Can (should) theories of crowding be unified?

Mehmet N. Agaoglu
School of Optometry, University of California, Berkeley,
Berkeley, CA, USA

Susana T. L. Chung
School of Optometry, University of California, Berkeley,
Berkeley, CA, USA

Objects in clutter are difficult to recognize, a phenomenon known as *crowding*. There is little consensus on the underlying mechanisms of crowding, and a large number of models have been proposed. There have also been attempts at unifying the explanations of crowding under a single model, such as the weighted feature model of Harrison and Bex (2015) and the texture synthesis model of Rosenholtz and colleagues (Balas, Nakano, & Rosenholtz, 2009; Keshvari & Rosenholtz, 2016). The goal of this work was to test various models of crowding and to assess whether a unifying account can be developed. Adopting Harrison and Bex's (2015) experimental paradigm, we asked observers to report the orientation of two concentric C-stimuli. Contrary to the predictions of their model, observers' recognition accuracy was worse for the inner C-stimulus. In addition, we demonstrated that the stimulus paradigm used by Harrison and Bex has a crucial confounding factor, eccentricity, which limits its usage to a very narrow range of stimulus parameters. Nevertheless, reporting the orientations of both C-stimuli in this paradigm proved very useful in pitting different crowding models against each other. Specifically, we tested deterministic and probabilistic versions of averaging, substitution, and attentional resolution models as well as the texture synthesis model. None of the models alone was able to explain the entire set of data. Based on these findings, we discuss whether the explanations of crowding can (should) be unified.

Introduction

Compared with foveal vision, peripheral vision is limited because of the reduced number of photoreceptors in the retina and the smaller population of cortical and subcortical neurons that are devoted to peripheral information processing. This is evident in, for instance, contrast sensitivity and letter acuity changes as a function of eccentricity (Herse & Bedell, 1989; Jacobs, 1979; Virsu & Rovamo, 1979; Weymouth, 1958). Recognition/identification of a target object is further impaired when it is presented in visual clutter, known as *crowding* (Bouma, 1970; Toet & Levi, 1992). Crowding can be observed in the fovea; however, it is stronger in the periphery and persists even when the letter size is scaled to compensate for changes in acuity (Chung, Levi, & Legge, 2001; Levi, Hariharan, & Klein, 2002; Levi, Klein, & Hariharan, 2002). Because objects in the environment are almost never in isolation, crowding is a ubiquitous phenomenon, and therefore, it sets a bottleneck for object recognition, visual search, and motor actions (eye, hand, and body movements), tasks crucial for survival (Levi, 2008; Pelli & Tillman, 2008; Whitney & Levi, 2011).

When asked to identify a crowded object, observers make various types of response errors. For instance, when the orientation of a target bar (or a letter) is to be reported in the presence of other flanking bars, observers may report the average orientation of the ensemble (averaging) or the orientation of one of the flankers (substitution), or they may simply guess (Ester, Klee, & Awh, 2014; Ester, Zilber, & Serences, 2015; Freeman, Chakravarthi, & Pelli, 2012; Greenwood, Bex, & Dakin, 2009; Hanus & Vul, 2013; He, Cavanagh, & Intriligator, 1996; Parkes, Lund, Angelucci, Solomon, & Morgan, 2001). Different types of errors made under crowding conditions can be considered as indicators of multiple (and potentially overlapping) processes. Indeed, there is still an ongoing debate as to whether or not crowding results from a single mechanism or multiple processes (Levi, 2008; Pelli, 2008; Pelli & Tillman, 2008; Whitney & Levi, 2011). Recent work implicated pooling based on receptive field sizes (Balas, Nakano, & Rosenholtz, 2009; Freeman & Simoncelli, 2011; Freeman, Ziemba, Heeger, Simoncelli, & Movshon, 2013; Keshvari & Rosenholtz, 2016), cortical distance (Mareschal, Morgan, & Solomon, 2010; Pelli, 2008), and attentional

Citation: Agaoglu, M. N., & Chung, S. T. L. (2016). Can (should) theories of crowding be unified? *Journal of Vision*, *16*(15):10, 1–22, doi:10.1167/16.15.10.

doi: 10.1167/16.15.10 Received June 14, 2016; published December 9, 2016 ISSN 1534-7362

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

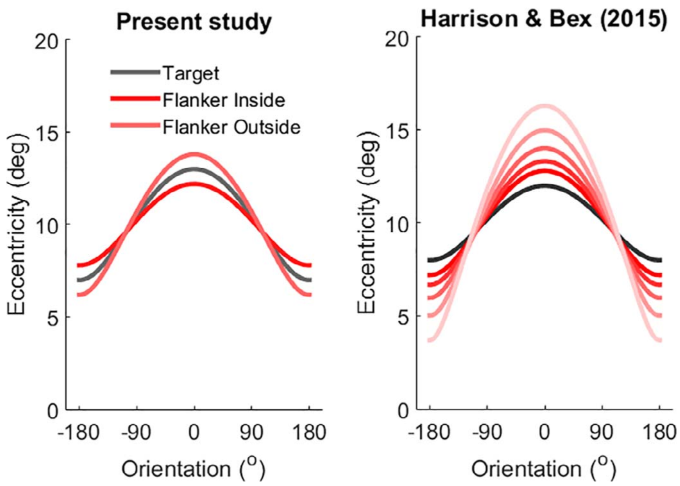

Figure 1. Eccentricity confound in the present study (left) and in Harrison and Bex's (2015) study (right). The eccentricity of the small gap in the C-stimuli as a function of their orientation is plotted. Gray lines represent the target object in both studies, whereas red curves represent the flankers. Different shades of red represent different sizes; the darker the smaller. Note that the smaller a flanker, the larger its weight on the final percept as based on the weighting field approach of the HB model. Note also that the flanker was always larger than the target in Harrison and Bex's study.

resolution (He et al., 1996; Intriligator & Cavanagh, 2001) as the causal factors for crowding. However, the diverse set of response errors made by observers cannot be fully explained by any of these proposed mechanisms alone.

Several attempts have been made to fully capture the statistics of the response errors and offer a "unifying" account for crowded object recognition (e.g., Balas et al., 2009; Cox & Riesenhuber, 2015; Harrison & Bex, 2015; Keshvari & Rosenholtz, 2016; Nandy & Tjan, 2012). Harrison and Bex (2015) proposed a probabilistic averaging model based on the population activity in the early visual areas and claimed that their model (hereafter referred to as the HB model) alone can reproduce the entire spectrum of response errors without resorting to multiple mechanisms. In short, their model consists of three stages: (a) a filtering stage, (b) a weighted-averaging stage, and (c) a decision-making stage. The essential property of the HB model, as the authors emphasized more recently (Harrison & Bex, 2016; Pachai, Doerig, & Herzog, 2016), is the weighting field centered on the target object. The contrast energy obtained from the first stage is weighted such that the contribution of a feature (regardless of its ownership) to the final percept decreases with its distance from the center of the target object. Note that there is no explicit feature extraction in this model. Harrison and Bex (2015) used concentric circles with a small gap, which defined their orientations (we hereafter refer to these as C-stimuli), with the flanker always surrounding the target, to test this model. As predicted, crowding of the target became weaker (indicated by smaller response errors) as the size of the flanker (hence, its distance from the target) increased. In addition, according to the HB model, the flanker in this paradigm should suffer from more crowding compared with the target, because the target, being closer to the center of the weighting field, always enjoys a higher weight. One way to test this is to ask observers to report the orientation of both the target and the flanker. In fact, Harrison and Bex (2015) asked observers to report both objects in their Experiment 2; however, the authors did not separately compare the response errors made for the target and the flanker.

Crowding literature is very rich in terms of the types of stimuli and configurations used to investigate certain aspects of crowded vision. Some studies used simple stimuli such as oriented bars, Gabor patches, or letters, whereas some others opted for more complex stimuli such as objects, tools, faces, or natural scenes. Using primitive stimuli allows one to remove or minimize the effects of irrelevant neural processes on the task at hand and to disentangle the contribution of several relevant processes on percepts. Using complex stimuli provides measurements more relevant to natural viewing conditions; however, it comes with the risk of involving contributions from multiple (potentially relevant or irrelevant) processes on the outcome measure. Harrison and Bex (2015) introduced a stimulus paradigm in which with only two concentric C-stimuli, one can potentially obtain all signature properties of crowding such as radial-tangential anisotropy, inner-outer asymmetry, and critical distance. Having only two objects also makes it convenient to simultaneously investigate the crowded representations of both the target and the flanker and provides a new means to pit different crowding models against each other. Although it seems promising, this stimulus paradigm has a crucial confounding factor. The eccentricity of the small gap, which defines the orientation of the Cs, varies strongly with orientation (Figure 1).

Inspired by an earlier texture synthesis model (Portilla & Simoncelli, 2000), Rosenholtz and colleagues recently introduced a texture synthesis model to account for crowding and several other visual phenomena (Balas et al., 2009; Keshvari & Rosenholtz, 2016; Rosenholtz, Huang, & Ehinger, 2012; Rosenholtz, Huang, Raj, Balas, & Ilie, 2012). According to their texture tiling model (hereafter referred to as the TT model), local as well as global image statistics (pixel statistics, correlation coefficients, magnitude correlations, and phase statistics) are extracted from any arbitrary stimuli by means of a set of oriented V1-like wavelets with various orientations, scales, and neighboring positions. The main proposition of the TT model is that peripheral information is jumbled up because of this textural representation. Recently, Keshvari and Rosenholtz (2016) presented a convincing case in which the TT model could reproduce the response errors (to a certain extent) in three different crowding studies with a wide range of stimuli (letters: Freeman et al., 2012; letter-like cross stimuli: Greenwood, Bex, & Dakin, 2012; Gabors with different spatial frequency, color, and orientation: Põder & Wagemans, 2007). Here, we tested the TT model with the novel stimulus paradigm introduced by Harrison and Bex (2015).

We had two aims in the present study. First, we tested two recently proposed models that have been claimed to be unifying accounts of crowding, as well as more conventional models such as averaging, substitution, and attentional resolution. More specifically, we tested the predictions of the HB model (particularly weighting field approaches used generally in pooling models of crowding) by adopting their concentric Cs paradigm (see the Methods section) but using flankers inner as well as outer with respect to the target C. Next, we generated thousands of "mongrels" (physically different but perceptually the same set of textures) and investigated the TT model with the help of a neural-network classifier. Second, we determined whether or not the eccentricity confounded in the concentric Cs paradigm significantly affects response errors. Finally, we discuss whether explanations of crowding can or should be unified.

Methods

Participants

Nine observers (including the first author) with normal or corrected-to-normal vision (20/20 or better in each eye) participated in the study. Except for the first author, all observers were naïve as to the purpose and the experimental details of the study. All observers gave written informed consent before the experiment started. The experimental protocols were approved by the Institutional Review Board at the University of California, Berkeley. All procedures were in accordance with the Declaration of Helsinki.

Apparatus

Visual stimuli were presented on a 32-in. Display++ display (Cambridge Research Systems, Rochester, UK) at a resolution of 1,920 × 1,080 and a frame rate of 120 Hz. The viewing distance was 76 cm, and each dimension of a single pixel at this viewing distance was 1.63 arcmin. Head movements of observers were minimized via a chin/head rest. Eye movements were monitored at 1000 Hz with an Eyelink 1000 (SR Research, Ottawa, Ontario, Canada) video-based eye tracker to ensure observers did not make any saccadic eye movements toward the C-stimuli. Each block of trials started with a standard nine-point calibration procedure. All visual stimuli were generated in MATLAB (R2012b; MathWorks, Natick, MA) with the Psychophysics Toolbox 3 (Brainard, 1997; Pelli, 1997) and its Eyelink extensions (Cornelissen, Peters, & Palmer, 2002). For some of the computations described below, we used the CircStat toolbox (Berens, 2009). Luminance measurements were performed via a Konica Minolta LS-110 photometer. Observers' responses were obtained via a keyboard.

Stimuli and procedures

Observers sat in a dimly lit room and were presented with two white (~143 cd/m^2) concentric C-stimuli on a gray (~80 cd/m^2) background at 10° eccentricity to the right of the center of the display. The stimulus parameters were similar to those used in Harrison and Bex's study. The diameter of the target C was always fixed at 3.0°, whereas the diameter of the flanker was either 2.2° or 3.8°. These two sizes for the flanker correspond to identical target-flanker distances. Note, however, that the reference to the "target" and "flanker" was solely for labeling purpose, because in all trials, observers were asked to report the orientation of both C-stimuli, referred to as the "inner" and "outer" C. Thus, observers did not know which one was the target and which one was the flanker. The stroke-thickness of each C-stimulus was 0.4°. The small gap that determined the orientation of a C-stimulus was inserted by means of an invisible (with the same color as the background) radial line centered on the C with a thickness of 0.27° (10 pixels). The orientation of the target ranged from 0° to 360° polar angles (in steps of 1°). The flanker's orientation ranged from ±45° from the orientation of the target, so that we would obtain more data where observers have been reported to make more errors (Harrison & Bex, 2015). Note that observers were not told about any of these specifics.

