Quantifying error distributions in crowding

Deborah Hanus

Department of Brain and Cognitive Science, Massachusetts Institute of Technology, Cambridge, MA, USA

Edward Vul

Department of Psychology, University of California, San Diego, San Diego, CA, USA

When multiple objects are in close proximity, observers have difficulty identifying them individually. Two classes of theories aim to account for this crowding phenomenon: spatial pooling and spatial substitution. Variations of these accounts predict different patterns of errors in crowded displays. Here we aim to characterize the kinds of errors that people make during crowding by comparing a number of error models across three experiments in which we manipulate flanker spacing, display eccentricity, and precueing duration. We find that both spatial intrusions and individual letter confusions play a considerable role in errors. Moreover, we find no evidence that a naïve pooling model that predicts errors based on a nonadditive combination of target and flankers explains errors better than an independent intrusion model (indeed, in our data, an independent intrusion model is slightly, but significantly, better). Finally, we find that manipulating trial difficulty in any way (spacing, eccentricity, or precueing) produces homogenous changes in error distributions. Together, these results provide quantitative baselines for predictive models of crowding errors, suggest that *pooling* and *spatial substitution* models are difficult to tease apart, and imply that manipulations of crowding all influence a common mechanism that impacts subject performance.

Introduction

When similar objects flank a target in peripheral vision, they interfere with the individuation and identification of the target object (Bouma, 1970, 1973; Ehlers, 1936; Korte, 1923; Strasburger, Harvey, & Rentschler, 1991). This interference is commonly called "crowding" (Ehlers, 1936, 1953; Stuart & Burian, 1962; Woodrow, 1938) but has also been called "lateral inhibition" (Townsend, 1971), "lateral interference" (Chastain, 1982; Estes, Allmeyer, & Reder, 1976; Estes & Wolford, 1971; Wolford, 1975), and "lateral masking" (Geiger & Lettvin, 1986; Monti, 1973; Taylor & Brown, 1972; Wolford & Chambers, 1983) (for reviews see: Levi, 2008; Pelli & Tillman, 2008; Strasburger, 2005; Strasburger et al., 1991; Whitney & Levi, 2011). Crowding persists when cortical magnification is accounted for (Woodworth, 1938), when the observer is given unlimited exposure time (Wilkinson, Wilson, & Ellemberg, 1997), and for a wide range of stimuli (letters: Bouma, 1970; Woodworth, 1938; vernier targets: Westheimer & Hauske, 1975; Landholt Cs: Flom, Weymouth, & Kahneman, 1963; Levi, McGraw, & Klein, 2000). Crowding even persists when the target and flankers are presented to different eyes (Flom et al., 1963), indicating that it has a binocular cortical mechanism. Crowding-induced deficits in visual perception have been implicated in limiting reading speed (Pelli et al., 2007), visual search (Vlaskamp & Hooge, 2006), and object perception (Levi, 2008; Pelli & Tillman, 2008). In the midst of these observations, a central question remains: Why does the visual system suffer from crowding?

Here we consider two classes of mechanisms thought to account for crowding that traditionally have been thought to be at odds with one another: spatial pooling and spatial substitution. The first class, spatial pooling, suggests that the visual system pools information from fixed regions of space at a given scale, so if two objects appear in this region their features are somehow combined into a single percept, which cannot be reliably identified as either of the original objects. These accounts originally were thought to result from a single "processor," processing information from "parallel visual channels" (Wolford, 1975), or "compulsory averaging" of signals (Parkes et al., 2001). Later, it was described in terms of information pooling into "integration fields" (Pelli, Palomares, & Majaj, 2004) and has more recently been recast as probabilistic inference under the assumption that at certain spatial scales in

Citation: Hanus, D., & Vul, E. (2013). Quantifying error distributions in crowding. *Journal of Vision*, *13*(4):17, 1–27, http://www.journalofvision.org/13/4/17, doi:10.1167/13.4.17.

doi: 10.1167/13.4.17 Received June 12, 2012; published March 22, 2013 ISSN 1534-7362

the periphery, the visual system calculates a summary "texture" (Portilla & Simoncelli, 2000) of the visual world (Balas, Nakano, & Rosenholtz, 2009; Freeman & Simoncelli, 2011) and can only use this texture information to guess which individual objects were present in that display. Support for these accounts can be found in experiments demonstrating that subjects have similar error rates when they are asked to identify stimuli in the periphery, when they are asked to identify a model-based texture summary of those same stimuli at the fovea (Balas et al., 2009), and from experiments that show that subjects fail to detect changes in the periphery as long as their textural representation is matched (Freeman & Simoncelli, 2011). On these spatial pooling accounts, the percept of a crowded object corresponds to a "mongrel" (Balas et al., 2009), which could arise from a number of possible component objects, thus yielding undifferentiable object metamers (Freeman & Simoncelli, 2011) from which one cannot infer the individual component objects. Thus, when two letters are presented within one pooling region, the features of both letters are integrated to form one synthesized percept that is neither the first nor the second letter (and need not look like any real letter at all).

The second class of mechanisms, spatial substitution, suggests that the visual system has limited resolution in its ability to pick specific objects, and thus adjacent objects may be substituted for one another. For instance, when attentional resolution (He, Cavanagh, & Intriligator, 1996; Treisman & Gelade, 1980) is limited, observers cannot correctly identify the crowded target because they cannot attend to a region sufficiently precise to isolate that object. When the attended region includes multiple objects, any one of them might be reported. Crucially, however the individual objects' representations are not pooled into a common representation, they are just spatially confused. Support for this account arises from experiments that show that while observers cannot report the orientation of crowded Gabors, they can adapt to their orientation (He et al., 1996) and from experiments that demonstrate that crowding-like interference between objects is not dependent on retinotopic position but on their position within a moving focus of attention (Cavanagh & Holcombe, 2007). This account can be cast in terms of spatial uncertainty (Strasburger et al., 1991) and sampling (Vul, Hanus, & Kanwisher, 2009): Observers are uncertain about the spatial location of the target and the fact that subjects can make spatially independent errors when asked to make two guesses about one target (Vul & Rich, 2010; Vul et al., 2009). Consequently, observers can identify and report the objects around the target, but they do not know which of those items is the target and thus tend to report items around the target (Strasburger et al., 1991).

The crucial distinction between spatial substitution and pooling models is this: Can crowded perception be explained by an additive mixture of responses driven by independent adjacent stimuli (spatial substitution models)? Or can perception of a given ensemble not be predicted from the individual stimuli independently and instead require interactions between adjacent stimuli (pooling models)? Although conceptually these two classes of models differ substantially from one another, these classes are difficult to distinguish without strong constraints on the space of textures used in texture synthesis models or individual item confusability in substitution models. For instance, if the space of textures includes high-level features resembling entire letters, then texture synthesis can yield behavior quite similar to spatial substitution. Likewise, if individual letters are not reported perfectly but instead may be confused for similar-looking letters, then independent spatial substitution models can yield behavior that seems like it requires spatial pooling. For instance, the observation that flankers similar to the target tend to be reported more often than dissimilar flankers (Bernard & Chung, 2011) has been used to argue that adjacent items must be pooled into a joint representation (Freeman & Simoncelli, 2011). The argument states that this joint representation must exist because independent substitution of adjacent letters could not yield such an effect, and the behavior could only be produced if adjacent letters were not considered independently. However, as we show later, this observation can be explained by independent spatial intrusions with letter confusion.

Here we aim to characterize the kinds of errors that observers make when reporting crowded letters while varying the distance between flankers (Experiment 1; cf. Bouma, 1970; Strasburger et al., 1991), eccentricity (Experiment 2; cf. Andriesen & Bouma, 1976; Strasburger et al., 1991), and the allocation of attention (Experiment 3; cf. Eriksen & Collins, 1969; Posner, 1980; Strasburger, 2005), using a stimuli arrangement reminiscent of that used by Eriksen and Collins (1969), Eriksen and Eriksen (1972), Eriksen and Hoffman (1974), and Eriksen and Rohrbaugh (1970). We compare these error distributions to predictions from eight statistical models of crowding that formalize (albeit with simplifications) variations of possible mechanisms of spatial substitution and spatial pooling.

The error models we consider differ along several dimensions that allow us to ask specific questions about the properties of crowded perception. To what extent do spatially adjacent letters influence the kinds of errors observers make (rather than just the error rate)? Do observers tend to report letters exactly or guess randomly, or do they tend to confuse some letters for others? How do observers combine adjacent errors to

make responses? Do different stimuli contribute independently (additively) to errors? Do errors reflect nonlinear combinations of adjacent stimuli? Furthermore, by quantifying separate characteristics of the error distributions via a few parameters, we can ask how varying crowding manipulations influence perception. Do different manipulations of crowding influence perception in qualitatively disparate ways, or do all manipulations qualitatively effect perception in the same way? We find that error distributions in crowding reflect both individual graphemic confusion of letters as well as a spatially dependent influence of nearby items, but our data cannot discern whether errors arise from additive independent combinations of adjacent items or from multiplicative pooling of those items. Furthermore, our data suggest that varying eccentricity, precueing, and letter spacing all influence crowding errors in the same way.

