please try to see the whole image and try to make a reasonable guess of how many groups of dots were presented. Click on the linear scale on the bottom of the screen in order to indicate how many groups of dots were presented. If you are not sure, please make guess. There is no

correct answer for this task, so you can count the groups in whatever way you feel the most comfortable and natural. If you feel like none of the dots belongs in a group, each dot can be counted as an independent group.”

The participants received 10 trials for practice and 180 trials for test. Since there was no right or wrong answer, no feedback was provided.

2.2. Modeling

In order to formally assess the perceptual groups in visual images, we used a modified version of the k -means clustering algorithm, which is one of the popular techniques for cluster analysis in computer vision. The k -means clustering algorithm has been applied to various domains such as computer vision (Ray & Turi, 1999), market segmentation (Chaturvedi, Carroll, Green, & Rotondo, 1997), and geostatistics (Honarkhah & Caers, 2010). The basic algorithm for k -means clustering is as follows (see also Fig. 2a):

- (1) k -Initial mean locations (i.e., centroids) are randomly generated by the algorithm (e.g., in Fig. 2a, three means are randomly generated within the data domain).
- (2) Each observation (e.g., each dot) is assigned to belong to one of the centroids by determining which yields the nearest distance, creating k -clusters.
- (3) The centroid location of each of the k -clusters is recalculated by the locations of its elements and replaced by the new mean, aiming to partition the dots into k -clusters to minimize the within-cluster sum of squares (WCSS).
- (4) Steps (2) and (3) are iterated until convergence has been reached such that the WCSS has been minimized.

Beginning from this basic algorithm of k -means clustering, we made an important modification for a window-based clustering algorithm (Fig. 2b). While the conventional k -means algorithm engages the number of clusters k as a free parameter, our window-based clustering algorithm returns k as a result of the algorithm. Intuitively stated, this change allows us to ask of each image “how many clusters are in this image, given a grouping window of size of T_d .” In our window-based clustering algorithm, each dot is initially in its own cluster (i.e., $k = N$) since the size of the

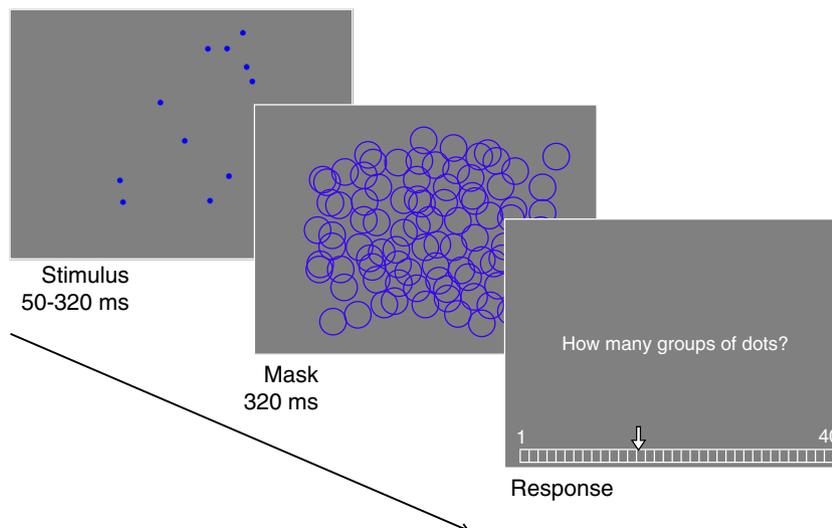


Fig. 1. Trial sequence and stimuli in Experiment 1. The stimulus containing multiple dots was presented for varying durations (50–320 ms) and followed by a mask stimulus that lasted for 320 ms. Participants were asked to report how many groups of dots were presented in the stimulus array by choosing a value between 1 and 40 on the continuous linear response scale, using a mouse cursor. The response array was presented until participants made a response.

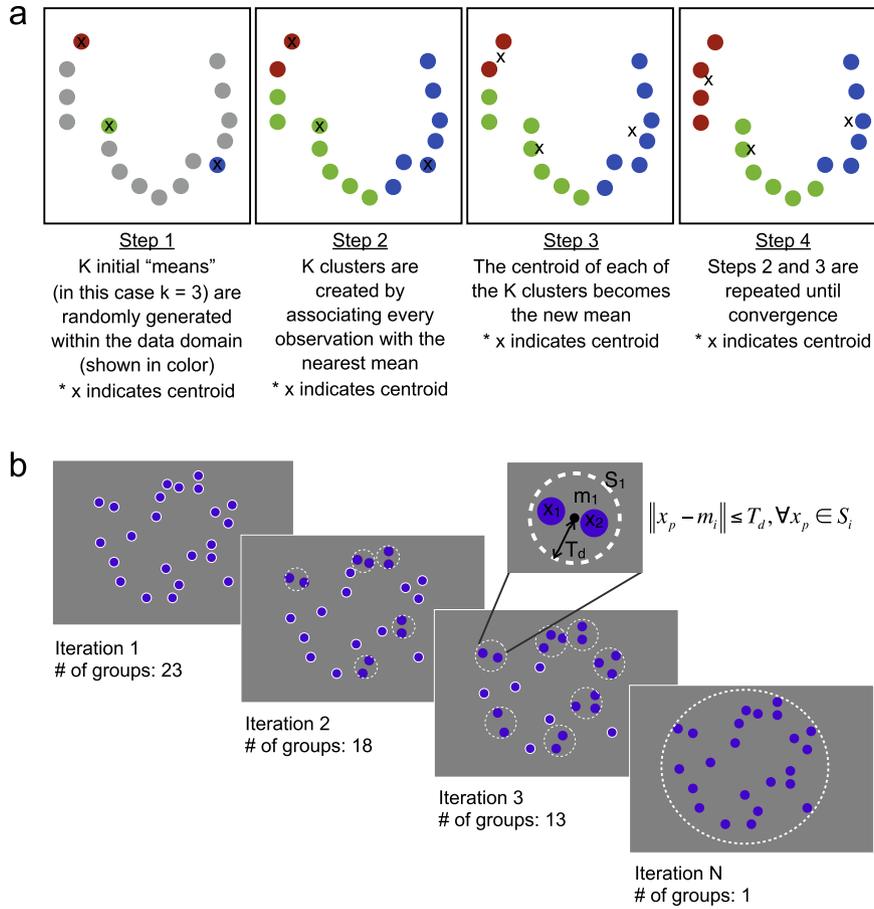


Fig. 2. Schematics of clustering algorithms. (a) The demonstration of the standard k -means algorithm. (b) The current algorithm of iterative clustering with a free parameter for the window size T_d . For each iteration, a modified k -means clustering algorithm was applied to predict the best estimates of the number of clusters given T_d . To fit each human observer's response, the model provides the best estimate among the T_d 's from the N iterations of clustering.

clustering window is as small as the size of each dot. Next, the algorithm increases the size of clustering window with each iteration, and this will have the result that the number of clusters is reduced (or remains the same) after each iteration. The window-based clustering algorithm partitions the N dots into k -clusters ($k \leq N$) by increasing the size of clustering window, until the corresponding partition fulfills the constraint that all of the dots are assigned to one of the clusters, as can be seen in Eq. (1):

$$\|x_p - m_i\| \leq T_d, \quad \forall x_p \in S_i, \quad (1)$$

Here, x_p is the location of each dot that belongs to a given cluster i ; m_i is the mean location of the cluster i ; T_d is the clustering radius in which all the dots closer to the centroid of a cluster with T_d are assigned to the cluster; and S_i indicates each cluster with an index of i . T_d is a constant, varying every iteration such that the size of clustering windows remains the same across all the k clusters per iteration. During this step, each dot can be assigned to one of the clusters only when the distance between the center of the dot and the centroid of the clustering window is smaller than T_d , and the algorithm varies the number of clusters k until all the dots in the visual array is assigned to one of the clusters while fulfilling the window constraint.

Similar to the conventional k -means algorithm (e.g., Steinhaus, 1956), our window-based clustering algorithm proceeds by alternating between assignment step and update step:

- (1) Assignment step: Assign each dot to the cluster with the closest mean,

$$S_i^{(t)} = \left\{ x_p : \|x_p - m_i^{(t)}\| \leq \|x_p - m_j^{(t)}\|, \quad \forall i \neq j, \quad 1 \leq j \leq k^{(t)} \right\} \quad (2)$$

- (2) Update step: Calculate the new means to be the centroid of the dots in the cluster,

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j \quad (3)$$

where i and j indicate indices of clusters, and t indicates each iteration step. The clustering threshold T_d is the clustering radius, and the model determines the best estimate of the number of clusters k at a given T_d value. The clustering threshold can be directly transformed to a clustering window diameter: $W_d = 2 \times T_d$. We find it intuitive to discuss results in terms of the clustering window diameter W_d , because W_d can directly index grouping scale: if W_d is large, more and more disparate items will be grouped together while if W_d is small, fewer and fewer items will be grouped.

Although the assignment step and update step that we implemented in our window-based clustering algorithm mostly resemble those in the conventional k -means algorithm, our window-based clustering algorithm has a crucial difference from the standard k -means algorithm in that the fit we determine to the input data does not depend on "random" choices each time the algorithm returns a k value. The critical difference comes from the fact that our window-based clustering algorithm does not fit the k value as

a free parameter, but returns a k value as an objective result of the varying T_d value. This allows us to determine a consistent and robust model prediction for the number of clusters in each stimulus image.

The steps above describe how our algorithm determines the dot assignments and cluster centroids that result in the number of clusters for any specified W_d . One can carry out this type of process for a wide range of values for W_d , or one can search for the W_d that provides a match to some goal state. Because our human data provides us with human ratings for the number of clusters that the subjects subjectively experienced in each image, we sought to obtain the value of W_d that minimized error between human judgments and model predictions – we will report results from fitting this algorithm to individual trial data as well as to the entire set of each participant's judgments. For each stimulus image, we ran our algorithm at varying W_d values in a wide range and obtained the estimated number of clusters at a given W_d value. We then compared the model estimation to the number of clusters that the participants had estimated for the image by calculating the model prediction error as an absolute difference between the model prediction and human estimation. When a range of different values of W_d yielded the same prediction for the participant's response for a given image, the algorithm returned the mean value from this range as its final estimate for W_d for the image. The range of the W_d values to be tested and the step size of the W_d increment within the range were determined by a random walk search procedure to improve the efficiency.