Figure 2A illustrates the stimuli and procedures. Each trial began with a white fixation cross at the center of the display. After a random delay (uniformly sampled from 1 to 1.5 s), the stimuli were shown for 500 ms. Roughly 300 ms after the stimuli disappeared, a probe, also in the shape of a C, was presented at the center of the display, and observers reported the orientation of each C-stimuli by adjusting the orientation of this central C-probe. If two C-stimuli (target and flanker) were presented, the order of responses was randomized across trials: In some of the trials, observers reported the orientation of the target first, and in the remaining trials, they reported the flanker first. Which C to report first was instructed to the

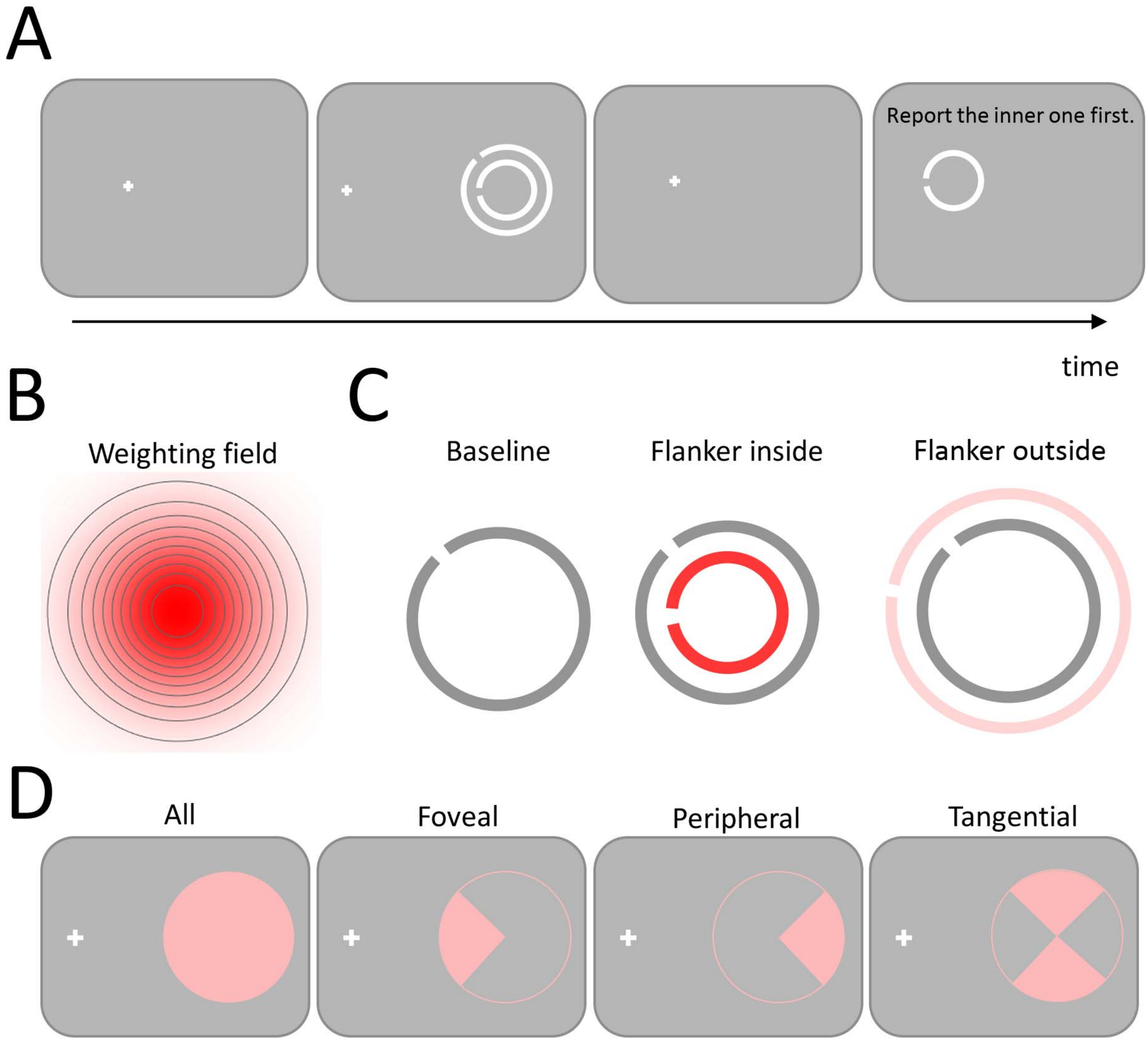

Figure 2. (A) The stimuli and procedures. Each trial started with a fixation point at the center of the display (note that each schematic shows mostly the right half of the display for clarity). A single C-stimulus (baseline condition) or two concentric C-stimuli (flanked condition) were shown for 500 ms at 10° eccentricity, right of the fixation point. Observers were asked to report the orientation of the Cs. The order of report (when there were two Cs) was randomized, and text instructions to observers appeared above fixation at each trial, after the stimuli disappeared from the display. (B) The weighting field proposed by Harrison and Bex (2015). Each gray circle represents an isoline with identical weights. (C) Stimulus conditions in the present study. In the baseline condition, only one C-stimulus was shown. In the crowded conditions, the flanker could be either inside or outside the target. Note that because observers were asked to report both objects, both objects were in fact "targets" for observers. Here the flanker-target distinction is used to simplify presentation of results. To be consistent, the target C always had the same size (gray circles), whereas the flanker could be smaller or larger than the target (red circles). Also note that the tone of red indicates the amount of weight enjoyed by the flanker in each condition. (D) Spatial binning of trials based on the orientation of the Cs.

observers with a text message (e.g., "Report the inner circle first") presented 5.4° above fixation, after the C-stimuli disappeared from the display. Moreover, the size of the central C-probe matched the size of the stimulus to be reported (target or flanker) to facilitate this process. Initial orientations of the central C-probe were randomly chosen from 0° to 360°. Note that both Cs were targets from the perspective of observers because observers had to report the orientations of both C-stimuli. Target-flanker distinction is used to organize the data and to facilitate the presentation of the results. This distinction does not affect (or is not affected by) the predictions of the HB model.

There were three stimulus configurations with equal probability within a block of trials: (a) target alone (baseline), (b) flanker inside, and (c) flanker outside. In a single block, observers completed 150 trials (50 trials for each condition). Observers could take a break anytime they wished, and if they did, the experiment continued after recalibrating the eye tracker. In total, each observer completed at least 900 trials (seven observers 900, one naïve observer 928, and another naïve observer 1,130 trials). Each observer completed practice trials (more than 50, less than 100) to become familiar with the task and the apparatus. Data from practice trials were excluded from all analyses.

Data analysis

The trials during which gaze position deviated more than 2° from the center of the display were discarded from further analyses. In total, only 5% of the trials were rejected according to this criterion, and on average 881 (±48) trials per observer were used in the following analyses.

Modeling response errors

We computed the response errors by subtracting the reported orientations from the actual stimulus orientations. We then used maximum-likelihood estimation to fit a single von Mises distribution (referred to as the vM model), as in the study of Harrison and Bex (2015). However, as we will show later, the vM model fit to our empirical data does not fully capture the observers' performance characteristics. Specifically, we observed more errors toward the tails of the distribution than the vM model would predict. Also, in some cases, we observed a sizeable number of 180° response errors that could not be fitted by the vM model. Consequently, we also tried two other models to fit our data: a von Mises plus a uniform distribution (referred to as the vM+U model) and two von Mises distributions (referred to as the vM+180vM model), one centered on the target's orientation and another one centered at 180° opposite orientation. We could have potentially tried other statistical models representing different crowding hypotheses such as averaging and substitution; however, our aim here was not to determine sources of errors. Our aim was to capture the shape of the distribution of response errors so that we can estimate the change in observers' performance by the presence of a flanker.

The vM model is defined as follows.

$$p_{\mathrm{vM}}(\in) = \frac{e^{k\cos(\in-\mu)}}{2\pi I_0(k)}, \quad (1)$$

where $p(\in)$ is the probability of response error $\in$, μ is the mean error, k is the concentration constant, and I_0 is a Bessel function of order zero. Note that $\in$ and μ are in radians. The vM+U model is defined as follows.

$$p_{vM+U}(\in) = wU(-180, 180) + (1-w)\frac{e^{k\cos(\in-\mu)}}{2\pi I_0(k)}, \quad (2)$$

where w represents the weight of the uniform component (i.e., guess rate) and U represents the uniform probability density within a range of (–180, 180). Finally, the vM+180vM model is defined as follows.

$$p_{vM+U}(\in) = w\frac{e^{k\cos(\in-\mu)}}{2\pi I_0(k)} + (1-w)\frac{e^{k\cos(\in-\mu+\pi)}}{2\pi I_0(k)}, \quad (3)$$

where w represents the weight of the primary von Mises distribution centered on the target's orientation. Note that the concentration coefficients of both von Mises distributions are identical for convenience. Therefore, the vM model has only one free parameter, whereas the vM+U and the vM+180vM models have two free parameters. The uniform component in the vM+U model captures the random guessing behavior of observers. This is an important point, especially for naïve observers who may not pay full attention to the task at all times and occasionally randomly guess the orientation of the target and flanker. These random guesses cannot be fully captured by a single von Mises distribution, and even if a single von Mises is forced to capture the variability in responses due to guessing, this results in artificially low concentration constants, k (i.e., very large standard deviations), for von Mises distribution. In fact, several studies reported that models containing a random guessing component perform better than those without it (Ester et al., 2014; Ester et al., 2015; Hanus & Vul, 2013; also see Põder & Wagemans, 2007).

We used Bayesian model comparison (BMC) technique (MacKay, 2003; Wasserman, 2000) to select the best statistical model to capture the data (for a step-by-step derivation of this technique, see Agaoglu, Agaoglu, Breitmeyer, & Ogmen, 2015, or Ester et al., 2015). In short, this technique computes the likelihood of data given a certain model, penalizes models with more free parameters, and assigns proportionately more weight to the penalizing factor for larger data sets in estimating the model performance. Assuming equal priors for all parameters (k for vM model, and k and w for vM+U model), the final form of the BMC metric is as follows:

$$\ln L(m_j) = \ln L_{\max}(m_j) - \sum_i^k \ln(R_i) + \ln\left[\int \exp\left(\ln L(m_j|\theta) - \ln L_{\max}(m_j)\right) d\theta\right], \quad (4)$$

where m_j represents the jth model, R_i represents the size of the range for ith free parameter, $L_{\max}(m_j)$ represents the maximum likelihood for the jth model, and finally ln $L(m_j)$ represents the BMC metric for the jth model. Parameters that correspond to $L_{\max}(m_j)$ can be regarded as the maximum likelihood estimation of the model parameters and are treated as the best parameters in the current study. We approximated the integral given in Equation 4 by a Riemann sum with 2,400 bins for k (from 0.025 to 60 in 0.025 steps) and 101 bins for w (from 0 to 1 in 0.01 steps). We set μ to zero for both models. In other words, the von Mises distribution was always centered on the actual orientation of the stimulus. To compare models, we first combined the BMCs in each condition by summing them (i.e., multiplying model likelihoods) and then computed the difference between the combined BMC values for each model. A better performing model will have a larger BMC based on Equation 4. A BMC difference of x between Model A and Model B corresponds to e^x-to-1 odds favoring Model A.

Perceptual error comparisons

We converted concentration coefficients of the best-fitting von Mises distributions to degrees via $\sigma = (\sqrt{1/k})(180/\pi)$, where k is the concentration coefficient and σ is defined as the perceptual error. In addition, the weight of the uniform component in the vM+U model and the weight of the second von Mises distribution in the vM+180vM model were also defined as metrics of perceptual error. We tested the predictions of the HB model by preplanned paired t tests.

Probability density estimation

We computed the reported difference between the flanker and target orientations and estimated the probability density of reported difference as a function of the actual physical difference between them. We used a fast and accurate state-of-the-art bivariate kernel density estimator (Botev, Grotowski, & Kroese, 2010), which does not assume any parametric model for the underlying data. Each density map was normalized so that the volume under the surface sums to unity. This enabled us to directly compare the density maps from different conditions against each other by simply subtracting one from another.