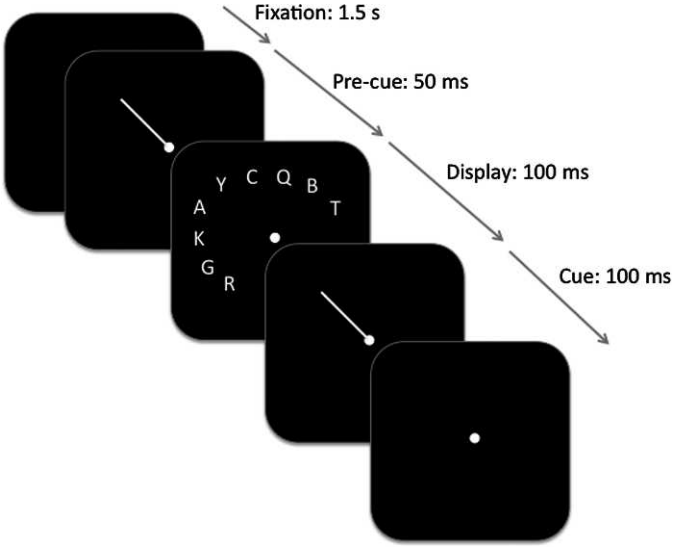

Figure 1. General experiment design: Subjects saw an array of nine letters arranged on the arc of a circle separated by a particular angle along the arc and had to report the central letter (also cued by a line extending from fixation). The fixation appears on screen for 1.5 s, followed by a precue (for 50 ms in Experiments 1a and 2, 200 ms in Experiment 1b, and variable in Experiment 3), then the letter array appears for 100 ms, and is followed by the cue for 100 ms. In different experiments we manipulated the angular spacing of the letters (Experiment 1a & 1b), the eccentricity of the arc (Experiment 2), or the precue duration—how far in advance of the letters the cueing line appeared (Experiment 3).

General methods

Participants

Sixty subjects between 18 and 45 years of age were recruited from the Massachusetts Institute of Technology subject pool and paid $10 for participation. Eighteen of these participated in Experiment 1a, 14 in Experiment 1b, 15 in Experiment 2, and 13 in Experiment 3.

Materials and procedure

On each trial in every experiment, subjects saw nine capital letters randomly selected from the Latin alphabet (26 letters total) arranged along an arc of a circle centered on the fixation point. The central letter of the array was pre- and post-cued by a line extending from fixation. Subjects had to report this central letter of the array by pressing the corresponding key on the keyboard. Because we know where each letter appeared on every trial, we could identify both the identity and the position of the reported letter: It was either one of the nine letters presented (and its corresponding position) or one of the 17 letters that was not present on screen. We used these responses to characterize the kinds of errors that subjects made. In Experiments 1a and 1b, we manipulated the spacing of the letters along the arc. In Experiment 2, we manipulated the eccentricity of the stimulus arc. In Experiment 3, we manipulated the precueing time.

Each trial began with a 1.5 s fixation display, followed by the presentation of the precueing line that remained on screen for an amount of time ranging between 50 ms and 200 ms (varying across experiments and explicitly manipulated in Experiment 3). Following the precue presentation, the letter array appeared for 100 ms. The letters were arranged on an arc of a circle centered with a regular spacing specified by the arc-angle between letter positions (see Figure 1). The center of the arc was chosen randomly (uniformly from the full perimeter of the circle) on each trial. After the offset of the letter array, the cueing line appeared again as a post-cue for another 200 ms.

Each experiment began with two familiarization trials; the results of these trials were discarded. Following these trials, each participant completed approximately 400–450 trials (the exact number varied slightly across participants). Conditions were pseudorandomly assigned to trials so that we had roughly equal numbers of trials in each condition for each subject (about 55–65, varying across experiments).

Experiments were programmed in PsychToolbox (Brainard, 1997) in Matlab 7 on a Windows XP computer. Stimuli were displayed at a resolution of 1024×768 on a 19 in Viewsonic G90f monitor at a viewing distance of roughly 50 cm.

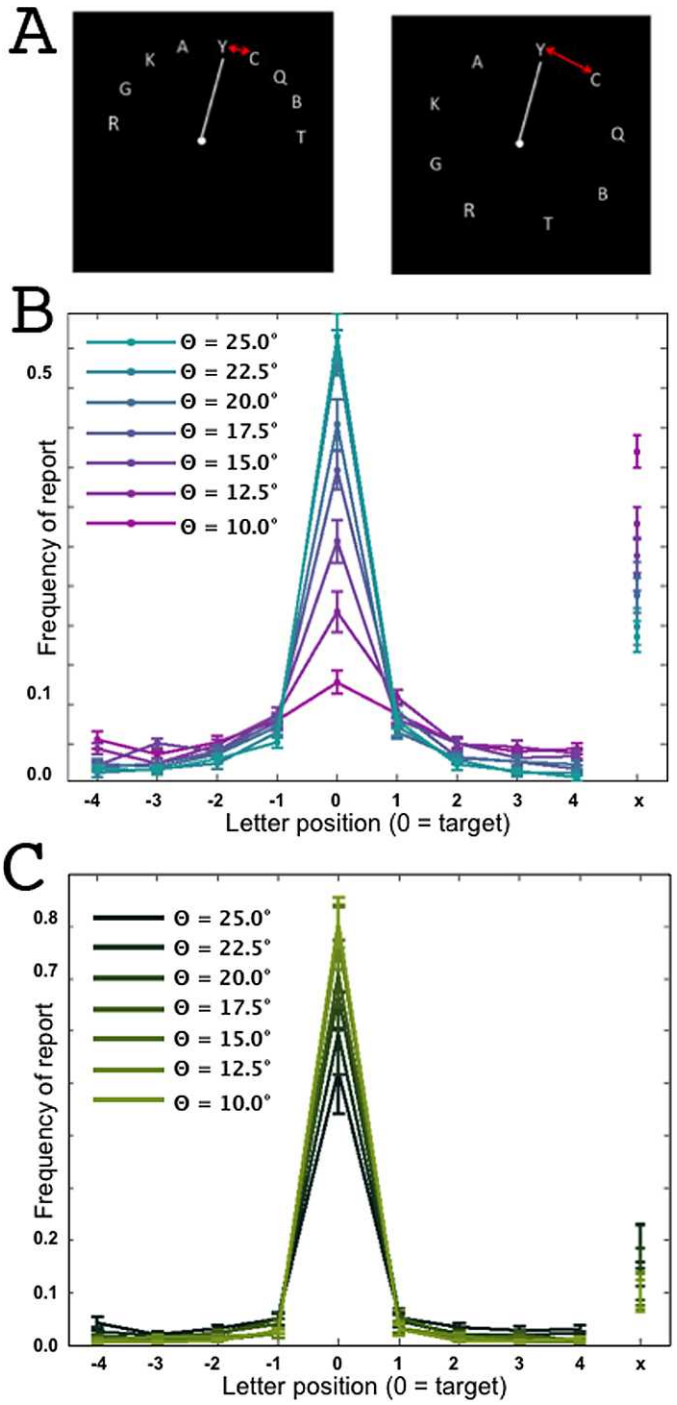

Figure 2. (a) Experiment 1a manipulated letter spacing, measured in angle of arc, while holding eccentricity and precueing duration constant. (b) The frequency (*y*-axis) with which different letter positions (*x*-axis) are reported in Experiment 1a. Along the *x*-axis, zero corresponds to the target, negative one and one are the immediate flanking letters, etc. (the disjointed points to the right indicate reports of letters that were not presented in the display). As letters were presented closer together, subjects performed worse (the probability of reporting the target decreased), and the rate at which adjacent letters, or even non-presented letters were reported increased. (c) The frequency (*y*-axis) with which different letter positions (*x*-axis) are reported in Experiment 1b.

Experiments 1a and 1b: Letter spacing

In Experiment 1a, we varied the angle between the presented letters to be 10°, 12.5°, 15°, 17.5°, 20°, 22.5°, or 25° of arc (see Figure 2) or 1.22°, 1.52°, 1.84°, 2.17°, 2.50°, 2.85°, or 3.20° of visual angle, while the other parameters were held constant. The eccentricity of the arc was fixed at 7° of visual angle, the precue appeared 50 ms prior to the letter array, and the individual letters' widths extended about 0.68° of visual angle each.

Experiment 1b differed from Experiment 1a only in that the precue was 200 ms rather than 50 ms so as to create a family of conditions where cue uncertainty played a lesser role. Because performance was near ceiling for a number of conditions, and the goal of this experiment is to assess whether our conclusions hold when spatial uncertainty is eliminated, we will exclude this experiment from the primary analyses and analyze it separately (see Appendix C).