Our modeling approach allows us to provide a fit to human grouping behavior. For example, perceptual groups can be organized at different scales (e.g., lower levels of a hierarchy can encode more specific details of an image while higher levels may organize these details into more global structural units; Palmer, 1975). In Fig. 3, for example, an observer may report experiencing only one cluster including all the dots in an array. But there are other grouping possibilities: a different observer might experience two perceptual groups – one group of 10 dots and another with 7. Or another observer might report seven groups of 2 or 3 each, when they “zoom-in” – that is, when they parse finer-grained details as forming perceptual groups. Likewise, in computer modeling, perceptual grouping by proximity can operate under a

“tight” grouping criteria – with only the closest elements being grouped – or under a “loose” grouping criteria – with elements grouped over longer distances. The free parameter W_d of our model – which estimates grouping window size for each observer for each image – provides an estimate that quantitatively describes human observers' grouping patterns (e.g., how tightly or loosely they group items in a visual array).

2.3. Results

In Experiment 1, participants were instructed to group dots in whatever way they felt the most natural, easy, and comfortable for them. We first looked at consistency in the behavioral judgments across observers. Since there was no correct answer for the images, we compared responses from the 10 subjects who were tested on the same 180 images and examined whether participants' responses agreed with one another. Fig. 4 demonstrates example scatter plots comparing the first two and the last two observers, and Table 1 reports the R^2 values for all pairwise correlations between observers. The x - and y -axes in Fig. 4 indicate different participants' estimations of the number of groups from the same images. The diagonal line indicates where the two participants' estimations are perfectly in agreement with each other. As can be seen by the positive correlations in the scatter plots, the pairs of individual participants highly agreed with each other in their estimation of the number of groups in each image (Fig. 4). Table 1 shows all of the pairwise comparisons across participants (i.e., each subject compared to each). The R^2 values in Table 1 suggest fairly uniform agreement in responses across subjects. Thus, despite the open-ended nature of the task, and the fact that the subjects could choose their own criterion for how to group items, the grouping patterns were highly consistent across individuals. This suggests that all observers found this to be a natural and intuitive task, and it also suggests that subjects may be using similar grouping criteria.

The agreement across observers raises the possibility that there is some common metric that observers are using to group items into groups – perhaps describable as a default grouping window size, W_d . If participants rely on a consistent and shared grouping window size (W_d) across trials, we may be able to identify this agreement by fitting a grouping window to each judgment. We fit our window-based k -means clustering model (with the single free parameter for grouping window size, W_d) to the participants' responses. For this first modeling effort, we allowed every human response to be fit by its own W_d . That is, the best-fit value of W_d , which minimizes the deviance between model prediction and human estimation of the number of groups on each stimulus image, was determined for each image and for each individual subject. Fig. 5a displays a histogram of all the best-fit grouping window sizes from all the presented stimulus images and from all the subjects (≈ 1800 trials). There is a clear peak at approximately 4° of visual angle. This means that most human responses seemed to emerge from a grouping window size of around $W_d = 4^\circ$. Because the stimulus arrays involved randomly positioned dots, it is not obvious from the stimuli that this distribution would be normally distributed around 4° . With this in mind, it is noteworthy that the majority of the best-fit estimates of grouping window size fell around 4° of visual angle (mean = 3.91, median = 3.93). This suggests that there may be a psychological default grouping window size that operates across our images of randomly positioned dots.

We next assessed the consistency of a default grouping window size across observers. For each observer, we varied the value of W_d for all images in the stimulus set and we calculated the unsigned error between the predicted number of clusters and the observer's response for each image. This is equivalent to searching for the single value of W_d that provides the most accurate fit to the entire set

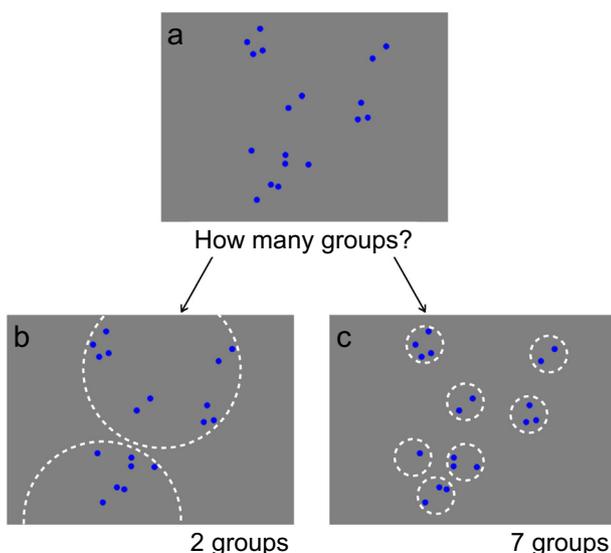


Fig. 3. Schematic example of the hierarchical structure of perceptual groups. For the same image (a), perceptual groups can be organized at different levels into either two larger groups (b) or 7 smaller groups (c).

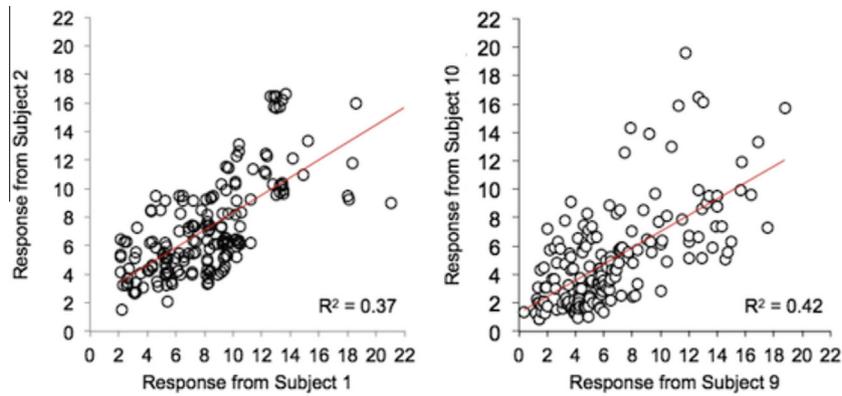


Fig. 4. Scatter plots of responses from three representative participants. The estimation of the number of groups of dots was highly consistent across the individuals.

Table 1

R² values for all pairwise correlations between observers' estimation of the number of groups in Experiment 1.

| | S1 | S2 | S3 | S4 | S5 | S6 | S7 | S8 | S9 | S10 |
|-----|----|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| S1 | | 0.37** | 0.48** | 0.18* | 0.45** | 0.40** | 0.36** | 0.50** | 0.45** | 0.22** |
| S2 | | | 0.69** | 0.26** | 0.53** | 0.39** | 0.37** | 0.22** | 0.42** | 0.33** |
| S3 | | | | 0.12* | 0.24** | 0.14* | 0.38** | 0.19* | 0.34** | 0.32** |
| S4 | | | | | 0.43** | 0.11* | 0.13* | 0.10 | 0.26** | 0.32** |
| S5 | | | | | | 0.38** | 0.31** | 0.45** | 0.33** | 0.36** |
| S6 | | | | | | | 0.15* | 0.23** | 0.28** | 0.21** |
| S7 | | | | | | | | 0.32** | 0.34** | 0.20** |
| S8 | | | | | | | | | 0.21** | 0.29** |
| S9 | | | | | | | | | | 0.42** |
| S10 | | | | | | | | | | |

* Indicates $p < .05$.

** Indicates $p < .01$.

of responses for a subject (i.e., one free parameter per subject). In Fig. 5b, we display the average unsigned error for each subject (with different colored lines) across the stimulus set for a range of values of W_d . As shown in Fig. 5b, the match between the model-predicted number of clusters and the observer's responses showed a marked decrease in error at a grouping window size of around 4° of visual angle for each observer. This means that all 10 observers appeared to have similar grouping window sizes as their default setting. It is perhaps surprising that all 10 observers showed similar grouping patterns given that they were not provided with any specific instructions about how to group. It may be that the grouping pattern that we quantify here as a default grouping window size of roughly 4° of visual angle for each observer reflects a universal feature of perceptual grouping. In support of this conjecture, if we assume that every observer had a grouping window size of 4° for every image we get a strong linear relationship between the model-predicted number of clusters and the human responses for number of clusters (Black dots in Fig. 5c). The R^2 for the regression between the average of human responses and model prediction was 0.99 ($p < .01$). In Fig. 5c, gray dots indicate the mean values for human participants' responses for each of the 180 images – to illustrate the variance across images that the model classified as having the same number of clusters. We found that, while there was some variance among the human judgments for different images, this variance was small relative to the good fit between the model and the human judgments – e.g., low spread of the gray dots (i.e., 180 individual images) around the regression line. This approach also allowed us to look at which images returned human judgments with the greatest deviation from the model-predicted number of clusters. In Fig. 5c, we present two stimulus images that resulted in differing human judgments (i.e., higher and lower than model prediction) while the model predicted the same number of clusters for these images. Human

participants estimated 5.5 clusters for the image with a red outline and 3.0 for the image with a blue outline, whereas the model predicted that both images have four clusters. It is clear simply from first impressions that the clusters in these images differ – even though our current k -means algorithm classified them similarly. This suggests that there will still be interesting variance to capture beyond our model fits.

The design of Experiment 1 also allowed us to ask whether grouping criterion changes over time. We compared estimates of grouping window size for each participant at each stimulus duration (from 50 ms to 320 ms). That is, the trials at each of 4 stimulus durations were sorted and then each group of trials was fit with a different value for W_d for each observer (i.e., 4 best-fit values of W_d per observer). Fig. 5d shows the time course of the best-fit estimates of grouping window size, averaged across participants. The estimates of grouping window size remained stable with duration as brief as 50 ms with little change over time, suggesting that a stable grouping was possible from the perceptual evidence contained in within a 50 ms exposure. This also suggests that participants' grouping criteria was not highly modulated by a feedback loop between active perception and grouping, active attention or response biases that might require perceptual evidence after 50 ms. This suggests that participants used a stable grouping strategy from as little as 50 ms of evidence, and they did not change their strategy with increased viewing times.