Simulations of alternative crowding models

We sought to determine whether or not our data can be explained by several alternative crowding models. Namely, we simulated the averaging, substitution, and attentional resolution accounts of crowding. For all of the following simulations, we took the actual orientations used in our experiments (>5,000 combinations of target and flanker orientations across all observers) and added a zero mean Gaussian noise with a standard deviation set to the standard deviation of the vM distribution in the winning statistical model (averaged across observers) for the baseline condition. This resulted in a distribution of orientations for each C-stimulus (target and flanker, or inner and outer C). For the averaging account, we randomly sampled an orientation from each distribution and took a weighted average of the two samples. Pure averaging occurs when samples have equal weights (i.e., 0.5). We also implemented a probabilistic averaging account in which both the weights and the probability of averaging were varied systematically. For the substitution account, we took a random sample from each distribution and randomly assigned it to the target or the flanker regardless of the actual origin of the sample. Previous studies showed that for certain tasks and stimuli, probabilistic substitution captures best the response errors (Ester et al., 2014; Ester et al., 2015; Hanus & Vul, 2013; Põder & Wagemans, 2007). Therefore, we have also independently varied the probability of substitution. For instance, for a substitution probability of one, observers *always* report the orientation of the flanker for the target and vice versa. On the other hand, for a substitution probability of zero, observers *never* substitute. The attentional resolution account posits that individual features are extracted by the visual system, and hence they are not lost; however, observers have difficulty at correctly attributing features to correct objects. We simulated this model by getting two samples either from the target distribution or from the flanker distribution or getting one sample from each distribution. We also implemented a probabilistic attentional resolution account.

The texture tiling model

Accumulating evidence suggests that crowding could arise due to textural (or summary statistics) representation of the peripheral information (Balas et al., 2009; Freeman et al., 2013; Freeman & Simoncelli, 2011; Keshvari & Rosenholtz, 2016; but also see Wallis, Bethge, & Wichmann, 2016). To determine whether a texture synthesis model inspired by the information processing in the early visual cortices can account for our data, we used the Portilla and Simoncelli (2000) texture synthesis model. This model is based on a series of band-limited filters with various scales and orientations. One can extract hundreds of local as well as global image statistics within a hierarchy of pooling regions from any arbitrary stimulus/image. A unique set of texture representations that are physically different but supposed to be perceptually similar (i.e., "mongrels," Balas et al., 2009; or "metamers," Freeman & Simoncelli, 2011) can be synthesized by feeding a random noise pattern and iteratively forcing the image statistics of the input image to match those of the target image. As commonly used in the literature, we used four scales, four orientations, and spatial neighborhood of nine pixels (Balas et al., 2009; Freeman & Simoncelli, 2011; Keshvari & Rosenholtz, 2016; Wallis et al., 2016). Because this model produces space-invariant textures, it may not preserve the general structure of a stimulus in its original form. To avoid deformations to the general structure of the stimuli, we used a noisy and very blurry version of the target image as the input to the texture synthesis process (Balas et al., 2009; Keshvari & Rosenholtz, 2016). Blurring was done by filtering the original images (128 × 128 pixels) with a Gaussian kernel with standard deviation of 0.5 pixel. The inner and outer diameters of the target ring were 73 and 87 pixels, respectively. This corresponds to a stroke width of seven pixels. We also added a Gaussian noise with zero mean and 0.1 variance. This preserved the general structure of the stimulus and introduced local textural changes only. We generated 1,600 synthetic images for the baseline condition and 8,000 images for the flanker outside condition. We did not attempt to compare the flanker inside and flanker outside conditions because it is not a "foveated" model (i.e., it does not account for eccentricity dependent tiling of receptive fields), and because the pooling regions do not overlap, it is not intended to account for peripheral vision or specific properties of crowding. Our aim in simulating this model was to investigate the types and frequencies of errors predicted by a texture synthesis model with our stimuli. Ideally, one would need to verify these two aspects by actually having human observers to foveally view these synthetic images and report the orientations. However, Balas et al. (2009) showed that a machine-learning algorithm trained with several principal components obtained from the hundreds of image statistics extracted by the texture model performed equally well compared with human observers. Here, neither the exact choice of the pattern classification algorithm nor its performance in classifying orientations is crucial. The important point is that the algorithm should be able to perform on par with human observers. We first computed the first 10 principle components from all extracted image statistics and then used a feed-forward neural network with three layers (10 units in the hidden layer, 360 sigmoidal units in the output layer, each of which represents orientations from 1° to 360°). We trained the network using the scaled conjugate gradient backpropagation algorithm ("trainscg" training function in MATLAB) with the blurry versions of the original images and made sure that the network achieved 100% classification performance for identifying the orientation of the target images, analogous to a human observer viewing the stimuli foveally with unlimited time. We trained separate networks for identifying the orientation of the target and flanker (i.e., inner and outer C) for the target images that contained both objects. Next, we obtained the predicted responses of the networks by feeding the synthesized textures as inputs. From predicted orientations, we computed the response errors for each C-stimulus independently.

Results

Modeling perceptual errors

To estimate how well observers did in reporting orientations, we fitted a set of mixture models to the distribution of response errors. We examined the distribution of response errors by using the BMC technique for each observer separately (see the Methods section). Figure 3 shows BMC differences with respect to the vM model for all observers. Figure 4A shows the best fits of each model in different conditions for a representative observer. For seven of the nine observers, the vM+U (a von Mises plus a Uniform distribution) model performed better. For one observer (S1), the vM performed slightly better than other models. Interestingly, for another observer (S9), the vM+180vM (a von Mises distribution centered on the actual orientation and another von Mises distribution centered on 180° opposite orientation) model performed the best, indicating that this observer reported completely opposite orientations in a considerable amount of trials (Figure 4B). On average, the ΔBMC is ~100, which corresponds to e^{100}-to-1 odds favoring the vM+U model over the vM model. According to Jeffrey's scale of interpretation (Jeffreys, 1998), this difference corresponds to a "decisive evidence" against the vM model. Therefore, the following results are based on the vM+U model. Note that the vM model overestimates the spread of response errors, as evident in Figure 4 (compare solid lines with dashed or dotted lines).

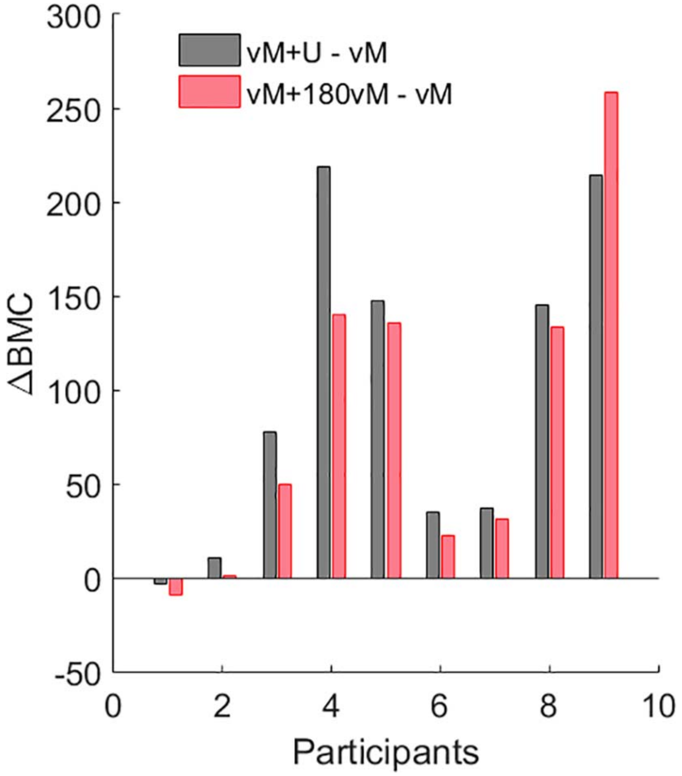

Figure 3. BMC differences from the vM model of all participants. Gray bars represent the BMC difference between the vM+U and vM models, whereas red bars represent the BMC difference between the vM+180vM and vM models. Positive values represent better model performance than the vM model.

Testing the weighting-field model

Figure 5A illustrates the predictions of the HB model, and Figure 5B shows the empirical results. As expected, when the target was accompanied by another C-stimulus, observers always made more (i.e., larger) errors in reporting either one compared with the case when the target was presented alone (baseline, Figure 5B horizontal line), the well-known crowding effect. Because of the shape of the weighting field, the HB model predicts weaker crowding when reporting the orientation of the inner (smaller) C (because its weight will be larger, Figure 2B, C). More specifically, the HB model predicts larger perceptual errors for the target than the flanker when the flanker is inside the target. Likewise, the HB model also predicts larger errors for the flanker when it is outside the target. However, as can be seen from Figure 5B, the perceptual errors of nine observers in the present study follow completely opposite trends. Although the model predicts stronger crowding (larger perceptual errors) for the target when the flanker is inside it, empirical data show a larger crowding effect when the flanker is outside the target ($t_8 = -2.271$; $p = 0.026$). Moreover, the model predicts weaker crowding for the flanker when it is inside the target (because it has a higher weight), but our data show the completely opposite trend ($t_8 = 5.540$; $p < 0.001$). Perceptual errors, or changes in performance due to crowding, can alternatively be measured by the guess rate (i.e., the weight of the uniform component in the vM+U model). We found a significant increase in random guessing in all crowded conditions ($p < 0.05$). However, we did not find any significant difference between the two flanker conditions for either the target or the flanker ($p > 0.64$; Figure 5C).

To examine how the presence of a flanker affected the distribution of response errors for the target, we computed the probability density of the reported difference between the target and flanker orientations as a function of the actual physical difference between them. Figure 6A shows the probability densities computed for the data pooled across observers. The leftmost panel shows the probability density for the entire data set, whereas the second and the third panels show the probability density for the flanker inside and flanker outside conditions separately. Several observations can be made about the origins of observers' behavior. First, the reported difference between the target and flanker is mostly close-to-veridical values, indicated by a strong probability along the 1:1 line in all panels. Second, there is a strong "cross" pattern formed by the identity line and the negative identity line (–1:1 line) in all panels. In other words, for a substantial amount of trials, the reported difference was –x when it was actually x. This pattern is a signature of the substitution errors and was evident in the probability density of reported differences in each and every observer. Interestingly, this pattern was not apparent in Harrison and Bex's (2015) study. Third, because of our experimental design (see the Methods section), although the actual difference ranged from –45° to 45°, there are quite a few cases in which the reported difference fell outside the (–45, 45) window. This can be a result of random guessing the orientation of one or both of the Cs.

To determine whether or not the probability density of reported differences was different across different flanker conditions, we subtracted the density in the flanker inside condition from that of the flanker outside condition. The resultant difference density is shown in Figure 6B with a different color map to facilitate comparisons. Here, red color represents a higher probability density for the flanker outside condition, whereas blue color represents a higher probability for the flanker inside condition. Moreover, greenish colors represent equal probability density. One clear pattern emerging from this figure is that the probability density along the identity line is larger in the flanker outside condition (indicated by the dark red region along the 1:1 diagonal in Figure 6B).