Experiment 2: Eccentricity

In Experiment 2 we varied the eccentricity of the letters to be 5°, 6°, 7°, 8°, 9°, or 10° of visual angle, while the spacing between the letters was fixed to be a constant 13.125° of arc on the circle. The precue appeared 50 ms before the onset of the letter array (as in Experiment 1a; see Figure 3). In an attempt to correct for crude cortical magnification effects (Rovamo & Virsu, 1979; Rovamo, Virsu, & Näsänen, 1978), the letter size scaled with eccentricity according to the following equation:

$$w = 0.171 * \left(1 + (0.42 * E) + (.0000875 * E^3)\right), \quad (1)$$

where w is the width of the letter in degrees visual angle, and E is the eccentricity in degrees visual angle. The parameters for this equation were obtained by averaging the superior and inferior visual field expressions reported in Carrasco and Frieder (1997) (which were in turn obtained by averaging estimates from Rovamo & Virsu, 1979 and Virsu & Rovamo, 1979). This yielded letter widths of roughly 0.53°, 0.61°, 0.68°, 0.75°, 0.83°, and 0.90° visual angle for eccentricities of 5°, 6°, 7°, 8°, 9°, and 10° visual angle, respectively. Thus, further in the periphery the letters were larger, and their spacing in degrees of visual angle increased (13.125° of arc for eccentricities of 5°, 6°, 7°, 8°, 9°, and 10° visual angle corresponds to arc lengths of 1.2°, 1.4°, 1.6°, 1.8°, 2.1°, and 2.3° of visual angle, respectively). The increasing spacing and increasing letter size roughly canceled out, such that letters subtended roughly 40% of the center-center distance between letters.

Experiment 3: Precueing

In Experiment 3, we varied how far in advance of the letter display the precue appeared: This precueing

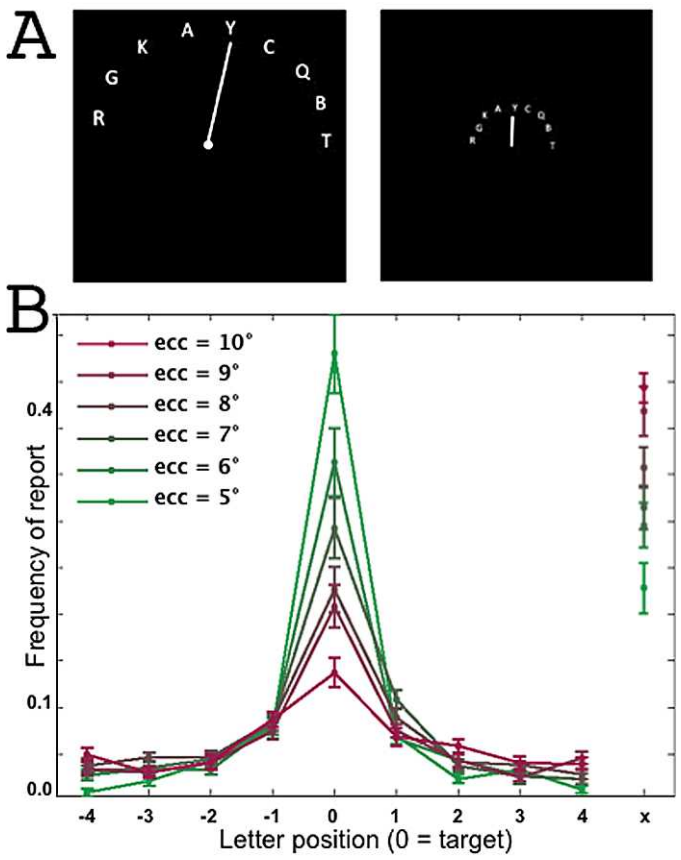

Figure 3. (a) Experiment 2 manipulated stimulus eccentricity, measured in degrees visual angle, while holding spacing (in terms of angle of arc) and precueing duration constant. Letter size was scaled to compensate for cortical magnification. (b) The frequency (*y*-axis) with which different letter positions (*x*-axis) are reported. At greater eccentricities, subjects reported the target less often, and flankers and non-presented letters more often.

duration was 0, 25, 50, 75, 100, 150, or 200 ms. The letters were presented at a constant eccentricity of 7° of visual angle with a constant spacing of 15° of arc (yielding a center-center arc length between adjacent letters of 1.8° visual angle; see Figure 4).

Data

In this section we briefly introduce plots of the raw response histograms to demonstrate the efficacy of our crowding manipulations before we undertake a more thorough model-based analysis of errors.

Experiments 1a and 1b: Letter spacing

In Experiment 1a, we varied the angle between the presented letters to be 10°, 12.5°, 15°, 17.5°, 20°, 22.5°, or 25° of arc, while the eccentricity was 7° of visual angle and precue duration was 50 ms. Figure 2 illustrates this manipulation and shows the distribution of spatial errors for each spacing condition averaged over subjects. As letters are presented closer together, the target letter is reported less often. Instead, letters adjacent to the target are reported more often and so are letters that were not presented in the display. We will later offer numerical summaries of these results evaluated in the context of different error models. Figure 2 shows the results of Experiment 1b, which is identical to Experiment 1a with the exception of a longer (200 ms) precue duration. Although large spacings with a long precue make the task quite easy (and participants are nearly at ceiling), we see the same pattern at smaller letter spacings: In place of targets, flankers and non-presented letters are reported more often when spacing decreases.

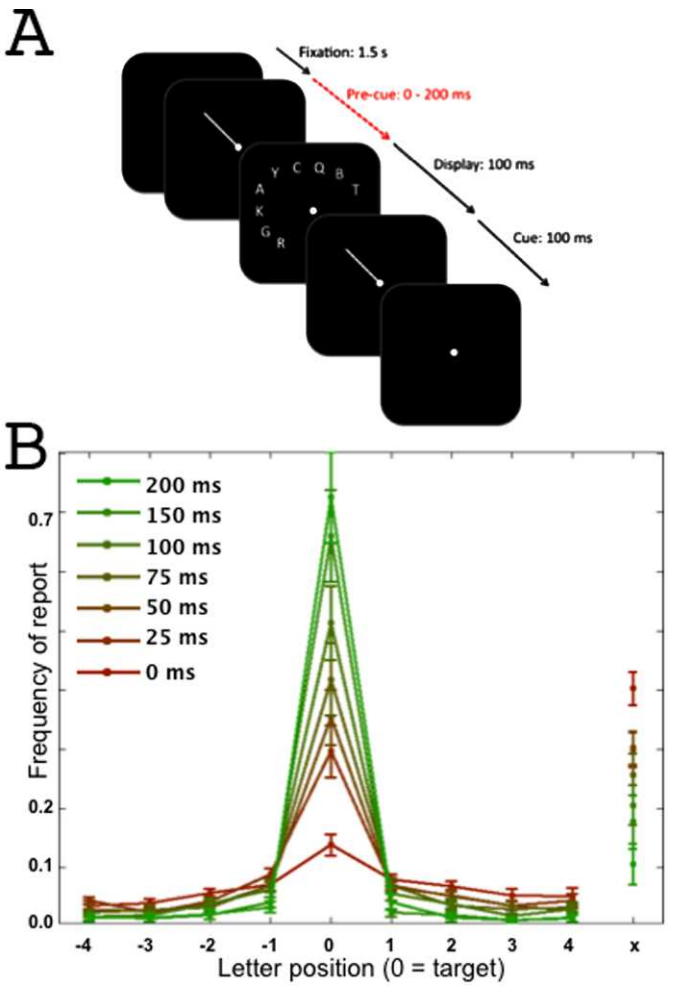

Figure 4. (a) Experiment 3 manipulated precueing duration while holding eccentricity and spacing constant. (b) The frequency (*y*-axis) with which each letter position (*x*-axis) is reported. Shorter precueing durations cause subjects to report the target less often and flankers and non-presented letters more often.

Experiment 2: Eccentricity

Because crowding increases in the periphery (Bouma, 1973; Levi, 2008; Pelli et al., 2004; Strasburger et al., 1991), in Experiment 2 we varied the eccentricity of the letters to be 5°, 6°, 7°, 8°, 9°, or 10° of visual angle. Letter spacing and size scaled with eccentricity (see methods, Figure 3). Figure 3 also shows the distribution of errors for each condition, and again, we find our crowding manipulation to be effective: As eccentricity increases, the target is reported less often, and in its place adjacent letters and non-presented letters are reported more often.

Experiment 3: Precueing

In Experiment 3, we varied how far in advance of the letter display the cue appeared: This precue duration was 0, 25, 50, 75, 100, 150, or 200 ms (see Figure 4). As in the first two experiments, making the task harder—in this case by providing a shorter precue interval—decreased the rate at which the target was reported while increasing the rate of report of adjacent and non-presented letters.

Experiment summary

Across the three experiments a simple pattern is evident: As the task becomes harder, by manipulating spacing, eccentricity, or precueing, (a) targets are reported less often, (b) adjacent items are reported more often, and (c) non-presented letters are reported more. Thus, we are confident that our manipulations make the identification of crowded letters more or less difficult. We now turn to the main goal of our paper: quantifying these error distributions in terms of different error models. By formally characterizing these patterns of errors, we aim to quantify what factors contribute to human errors in these crowding tasks and whether the three qualitatively different manipulations of difficulty affect errors differently or whether they manipulate difficulty in the same manner.