The above results suggest that our k -means clustering algorithm provides an accurate and stable fit to human grouping judgments. To test the robustness of our estimate, we next controlled for two measures of visual density and display area in our stimuli. These are important controls as there are already published effects of density and display area on perceived numerosity (e.g., Durgin, 1995; Tibber et al., 2012; Tokita & Ishiguchi, 2010). In order to ensure that the current k -means clustering estimate does

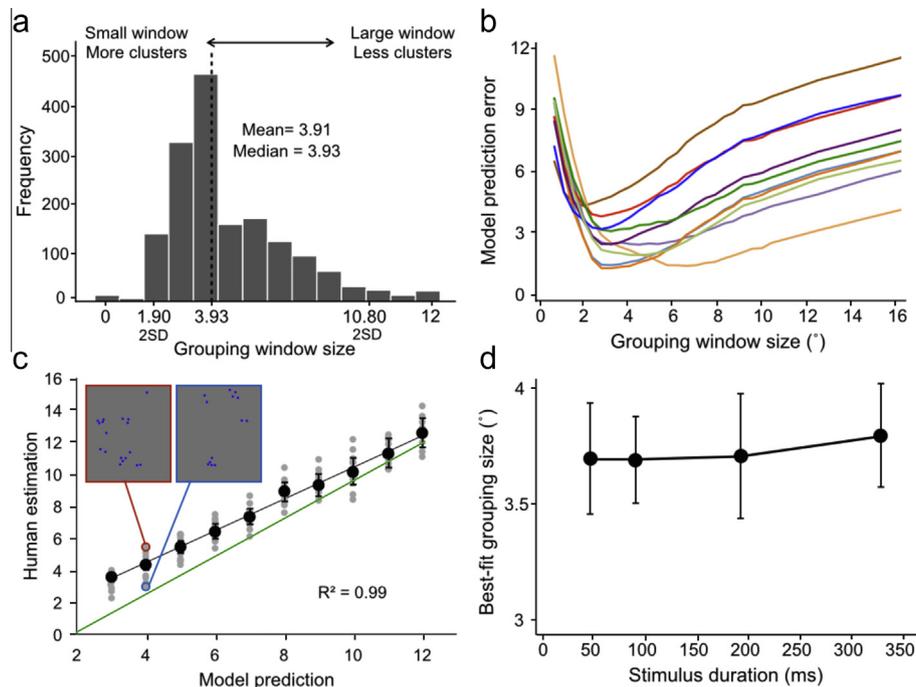


Fig. 5. Modeling results. (a) Histogram of the best-fit window size for all the images presented to all the participants, predicted by the algorithm to predict participants' grouping pattern. The mean and the median of the best-fit grouping window size was 3.91 and 3.93, respectively. (b) The model prediction error as a function of the grouping window size. Each curve represents model prediction error for each participant (color-coded by participant). For most participants, the least model prediction error was yielded at the grouping window size of about 4° of visual angle. (c) The correlation between the average of human estimates for the number of clusters and the model prediction. The black dots indicate the overall average of human responses as a function of the model-predicted number of clusters. The gray dots indicate mean responses of human participants for each of the 180 images. Two example images show the displays which yielded the most over- and under-estimation from the model prediction. The red outlined image shows a trial where humans judged the display to have more clusters than the model predicted, and the blue outlined image shows a trial where humans judged the display to have fewer clusters than the model predicted. The green line indicates equality, $y = x$. (d) Best-fit grouping window size as a function of stimulus duration (50–320 ms), provided by the current clustering algorithm to explain participants' grouping patterns. The absence of change over time suggests that the best-fit grouping window was stable, and that grouping pattern did not change over time. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

not simply parasitize on existing effects of density or area we took the statistical approach of controlling for density and display area in an analysis of the correlation between model predictions (for the number of clusters in each stimulus given a W_d of 4°) and human responses for the number of clusters.

Since density can be quantified by dividing the number of items by the display area (e.g., Durgin, 1995, 2008; Tibber et al., 2012), we first computed the number of items per visual degree for each image. These values entered into a linear regression with human estimation of the number of clusters. We saved the residual values from this regression, thereby creating a measure of human responses that is controlled for density. We next entered the number of items per visual degree into a linear regression with the model-predicted number of clusters; we saved the residuals from this regression; thereby creating a measure of model-estimates of number of clusters that is controlled for density. We were then able to ask whether there remained any significant relationship between the residuals for human responses (controlled for density) and the residuals for model-estimates of number of clusters (controlled for density). A linear regression on these residuals revealed that there was still a significant relationship between human number estimates and model-estimated number of clusters, after controlling for visual density (Fig. 6a).

Although the number of items per visual degree within the display area has been frequently used in the literature as a measure of density, it may not be an ideal measure for visual number. Because the experience of number can be affected by display area (Hurewitz, Gelman, & Schnitzer, 2006; Vos, Van Oeffelen, Tibosch, & Allik, 1988), it may be that a measure of density that is responsive to the visual dispersion of items would be preferred.

Thus, using the minimal area encompassing the target items as the area of the display may be a better approach than using the whole display region as an estimate of area. One standard way of estimating the size of the minimal area encompassing the display items is the alpha shape (Edelsbrunner, Kirkpatrick, & Seidel, 1983). An alpha shape is a concrete geometric object that is uniquely defined for a particular point set. Fig. 6b shows examples of the alpha shapes defined for our stimulus images. For this second analysis, we computed the area of the alpha shape for each stimulus image and then calculated the density of items within this alpha shape by dividing the number of items by the area of the alpha shape. We then examined whether there still remained a significant relationship between the residuals for human responses (controlled for density within the alpha shape) and the residuals for model-estimated number of clusters (controlled for density within the alpha shape). A linear regression on these residuals revealed that there was still a significant relationship ($R^2 = .90$, $p < .01$) between human number estimates and model-estimated number of clusters, after controlling for visual density within the alpha shape (Fig. 6c). Therefore, we conclude that the model-estimated number of clusters from the k -means clustering algorithm is not wholly dependent on the effects of visual density or display region. Next, we turn to comparing the k -means clustering model to a known model of grouping by proximity.

2.4. Comparison to a relevant model of perceptual grouping by proximity: the CODE algorithm and its extensions

The goal of Experiment 1 was to test our modeling approach for estimating the number and size of the perceptual groups human

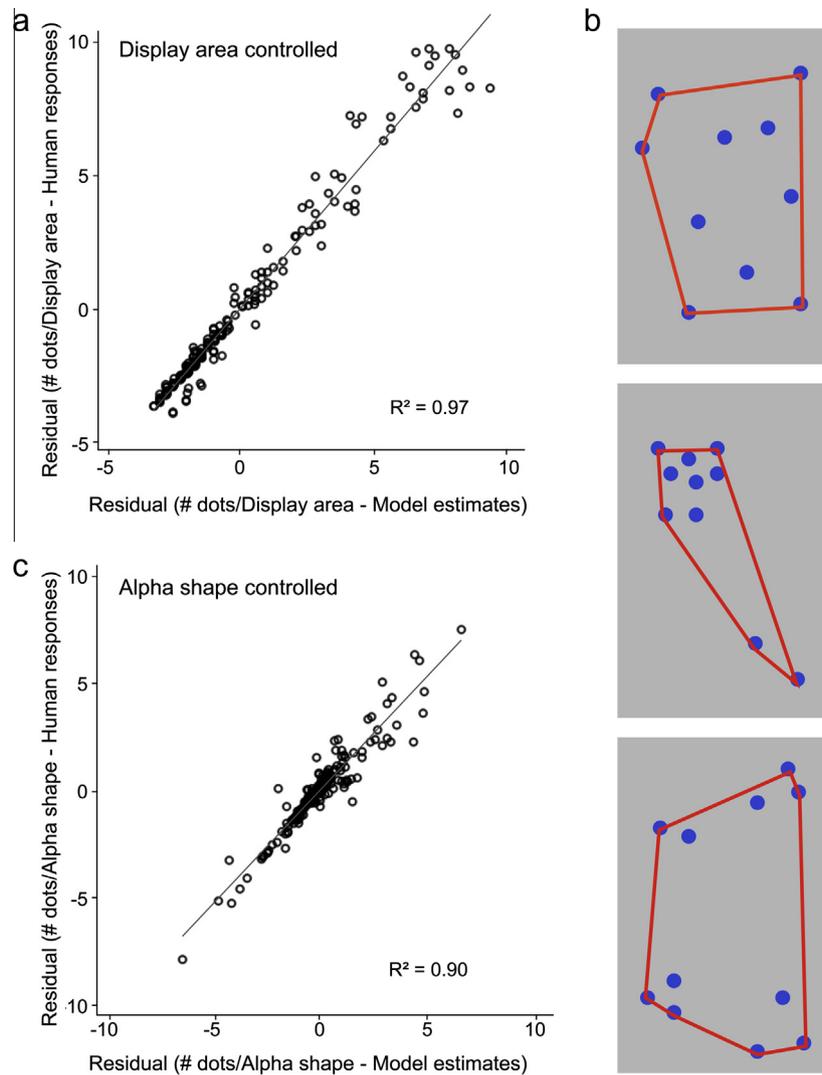


Fig. 6. Controlling for density. (a) Correlation between the residuals for human responses (controlled for density) and the residuals for model cluster ratings (controlled for density). (b) Example stimulus displays with alpha shape (red outlines). (c) Correlation between the residuals for human responses (controlled for alpha shape density) and the residuals for model cluster ratings (controlled for alpha shape density). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

observers report seeing in arrays of randomly scattered dots. The results of Experiment 1 suggest that our modified *k*-means clustering algorithm provides an accurate estimate of the number of groups reported by human observers, that this algorithm can translate human responses into an estimate of the critical grouping window size that underlies human observers' extraction of perceptual groups from these images, and that these estimates are stable from as early as 50 ms of display duration. We next compare our *k*-means clustering algorithm to another formal model of perceptual grouping by proximity, CODE (van Oeffelen & Vos, 1982). Our goal was to determine which of these formal approaches more accurately reflects the judgments of human observers for the number of perceptual groups in each image.