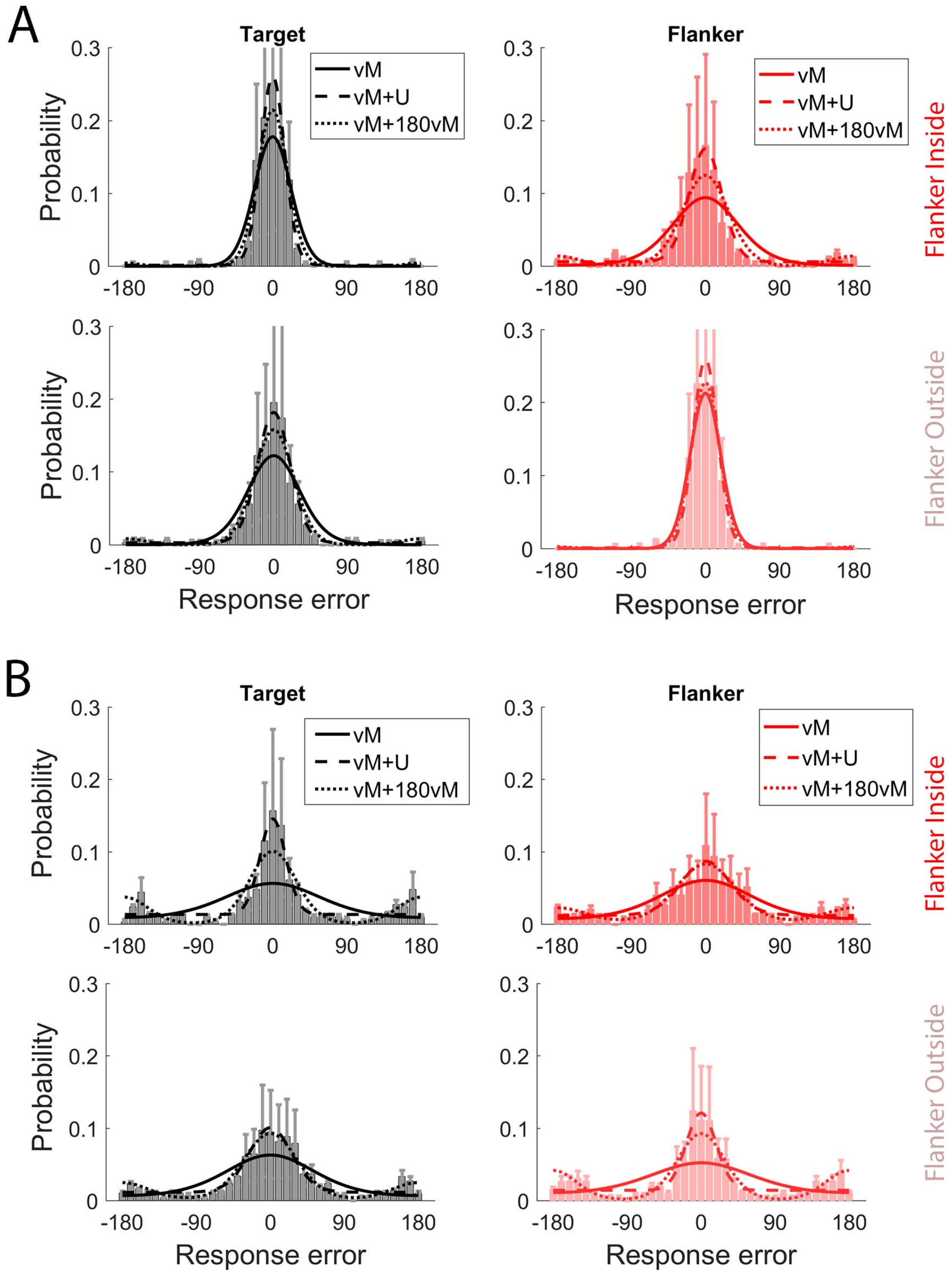

Figure 4. Distribution of response errors made when reporting the target (gray) and the flanker (red) and model fits for observer (A) S4 and (B) S9. Panels in the top row represent the flanker inside condition, whereas the bottom row represents the flanker outside condition. Note that the vM model (solid lines) overestimates the standard deviation of the error distributions. Error bars represent 95% confidence intervals obtained from bootstrapping (1,000 iterations, randomly sampling data with replacement).

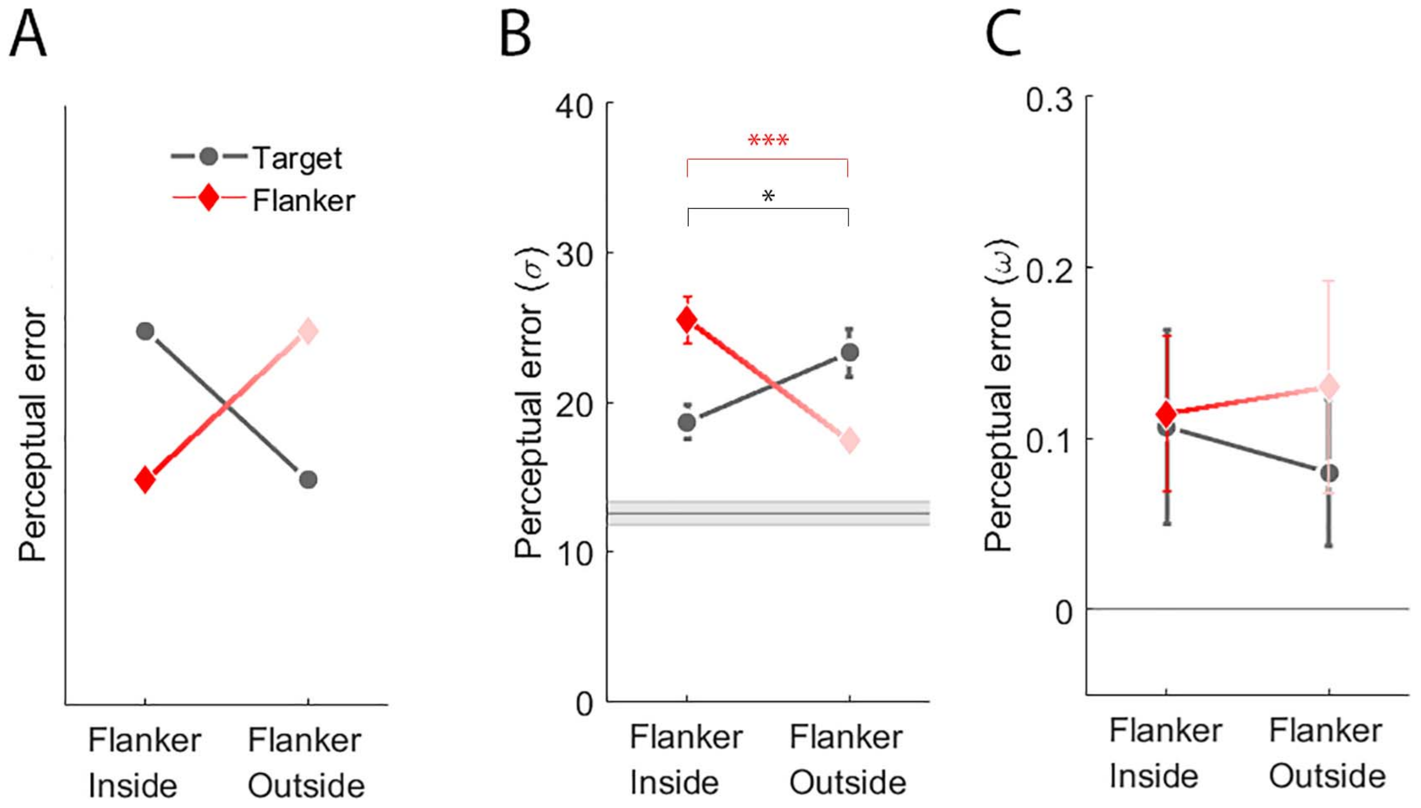

Figure 5. Perceptual errors in all experimental conditions. (A) Predicted pattern of results from the HB model simply based on the weighting field. Because of the shape of the weighting field, the HB model predicts weaker crowding when reporting the orientation of the inner (smaller) C. More specifically, the HB model predicts larger perceptual errors for the target than the flanker when the flanker is inside the target. Likewise, the HB model also predicts larger errors for the flanker when it is outside the target. (B, C) The results of the present study averaged across nine observers. (B) The standard deviation of vM in the best-fitting vM+U model. The horizontal gray line represents the perceptual error for the baseline (uncrowded) condition. (C) The weight of uniform in the best-fitting vM+U model (i.e., guess rate). The gray line at 0 represents the guess rate for the baseline condition; observers never guessed in this condition. Error bars and shaded region represent ±*SEM* (*$p < 0.05$, ***$p < 0.001$).

Testing the texture tiling model

Recent evidence suggests that crowding might be a result of textural representation or ensemble averaging of peripheral information in the visual system. Balas et al. (2009) and Keshvari and Rosenholtz (2016) presented a strong case supporting this view. More specifically, Keshvari and Rosenholtz (2016) recently used a slightly modified version of a texture model (Portilla & Simoncelli, 2000) and generated hundreds of mongrels that are physically different but perceptually similar to the original stimuli, for the stimuli used in three different studies (Freeman et al., 2012; Greenwood et al., 2012; Põder & Wagemans, 2007). The authors showed that the errors made when observers foveally viewed the mongrels and the errors made when observers peripherally viewed the original stimuli were similar. Remarkably, this similarity holds true for letter recognition, for orientation or position judgments of a letter-like stimuli, and for differentiating objects defined by conjunction of multiple features. We sought to determine whether the pattern of errors reported here can be predicted by the aforementioned texture model. To this end, we used 10 principle components[1] of the image statistics extracted by the texture model from our stimuli to classify the orientation of each C-stimulus by an artificial neural network (see the Methods section). Our results suggest that although the texture model can predict close to veridical reports, 180° errors, and random guesses, it cannot fully capture the frequency of different types of errors (Figure 7, rightmost panel; Figure 8, top-right panel). More specifically, it significantly overestimates the occurrence of 180° errors and very rarely produces substitution errors.

Testing conventional models of crowding

Harrison and Bex (2015) claimed that neither averaging of target and flanker features, nor reporting the flanker for the target and vice versa, nor limited attentional resolution can account for the *average* reported target-flanker differences in their study, although

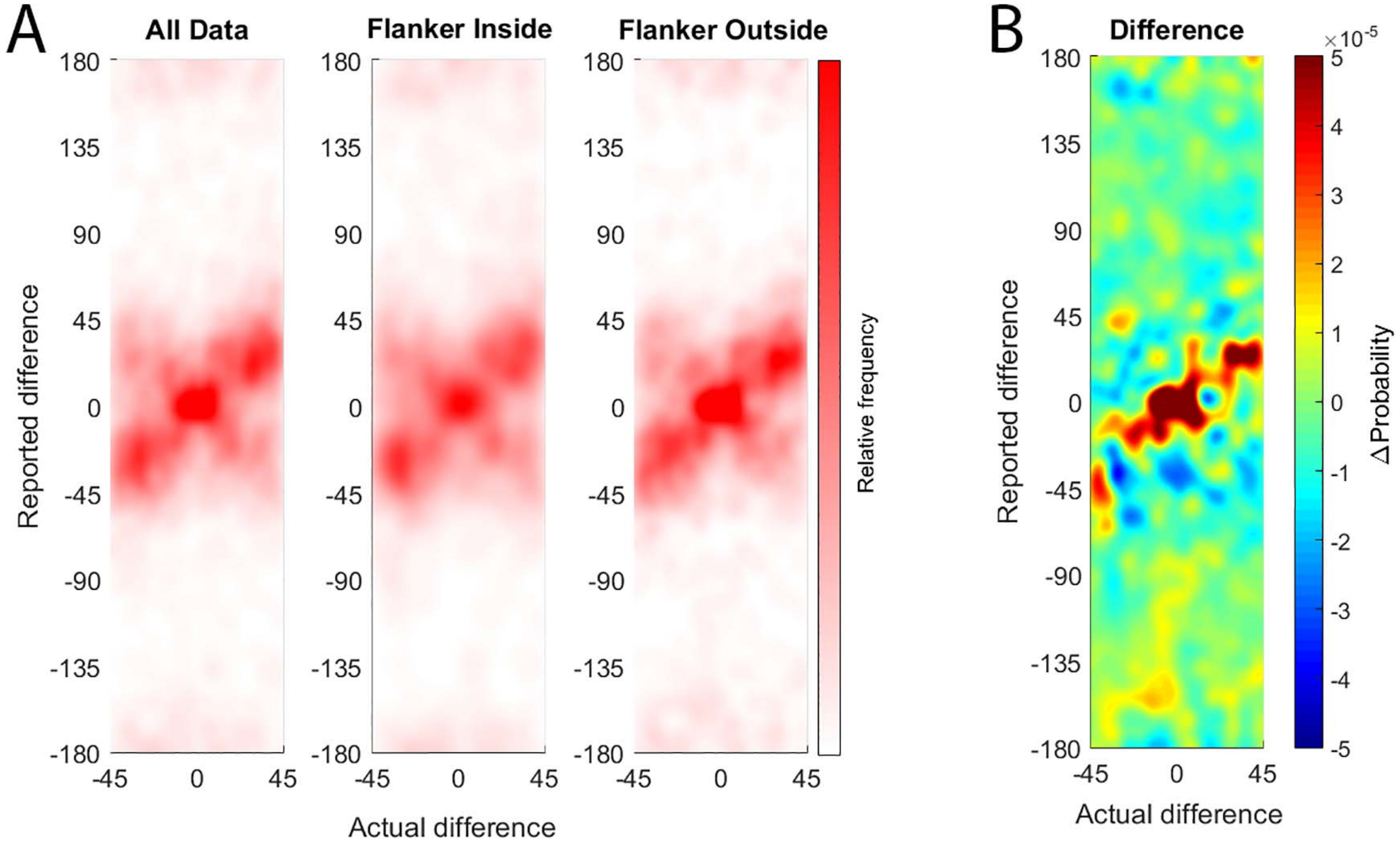

Figure 6. (A) Probability density maps for reported target-flanker difference as a function of actual differences. The leftmost panel shows all data combined, whereas the middle and the rightmost panel in A show data from the flanker inside and flanker outside conditions, respectively. (B) The difference between density maps in the flanker outside and flanker inside condition.

they noted that each of these accounts predicts different *distribution* of reported target-flanker differences (see supplementary figure S4 in Harrison & Bex, 2015). We think that asking observers to report the orientation of both objects in a crowded display and analyzing the reported differences against the actual differences is a very elegant and powerful way to test alternative models of crowding in a more decisive way. However, we also think