Error models

To characterize the errors that observers make under crowded conditions we need to consider different classes of error models that embody different mechanisms by which crowding flankers might obscure the target letter. The models we consider rely on two weighting/similarity functions: the *spatial weighting function*, describing how much flankers at different distances from the target influence target reports and the *letter confusion matrix*, describing which letters tend to be more or less orthographically confusable with which other letters. Once we have the spatial weighting function and letter confusion matrix, we consider different response models that combine spatial and orthographic confusion differently to yield responses. We end up with eight error models that capture different possible crowding mechanisms.

Spatial weighting function

All models of crowding assume some sort of spatial weighting function. In pooling models, the spatial weighting amounts to the spatial breadth of the texture summaries used to infer the presented letters. In spatial substitution models, the spatial weighting amounts to the probability that each of the adjacent letters will be reported in place of the target. Thus, the statistical models of human errors that we consider are all built around a spatial weighting function: How much do letters at different distances from the target influence responses?

Given the shape of the histograms of spatial intrusions in our data (e.g., Experiment 1a in Figure 2)—that is, how often each of the distracters are reported—we adopted a Laplacian probability distribution as our spatial weighting function. The Laplacian probability distribution is peaked at the target with log probability falling off proportionally to inverse distance (we measure distance simply as letter position).[1] The Laplacian distribution is analogous to a Gaussian probability distribution, but it uses the absolute value rather than the square of distance. As we use it, this weighting function has one parameter: the Laplacian scale parameter, **b**, which determines how steeply the weighting function drops off with distance (we write **b** in bold to indicate that it will be a free parameter for a number of error models).

We write the weight given to a position x (with zero being the target, negative one and one being the two immediate flankers, and so on) as $P_L(x|\mathbf{b})$, where **b** is the Laplacian scale parameter. Because we are interested in discrete positions, this weighting function is defined in terms of the Laplacian cumulative density function (Ψ_L), assigning each letter position a weight based on the probability mass in the interval around that letter position:

$$P_L(x|\mathbf{b}) = \Psi_L(x + 0.5|\mathbf{b}) - \Psi_L(x - 0.5|\mathbf{b}). \quad (2)$$

The Laplacian cumulative density function (with a fixed mean of zero, centered on the target), is given by:

$$\Psi_L(x|\mathbf{b}) = \begin{cases} 0.5\exp[x/\mathbf{b}] & \text{if } x \leq 0 \\ 1 - 0.5\exp[-x/\mathbf{b}] & \text{if } x \geq 0. \end{cases} \quad (3)$$

If the scale parameter is particularly small ($\mathbf{b} \leq 0.1$), the Laplacian spatial weighting function will place all weight on the target ($P_L(0|\mathbf{b} \leq 0.1) > 0.99$). If the scale parameter is particularly large ($\mathbf{b} \geq 10$), this function distributes weight nearly uniformly over the presented items ($P_L(0|\mathbf{b} \geq 10) < 0.135$. For a perfectly uniform distribution, $P_L(0) = 1/9 = 0.\overline{11}$).

Letter confusion matrix

To account for the rich structure of errors on individual trials, we must consider not only how letters at differing distances from the target influence responses but also how the identification of the target may influence responses. In other words, we must consider a letter confusion matrix: How likely is a letter i to be confused with an alternate letter j?

We obtained a letter-letter confusion matrix empirically by tabulating all intrusions from all trials from all conditions, all subjects, and all three experiments. Each time the target was letter i, and the response was letter j we incremented by one the (i, j)th entry of the confusion matrix. Thus, the ith row of the empirical confusion matrix corresponds to the empirical counts of intrusions every time the target was letter i. This procedure yielded a total of 17,571 observations, which were roughly equally distributed per target letter (mean count per target letter: 676, standard deviation: 49, minimum count: 605, maximum count: 776). These raw counts were adjusted by adding one to every cell (thus eliminating zeros from the probabilities); in practice only two cells contained zero observations: reporting the letter "I" when the target was a "Q," and reporting "I" when the target was a "Y." We normalized these counts so that each row summed to one, thus obtaining a confusion matrix in which each row reflects the conditional probability of reporting each of 26 letters given a specific target letter. This confusion matrix is shown in Figure 5 and is available online.[2]

As we will show in later sections, this confusion matrix confers a considerable advantage to fitting letter intrusions across conditions and experiments, indicating that the confusion matrix captures several important aspects of human responses. We do not believe that this advantage arises from overfitting the confusion matrix to the real data because we used all trials, all conditions, and all subjects in all experiments to define it; with that much data, any unsystematic errors will average out, yielding no substantial benefit in fitting specific trial idiosyncrasies. We confirmed that we are not overfitting specific trial errors by using an independently defined cross-validation matrix: For each subject, we define a confusion matrix using data from all the other subjects (thus using roughly 17,200 data points for each subject-specific confusion matrix).

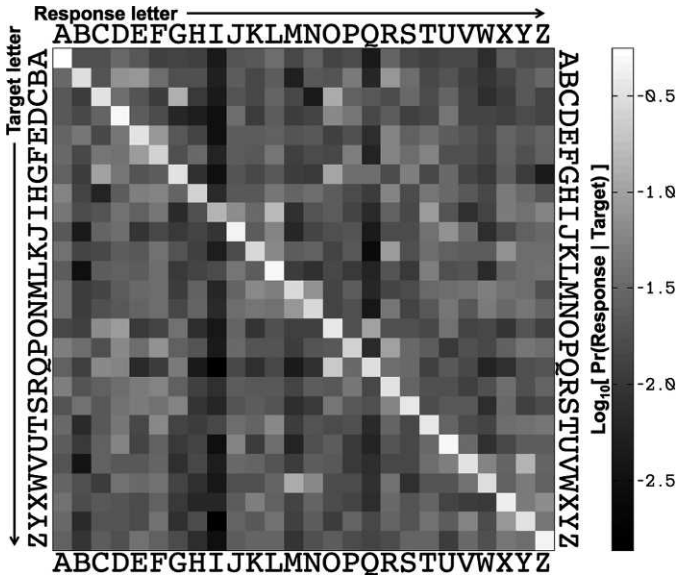

Figure 5. Empirically calculated letter-letter confusion matrix, available online at http://www.edvul.com/crowding-models/.

Results using such cross-validated confusion matrices differed numerically only slightly and all qualitative and quantitative patterns remained unchanged. Because none of our results or conclusions change using this more complicated procedure, we will describe model results using a single confusion matrix as described.

Instead of overfitting, we believe the advantage of using a confusion matrix (rather than assuming uniform random guessing) in our error models arises from two systematic deviations from random guessing. First, some letters are harder to report than others—for instance, when the letter "I" is the target, it is correctly reported only 15% of the time, while the letter "A" is reported correctly 56% of the time; the variation in proportion of correct reports across different target letters is highly significant (via chi-squared test for independence: $\chi^2[25] = 670$, $p < 0.001$). Second, when the correct target is not reported, observers do not guess randomly—some letters tend to be confused with specific other letters. For instance, when the correct target is a "C," then "G" and "O" are reported about 13% and 12% of the time, as compared to "I," which is only reported 0.5% of the time. For every target letter, the frequency with which the 25 other letters were erroneously reported was significantly different from a uniform distribution (smallest $\chi^2[24] = 58.6$, $p < 0.001$—for "A"—all other letters had larger chi-squared values and correspondingly smaller p values).

What aspects of letter shape give rise to these systematic properties of the confusion matrix? Unfortunately, we cannot offer an answer here (see Mueller & Weidemann, 2012, for a recent review of the field). We can say that a simplistic confusion model based on

pixel-wise correlations of smoothed letter images fails to capture the rich structure of our empirical confusion matrix. This simple model can capture a small but significant portion of the off-diagonal structure—that is, how often each possible error letter is reported for a given target letter, an average correlation of 0.476 between the predicted and empirical probabilities of all the nontarget letters for each target letter. However, this pixel-wise correlation model cannot capture the relative difficulty of some target letters compared to others: a correlation of only 0.1 on the diagonal (the probability of reporting the target letter across different targets). A predictive model of this confusion matrix would seem an ideal target for sophisticated models of texture-based letter inference (with classic precedents from Gervais, Harvey, & Roberts, 1984 and Townsend, 1971).

Response models

The crucial distinction among the error models we consider is the response model. How are the weighted letters used to produce guesses? We consider three classes of response models: (a) *direct spatial substitution*: each letter is reported with a frequency proportional to its weight, (b) *letter confusion*: each letter may be confused with similar-looking letters, and (c) *multiplicative combination*: reported letters tend to be those that are similar to multiple adjacent letters (a naïve pooling model). For each class of response models, we write out a probability model describing the probability that a letter j will be reported given the presented letter array $\mathcal{L}$.