The CODE algorithm was proposed by van Oeffelen and Vos (1982) and has been used by later investigators (e.g., Allik & Tuulmets, 1991; Compton & Logan, 1999; Logan, 1996; Smits, Vos, & Van Oeffelen, 1985; van Oeffelen & Vos, 1984; Vos & Helsen, 1991; Vos et al., 1988). In the CODE algorithm, each dot in the stimulus array exerts an influence on its neighboring region. This influence of neighboring dots is stronger when they are close to each other, and decreases as they are separated further (van Oeffelen & Vos, 1982). To implement this effect, the CODE

algorithm uses a spread function in the shape of a normal distribution, centered on each element's location in the visual array. The dispersion is the standard deviation of each spread function, which is defined as half of the distance between every element and its specific nearest neighbor. Once the spread function is built, it is rescaled so that the height of its peak equals f_0 . Then, the spread functions of all dots in the stimulus image are summed, to create a strength gradient for the stimulus image as a whole. In this model, the clustering of elements causes the strength of grouping for some regions of the stimulus array to surpass a threshold value f_0 . When this occurs, the elements that are included within that region are identified as belonging to a single group.

Compton and Logan (1993) assessed the CODE model and claimed that the original CODE might overemphasize the extent of interactivity among dots as they are being grouped. They made several extensions of the CODE model (Compton & Logan, 1993). The extension that is most comparable to our clustering algorithm used a single variable threshold. As in one of our applications of our clustering model, this CODE model determines the threshold values separately for each stimulus image and calculates the model prediction error by subtracting the actual human rating and the model predicted estimation. Then this CODE model chooses the

best single value for the threshold for each individual participant as the one that matches the greatest number of the participant's judgments.

For the purpose of comparison to our clustering algorithm, we exploited both the original single threshold version of CODE (with the threshold set to 1, which is the maximum height of each spread function; van Oeffelen & Vos, 1982) and an extension of the CODE with the best threshold value that was determined from the participants' responses (Compton & Logan, 1993). To compare model performance, we calculated the correlation coefficient and the prediction error. The correlation coefficient was calculated based on the human responses for the number of clusters and the model prediction for the number of clusters in each image. The prediction error was calculated as:

$$\text{Prediction error (\%)} = 100 \times \frac{(\text{Model prediction} - \text{Human response})}{\text{Human response}} \quad (4)$$

In order to generate the model predictions, we used the fixed threshold value of 1 for the original CODE. For the extension of CODE, we first varied the threshold values for all the images and searched for the best-fit parameter which provides the least prediction error for each participant. For our clustering model, we applied the clustering window size of 4° . Fig. 7a and b demonstrate the correlation coefficients between human responses and model prediction and the percentage of model prediction errors from the three different models. Both the prediction error and correlation coefficient reveal that our clustering model provides a more accurate prediction of human observers' clustering pattern than the grouping algorithms based on CODE.

It is important to note that both the original CODE and the extension of the CODE algorithm with the best threshold behave reasonably well in most cases. As shown in Fig. 8a, both models based on CODE behave similar to our clustering algorithm and provide a prediction that resembles the human grouping response (although the CODE with the best threshold made a larger prediction error by parsing some clusters into smaller clusters for this image). However, the CODE algorithms seem to fail in other critical cases. For example, when individual dots are loosely distributed in the visual array as in Fig. 8b, the spread function of the original CODE algorithm exerts its influence over a very large area of the stimulus field, clustering all the members into large groups together. As a result, the original CODE algorithm predicts two groups, while the actual human response was 13.6 on average. And, on the other hand, when the distance between nearest neighbors decreases and collapses to approximately zero, which is the extreme case when two dots

occupy the two closest positions in space (Fig. 8c), the spread functions of both the original CODE and the CODE with the best threshold severely shrink. The resulting narrow spread functions cause these models to consider most of the dots as separate groups, providing the prediction of 6 and 10 groups respectively (c.f., human response was 2.02 on average). Contrary to the CODE algorithms, our *k*-means model accurately predicts the human grouping patterns for all three cases (Fig. 8a–c).

2.5. Discussion

Our modeling approach with one free parameter for the grouping window size provides a formal description of human observers' grouping pattern, suggesting a viable algorithm for human perceptual grouping. The model estimates of participants' grouping window sizes reveal that participants' grouping pattern was highly consistent across individuals. Our clustering model seeks to formally describe human observers' impression of perceptual groups when they were free to group items in their own way, and it may, in this sense, provide a quantitative measurement for "common sense" in human observers' impression of perceptual groups. The agreement across observers suggests that participants tend to exploit similar rules for determining whether items are grouped together or not, despite a visual stimulus set that involved displays that varied in number and where individual items were entirely unconstrained in their position within the display window.

The estimates of participants' grouping window sizes were constant from as early as 50 ms, suggesting that such a grouping scheme may be pre-attentive and non-iterative. This suggests that the fast, obligatory processing of perceptual groups could support the rapid read-out of approximate number from a visual array. It may be that perceptual groups that are available in a pre-attentive and robust manner from as little as 50 ms of perceptual evidence can support or affect the rapid extraction of the approximate number of individual dots, and empower or perhaps bias human observers' estimation of approximate number. In Experiment 2, we test these possibilities by investigating whether human observers' underestimation tendency in numerosity estimation can be captured by our clustering model.

3. Experiment 2

A fundamental requirement of human and animal number estimation is the ability to estimate the approximate number of elements in collections, irrespective of differences in sizes, shapes, dispersion or organization in the scene (Dehaene & Changeux,

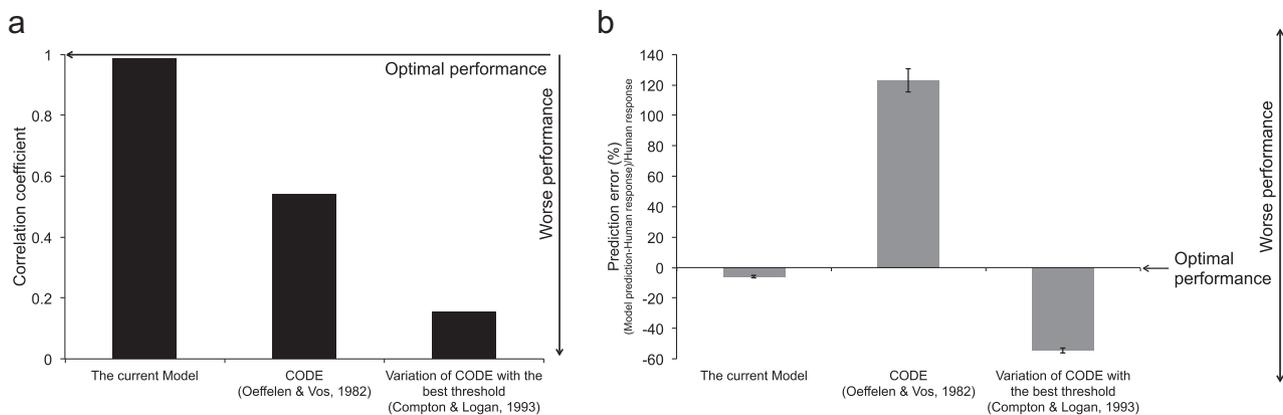


Fig. 7. Model comparison. (a) Correlation coefficient between the participants' responses and model prediction by the current algorithm, the original CODE algorithm and its variation. (b) The percentage of model prediction error for each algorithm, defined by the average deviations between the participants' responses and model prediction, divided by the participants' responses.