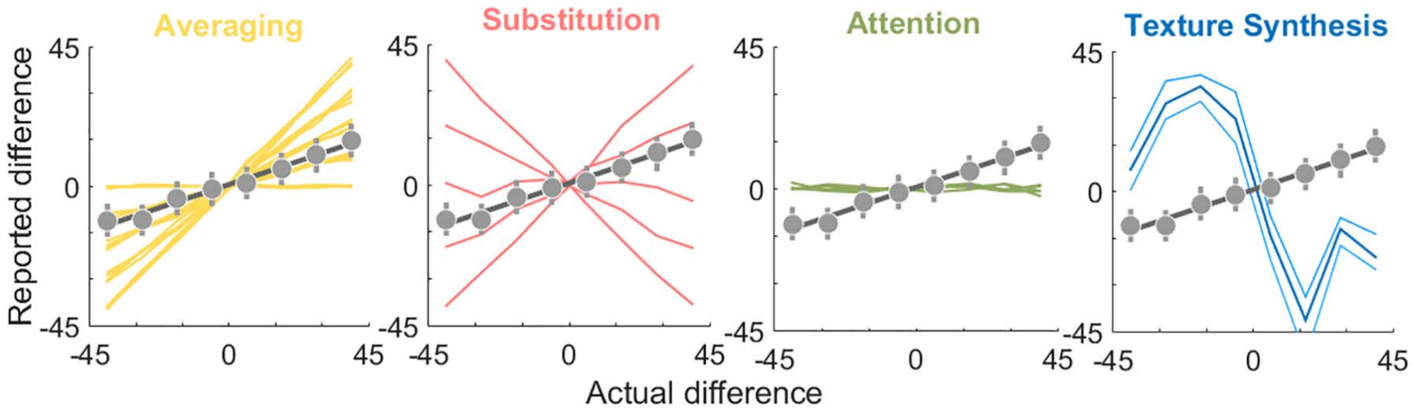

Figure 7. Predictions of alternative models for the mean reported target-flanker difference as a function of actual differences. Markers and the gray solid line represent empirical data and the best-fitting line with a slope of 0.334 ($R^2 = 0.977$, $p < 0.0001$). Error bars represent 95% confidence intervals obtained from bootstrapping across observers. All models (except the texture synthesis model) were simulated with a series of parameters to demonstrate the degree of freedom of each model. As can be seen, our results suggest that averaging and substitution models can account for the mean reported differences. However, the averaging model fails to account for the distribution of reported differences (see Figure 8).

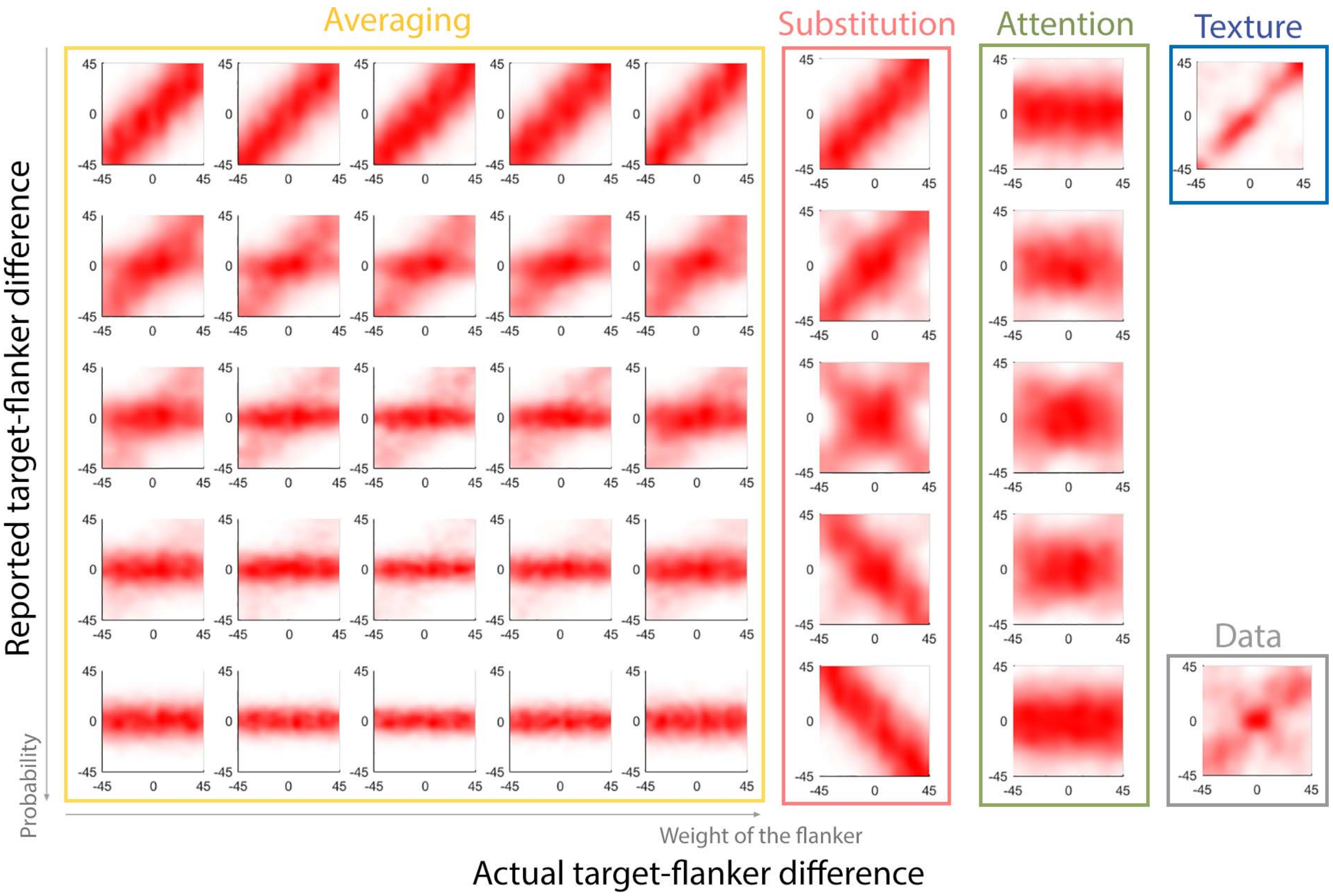

Figure 8. Empirical and simulated probability density maps for reported target-flanker difference as a function of actual differences. The empirical density map is shown in the bottom-right panel (gray outline). Each row represents a different probability (0 and 1 for the first and last rows, respectively) of a given mechanism. For instance, for the third row, the probability of averaging (or substitution, or attentional resolution) to occur is 0.5. For the averaging account, different columns represent different weights given to the flanker (0 and 1 in the leftmost and rightmost columns, respectively). The texture synthesis model predicts frequent occurrence for 180° errors, and because we focus on the (−45, 45) intervals in this figure, these errors are not shown here.

that comparing the average reported differences between the data and the predictions of straw man proposals on crowding is not very informative. More elaborate simulations should be performed to fully refute a hypothesis. To demonstrate that at least two of these alternatives, namely, averaging and substitution, can in fact account for the *average* reported target-flanker differences, we simulated these alternative models using the actual target-flanker orientation pairs used in our experiments (see the Methods section). Previous research showed that probabilistic substitution performed equally well or better than any variation of the averaging account (Ester et al., 2014; Ester et al., 2015; Hanus & Vul, 2013; Strasburger & Malania, 2013). Moreover, at least a handful of studies showed that weighted averaging, rather than pure averaging with equal weights, can account for errors made in crowding (Dakin, Bex, Cass, & Watt, 2009; Freeman et al., 2012; Greenwood et al., 2009, 2012). Probabilistic reports need not be limited to the substitution account and can take place in other accounts as well. Here, we investigated the degree of freedom of these models in explaining the reported target-flanker differences as a function of the actual differences. Figure 7 illustrates the behavior of each model along with the averaged reported target-flanker difference data obtained in the present study. Our simulations suggest that probabilistic weighted averaging and probabilistic substitution can separately account for the *average* reported target-flanker difference, although only the latter can partially account for the *distribution* of reported target-flanker difference (see Figure 8). The attentional resolution model always predicts zero mean for reported difference even with a probabilistic implementation. Figure 8 shows probability maps rather than just average values for all simulated lines in Figure 7. Note that none of the models discussed so far (averaging, substitution, and attentional resolution) can predict occasional guesses or 180° errors. In short, our results suggest that (a) testing

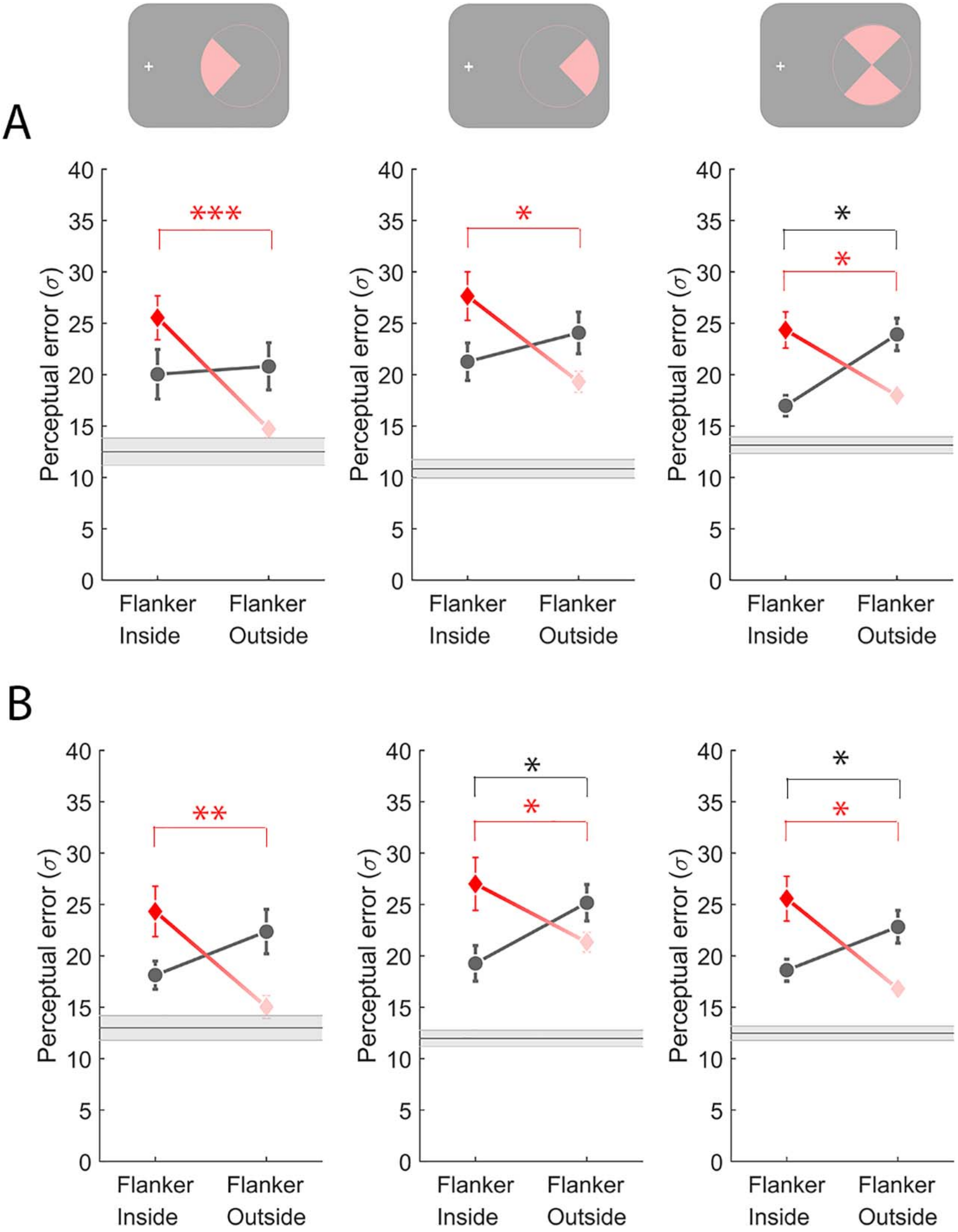

Figure 9. The effect of eccentricity confounding on the general pattern of results. Trials were categorized based on (A) the target's and (B) the flanker's orientation. The leftmost column represents the foveal trials, the middle panels represent the peripheral trials, and the rightmost column represents the tangential trials. See text for definition of each bin. All color/marker conventions are identical to those used in Figure 5. $^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