Direct spatial substitution

The simplest response model is that letters are guessed with a frequency proportional to their weight with some probability of random guessing. We consider this to be the default spatial substitution model: errors arise from participants reporting adjacent items in place of the target. Under this response model, the probability of a subject reporting letter j (indexed from one to 26) on a given trial can be written:

$$P_S(j|\mathcal{L}) = \mathbf{p}\frac{1}{26} + (1-\mathbf{p})\sum_{x=-4}^{4} P_L(x|\mathbf{b})\delta(\mathcal{L}(x)\equiv j), \tag{4}$$

where $\mathcal{L}(x)$ is the letter in location x ($x = 0$ for the target, and the presented array spans $x = -4$ to $x = 4$), and $\delta(\mathcal{L}(x) \equiv j)$ is 1 when $\mathcal{L}(x)$ is the letter j (that is, when position x contains the letter j) and zero otherwise.

This direct spatial substitution model has two parameters: (a) the probability of random guessing (**p**) and (b) the spread of the spatial weighting function (**b**). This implicit model is adopted by proponents and opponents of spatial substitution accounts of crowding. When errors are made, they are likely to be simple substitutions of flankers for the target.

There are two noteworthy limiting cases of the direct spatial substitution model that serve as useful baselines. First, when the probability of random guessing is one ($\mathbf{p} = 1$), this model reduces to completely random guesses. We treat this pure random guessing case as the baseline for all model fits: It predicts a constant likelihood of 1/26 for every trial. Second, when the spread of the spatial weighting function approaches zero ($\mathbf{b} < 0.01$, i.e., spread is very small), this model reduces to a simple mixture of **p** proportion of random guesses and (1 − **p**) proportion of correct answers. We consider this mixture of correct and random to be our second baseline, which captures variation in difficulty/accuracy across conditions/subjects but assigns equal probability to every error.

Spatial substitution plus letter confusion

A more sophisticated response model accounts for the fact that even when a single letter is to be reported, subjects do not always report it because the letter may be confused with other similar-looking letters. We capture such errors by obtaining a letter-letter confusion matrix ($Q[i, j]$, described in the previous section), which describes how often a presented letter i is confused with letter j. Under this model, the spatial weighting function determines the linear combination of presented letters, and the response distribution is given by a weighted mixture of the corresponding rows from the confusion matrix.

Thus, adding letter confusion to the spatial substitution model replaces the $\delta(\mathcal{L}(x) \equiv j)$ expression (saying that only the target and flankers themselves may be reported) with $(Q\mathcal{L}(x), j)$, which says that not only may flankers be reported but also letters similar to the target and flankers may be reported.

The full letter confusion model is this: Each response is a random guess with probability p; otherwise, with probability $(1 - p)$, it is a response drawn from an additive mixture of the rows corresponding to the nine presented letters (weighted by the spatial weighting function parameterized by **b**). The rows of the letter confusion matrix are raised to an exponent **q** (between 0.3 and three). This exponent **q** can be thought of as the Luce-choice exponent for the confusion matrix (Luce, 1959). When it is large, predicted confusions are minimal and subjects tend to report the intended letter. When it is small, the probability of similar letters being reported increases.

$$P_C(j|\mathcal{L}) = \mathbf{p}\frac{1}{26} + (1-\mathbf{p})\sum_{x=-4}^{4} P_L(x|\mathbf{b}) \times \frac{Q\left(\mathcal{L}(x), j\right)^{\mathbf{q}}}{\sum_{j'=1}^{26} Q\left(\mathcal{L}(x), j'\right)^{\mathbf{q}}}. \quad (5)$$

Note that if a given row from the letter confusion matrix is exponentiated, then this conditional probability must be renormalized by dividing it by the sum of all possible letters that might have been reported (j').

It is worth considering a few limiting cases of this model. If $\mathbf{p} = 1$, then we again obtain the purely random guessing model. The same random guessing model may be recovered if the confusion matrix exponent is particularly low, thus making the confusion matrix effectively uniform (e.g., $\mathbf{q} < 0.01$; even for $\mathbf{q} < 0.3$ deviations from uniformity are insubstantial). As $\mathbf{q}$ becomes particularly large (larger than three or so), this model approximates the direct spatial substitution model described above, because in this case, only the diagonal of the confusion matrix will have any substantial report probability.

If $\mathbf{b} < 0.01$ (the scale of the spatial weighting function is small) and $\mathbf{p} = 0$ (the proportion of random guessing is zero), then we obtain a simple target-letter confusion model in which flankers play no role in determining errors, and all errors correspond only to orthographic confusions with the target letter. Thus, the effect of crowding under such a model is to exponentiate the independent target-letter confusion matrix. We consider this the simple target-letter confusion error model. We may also consider a model where $\mathbf{b} < 0.1$ (so flankers have no influence) but $\mathbf{p} > 0$ (so there is some chance of uniform random guessing). This yields a model that is a mixture of random guesses and confusions with the target letter itself. We find that there is no noticeable advantage (in terms of fit to our data) gained by including this random guessing parameter, so we prefer the simple target-letter confusion model. In general we find that the random guessing parameter may be dropped from the full substitution/confusion model because a model with $\mathbf{p} = 0$ does as well as models where $\mathbf{p}$ is variable. We suspect this is because a confusion matrix raised to a low exponent captures uniform random guesses as effectively as a separate random guessing parameter; thus, there is no need for the additional complication.

Multiplicative combination

Finally, we consider a naïve pooling model, in which the response distribution is not an additive independent mixture of the presented letters but is instead a multiplicative mixture. On this model, the probability of a letter being reported is proportional to the product of the similarities between that letter and all the presented letters, weighted by their spatial position.

This is an implicit pooling model, because it combines information from multiple letters to make a decision. This error model favors reporting letters that are similar to multiple letters in the display, while it disfavors letters that are merely similar to only one presented letter. However, this is a naïve pooling model because letters are combined by merely multiplying their respective entries from the confusion matrix. This may be an accurate summary of some pooling models but not most. In general, without specifying both the textures used to summarize the display and the decision rule by which letters are selected given the pooled texture representation, pooling models may take on many behaviors.

To make this multiplicative combination model clear, we first specify how the weighted multiplication is carried out:

$$\Phi(j|\mathcal{L}) = \prod_{x=-4}^{4}\left(\frac{Q\left(\mathcal{L}(x), j\right)^{\mathbf{q}}}{\sum_{j'=1}^{26} Q\left(\mathcal{L}(x), j'\right)^{\mathbf{q}}}\right)^{P_L(x|\mathbf{b})}. \quad (6)$$

Note that the spatial weights are used as exponents on the confusion probabilities associated with each of the presented letters (these can be seen as additive weights when considering log probability of the confusion matrix). This means that items far away from the target will not noticeably influence the multiplicative combination, because an exponent near zero results in a uniform distribution that has no effect when multiplied with other distributions.

This multiplicative mixture is not normalized, so it requires an additional normalization step to produce a probability, which we then combine with a simple random guessing process:

$$P_{\mathcal{M}}(j|\mathcal{L}) = \mathbf{p}\frac{1}{26} + (1-\mathbf{p})\frac{\Phi(j|\mathcal{L})}{\sum_{j'=1}^{26}\Phi(j'|\mathcal{L})}. \quad (7)$$

Altogether, this multiplicative combination model has three parameters: the probability of random guessing ($\mathbf{p}$), the spread of the spatial weighting function ($\mathbf{b}$), and the exponent of the confusion matrix ($\mathbf{q}$). As in all other models, as the probability of random guessing goes to one, this model produces simple random guessing. This is also the limit reached if the confusion matrix exponent ($\mathbf{q}$) falls below 0.01. As the spatial spread parameter falls below about 0.1, this model is indistinguishable from the target-letter confusion

model because the nontarget letters have effectively zero weight and so they do not contribute to the multiplicative mixture.

We consider this to be a naïve pooling model because it predicts errors that are similar to multiple presented letters; flanker identities will multiplicatively interact with the target identity to yield errors. We believe this was also Freeman, Chakravarthi, and Pelli's (2012) intuition when they argued that preferential errors due to flankers that are more confusable with the target are evidence against a simple spatial intrusion model. Nonetheless, this model is only one of infinitely many potential pooling models; thus, its performance cannot be used to make claims about pooling models in general.

Complete error models

Altogether, we consider eight unique error models, some of which are limiting cases of one or more response models.