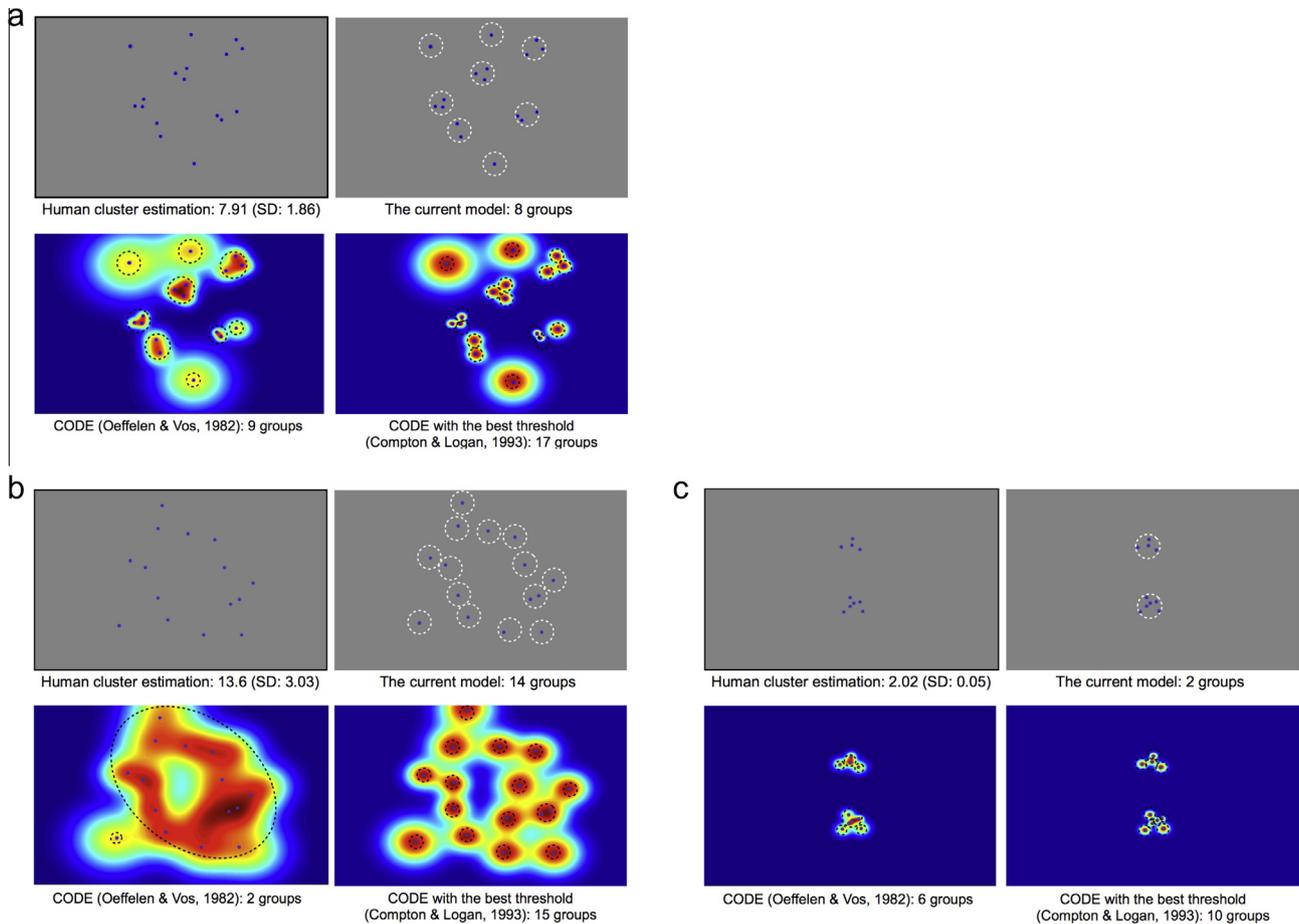


Fig. 8. Comparing three different grouping algorithms: our clustering model, the original CODE, and the extension of the CODE with the best threshold value. In each image group (a–c), the “Human cluster estimation” image shows the original display image and the average of the humans’ behavioral judgments for the number of clusters in that image; the “current model” image shows the current k -means clustering algorithm’s response for the number of clusters in the image and those clusters are depicted with dashed circles correctly sized to reflect the critical grouping window size of 4° ; the “CODE” image shows the predicted groups from the CODE algorithm along with a heatmap generated by CODE to reflect the influence of each element, groups are shown with blue dashed ellipses whose sizes are drawn only for illustration; the “CODE with the best threshold” image show the predicted groups from this algorithm along with a heatmap generated by the algorithm that reflects the influence of each element, groups are shown with blue dashed ellipses whose sizes are drawn only for illustration (a): one example in which the current model and CODE perform reasonably well in describing human grouping patterns while CODE with the best threshold overestimates the number of clusters. (b) One extreme case in which the original CODE algorithm grouped all but one of the dots into one single cluster. The participants’ estimation of the number of groups for this stimulus image was 13.6, on average. The current model and CODE with the best threshold predicted fairly well. (c) Another extreme case where the current algorithm successfully predicted participants’ cluster estimation behavior while CODE and the CODE with the best threshold algorithms both failed. Although participants’ cluster estimation was 2.02 on average, the original CODE algorithm and the extension of CODE provided 6 and 10 groups respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

1993). However, the mechanism supporting the extraction of approximate number remains unclear. Overall density in a visual image can affect numerical judgments (Dakin et al., 2011) as can overall display size (Tokita & Ishiguchi, 2010). These perceptual effects may suggest that numerical estimation does not involve processing beyond early visual stages. However, other researchers have found that object connectedness can also affect numerical estimation (Franconeri et al., 2009; He et al., 2009), suggesting that further processing may indeed be involved before number extraction.

In Experiment 1, we acquired human ratings for the number of groups experienced in randomly scattered dot arrays. Based on the human responses, we also provided an algorithm for estimating the perceptual groups present in these arrays. In Experiment 2, we asked a new group of naïve subjects to estimate the number of individual dots in the same arrays. We measured human performance at this dot estimation task, and we assessed observers’ tendencies to over or underestimate the number of elements in these displays and observers’ precision of dot estimation judgment. Using human judgments, along with our model-based estimates of

the number of perceptual groups in each display, we investigated the extent to which early perceptual grouping may affect both the bias and precision of number judgments. We hypothesize that, if perceptual clusters in a stimulus affect the extraction of the number of dots, then both human observers’ tendency to over- or underestimate the numerosity and the precision of observers’ numerosity estimation will be well captured by their grouping pattern for each stimulus image. Furthermore, if we can identify the functional relationship between the number of clusters and these estimates of the bias (e.g., over- or underestimation) and of the precision, then we may be able to effectively predict human observers’ dot estimation performance simply based on the number of clusters, predicted by our clustering model.

3.1. Method

3.1.1. Subjects

In Experiment 2, a different group of 11 naïve participants participated for course credit. All of the subjects had normal or corrected-to-normal vision.

3.1.2. Apparatus and stimuli

All the aspects were identical to those in Experiment 1.

3.1.3. Procedure

Participants were presented with exactly the same 180 stimulus images that were used in Experiment 1. The order of presentation of each image was randomized for each individual subject. On each trial, a stimulus array containing multiple dots was presented for 320 ms, followed by a mask array and response array. The mask and response screens were identical to Experiment 1.

Participants were asked to estimate the total number of individual dots in the image. No feedback was given to the participants.

3.2. Results

In order to examine the relationship between the grouping patterns and participants' estimation of the approximate number of dots, we first calculated a clustering index (CI) by dividing the actual number of dots present in an image by the number of clusters that our clustering algorithm predicted. The prediction of the number of clusters by the model was made at a grouping window size of 4° of visual angle, as we estimated this to be near optimal in Experiment 1. The CI indicates how much the dot elements are clustered together in the stimulus image. In other words, a large value of CI means that the elements are more grouped together (densely clustered) and a small value of CI means that the elements are not tightly grouped together (i.e., at a CI = 1, the number of individual dots and the number of clusters are the same). We then binned the stimulus set as a function of CI's by a sliding window with a size of 1 and with an overlap of 0.2.

We first looked for a relationship between the level of CI and the magnitude of underestimation. The magnitude of underestimation was quantified as the slope (β) from the linear regression of the actual numerosity and human estimation. Fig. 9a plots the slope β as a function of CI. The negative correlation between β and CI ($R = -0.86$) was significant ($p < .01$), suggesting that as more dots are clustered together (e.g., higher CI), the numerosity of the dots are more underestimated (e.g., lower β). Together, these results suggest that when participants tend to group more items together, they also tend to underestimate the number of elements more.

This result suggests a systematic relationship between the magnitude of underestimation of numerosity (Experiment 2) and the grouping patterns of human observers (Experiment 1). We further explored this connection between the estimates from our cluster model and human observers' numerosity estimation by determining how well the CI for each stimulus can be used to predict the mean and variability of human responses for the number of elements.

As noted above, in considering the expected mean of responses, we first investigated β for each level of CI. Here we found a strong negative relationship between CI and β : as dots became more clustered (resulting in a higher CI) β became smaller (meaning that observers tended to underestimate the number of dots more) (Fig. 9a). We next considered predicting the variability of human responses. Recall that variability in human responses will be reflected in the coefficient of variation, CV. This is the SD of responses divided by the mean. Elsewhere we have shown how to calculate this estimate across all stimulus values using maximum-likelihood estimation (the PsiMLE method, Odic, Im, Eisinger, Ly, & Halberda, in press). We used PsiMLE to compute a CV for each level of CI. We found a strong positive relationship between CI and CV – as dots became more clustered (resulting in a higher CI) CV became larger (indicating higher variability in human responses), $R = 0.81$, $p < .01$ (Fig. 9b). These results suggest that both β and CV are systematically related to the grouping pattern within the stimulus such that when participants tend to group

more items together they also tend to underestimate and generate more variable estimations of the dot numerosity.

As a further test of the relationship between clustering and number estimation, we attempted to use the k -means model to predict human number responses for all 1980 trials in the data set (i.e., 180 trials \times 11 subjects). To predict human number responses from the k -means cluster model we relied on a clustering window size (W_d) of 4° , and we took account of the group-wide relationships between clustering index (CI) and β and CV demonstrated in Fig. 9a and b. For each of our 180 image, we identified the number of clusters in the image by running the k -means model at $W_d = 4^\circ$. Dividing the actual number of dots in the image by the number of clusters derived a clustering index (CI) for each image. We then used each clustering index to determine a predicted CV and β for each trial using the regression equations found in our group-wide analyses (e.g., the appropriate CV and β can simply be read off Fig. 9a and b once you have the CI for a particular image). Finally, by combining these with the actual number of dots on the trial, it becomes possible to specify a Gaussian distribution that is a model-based prediction of the mean and spread of human responses for each trial. To compare these model predictions to actual human responses, we took a single estimate from the predicted Gaussian for each trial, for each subject (e.g., 8.3291). This number could then be compared to the actual behavioral response of the human subject on that trial – a click on a continuous number line (e.g., 9.1742).

In Fig. 9c, each dot is one of the 1980 trials in our dataset. The x-axis for each dot is the single number that was the cluster model's prediction for that trial (e.g., 8.3291) and the y-axis for each dot is the single number that was clicked by the participant on the number line (e.g., 9.1742). As can be seen in Fig. 9c, our clustering model accurately predicted human participants' estimation of dot numerosity ($R = .75$, $p < .01$). Next, we turn to comparing our model predictions for human number estimation to a known model of number estimation.

3.3. Comparison to a relevant model of perceived numerosity: Occupancy model

The occupancy model, proposed by Allik and Tuulmets (1991), was designed as a model of human number estimation in random dot arrays. The model assumes that each dot occupies a circular territory with a fixed subjective amount of area (an occupancy index) of radius r centered at each dot. The model then outlines each item with the circular territory and each dot is assumed to have an impact upon its neighborhood in the radius r . The occupancy model postulates that the total area in a visual array that is occupied by the region of dots provides the basis for numerosity estimation. Specifically, if two dots are apart less than distance r , their territories overlap each other and the contribution of those dots to numerosity estimation decreases. With only one parameter (r), the Occupancy model can predict number perception illusions where clustered displays tend to be perceived as containing fewer objects (e.g., Ginsburg & Goldstein, 1987). This is because dots in clustered displays will produce more overlaps. While being relatively simple, this elegant model can also predict human observers' number estimations accurately based on the ratio of the total area occupied by the dots in a stimulus image, accounting for the overlaps and the sum of the number of pixels of the dots. In order to compute this total area, the visual system could distribute its resources over a large area without the need to access any of the dots beyond computing the filled area.

To test whether the occupancy model also generalizes to our current dataset, and to test whether the occupancy radius r is a relatively universal parameter across studies, we exploited the occupancy model to find the best occupancy radius r for our data.