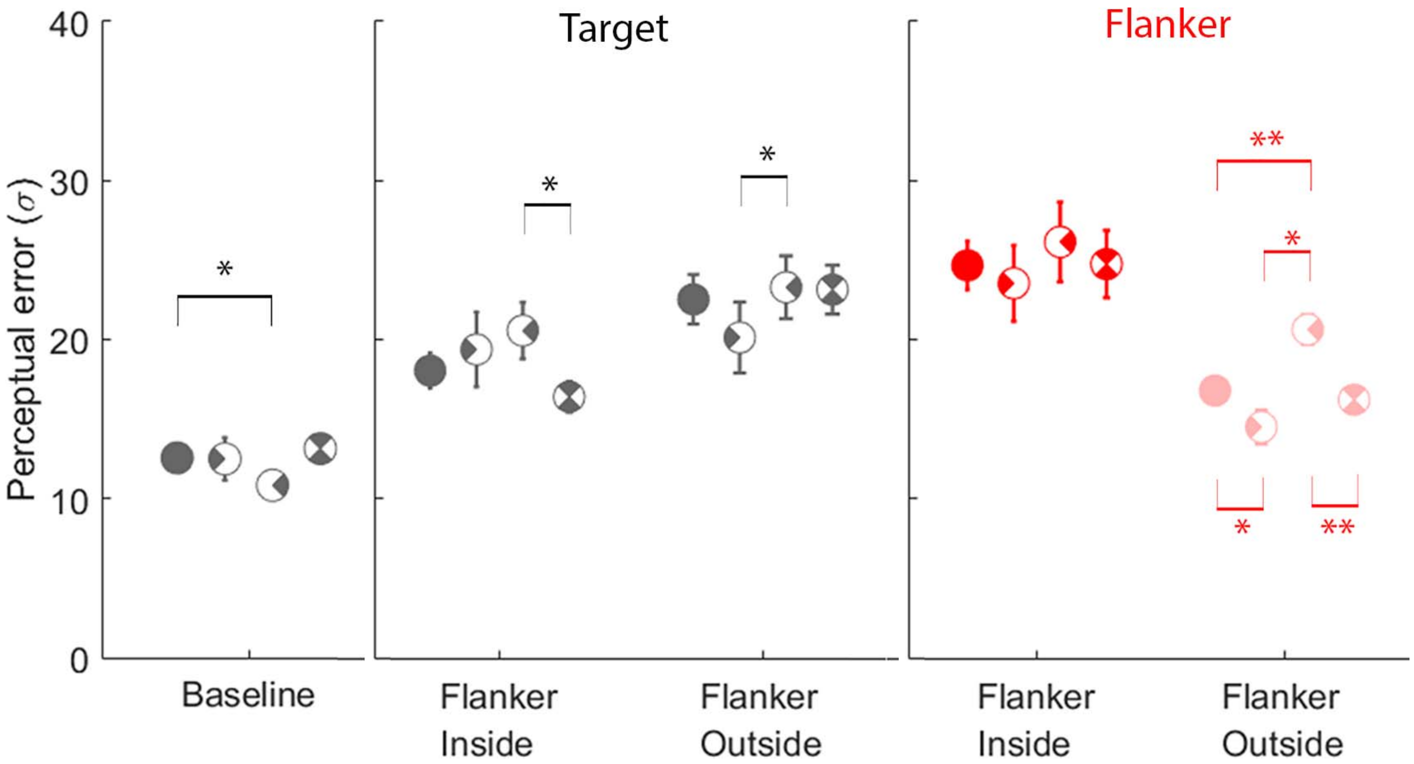

Figure 10. The effect of eccentricity on perceptual errors made when reporting each item on the display in all conditions. The leftmost panel shows data from the baseline condition, and the middle (rightmost) panel shows perceptual errors made when reporting the target (flanker) in the crowded conditions. Error bars indicate ±*SEM* ($n = 9$).

only the pure averaging or pure substitution is not informative and might be misleading and (b) using the mean responses rather than the distribution of responses may lead to incorrect conclusions regarding the underlying mechanism of crowding.

Testing the stimulus paradigm

The stimuli used by Harrison and Bex (2015) is powerful in identifying the types and the sources of response errors made by observers; however, it has a crucial confounding factor that limits its usage to a very limited set of eccentricities and stimulus sizes. Figure 1 plots the eccentricity from fixation of the small gap in each C-stimulus, which determines its orientation, as a function of orientation in both the present study and in Harrison and Bex (2015). As can be seen, the eccentricity of the one and only feature in the stimuli changes drastically with orientation of the gap as well as with the size of the flanker C-stimulus. For instance, for the largest flanker size used in Harrison and Bex (2015), the eccentricity of the small gap in the flanker ranged from roughly 4° to as large as 17°. Because recognition in both isolation and crowded displays depends strongly on eccentricity, it is obvious that one cannot pool the data obtained with different orientations of a flanker at this size.

To determine whether or not and how this confounding factor affected our results, we separated the trials into three spatial bins based on the orientation of the target (see Figure 2D). We defined these three spatial bins as follows. A trial was categorized as a "foveal" trial when the target was oriented between 135° and 225°. The trials in which the target was oriented between –45° and 45° were categorized as "peripheral" trials. Finally, all other trials were categorized as "tangential" trials. Because the orientation of the flanker was restricted to be within (–45°, 45°) around the orientation of the target, a foveal trial for the target may not necessarily be a foveal trial for the flanker. Therefore, we repeated the categorization of trials based on the flanker's orientation. We performed the same analyses as we reported for the combined data, and Figure 9 summarizes the perceptual errors for all conditions and spatial bins.

A two-way repeated-measures analysis of variance (ANOVA) for the target with spatial bins (foveal, peripheral, and tangential) and stimulus conditions (flanker inside and flanker outside) as factors revealed no significant main effect: spatial bins, $F(2, 16) = 1.765$, $p = 0.203$; stimulus conditions, $F(1, 8) = 2.238$, $p = 0.173$). There was no significant interaction between the main factors, $F(2, 16) = 3.471$, $p = 0.056$. A separate 2-

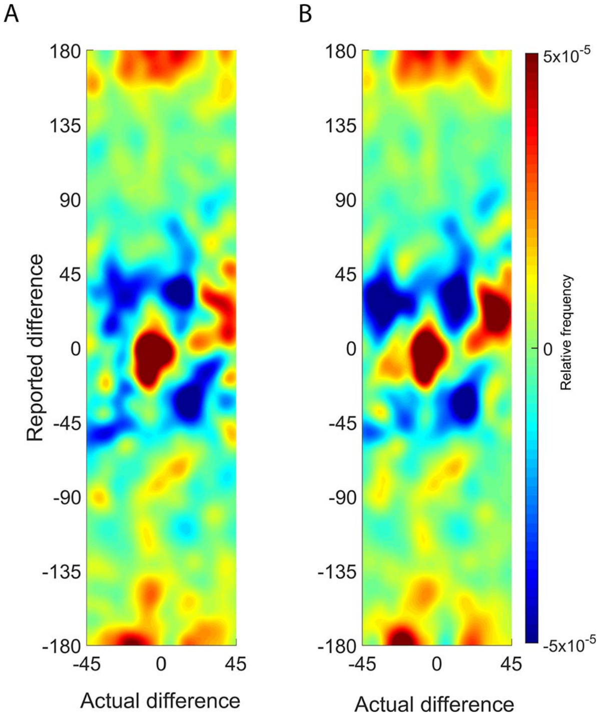

Figure 11. The difference between probability density maps (foveal-peripheral) of reported versus actual target-flanker difference in the foveal and peripheral conditions when binning was done based on the (A) target and (B) flanker.

way repeated-measures ANOVA for the flanker revealed a significant effect of main factor stimulus conditions, $F(1, 8) = 29.333$, $p < 0.001$. However, neither spatial bins, $F(2, 16) = 3.404$, $p = 0.059$, nor their interaction with stimulus conditions, $\mathrm{F}(2, 16) = 2.306$, $p = 0.132$, had a significant effect.

Next, we performed a series of preplanned paired *t* tests to determine whether or not the same relationship we found in Figure 5B is present for all spatial bins. In all cases, the crowding of flanker by the target was significantly stronger when the flanker was inside (red markers and lines in Figure 9). However, the crowding of target by the flanker was significantly influenced by the orientation of the Cs. When categorized based on the orientation of the target, perceptual errors in the flanker inside and flanker outside conditions did not differ from each other ($p > 0.05$) in the foveal and peripheral trials (Figure 9A, the leftmost and middle panels). Perceptual errors made while reporting the target were significantly larger in the flanker outside condition only in the tangential trials (Figure 9A, rightmost panel). Categorizing trials based on the orientation of the flanker produced similar results (Figure 9B).

Figure 9 allows us to see how the general pattern of results changes based on the eccentricity of the small gap in the Cs. Next, we performed preplanned paired *t* tests across different spatial bins for the target and flanker separately in all three stimulus configurations (baseline, flanker inside, and flanker outside) to determine whether or not perceptual errors for each item on the display was significantly affected by eccentricity. Figure 10 replots the results from Figure 9 in a way that is easier to visually make pairwise comparisons. The presence of at least one pairwise statistical difference would suggest that the concentric C paradigm cannot be used to investigate crowding with the stimulus parameters used in the present study. As indicated by stars in Figure 10, we found several pairwise differences in perceptual errors across different spatial bins.

We also computed the probability density functions for reported differences as a function of actual differences in all three spatial bins separately. We were specifically interested in the difference of probability densities of the foveal and peripheral trials, because the effect of eccentricity would be manifested more in the difference between these two bins. Figure 11 shows the probability density difference between the foveal and peripheral trials (Figure 11A based on the target, Figure 11B based on the flanker). Had there been no confounding factor in the stimulus paradigm, one would expect to see a uniform green difference map in Figure 11. However, there is a distinct and interesting pattern in the density difference maps. When the actual orientation difference between the target and flanker was within (−15, 15) and (30, 45), observers' responses were more veridical in the foveal trials (indicated by dark red regions in Figure 11). Interestingly, the reported differences were more spread out beyond the identity line (i.e., veridical perception) in the peripheral trials, as indicated by the blue regions in Figure 11. This also suggests that observers made more substitution errors in the peripheral trials compared with the foveal trials. Finally, the red regions in the upper and lower boundaries in the difference maps suggest that observers tend to report completely opposite (~180°) orientations for the target and flankers more often in the foveal trials than in the peripheral trials.

Discussion

Summary

Our aims in this study were (a) to evaluate the two prominent models as well as more conventional hypotheses of crowding by using the same stimulus paradigm and (b) to determine how crucial the eccentricity confound in the stimulus paradigm in Harrison and Bex (2015) was in affecting the data. We found that (a) the HB model, and hence the weighting field approach, and the TT model fail to account for perceptual errors for peripherally viewed concentric Cs, (b) the eccentricity confound significantly affects both the average and the distribution of response errors and therefore limits the usefulness of this stimulus paradigm to very small size/eccentricity ratios, and (c) although probabilistic weighted averaging and probabilistic substitution models can separately account for the average reported target-flanker differences, only the latter can partially account for the distribution of reported differences. In the following, we discuss our findings.

Failure of the weighting-field approach

According to the weighting field centered on a target stimulus, the inner C should always have a larger weight, and therefore, it should be a stronger "crowder" *and* should be less prone to crowding compared with the outer C. To test these predictions, we fixed the size of one C and varied the size of the other one. Here, to facilitate comparison of data and predictions, we termed the one with the fixed size across all conditions as the *target* and the other as the *flanker*, although observers were asked to report the orientation of both objects. Our results show patterns completely opposite from the predictions of the HB model.

Previously, it has been shown that crowding can be reduced or completely eliminated when flankers are grouped or when they form a good "Gestalt" (Manassi, Sayim, & Herzog, 2012, 2013). In a recent critique, Pachai et al. (2016) demonstrated that crowding is reduced when the target C is surrounded by multiple flankers with the same orientation rather than just one, and the HB model fails to account for this reduction in crowding. In their reply, Harrison and Bex (2016) claimed that Pachai et al. (2016) overlooked the critical aspect of their model, the weighting field. Moreover, the authors proposed that the apparent reduction of crowding with multiple surrounding flankers (having the same orientation) can be accounted for by a simple change in the front end of their model. Instead of using a bank of orientation-tuned filters in the first stage, Harrison and Bex (2016) used a filter-rectify-filter model of early visual texture processing (Bergen & Landy, 1991) and claimed that this updated version of their model can account for the data put forth by Pachai and colleagues (2016). Although Harrison and Bex (2016) did not perform any quantitative comparisons, here our aim was not to determine what type of front end would be best in a crowding model. Instead, we focused on the backbone of the HB model, the weighting field, and showed that regardless of the type of front end used, the weighting field approach fails to predict perceptual errors made for inner and outer Cs.

The proposed model also fails to explain radial-tangential anisotropy and inner-outer (with respect to fovea) asymmetry (Bouma, 1970; Toet & Levi, 1992), which are considered as litmus tests for crowding (Levi, 2008; Whitney & Levi, 2011). Although the former property can be explained by a change of distance metric (e.g., cortical distance), the latter cannot be accounted for by a weighting-field approach, a key component of their model, because which object's small gap is closer to the fovea also varies with their orientation, but the model assigns identical weights to all orientations.