(0) Random guessing

This error model is our null baseline, assigning a uniform probability to every possible response. It arises from any of our response models (e.g., Equation 4) when the probability of random guessing is one ($\mathbf{p} = 0$).

$$P(j|\mathcal{L}) = \frac{1}{26} \tag{8}$$

(1) Correct/random mixture

This error model is our second baseline, predicting a correct response with probability $(1 - \mathbf{p})$ and a random (uniform) guess with probability $\mathbf{p}$. Although several response models collapse to this pure mixture of correct and random guesses, we define this error model as the limiting case of the direct spatial substitution model (e.g., Equation 4) when the scale of the spatial weighting function is near zero ($\mathbf{b} = 0.01$).

$$P(j|\mathcal{L}) = \mathbf{p}\frac{1}{26} + (1 - \mathbf{p})\delta\Big(\mathcal{L}(0) \equiv j\Big) \tag{9}$$

(2) Direct spatial substitution

This model is a response that is a random guess with probability $\mathbf{p}$ and otherwise is an exact report of one of the presented letters, with probability proportional to the spatial weighting function. We define this model as Equation 4 with $\mathbf{p} < 1$, and $\mathbf{b} > 0.1$ but note that either of the response models using the letter confusion matrix models reduce to direct spatial substitution when $\mathbf{q} > 3$.

$$P(j|\mathcal{L}) = \mathbf{p}\frac{1}{26} + (1 - \mathbf{p})\sum_{x=-4}^{4} P_L(x|\mathbf{b})\delta\Big(\mathcal{L}(x) \equiv j\Big) \tag{10}$$

(3) Target letter confusion

This model is a response that is based only on the target letter, but that target letter may be misreported according to the confusion matrix exponentiated by $\mathbf{q}$. We define this as the limiting case of the spatial substitution plus letter confusion (Equation 5) model when there is no random guessing ($\mathbf{p} = 0$) and the spatial scale is near zero ($\mathbf{b} = 0.01$) (note that this is also the limiting case of the multiplicative combination model when $\mathbf{p} = 0$ and $\mathbf{b} < 0.01$).

$$P(j|\mathcal{L}) = \frac{Q\Big(\mathcal{L}(0), j\Big)^{\mathbf{q}}}{\sum_{j'=1}^{26} Q\Big(\mathcal{L}(0), j'\Big)^{\mathbf{q}}} \tag{11}$$

(4) Spatial substitution/confusion

This model is a response that is either a random guess or a guess based on one of the presented letters considered independently but potentially misreported given the confusion matrix. This is given by Equation 5 where $\mathbf{p} < 1$, $\mathbf{b} > 0.1$, and $3 > \mathbf{q} > 0.3$.

$$P(j|\mathcal{L}) = \mathbf{p}\frac{1}{26} + (1 - \mathbf{p})\sum_{x=-4}^{4} P_L(x|\mathbf{b}) \times \frac{Q\Big(\mathcal{L}(x), j\Big)^{\mathbf{q}}}{\sum_{j'=1}^{26} Q\Big(\mathcal{L}(x), j'\Big)^{\mathbf{q}}} \tag{12}$$

(5) Spatial substitution/confusion ($\mathbf{p} = 0$)

This model is a response that is a guess based on one of the presented letters considered independently but potentially misreported given the confusion matrix (without random guesses; Equation 5 where $\mathbf{p} = 0$, $\mathbf{b} > 0.1$, and $3 > \mathbf{q} > 0.3$).

$$P(j|\mathcal{L}) = \sum_{x=-4}^{4} P_L(x|\mathbf{b}) \frac{Q\Big(\mathcal{L}(x), j\Big)^{\mathbf{q}}}{\sum_{j'=1}^{26} Q\Big(\mathcal{L}(x), j'\Big)^{\mathbf{q}}} \tag{13}$$

(6) Multiplicative combination

This is our naïve pooling model, predicting either a random guess or a guess based on the multiplicative combination of the presented letters. This is given by Equation 7 where $\mathbf{p} < 1$, $\mathbf{b} > 0.1$, and $3 > \mathbf{q} > 0.3$.

$$\Phi(j|\mathcal{L}) = \prod_{x=-4}^{4} \left(\frac{Q\left(\mathcal{L}(x), j\right)^{\mathbf{q}}}{\sum_{j'=1}^{26} Q\left(\mathcal{L}(x), j'\right)^{\mathbf{q}}} \right)^{P_L(x|\mathbf{b})}$$

$$P(j|\mathcal{L}) = \mathbf{p}\frac{1}{26} + (1-\mathbf{p})\frac{\Phi(j|\mathcal{L})}{\sum_{j'=1}^{26}\Phi(j'|\mathcal{L})} \quad (14)$$

(7) Multiplicative combination ($p = 0$)

This is a simplification of our naïve pooling model, predicting a guess based on the multiplicative combination of the presented letters (without random guesses; Equation 7 where $\mathbf{p} = 0$, $\mathbf{b} > 0.1$, and $3 > \mathbf{q} > 0.3$).

$$P(j|\mathcal{L}) = \frac{\Phi(j|\mathcal{L})}{\sum_{j'=1}^{26}\Phi(j'|\mathcal{L})} \quad (15)$$

Model fitting

We obtained the maximum likelihood parameters for each model for each subject in each condition. Thus we estimated between one and three parameters (depending on the model) to assign the highest aggregate probability to the specific responses issued by each subject in the 50–65 trials they had in each condition.

Each model specifies a likelihood function for the data. For instance, if the letter array on a single trial were "A", "B", "C", "D", "E", "F", "G", "H", or "I", (the target being the central letter, "E") and the subject reported "E", the likelihood of this trial given under the spatial substitution/confusion model (Equation 12) with parameters $\mathbf{p} = 0.2$, $\mathbf{b} = 0.5$, $\mathbf{q} = 1.2$ would be 0.2535 (see Appendix A for a more thorough description of how this number is calculated). We multiply all trial likelihoods obtained this way for each trial for a given subject in a given condition to obtain the net likelihood of a given set of parameters for a given subject in a given condition. We then find (using numerical optimization) the maximum likelihood set of parameters for a given subject:

$$\left[\hat{\mathbf{b}}, \hat{\mathbf{p}}, \hat{\mathbf{q}}\right] = \underset{\mathbf{b},\mathbf{p},\mathbf{q}}{\operatorname{argmax}} \left[\prod_{t\in T} \sum_{x=-4}^{4} P_L(x|\mathbf{b}) \frac{Q\left(\mathcal{L}^{(t)}(x), j^{(t)}\right)^{\mathbf{q}}}{\sum_{j'=1}^{26} Q\left(\mathcal{L}^{(t)}(x), j'\right)^{\mathbf{q}}} \right], \quad (16)$$

where T is the set of all trials in the condition of interest for the particular subject, t is a specific trial, $\mathcal{L}^{(t)}$ is the set of letters presented on that specific trial, and $j^{(t)}$ is the response given on that trial.

By obtaining model fits to each subject in each condition we avoid aggregating across subjects (thus respecting the across-subject variation in performance); furthermore, individual subject model fits allow us to perform statistical inference on model parameters by considering their across-subject variation in each condition. The rest of this paper considers how well the different models can describe human error patterns and how the frequency and quality of different types of errors (as estimated by various model parameters) change as we manipulate difficulty.

Accuracy as described by model parameters

In the Data section, we showed the frequency with which each of the presented letter positions was reported on each trial, showing that our manipulations, as intended, made the task correspondingly harder or easier. Here we aim to demonstrate how these performance variations can be captured by changes in model parameters of two of our simple single-parameter models: the correct/random mixture model and the target-letter confusion model. For the correct/random mixture model, accuracy corresponds to $1 - \mathbf{p}$; that is, the probability that the correct letter will be reported. For the target-letter confusion model, accuracy is a function of the confusion matrix exponent ($\mathbf{q}$). When $\mathbf{q}$ approaches 3 ($10^{0.5}$), the model describes responses that are nearly always the correct letter. When $\mathbf{q}$ approaches 0.3 ($10^{-0.5}$), the model describes responses that are effectively uniform random guessing.

Figure 6 shows the maximum likelihood model parameters for individual subjects in each of the conditions in each experiment, as well as ellipses with height and position corresponding to the across-subject mean ±1 standard error. Both ways of measuring accuracy capture the obvious result: Manipulating difficulty (either via spacing, eccentricity, or precueing)