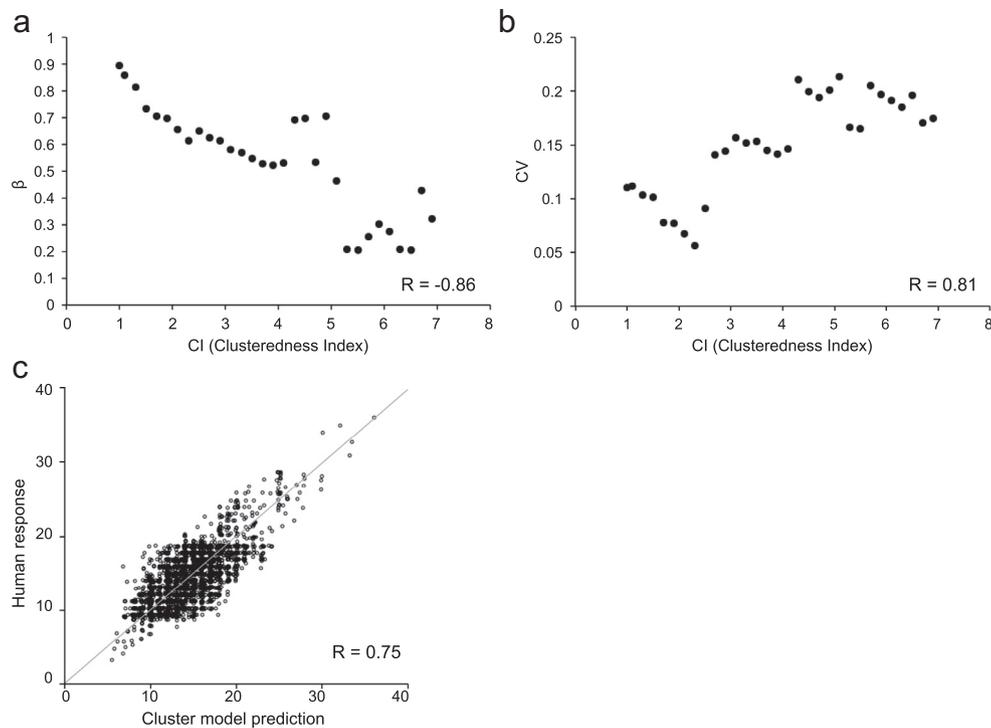


Fig. 9. Model prediction for data from Experiment 2. (a) Model prediction for β which reflects the magnitude of underestimation of dot numerosity as a function of CI (clustering index) a measure of how much dots are clustered together. (b) Model prediction for CV which inversely reflects the precision of numerosity estimation, as a function of CI (clustering index). (c) Model predicted number of dots based on the CI, β and CV values (x -axis) and the actual human responses (y -axis) for the stimulus images in Experiment 2. The gray line indicates equality, $y = x$.

We first generated predictions for the occupancy model by varying the occupancy radius r (between 0.2° and 2° of visual angle) and searched for the occupancy radius r that provided the minimum prediction error for each participant of Experiment 2. We also computed a group average r by fitting the entire dataset as a single group. These agreed well, and the best-fit occupancy radius r for our dataset was about 0.34° of visual angle, which is very close to the occupancy radius r ($18\text{--}22'$ arc corresponding to $0.30\text{--}0.37^\circ$ of visual angle) that was found in the original paper on the occupancy model by Allik and Tuulmets (1991), thereby replicating their study. Although experimental settings and parameters were different between Experiment 2 and the methods of Allik and Tuulmets (e.g., the size of the stimuli and the display areas involved), the human responses during the dot enumeration task in Experiment 2 can be explained by an occupancy radius r quite similar to the original occupancy model.

The occupancy model was designed to generate predictions for human number estimation. Using the best-fit value for the occupancy radius r (0.34°) we next found that the occupancy model predictions correlated with human number estimations ($R = .485, p < .01$). This is somewhat less than the correlation noted earlier for our cluster model predictions of human number estimation ($R = .75, p < .01$). Next, we sought to compare the predictions of the occupancy and cluster models directly via a partial correlation. This analysis allows us to determine how well the occupancy model predictions correlate with human judgments once the predictions of the cluster model are controlled for, and how well the cluster model predictions correlate with human judgments once the predictions of the occupancy model are controlled for. In this partial correlation, the predictions of the cluster model still correlated well with human judgments even after controlling for the predictions of the occupancy model ($R = .643, p < .01$). The predictions of the occupancy model, while significant, correlated

less well with human judgments after controlling for the predictions of the cluster model ($R = .15, p < .01$).

Rather than claiming that the cluster model is superior to the occupancy model as a model of number estimation behavior, we would like to highlight the similarity in the approach of these two models – both focus on the overlap and relationships across elements within a display. We suggest that the results of Experiment 3 might motivate a continued and deeper exploration into the effects of underestimation and increased variability in number judgment tasks as affected by grouping and overlap in stimulus arrays.

3.4. Discussion

Although our k -means clustering model was aimed at estimating the number and size of perceptual groups in random dot arrays (Experiment 1), in an extension of the model we found that its estimate of clustering index (CI) can be used to effectively predict human estimations of the number of total dots within dot arrays (Experiment 2). In Experiment 3 we explored whether observers could learn to ignore this bias from perceptual grouping or whether, even with feedback, the underestimation bias from perceptual grouping would be observed in number estimation.

4. Experiment 3

Since perceptual groups can be achieved from a brief presentation (e.g., 50 ms) and pre-attentively, relying on perceptual groups might provide rapid access to multiple dots for the extraction of approximate number. And, if extraction of approximate number is dependent on higher-order groupings of elements, then the effects of clustering on human number estimates might persist even in the face of clear and consistent feedback to counteract

these effects. In Experiment 3, we investigated whether observers can learn to calibrate their responses and adjust their reliance on perceptual groups when estimating the approximate number of dots. We provided explicit feedback on each trial for the correct answer. If participants can calibrate their responses and reduce the tendency to rely on perceptual groups in estimating dot numerosity according to feedback (e.g., Izard & Dehaene, 2008), their responses will be less-influenced by the clustering index of the images. However, if reliance on perceptual groups in estimating the number of dots is automatic and resistant to change, then participants' underestimation bias for the more clustered images would still be observed.

4.1. Method

4.1.1. Subjects

A different group of 19 naive undergraduate students from Johns Hopkins University participated in the experiment in exchange for course credit. All of the subjects had normal or corrected-to-normal vision.

4.1.2. Apparatus and stimuli

All aspects of stimuli were the same as those of Experiments 1 and 2, except that a new set of 198 stimulus images containing randomly positioned dots (5–35 dots) were generated for Experiment 3.

4.1.3. Procedure

The procedure of Experiment 3 was the same as that of Experiment 2, except that participants were given feedback on every trial. After participants made their response by clicking on the response scale, a small vertical bar appeared at the location of the correct answer. Participants were made aware of this bar and told that it showed the position of the correct answer.

4.2. Results

As in Experiment 2, we calculated the CI by dividing the actual number of dots by the model-predicted number of clusters and used a sliding window of size 1 for binning. We then calculated the β , which is a slope of participants' numerosity estimation as a function of the actual numerosity. The relationship between the β and CI is shown in Fig. 10, with gray dots, along with the reprint of the dataset from Experiment 2 for comparison. A first observation is that the β values from Experiment 3 are higher than the β values from Experiment 2. Ideal performance would occur at $\beta = 1$, no underestimation, and the further β is below 1 the more underestimation the observer has in their behavior. The gray dots (Experiment 3) are above the black dots (Experiment 2) and the gray dots are closer to the ideal value of 1. This means that providing feedback did help the observers reduce their tendency to underestimate dots. This was true across all levels of clustering index (CI).

A second effect is also important to consider. That is, even while feedback was successful in helping observers reduce their underestimation bias, this feedback was not uniformly successful in eliminating this bias. We observed a weak negative correlation between the CI and β , which was significant ($R = -0.43$, $p < .05$). The negative correlation between the CI and β suggests that when the dots are more clustered together (e.g., higher CI), they are perceived to be less numerous (lower β). However, it is true that the negative correlation was considerably reduced compared to Experiment 2 where participant did not receive any feedback. Therefore, these results are somewhat mixed. That the β values in Experiment 3 are closer to the ideal value of 1 suggests that subjects can learn to overcome their underestimation bias. But, that we still observe a

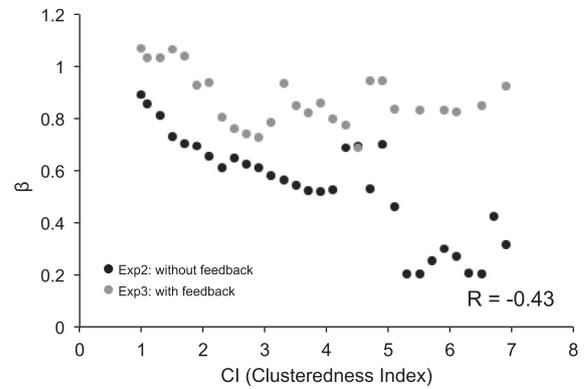


Fig. 10. Model prediction for β , which reflects the magnitude of underestimation of dot numerosity as a function of CI (clustering index) a measure how much dots are clustered together. For ease of comparison, black dots reprint the data from Experiment 2 in which participants did not receive any feedback, and gray dots indicate the data from Experiment 3 in which participants received feedback about the actual answer on each trial.

relationship between β and CI reveals that, even with feedback, the observers still display the cost of increased clustering index and a persistent underestimation bias with increasing clustering index in the displays. This suggests that reliance on perceptual groups for numerosity estimation cannot be completely eliminated, because perceptual groups are extracted automatically and rapidly and affect estimation of dot numerosity.

5. General discussion

We all share the remarkable ability to rapidly perceive the approximate number of items in visual collections. Results of previous studies suggest that extracting approximate number information may not require the sampling and segregating of individual items from a collection, rather some holistic process may be involved (Barth et al., 2003; Tibber et al., 2012). However, even if each individuated member is not segmented, a sense of “how many things there are” must included a description of the thing that is to be counted (e.g., “dots of this size”), and the extraction of approximate number might rely on this or some other unitizer to group continuous perception into chunks for further processing. We propose that perceptual groups extracted in an automatic and pre-attentive manner can serve as such a unit for approximate number estimation. In Experiment 1, we found that participants' perception of groups in random dot arrays could be fit by a k -means clustering algorithm and that the predictions of this algorithm more closely matched human responses than two versions of the popular CODE algorithm (in fact, we tested all available versions of CODE and our k -means clustering algorithm was superior to all versions of CODE – we presented the two most relevant versions of CODE here, the original and the single parameter version).