The following point made by Harrison and Bex (2015) is a valid one, however. The same population of neurons might be responsible for partial representations of multiple objects, leading to contamination of encoded features of one from another. This is essentially a pooling model based on a spatially limited integration region. However, rather than using receptive field– or attentional spotlight–based pooling regions, the authors opted for an arbitrary pooling region. Is it possible to account for the data presented here by using a different spatial profile for the weighting field? How about a nonmonotonic weighting field? For instance, the weights might be largest on and around the contours of the target and fall off with distance from its contours. Alternatively, the weighting field could have both suppressive and facilitative regions, analogous to the "doughnut" model described by Strasburger and Malania (2013). Even if one can come up with an arbitrary weighting field profile to account for the data presented here, the biological underpinnings or implications of such a field will be questionable. In fact, although the bank of filters in the front end of the HB model is biologically inspired, the weighting field in the second stage remains at most an abstract concept, no different than other probabilistic models of crowding such as averaging or substitution.

New constraint for crowding models?

There is no shortage of models for crowding (see reviews: Levi, 2008; Whitney & Levi, 2011). These models can be categorized by the level of prior information needed for the model to operate. Some models require the statistics of features or stimulus dimensions to come up with estimates on the statistics of crowded percepts (Dakin et al., 2009; Ester et al., 2014; Ester et al., 2015; Freeman et al., 2012; Greenwood et al., 2009, 2012; Hanus & Vul, 2013; He et al., 1996; Parkes et al., 2001; Põder & Wagemans, 2007; van den Berg, Johnson, Anton, Schepers, & Cornelissen, 2012). Averaging (i.e., pooling), substitution, and attentional resolution models belong to this category.

Other models operate on image statistics and can be used with any arbitrary stimuli (Balas et al., 2009; Freeman & Simoncelli, 2011; Keshvari & Rosenholtz, 2016; see a recent critique: Wallis et al., 2016). Here, we sought to determine whether averaging, substitution, attentional resolution, and a variant of the Portilla and Simoncelli (2000) texture synthesis model (or the texture tiling model, Balas et al., 2009; Keshvari & Rosenholtz, 2016; Rosenholtz, Huang, & Ehinger, 2012; Rosenholtz, Huang, Raj, et al., 2012) can account for the probability density maps for reported target-flanker differences. Our results suggest that only a probabilistic substitution model can partially explain our findings.

Some studies on crowding favored pooling or averaging accounts, whereas some others reported predominantly substitution errors. Although one account can explain a set of data, it fails to generalize to another set. We suggest that because the exact stimulus characteristics and experimental procedures in different studies differ from each other extensively, any model of crowding based on a single mechanism is doomed to fail because it is expected to get predominantly averaging errors in one study and perhaps more substitution errors in another because of these differences. The success of the probabilistic substitution model in this study is no different. Although the target and flanker could have any orientation from 1° to 360°, the substitution account was the most promising one among all others. However, this might be simply due to the presence of only two objects (or features) and the requirement of the task that both objects should be reported. Observers might be making mistakes in report order, even though which item to be reported first was clearly instructed by several means (with a text and size-matched C on the display during the response phase). Support for this view comes from very few substitution errors reported by Harrison and Bex (2015). In their study, observers always reported the orientation of the inner C first, which was always the target. The apparent discrepancy between the two studies, therefore, might be explained by the differences in the experimental procedures. Additional experiments in which observers are asked to report only one or more than two items can be performed to fully resolve this issue.

Consistent with a recent report on texture models, our results cannot be explained by texture synthesis. We generated thousands of mongrels (i.e., perceptually similar but physically different set of textures) and computed predicted response errors by means of machine classification (see the Methods section). The texture model failed to account for our data. First, it almost never produced mongrels with substituted features. Second, it overestimated the frequency of 180° errors in both the baseline and flanked conditions. Figure 12 shows several example mongrels generated from the same original image (top row, baseline; bottom row, flanker outside). Note that in some cases, the Cs in the mongrels are completely opposite from their actual orientations, and in some others, a single C has multiple gaps.

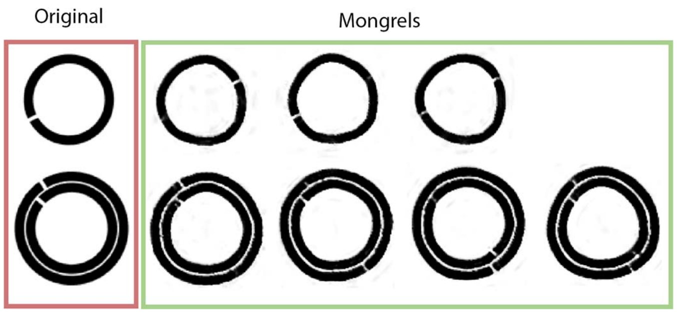

Figure 12. A set of mongrels generated by the texture synthesis model of Portilla and Simoncelli (2000) as implemented in Balas et al. (2009).

Concentric Cs paradigm

One powerful feature of the simple stimulus paradigm used by Harrison and Bex (2015) is that very strong crowding can be induced by only one flanker with only one feature. This simplicity allows us to examine the several potential ways crowding may occur. For instance, the orientation of the flanker might be reported for the target (or vice versa) or the orientation of both Cs might be averaged. Moreover, this paradigm makes reporting both the target and flanker more feasible and allows one to investigate the representations of both objects when crowding occurs. As we show here (Figures 11 and 12), the reported orientation difference between the target and flanker provides a very strong test for crowding models. In fact, none of the candidate models considered here alone could account for the distribution of reported differences.

Although it proved to be very useful, the experimental paradigm used by Harrison and Bex (2015) has a confounding factor that limits its usefulness beyond certain stimulus sizes; the eccentricity of the small gap in the Cs covaries with their orientation (Figure 1). We found that eccentricity confound affects perceptual errors significantly even at 10° eccentricity with a 2.2° flanker. Therefore, we do not recommend this stimulus paradigm for investigating crowding or other eccentricity-dependent phenomena.

Response bias, substitution errors, and random guesses

Our data differ from Harrison and Bex's in several other ways. Harrison and Bex (2015) reported an interesting response bias in which observers were biased to report dissimilar orientations for the target and flanker even when the physical orientations of the target and flanker were very close (see their figure 3). This finding, in fact, led them to add another stage to their model, a decision stage, to fully account for their results. However, in the present study, none of the observers showed any response bias. Second, observers made more substitution errors in the present study, which was manifested as a cross pattern in probability density maps for reported versus actual target-flanker difference (see Figures 6 and 12). Third, observers randomly guessed in ~12% of the trials in the present study, whereas no guessing behavior was reported in Harrison and Bex's (2015) study. We hypothesize that this discrepancy might be due to the difference in the subject pools—eight naïve observers and one of the authors took part in our study, whereas the two authors and a naïve observer participated in Harrison and Bex's study. Additional support for random guesses comes from several studies that showed that stronger crowding results in higher guess rates and that mixture models with a random guess component perform equally well or better than those without it (Ester et al., 2014; Ester et al., 2015; Hanus & Vul, 2013; Põder & Wagemans, 2007).

Can (should) crowding be unified with a univariate model?

Admittedly, we did not exhaust all possible crowding models in the literature. Our aim was to specifically look at the recent models, which claim to unify the crowding phenomenon. One of these models, which we did not investigate in depth because a quantitative formulation for arbitrary stimuli is currently not available, was presented by Nandy and Tjan (2012). These authors' rather unique and elegant proposal was aimed at the roots of crowding (i.e., why peripheral vision is limited beyond its limited resolution, rather than the consequences of crowding, i.e., the exact statistics of response errors). Nandy and Tjan (2012) proposed that crowding results from learning of saccade-confounded image statistics during early development of the visual cortex. They claimed that the two characteristic properties of crowding, namely, the extent of the crowding zone and the inner-outer asymmetry, can be accounted for by a single assumption of constant-sized (~6 mm) lateral interaction zones in V1, which leads to the eccentricity scaling of 0.5, the so-called "Bouma's law" (also see Pelli, 2008). They further showed that a brief temporal overlap between saccade execution and the deployment of spatial attention may lead to more erroneous representations of the peripheral stimuli along the radial axes from the fovea than the tangential axes and therefore can explain the third signature property of crowding, the radial-tangential anisotropy.

Because the Nandy-Tjan (NT) model can account for all fundamental characteristics of crowding, it is expected to be able to explain the main findings in the present study. However, note that our results are not fully compatible with all characteristics of crowding. For instance, when both the target and flanker were oriented toward the 3-o'clock direction (i.e., the peripheral condition; see Figure 9), the small gap of the target is more inward (i.e., closer to the fovea) than that of the flanker when the flanker is bigger than the target (the flanker outside condition). In this case, because of the inner-outer asymmetry of crowding, one would predict a stronger crowding (larger perceptual errors) for the target than the flanker. Moreover, when the flanker is smaller than the target (the flanker inside condition), one would expect completely opposite results (i.e., more crowding for the flanker). Our results (Figure 9, middle panels) indeed confirm these predictions. However, when we look at the case in which both the target and flanker were oriented toward the 9-o'clock direction (i.e., the foveal condition), our results cannot be explained by the inner-outer asymmetry of crowding. For instance, when the flanker is inside the target, its small gap will be more outer with respect to the fovea and should be less crowded. However, we found the crowding of the flanker's gap to be much stronger (Figure 9, leftmost column, the flanker inside condition). Clearly, the NT model and any model of crowding that can explain the inner-outer asymmetry of crowding cannot account for these results. However, it is also not clear whether these results should be considered under the umbrella of the crowding literature or should be considered as telltale signs of unsuitability of this stimulus paradigm to investigate crowding.

The crowding literature shows huge variations in the scaling of crowding zones with different stimulus conditions. The size of the crowding zone is obviously affected by the threshold or performance criterion used. However, it has also been shown that even with similar criteria, the size of crowding zones can range from 0.1 times up to ~0.7 times the eccentricity (Chung et al., 2001; Kooi, Toet, Tripathy, & Levi, 1994; Toet & Levi, 1992; Tripathy, Cavanagh, & Bedell, 2014; Wolford & Chambers, 1984), which corresponds to roughly 1.6 mm to 8.5 mm cortical distances (Pelli, 2008), respectively. Thus, because the rationale of the NT model comes from the neurophysiological finding that the extent of lateral interactions in V1 is roughly 6 mm (Stettler, Das, Bennett, & Gilbert, 2002), it is difficult to reconcile the vastly different crowding zones reported in the literature with the NT model. Additional steps in the visual processing hierarchy might possibly lead to this complicated picture about crowding zones (Whitney & Levi, 2011). However, this would mean that a model with a *single* mechanism cannot explain the crowding phenomenon.

Our empirical data and modeling results also suggest that crowded percepts cannot be fully accounted for by a *single* mechanism (or model). A combination of multiple mechanisms can possibly capture the statistics of the crowded percepts. However, the question arises as to whether or not the attempts to unify an apparently "multivariate" and highly complex phenomenon such as crowding with a "univariate" mechanism or model are worth the effort. In other words, should we really unify crowding? The common sense answer would be positive because parsimony of hypotheses/models is always sought after in science. The part of the problem is that many seemingly similar but mechanistically different phenomena tend to be categorized under the same umbrella in an effort to organize the knowledge in the field. Therefore, constraints for theoretical models become inflated. For instance, both masking by light and metacontrast masking are considered as masking, but they stem from distinct neurophysiological processes. Similarly, for a long time, crowding was thought to be similar to masking (Mansfield, Legge, & Ortiz, 1998; Nazir, 1992; Polat & Sagi, 1993; Townsend, Taylor, & Brown, 1971; Wolford & Chambers, 1984) until subsequent efforts to dissociate the two (Chung et al., 2001; Levi, Hariharan, et al., 2002; Levi, Klein et al., 2002; Levi, 2008; Pelli, Palomares, & Majaj, 2004). However, even in recent studies in which reduction in performance or larger response errors can be due to not only crowding but also other factors (such as surround suppression or lateral masking), performance impairments were attributed solely to crowding. For instance, in their Experiment 1, Harrison and Bex (2015) also used a flanker without a small gap (i.e., a ring) and claimed that the increase in perceptual errors due to the presence of this flanker is due to crowding. The authors discredited the substitution, averaging, or attentional resolution accounts because none of them can account for the crowding effect in this condition. However, one can claim that the increase in perceptual errors in this condition is more due to lateral masking. Moreover, in studies on metacontrast masking, a surrounding ring or square is generally used as a good metacontrast mask.