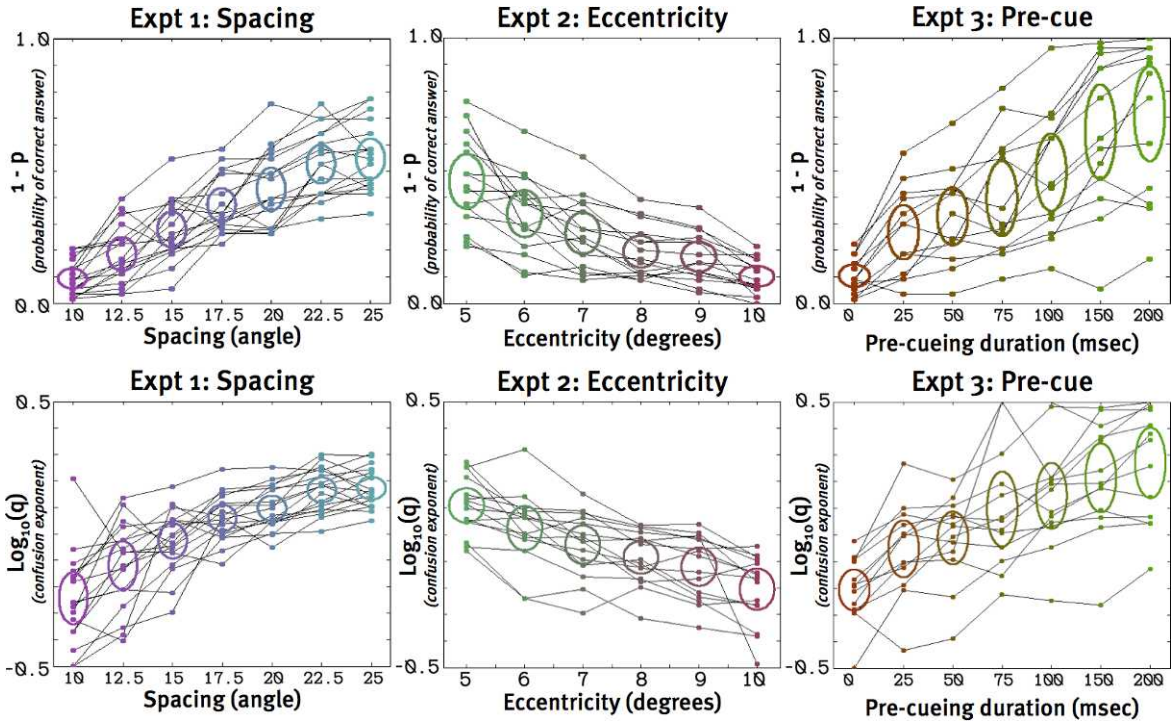

Figure 6. Model parameters describing subjects' performance as a function of condition for each experiment. (top) Performance as measured by proportion of correct responses obtained from the correct/random mixture model ($1 - \mathbf{p}$ in Equation 9) for each of the three experiments. (bottom) Performance as measured by the base-10 logarithm of the confusion matrix exponent from the target-letter confusion model (q in Equation 11). Each point in the plot corresponds to the parameter estimate for a subject in a given condition; because all experiments were within-subject, lines connect parameter estimates from individual subjects across conditions. Ellipses show the mean and standard error across subjects in each condition (center of ellipses is the mean with total ellipse height given by ± 1 *SEM*; width of ellipses in x is constant and meaningless, presented only for graphical purposes). For consistency with subsequent plots, different experiments/conditions are presented in slightly different hues: Within each experiment harder conditions are redder.

makes the task harder. To be thorough, we quantify these changes below.

First we consider accuracy as measured by the probability of reporting the correct answer (using the correct/random mixture model). In Experiment 1a, for every extra degree of arc on the circle, accuracy increases by 3.1%, 95% confidence interval on the slope: [0.0272, 0.0355], $t(107) = 15$, $p < 0.001$; $r^2 = 0.65$ for the full model repeated-measures regression including a random subject intercept. In Experiment 2, for every additional degree of eccentricity, accuracy decreases by 6.6%, 95% confidence interval on the slope: [–0.082, –0.052], $t(74) = 8.8$, $p < 0.001$; $r^2 = 0.47$ for the full repeated-measures regression. In Experiment 3, every additional millisecond of precueing duration increases accuracy by 0.3%, 95% confidence interval on the slope: [0.0023, 0.0037], $t(77) = 8.6$, $p < 0.001$; $r^2 = 0.45$ for the full repeated-measures regression. In short, accuracy as measured by the probability of reporting the correct answer changes as predicted, and the trends seen in Figure 6 are highly significant.

We obtained similar results when considering accuracy in terms of the $\log_{10}(\mathbf{q})$, i.e., the log of the confusion matrix exponent from the target-letter confusion model (Equation 11).[3] The target-letter confusion model has an advantage over a correct/random mixture model in capturing the specific kinds of errors that observers make. For the present analysis, we focus only on the fact that the confusion matrix exponent in the target-letter confusion model effectively captures changes in accuracy: When $\mathbf{q} = 1$ (i.e., $\log_{10}(\mathbf{q}) = 0$), accuracy is 35%, on average; when $\mathbf{q} = 2$ (i.e., $\log_{10}(\mathbf{q}) = 0.3$), accuracy is about 80%; when $\mathbf{q} = 0.5$ (i.e., $\log_{10}(\mathbf{q}) = -0.3$), accuracy is about 14%. In Experiment 1a, the $\log_{10}$ of the confusion matrix exponent ($\mathbf{q}$) increased by 0.028 for every extra degree of arc in spacing, 95% confidence interval on the slope: [0.024, 0.032], $t(107) = 14.75$, $p < 0.001$; $r^2 = 0.63$ for the full repeated-measures regression. In Experiment 2, for every additional degree of eccentricity, the $\log_{10}$ confusion matrix exponent decreases by 0.058, 95% confidence interval on the slope: [–0.072, –0.044], $t(74) = 8.2$, $p < 0.001$; $r^2 = 0.43$ for the full repeated-measures regression. In Experiment 3, every additional millisec-

ond of precueing duration increased the $\log_{10}$ confusion matrix exponent by 0.0024, 95% confidence interval on the slope: [0.0018, 0.0029], $t(77) = 8.6$, $p < 0.001$; $r^2 = 0.45$ for the full repeated-measures regression. To summarize: Accuracy as measured by the confusion matrix exponent changes as predicted, and the trends seen in Figure 6 are highly significant.

These plots and analyses are intended to convey both that (a) our manipulations had robust and predictable effects on accuracy, and (b) our simple error models can capture these changes. However, these plots also highlight a feature in our data: Accuracy varies considerably across subjects. This across-subject variability is one reason why fitting models to individual subject-condition pairings was an important analysis decision.

Model comparison

We fit each model to all trials in a given condition from each subject to obtain a set of maximum likelihood parameter estimates for each subject-condition combination. These maximum likelihood parameters also yield the maximized likelihood for each subject-condition ($L(s, c)$)—a measure of how well the model could account for the responses in all trials of a particular subject (s) in a given condition (c). Because different subject-condition combinations contained different numbers of trials (n), $L(s, c)$ is not comparable across subjects or conditions. To obtain an interpretable measure we calculate the log likelihood per trial ($\log_{10}L(s, c)/n$) for each subject condition. This log likelihood per trial is invariant to the number of trials that went into a particular subject-condition combination and simply reflects the average model fit for all trials in that condition (for comparison: a random guessing predicts a constant log likelihood per trial of −1.415, and a model that predicts the exact response on every trial has a constant log likelihood of zero). Even raw log likelihood per trial is minimally informative, because what is relevant are differences between models; thus, we consider the gain in average log likelihood per trial compared to different baselines. Figure 7 (top) uses completely random guessing as a baseline. Completely random guessing assigns each trial a log likelihood of −1.415, insofar as a model can better capture subjects' responses, the log likelihood per trial will be higher (less negative), and there will be some positive gain in log likelihood per trial. We calculate this gain above the random guessing baseline for each subject-condition combination (so as to factor out across-subject variability in performance). Figure 7 (top) shows how well the various models perform in terms of improvement in log likelihood per trial compared to a baseline of pure random guessing.

Because Figure 7 (top) shows a considerable amount of across-condition variability (some conditions are much easier than others), it is also relevant to compare our specific error models to the correct/random guessing model. Correct/random guessing can capture variation in accuracy but cannot predict which letters subjects will report when an error is made. Figure 7 (bottom) shows the log likelihood per trial gain of various models compared to a mixture of correct and random guessing (black points in the top panels) in which the target is given probability $(1 - \mathbf{p})$, and every other response has probability $\mathbf{p}/26$. This log likelihood per trial gain of a more complicated model over the correct/random mixture model indicates that the more complicated model can capture not only the changes in error rate as a function of condition but also what kinds of errors are made by participants.

From the log likelihood comparisons in Figure 7, we draw a number of conclusions.

(1) In particularly difficult conditions, all models perform poorly. The log likelihood per trial gain above a purely random guessing model is minimal (Figure 7, top). This occurs because when subjects make more mistakes, the entropy of their guesses increases, which causes all models to have relatively low likelihoods (even though they might predict specific errors a bit better than purely random guessing). As the conditions become easier, all models that can capture variation in difficulty exhibit roughly the same advantage over the random guessing baseline (including even the simple second baseline model, a mixture of correct and random guesses).

(2) More sophisticated error models provide a qualitatively constant advantage over a simple correct/random mixture model across experiments and conditions (Figure 7, top; although there are stable and significant numerical differences seen in the bottom panels of Figure 7).