In Experiment 2 we demonstrated that the output of the k -means clustering algorithm can predict number estimations in random dot arrays and that the predictions of this cluster-based number algorithm outperformed a leading model of number estimation (the Occupancy model). The number predictions generated from the k -means clustering algorithm matched more closely the number estimates of real human subjects, suggesting that perceptual groups may play an important role in determining our conceptual experience of approximate number. Specifically, as items are grouped into fewer perceptual clusters, the perception of approximate number becomes biased towards the items looking

less numerous. In Experiment 2 we found that this effect of clustering remained even controlling for the density of the display.

Lastly, in Experiment 3 we found that the impact of perceptual clustering on number estimation is not under voluntary control. Participants could reduce, but not eliminate, their underestimation bias by recalibrating their responses according to feedback provided on each trial – clustering still caused them to underestimate. This is consistent with the possibility that some of the effect of perceptual groups on numerosity estimation is not strategic but may be automatic.

Perceived groups in a visual image have been considered as a compulsory product from pre-attentive processes (Neisser, 1967). Such groups may serve as a form of primitive chunking – generating a hierarchical reorganization of items into groups with each group functioning as a unit. In this manner, perceptual grouping may allow for parsing multiple nested-levels of representation of the same stimuli from individuated items to one global scene perception. Halberda et al. (2006) suggested that, in a number estimation task, hierarchical coding of “group” and “individual” are both available for enumeration by the approximate number system. They further suggested that the level of a “group” may be necessary prior to enumeration of individuals by the approximate number system. The groups structured by perceptual grouping are a reasonable candidate for extraction of approximate number, and this may be especially so in displays of randomly scattered items where the first challenge for enumeration is to determine what aspects of the visual image are to be used for ensemble perception. It is therefore reasonable to expect that the elements are spatially grouped into clusters even before extraction of approximate number. Groups built from early automatic grouping processes may not be easily overridden or split into single individuals – e.g., such segmentation often requires focused attention towards each individual object (e.g., Burkell & Pylyshyn, 1997). Therefore, perceptual groups built from the visual image may serve as a unit for ensemble perception, allowing for a fast, effortless process in a global manner. When elements form a cluster, they will be experienced to be less numerous than items which join to form multiple clusters because they tend to be perceived as a single unit and to serve “as real as the organization of a homogeneous spot,” (Koffka, 1935).

Most of the classical demonstrations of perceptual grouping have focused on “effective” stimulus images that were generated in the manner that they apparently contain distinct groups (e.g., clusters of elements are obvious enough) so that the viewers are guided to experience similar patterns from these images. It is somewhat true that human observers’ grouping patterns have not been rigorously examined in ambiguous cases in which the goodness of grouping in the images is low (e.g., when items are randomly positioned and not clearly clustered).

In the current study, we presented randomly positioned dots in a visual array so that there was no clear cue for observers to group the dots in a specific way. In order to formally describe and compare observers’ grouping behavior we introduced a new *k*-means approach to describe grouping patterns based on proximity. This model relied on a modified *k*-means clustering algorithm, with a single free parameter for the size of grouping window. In the model, grouping window size functions as a threshold for grouping such that items that are closer to each other than the value of grouping window size would be grouped together. We found that this simple model can explain participants’ grouping behavior remarkably well and accurately predicted their estimation of the number of perceptual groups.

Our approach allowed us to quantitatively describe human grouping, to compare among participants’ grouping patterns using a single value of grouping window size from each individual. We found that the best-fit values of critical distance for determining

whether elements were grouped into one cluster was generally approximately 3.9° of visual angle, measured by the diameter of the best fit grouping window. The critical distance for grouping was highly consistent across individual subjects and across the stimulus images regardless of the actual number of items in each of the images. Note that we intended to use the least-constrained settings in which dots were randomly positioned in each visual array and participants were free to group them however they felt the most intuitive and natural for them. It is therefore intriguing to see that participants mostly agreed with each other and that their grouping pattern remained quite consistent across different images throughout trials.

Our window-based clustering algorithm provides a formal description of perceptual grouping of items by proximity and how observers’ grouping behavior may generate a bias for the rapid extraction of the approximate number of items. In particular, our approach suggests a viable mechanism of underestimation of approximate number: when elements are located within an observers’ grouping window size, they are underestimated while when they are further from one another than an observers’ grouping window size, they are not underestimated at all.

These findings on grouping and the underestimation bias in extraction of approximate number may also be consistent with the emerging evidence that vision is often limited by spacing between objects. ‘Crowding’ can occur when objects are too close together and features from multiple objects are mandatorily combined in perception. In many different studies, crowding has been observed: grating discrimination, object recognition such as faces, reading letters, visual search, and selective attention (for review, see Pelli & Tillman, 2008). Observers fail to discriminate and recognize items within the window where the spacing between items does not exceed a critical spacing and the size of the window limits the efficiency and speed of visual processing such as reading and search (Pelli & Tillman, 2008; Pelli et al., 2007; Reddy & VanRullen, 2007). Within the window, crowding occurs as a product of “faulty information pooling,” (Levi, 2008; Pelli, Palomares, & Majaj, 2004; Pelli & Tillman, 2008). A similar idea was also proposed using the concept of the minimum of attentional resolution (Intrilligator & Cavanagh, 2001). The degree of mandatory integration of items within a given region is also modulated by the grouping principle (Banks, Bodinger, & Illige, 1974; Banks, Larson, & Prinzmetal, 1979; Prinzmetal & Banks, 1977, 1983). For example, the greater the degree of grouping between target and distractors (modulated by similarity or proximity, etc.), the worse observers’ performance in detecting and discriminating the target item (Kooi, Toet, Tripathy, & Levi, 1994; Sayim, Westheimer, & Herzog, 2010, 2011).

Together, our approach to window-based clustering allows us to formally describe human observers’ grouping behavior based on the window for spatial integration and a pooling process, and our results show that the limit of the extent over which observers can integrate features of multiple items mandatorily modulates observers’ grouping pattern and the well-known underestimation bias in extraction of approximate number. Our results suggest that rapid extraction of approximate number from a visual array containing multiple items can be achieved by spatial parsing which relies on the fast and global processing of perceptual groups. Perceptual groups may serve as units for extraction of approximate number by providing fast read-outs from multiple sets of items over space.

References

- Allik, J., & Tuulmets, T. (1991). Occupancy model of perceived numerosity. *Perception & Psychophysics*, 49, 303–314.