All in all, although we applaud the attempts at unifying various types of response errors in crowding studies, we think that without a better taxonomy of crowding—instead of calling everything crowding, perhaps introducing types of crowding (as in the masking literature)—unifying attempts will remain unsuccessful.

Keywords: crowding, texture synthesis, weighted averaging, peripheral vision, probabilistic substitution

Acknowledgments

The authors thank Drew Wodecki for his help in data collection, Peter Bex and William Harrison for helpful discussions and clarifying several details of their model, and Haluk Ogmen, Michael Herzog, and Matthew Pachai for their invaluable comments and suggestions on an early draft of the manuscript. This study was supported by research Grants R01-EY012810 and R01-EY017707 from the National Institutes of Health.

Commercial relationships: none.
Corresponding author: Mehmet Naci Agaoglu.
Email: mnagaoglu@gmail.com.
Address: University of California, Berkeley, Berkeley, CA, USA.

Footnote

[1] The exact choice of the number of principle components is not crucial here. The only limitation and concern here is to obtain close-to-perfect reports for the original stimuli, which is analogous to a human observer performing the task while foveally viewing the stimuli. Our preliminary simulations showed that any number of principle components larger than 5 five suffices to obtain 100% classification performance with the original stimuli, and does not affect the distribution of errors predicted by the texture model.

References

Agaoglu, S., Agaoglu, M. N., Breitmeyer, B. G., & Ogmen, H. (2015). A statistical perspective to visual masking. *Vision Research*, *115*, 23–39, doi:10.1016/j.visres.2015.07.003.

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

Berens, P. (2009). CircStat: A MATLAB toolbox for circular statistics. *Journal of Statistical Software*, *31*(10), 1–21, doi:10.1002/wics.10.

Bergen, J. R., & Landy, M. S. (1991). Computational modeling of visual texture segregation. In M. S. Landy & J. A. Movshon (Eds.), *Computational models of visual processing* (Vol. 1, pp. 253–271). Cambridge, MA: MIT Press.

Botev, Z. I., Grotowski, J. F., & Kroese, D. P. (2010). Kernel density estimation via diffusion. *Annals of Statistics*, *38*(5), 2916–2957, doi:10.1214/10-AOS799.

Bouma, H. (1970). Interaction effects in parafoveal letter recognition. *Nature*, *226*(5241), 177–178, doi:10.1038/226177a0.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*, 433–436, doi:10.1163/156856897X00357.

Chung, S. T. L., Levi, D. M., & Legge, G. E. (2001). Spatial-frequency and contrast properties of crowding. *Vision Research*, *41*(14), 1833–1850, doi:10.1016/S0042-6989(01)00071-2.

Cornelissen, F. W., Peters, E. M., & Palmer, J. (2002). The Eyelink Toolbox: Eye tracking with MATLAB and the Psychophysics Toolbox. *Behavior Research Methods, Instruments, & Computers*, *34*(4), 613–617, doi:10.3758/BF03195489.

Cox, P. H., & Riesenhuber, M. (2015). There is a "u" in clutter: Evidence for robust sparse codes underlying clutter tolerance in human vision. *Journal of Neuroscience*, *35*(42), 14148–14159, doi:10.1523/JNEUROSCI.1211-15.2015.

Dakin, S. C., Bex, P. J., Cass, J. R., & Watt, R. J. (2009). Dissociable effects of attention and crowding on orientation averaging. *Journal of Vision*, *9*(11):28, 1–16, doi:10.1167/9.11.28. [PubMed] [Article]

Ester, E. F., Klee, D., & Awh, E. (2014). Visual crowding cannot be wholly explained by feature pooling. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(3), 1022–1033, doi:10.1037/a0035377.

Ester, E. F., Zilber, E., & Serences, J. T. (2015). Substitution and pooling in visual crowding induced by similar and dissimilar distractors. *Journal of Vision*, *15*(1):4, 1–12, doi:10.1167/15.1.4. [PubMed] [Article]

Freeman, J., Chakravarthi, R., & Pelli, D. G. (2012). Substitution and pooling in crowding. *Attention, Perception, & Psychophysics*, *74*(2), 379–396, doi:10.3758/s13414-011-0229-0.

Freeman, J., & Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature Neuroscience*, *14*(9), 1195–1201, doi:10.1038/nn.2889.

Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P., & Movshon, J. A. (2013). A functional and perceptual signature of the second visual area in primates. *Nature Neuroscience*, *16*(7), 974–981, doi:10.1038/nn.3402.

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

Greenwood, J. A., Bex, P. J., & Dakin, S. C. (2012). Crowding follows the binding of relative position and orientation. *Journal of Vision*, *12*(3):18, 1–20, doi:10.1167/12.3.18. [PubMed] [Article]

Hanus, D., & Vul, E. (2013). Quantifying error distributions in crowding. *Journal of Vision*, *13*(4):17, 1–27, doi:10.1167/13.4.17. [PubMed] [Article]

Harrison, W. J., & Bex, P. J. (2015). A unifying model of orientation crowding in peripheral vision. *Current Biology*, *25*(24), 3213–3219, doi:10.1016/j.cub.2015.10.052.

Harrison, W. J., & Bex, P. J. (2016). Reply to Pachai et al. *Current Biology*, *26*, R353–R354, doi:10.1016/j.cub.2016.03.024.

He, S., Cavanagh, P., & Intriligator, J. (1996). Attentional resolution and the locus of visual awareness. *Nature*, *383*(6598), 334–337, doi:10.1038/383334a0.

Herse, P. R., & Bedell, H. E. (1989). Contrast sensitivity for letter and grating targets under various stimulus conditions. *Optometry and Vision Science*, *66*(11), 774–781.

Intriligator, J., & Cavanagh, P. (2001). The spatial resolution of visual attention. *Cognitive Psychology*, *43*(3), 171–216, doi:10.1006/cogp.2001.0755.

Jacobs, R. J. (1979). Visual resolution and contour interaction in the fovea and periphery. *Vision Research*, *19*(11), 1187–1195.

Jeffreys, H. (1998). *The theory of probability*. Oxford, UK: Oxford University Press.

Keshvari, S., & Rosenholtz, R. (2016). Pooling of continuous features provides a unifying account of crowding. *Journal of Vision*, *16*(3):39, 1–15, doi:10.1167/16.3.39. [PubMed] [Article]

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*(2), 255–279, doi:10.1163/156856894X00350.

Levi, D. M. (2008). Crowding—An essential bottleneck for object recognition: A mini-review. *Vision Research*, *48*(5), 635–654, doi:10.1016/j.visres.2007.12.009.

Levi, D. M., Hariharan, S., & Klein, S. A. (2002). Suppressive and facilitatory spatial interactions in peripheral vision: Peripheral crowding is neither size invariant nor simple contrast masking. *Journal* *of Vision*, *2*(2):3, 167–177, doi:10.1167/2.2.3. [PubMed] [Article]

Levi, D. M., Klein, S. A., & Hariharan, S. (2002). Suppressive and facilitatory spatial interactions in peripheral vision: Foveal crowding is simple contrast masking. *Journal of Vision*, *2*(2):2, 140–166, doi:10.1167/2.2.2. [PubMed] [Article]

MacKay, D. J. (2003). *Information theory, inference, and learning algorithms* (Vol. 7). Cambridge, UK: Cambridge University Press.

Manassi, M., Sayim, B., & Herzog, M. H. (2012). Grouping, pooling, and when bigger is better in visual crowding. *Journal of Vision*, *12*(10):13, 1–14, doi:10.1167/12.10.13. [PubMed] [Article]

Manassi, M., Sayim, B., & Herzog, M. H. (2013). When crowding of crowding leads to uncrowding. *Journal of Vision*, *13*(13):10, 1–10, doi:10.1167/13.13.10. [PubMed] [Article]

Mansfield, J. S., Legge, G. E., & Ortiz, A. (1998). The role of segmentation in lateral masking. *Investigative Ophthalmology and Visual Science, 39*(Suppl.), S859.

Mareschal, I., Morgan, M. J., & Solomon, J. A. (2010). Cortical distance determines whether flankers cause crowding or the tilt illusion. *Journal of Vision*, *10*(8):13, 1–14, doi:10.1167/10.8.13. [PubMed] [Article]

Nandy, A. S., & Tjan, B. S. (2012). Saccade-confounded image statistics explain visual crowding. *Nature Neuroscience*, *15*(3), 463–469, doi:10.1038/nn.3021.

Nazir, T. A. (1992). Effects of lateral masking and spatial precueing on gap-resolution in central and peripheral vision. *Vision Research*, *32*(4), 771–777, doi:10.1016/0042-6989(92)90192-L.

Pachai, M. V., Doerig, A. C., & Herzog, M. H. (2016). How best to unify crowding? *Current Biology*, *26*(9), R352–R353, doi:10.1016/j.cub.2016.03.003.

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

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442, doi:10.1163/156856897X00366.

Pelli, D. G. (2008). Crowding: A cortical constraint on object recognition. *Current Opinion in Neurobiology*, *18*, 445–451, doi:10.1016/j.conb.2008.09.008.

Pelli, D. G., Palomares, M., & Majaj, N. J. (2004). Crowding is unlike ordinary masking: Distinguishing feature integration from detection. *Journal of Vision*, *4*(12):12, 1136–1169, doi:10.1167/4.12.12. [PubMed] [Article]

Pelli, D. G., & Tillman, K. A. (2008). The uncrowded window of object recognition. *Nature Neuroscience*, *11*, 1129–1135, doi:10.1038/nn.2187.

Põder, E., & Wagemans, J. (2007). Crowding with conjunctions of simple features. *Journal of Vision*, *7*(2):23, 1–12, doi:10.1167/7.2.23. [PubMed] [Article]

Polat, U., & Sagi, D. (1993). Lateral interactions between spatial channels: Suppression and facilitation revealed by lateral masking experiments. *Vision Research*, *33*(7), 993–999, doi:10.1016/0042-6989(93)90081-7.

Portilla, J., & Simoncelli, E. P. (2000). Parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, *40*(1), 49–71, doi:10.1023/A:1026553619983.

Rosenholtz, R., Huang, J., & Ehinger, K. A. (2012). Rethinking the role of top-down attention in vision: Effects attributable to a lossy representation in peripheral vision. *Frontiers in Psychology*, *3*(FEB), 1–15, doi:10.3389/fpsyg.2012.00013.

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

Stettler, D. D., Das, A., Bennett, J., & Gilbert, C. D. (2002). Lateral connectivity and contextual interactions in macaque primary visual cortex. *Neuron*, *36*(4), 739–750, doi:10.1016/S0896-6273(02)01029-2.

Strasburger, H., & Malania, M. (2013). Source confusion is a major cause of crowding. *Journal of Vision*, *13*(1):24, 1–20, doi:10.1167/13.1.24. [PubMed] [Article]

Toet, A., & Levi, D. M. (1992). The two-dimensional shape of spatial interaction zones in the parafovea. *Vision Research*, *32*(7), 1349–1357, doi:10.1016/0042-6989(92)90227-A.

Townsend, J. T., Taylor, S. G., & Brown, D. R. (1971). Lateral masking for letters with unlimited viewing time. *Perception & Psychophysics*, *10*(5), 375–378, doi:10.3758/BF03207464.

Tripathy, S. P., Cavanagh, P., & Bedell, H. E. (2014). Large crowding zones in peripheral vision for briefly presented stimuli. *Journal of Vision*, *14*(6):11, 1–11, doi:10.1167/14.6.11. [PubMed] [Article]

van den Berg, R., Johnson, A., Anton, A. M., Schepers,

A. L., & Cornelissen, F. W. (2012). Comparing crowding in human and ideal observers. *Journal of Vision*, *12*(6):13, 1–15, doi:10.1167/12.6.13. [PubMed] [Article]

Virsu, V., & Rovamo, J. (1979). Visual resolution, contrast sensitivity, and the cortical magnification factor. *Experimental Brain Research*, *37*(3), 475–494, doi:10.1007/BF00236818.

Wallis, T. S. A., Bethge, M., & Wichmann, F. A. (2016). Testing models of peripheral encoding using an oddity paradigm. *Journal of Vision*, *16*(2):4, 1–30, doi:10.1167/16.2.4. [PubMed] [Article]

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, *44*, 92–107, doi:10.1006/jmps.1999.1278.

Weymouth, F. W. (1958). Visual sensory units and the minimal angle of resolution. *American Journal of Ophthalmology*, *46*(1 Pt. 2), 102–113.

Whitney, D., & Levi, D. M. (2011). Visual crowding: A fundamental limit on conscious perception and object recognition. *Trends in Cognitive Sciences*, *15*(4), 160–168, doi:10.1016/j.tics.2011.02.005.

Wolford, G., & Chambers, L. (1984). Contour interaction as a function of retinal eccentricity. *Perception & Psychophysics*, *36*(5), 457–460, doi:10.3758/BF03207498.