- Anobile, G., Cicchini, G. M., & Burr, D. C. (2014). Separate mechanisms for perception of numerosity and density. *Psychological Science*, 25, 265–270.
- Banks, W. P., Bodinger, D., & Illige, M. (1974). Visual detection accuracy and target noise proximity. *Bulletin of the Psychonomic Society*, 2, 411–414.
- Banks, W. P., Larson, D. W., & Prinzmetal, W. (1979). Asymmetry of visual interference. *Perception & Psychophysics*, 23, 447–456.
- Barth, H., Kanwisher, N., & Spelke, E. S. (2003). The construction of large number representations in adults. *Cognition*, 86, 201–221.
- Bevan, W., Maier, R., & Helson, H. (1963). The influence of context upon the estimation of number. *American Journal of Psychology*, 76, 464–469.
- Binet, A. (1890). La perception des longueurs et des nombres chez quelques petits enfants. *Revue Philosophique de la France et de l'Étranger*, 30, 68–81.
- Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, 10, 433–436.
- Burkell, J. A., & Pylyshyn, Z. W. (1997). Searching through subsets: A test of the visual indexing hypothesis. *Spatial Vision*, 11, 225–258.
- Chaturvedi, A., Carroll, J. D., Green, P., & Rotondo, J. A. (1997). A feature-based approach to market segmentation via overlapping *k*-centroids clustering. *Journal of Marketing Research*, 34, 370–377.
- Choo, H., & Franconeri, S. L. (2014). Enumeration of small collections violates Weber's law. *Psychonomic Bulletin & Review*, 21, 93–99.
- Compton, B. J., & Logan, G. D. (1993). Evaluating a computational model of perceptual grouping by proximity. *Perception & Psychophysics*, 53, 403–421.
- Compton, B. J., & Logan, G. D. (1999). Judgments of perceptual groups: Reliability and sensitivity to stimulus transformation. *Perception & Psychophysics*, 61, 1320–1335.
- Cordes, S., Gelman, R., & Gallistel, C. R. (2001). Variability signatures distinguish verbal from nonverbal counting for both large and small numbers. *Psychonomic Bulletin & Review*, 8, 698–707.
- Crollen, V., Castronovo, J., & Seron, X. (2011). Under- and over-estimation: a bi-directional mapping process between symbolic and non-symbolic representations of number? *Experimental Psychology*, 58(1).
- Dakin, S. C., Tibber, M. S., Greenwood, J. A., Kingdom, F. A. A., & Morgan, M. J. (2011). A common visual metric for approximate number and density. *Proceedings of the National Academy of Sciences*, 108, 19552–19557.
- Dehaene, S. (2003). The neural basis of the Weber–Fechner law: a logarithmic mental number line. *Trends in Cognitive Sciences*, 7, 145–147.
- Dehaene, S., & Changeux, J. P. (1993). Development of elementary numerical abilities: A neuronal model. *Journal of Cognitive Neuroscience*, 5, 390–407.
- Durgin, F. H. (1995). Texture density adaptation and the perceived numerosity and distribution of texture. *Journal of Experimental Psychology: Human Perception and Performance*, 21, 149–169.
- Durgin, F. H. (2008). Texture density adaptation and visual number revisited. *Current Biology*, 18, R855–R856.
- Edelsbrunner, H., Kirkpatrick, D. G., & Seidel, R. (1983). On the shape of a set of points in the plane. *IEEE Transactions on Information Theory*, 29, 551–559.
- Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, 8, 307–314.
- Franconeri, S. L., Bemis, D. K., & Alvarez, G. A. (2009). Number estimation relies on a set of segmented objects. *Cognition*, 113, 1–13.
- Frank, M. C., Everett, D. L., Fedorenko, E., & Gibson, E. (2008). Number as a cognitive technology: Evidence from Pirahã language and cognition. *Cognition*, 108, 819–824.
- Frith, C. D., & Frith, U. (1972). The solitary illusion: An illusion of numerosity. *Perception & Psychophysics*, 11, 409–410.
- Gilmore, C., Attridge, N., Clayton, S., Cragg, L., Johnson, S., Marlow, N., ... Inglis, M. (2013). Individual differences in inhibitory control, not non-verbal number acuity, correlate with mathematics achievement. *PLoS One*, 8, e67374.
- Ginsburg, N. (1976). Effect of item arrangement on perceived numerosity: Randomness vs regularity. *Perceptual and Motor Skills*, 43, 663–668.
- Ginsburg, N., & Goldstein, S. R. (1987). Measurement of visual cluster. *American Journal of Psychology*, 100, 193–203.
- Haber, R. N. (1970). Note on how to choose a visual noise mask. *Psychological Bulletin*, 74, 373–376.
- Halberda, J., Sires, S. F., & Feigenson, L. (2006). Multiple spatially overlapping sets can be enumerated in parallel. *Psychological Science*, 17, 572–576.
- Hauser, M. D., Carey, S., & Hauser, L. B. (2000). Spontaneous number representation in semi-free-ranging rhesus monkeys. *Proceedings of the Royal Society of London B*, 267, 829–833.
- He, L., Zhang, J., Zhou, T., & Chen, L. (2009). Connectedness affects dot numerosity judgment: Implications for processing for configural processing. *Psychonomic Bulletin & Review*, 16, 509–517.
- Honarkhah, M., & Caers, J. (2010). Stochastic simulation of patterns using distance-based pattern modeling. *Mathematical Geosciences*, 42, 487–517.
- Hurewitz, F., Gelman, R., & Schnitzer, B. (2006). Sometimes area counts more than number. *Proceedings of the National Academy of Sciences*, 103, 19599–19604.
- Indow, T., & Ida, M. (1977). Scaling of dot numerosity. *Perception & Psychophysics*, 11, 5–9.
- Intrilligator, J., & Cavanagh, P. (2001). The spatial resolution of visual attention. *Cognitive Psychology*, 43, 171–216.
- Izard, V., & Dehaene, S. (2008). Calibrating the mental number line. *Cognition*, 106, 1221–1247.
- Jevons, W. S. (1871). The power of numerical discrimination. *Nature*, 3, 281–282.
- Koffka, K. (1935). *Principles of Gestalt psychology*. New York: Harcourt, Brace and World.
- Kooi, F. L., Toet, A., Tripathy, S. P., & Levi, D. M. (1994). The effect of similarity and duration on spatial interaction in peripheral vision. *Spatial Vision*, 8, 255–279.
- Krueger, L. E. (1982). Single judgments of numerosity. *Perception & Psychophysics*, 31, 175–182.
- Krueger, L. E. (1984). Perceived numerosity: A comparison of magnitude production, magnitude estimation, and discrimination judgments. *Perception & Psychophysics*, 35, 536–542.
- Le Corre, M., & Carey, S. (2007). One, two, three, four, nothing more: An investigation of the conceptual sources of the verbal counting principles. *Cognition*, 105, 395–438.
- Levi, D. (2008). Crowding—an essential bottleneck for object recognition: A mini-review. *Vision Research*, 48, 635–654.
- Logan, G. D. (1996). The CODE theory of visual attention: An integration of space-based and object-based attention. *Psychological Review*, 103, 603–649.
- Meck, W. H., & Church, R. M. (1983). A mode control model of counting and timing processes. *Journal of Experimental Psychology: Animal Behavior Processes*, 9, 320–334.
- Miller, L., & Baker, R. A. (1968). The effects of shape, size, heterogeneity, and instructional set on the judgment of visual number. *The American Journal of Psychology*, 81, 83–91.
- Neisser, U. (1967). *Cognitive psychology*. New York: Appleton-Century-Crofts.
- Odic, D., Im, H. Y., Eisinger, R., Ly, R., & Halberda, J. (in press). PsiMLE: A maximum-likelihood approach to estimating psychophysical scaling and variability more reliably, efficiently, and flexibly. *Behavioral Research Methods*.
- Palmer, S. E. (1975). The effects of contextual scenes on the identification of objects. *Memory & Cognition*, 3, 519–526.
- Pelli, D. G. (1997). The video toolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, 10, 437–442.
- Pelli, D. G., Palomares, M., & Majaj, N. J. (2004). Crowding is unlike ordinary masking: Distinguishing feature integration from detection. *Journal of Vision*, 4, 1136–1169.
- Pelli, D. G., & Tillman, K. A. (2008). The uncrowded window of object recognition. *Nature Neuroscience*, 11, 1129–1135.
- Pelli, D. G., Tillman, K. A., Freeman, J., Su, M., Berger, T. D., & Majaj, N. J. (2007). Crowding and eccentricity determine reading rate. *Journal of Vision*, 7, 1–36.
- Pomerantz, J. R. (1983). Global and local precedence: Selective attention in form and motion perception. *Journal of Experimental Psychology: General*, 112, 516–540.
- Pomerantz, J. R. (1981). Perceptual organization in information processing. In M. Kubovy & J. R. Pomerantz (Eds.), *Perceptual organization* (pp. 141–180). Hillsdale, NJ: Erlbaum.
- Ponzo, M. (1928). Urteilstauschungen über Mengen. *Archiv Für die gesamte Psychologie*, 65, 129–162.
- Prinzmetal, W., & Banks, W. P. (1977). Good continuation affects visual detection. *Perception & Psychophysics*, 21, 389–395.
- Prinzmetal, W., & Banks, W. P. (1983). Perceptual capacity limits in visual detection and search. *Bulletin of the Psychonomic Society*, 4, 263–266.
- Ray, S., & Turi, R. H. (1999). Determination of number of clusters in *k*-means clustering and application in colour image segmentation. In: *Proceedings of the 4th international conference on advances in pattern recognition and digital techniques* (pp. 137–143). Calcutta, India.
- Reddy, L., & VanRullen, R. (2007). Spacing affects some but not all visual searches: Implications for theories of attention and crowding. *Journal of Vision*, 7, 1–17.
- Sayim, B., Westheimer, G., & Herzog, M. H. (2010). Gestalt factors modulate basic spatial vision. *Psychological Science*, 21, 641–644.
- Sayim, B., Westheimer, G., & Herzog, M. H. (2011). Quantifying target conspicuity in contextual modulation by visual search. *Journal of Vision*, 11, 1–11.
- Schultz, D. W., & Eriksen, C. W. (1977). Do noise masks terminate target processing? *Memory & Cognition*, 5, 90–96.
- Simon, T. J., & Vaishnavi, S. (1996). Subitizing and counting depend on different attentional mechanisms: Evidence from visual enumeration in afterimages. *Perception & Psychophysics*, 58, 915–926.
- Smets, K., Gebuis, T., Defever, E., & Reynvoet, B. (2014). Concurrent validity of approximate number sense tasks in adults and children. *Acta Psychologica*, 150, 120–128.
- Smits, J. T., Vos, P. G., & Van Oeffelen, M. P. (1985). The perception of a dotted line in noise: A model of good continuation and some experimental results. *Spatial Vision*, 1, 163–177.
- Sperling, G. (1963). A model for visual memory tasks. *Human Factors*, 5, 19–31.
- Steinhaus, H. (1956). Sur la division de corps matériels en parties. *Bulletin of the Polish Academy of Sciences*, 4, 801–804.
- Stoianov, I., & Zorzi, M. (2012). Emergence of a “visual number sense” in hierarchical generative models. *Nature Neuroscience*, 15, 194–196.
- Taves, E. H. (1941). Two mechanisms for the perception of visual numerosness. *Archives of Psychology*, 37, 1–47.
- Tibber, M. S., Greenwood, J. A., & Dakin, S. C. (2012). Number and density discrimination rely on a common metric: Similar psychophysical effects of size, contrast, and divided attention. *Journal of Vision*, 12, 1–19.
- Tokita, M., & Ishiguchi, A. (2010). How might the discrepancy in the effects of perceptual variables on numerosity judgment be reconciled? *Attention, Perception, & Psychophysics*, 72, 1839–1853.
- Trick, L. M., & Pylyshyn, Z. W. (1993). What enumeration studies can show us about spatial attention: Evidence for limited capacity preattentive processing. *Journal of Experimental Psychology: Human Perception and Performance*, 19, 331–351.
- Trick, L. M., & Pylyshyn, Z. W. (1994). Why are small and large numbers enumerated differently? A limited-capacity preattentive stage in vision. *Psychological Review*, 101, 80–102.
- van Oeffelen, M. P., & Vos, P. G. (1982). Configurational effects on the enumeration of dots: Counting by groups. *Memory & Cognition*, 10, 396–404.

- van Oeffelen, M. P., & Vos, P. G. (1984). Enumeration of dots: An eye movement analysis. *Memory & Cognition*, *12*, 607–612.
- Vos, P. G., & Helsen, E. L. (1991). Modelling detection of linear structure in dot patterns. In I. Dvorak & A. V. Holden (Eds.), *Mathematical approaches to brain functioning diagnostics* (pp. 429–441). Manchester, UK: Manchester University Press.
- Vos, P. G., Van Oeffelen, M. P., Tibosch, H. J., & Allik, J. (1988). Area-numerosity interactions. *Psychological Research Psychologische Forschung*, *50*, 148–150.
- Wertheimer, M. (1924). Gestalt Theory. Address Before the Kant Society, Berlin, 7th December, 1924; reprinted in Wertheimer M (1950) in *A Sourcebook of Gestalt Psychology*, W.D. Ellis (ed.) (pp. 1–11), Humanities, New York.
- Whalen, J., Gallistel, C. R., & Gelman, R. (1999). Non-verbal counting in humans: The psychophysics of number representation. *Psychological Science*, *10*, 130–137.
- Xu, F., & Spelke, E. S. (2000). Large number discrimination in 6 month-old infants. *Cognition*, *74*, B1–B11.