(3) The direct spatial substitution model (which predicts intrusions from adjacent positions) provides some benefit over the correct/random mixture model, indicating that observers tend to substitute flankers directly for the target—this can be seen by comparing the purple data points to the baseline (zero) in Figure 7 (bottom). This is seen both in the $\log_{10}$ likelihood per trial gain averaged across all conditions (mean: 0.0458, SD: 0.0127) and in across-subject t tests within each condition (across the 20 conditions, the smallest t value is 3.07, with a corresponding $p = 0.009$; all other conditions had higher t values[4]). The overall advantage of direct spatial substitution over a simple correct/random mixture can be tested via a

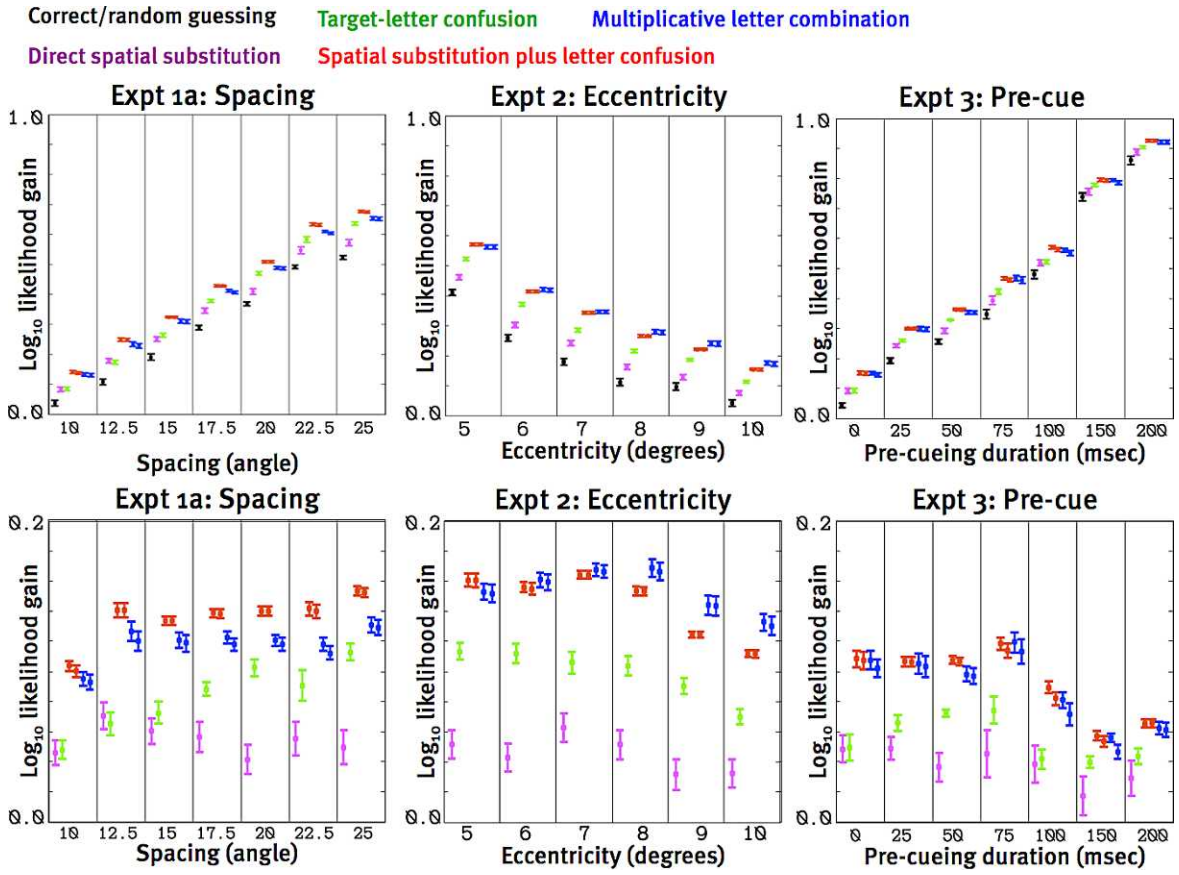

Figure 7. Improvements in $\log_{10}$ likelihood per data point for different error models compared to different baselines (higher is better) for all conditions in all three experiments. The spatial substitution plus letter confusion model and the multiplicative combination model each contain two points in the same color: one including the random guessing (**p**) and the other which is constrained to be zero. There is no difference between this variation in the models, so they are the same color in the plot. (top) Other error models compared to uniform random guessing. Each point corresponds to the average gain in log likelihood per trial over the fully random model (more is better). Plotted points are averages over subjects ± 1 *SEM* (for reference, random guessing predicts constant $\log_{10}$ likelihood $= -1.415$ per data point). These plots show that the average likelihood per data point is largely dominated by variation across conditions: On harder trials, responses are less predictable. (bottom) Average log likelihood per trial gain compared to the correct/random mixture model (black points in top panel). This baseline takes into account variation in difficulty across conditions, thus the differences between error models are easier to discern.

likelihood ratio test that aggregates likelihoods over all subjects and conditions and accounts for the extra parameters in the spatial substitution model. The spatial substitution model (total $\log_e$ likelihood $= -43{,}653$[5] with 614 parameters—two per subject per condition) has a higher likelihood than the correct/random mixture (total $\log_e$ likelihood $= -45{,}537$, with 307 parameters) with a very significant likelihood ratio test ($X^{(2)}(307) = 3768$, $p \approx 0$). This indicates that subjects' errors do indeed tend to come from adjacent letters, and a model that does not account for this cannot fit the error distributions as well as one that does.

(4) However, the advantage of the spatial substitution model over the correct/random mixture is quite small compared to the further improvement in fit gained by further taking the confusion matrix into account in either the spatial substitution plus letter confusion model (average across conditions $\log_{10}$ likelihood/trial gain: 0.0766, $SD = 0.024$; likelihood ratio test: $X^{(2)}(650) = 5822$, $p \approx 0$, the comparison between purple and red points in Figure 7, bottom), or in a multiplicative letter combination model (average $\log_{10}$ likelihood/trial gain: 0.07, $SD = 0.028$; likelihood ratio test: $X^{(2)}(650) = 5822$, $p \approx 0$, the comparison between purple and blue points in Figure 7).

(5) Similarly, taking the letter confusion matrix into account provides considerable advantage over a simple correct/random mixture model (average $\log_{10}$ likelihood/trial gain: 0.078, $SD = 0.0254$—comparison between green points and baseline in Figure 7, bottom). But again, there is a considerable further benefit of adding spatial weighting to the

model, which means that responses are not based on the target alone but incorporate influences from adjacent letters. Again, this is the case both for the spatial substitution plus letter confusion model ($\log_{10}$ likelihood/trial gain: mean = 0.044, SD = 0.014; likelihood ratio test: $X^2(307) = 3662$, $p \approx 0$—comparison between green and red points) and multiplicative combination models ($\log_{10}$ likelihood/trial gain: mean = 0.038, $SD = 0.017$; likelihood ratio test: $X^2(307) = 3126$, $p \approx 0$—comparison between green and blue points).

The previous two findings indicate that neither spatial substitution nor letter confusion alone are sufficient to account for the error distributions; a successful model must incorporate both.

(6) For models that employ both spatial substitution and letter confusion, there is no substantial difference between those with and without a random guessing parameter. Comparing the spatial substitution plus letter confusion model with a random guessing parameter to that without (comparison between the pairs of red points in each condition in Figure 7), we find little difference between them. The average $\log_{10}$ likelihood/trial gain is 0.0014, SD: 0.0020 from adding a random guessing parameter (although the pairwise t test in one condition is significant, the log-likelihood ratio test shows no effect; $X^2(307) = 101$, $p \approx 1$). Comparing the multiplicative combination model with a random guessing parameter to one without yields similar results (comparison between pairs of blue dots in each condition in Figure 7); the average $\log_{10}$ likelihood/trial gain from the random guessing parameter is 0.0034, SD: 0.0027, and the likelihood ratio test shows no effect ($X^2(307) = 266$, $p = 0.95$). This result captures the intuition that the confusion matrix, as we have defined it, captures systematic intrusions and can capture uniform random guessing to the degree that it exists.

(7) There is no stable difference between the spatial substitution plus letter confusion model and the multiplicative combination model (comparison between red and blue points in Figure 7). While in some conditions in Experiment 2 the multiplicative model better describes the error data (in 3/6 conditions pairwise $t > 2$ and $p \leq 0.05$), in most conditions of Experiment 1a the spatial substitution plus letter confusion model is a better description (in 6/7 conditions $t > 2$ and $p \leq 0.05$). The average log-likelihood/trial gain of the spatial substitution plus letter confusion model is small and variable (across condition mean: 0.0063, SD: 0.0133). That said, if we pool likelihoods over all conditions and all experiments to compare these models via a likelihood ratio test with 1° of freedom, we find there is a highly significant advantage to the spatial substitution plus letter confusion model ($X^2(1) = 535$, $p \approx 0$). However, because this advantage is neither large nor stable across conditions, we do not believe it should be taken seriously. These results reinforce the point made in the Introduction: When put on an equal footing, spatial substitution and pooling models are difficult to tease apart.

Together, these results suggest that a good model of human error distributions must take into account the spatial weighting function (the degree to which flankers at different distances from the target influence the response distribution) and the pairwise letter confusions (the propensity to misreport some characters as specific other letters). As long as both of these factors are incorporated in the model, even qualitative differences of implementation will make little difference in the goodness of fit.

Effects of target-flanker similarity

Freeman et al. (2012) argued that because flankers that are similar to the target are reported more often than those that are dissimilar, a spatial substitution model cannot be right, but instead a pooling model must be adopted. This reasoning makes sense if one considers a model such as the one we call direct spatial substitution model (with random guessing): When a substitution occurs, that letter is reported exactly. However, as we show below, this logic does not apply to a spatial substitution model with letter confusion.

We considered only those trials in which only one immediate flanker was similar to the target. Similarity was defined as the flanker being in the top five letters that tend to be reported for a given target, as calculated via our confusion matrix. Figure 8 shows that if we only consider such trials, we find the same effects Freeman et al. (2012), Huckauf and Heller (2002), and Krumhansl and Thomas (1977) reported: Similar flankers are reported more often. However, no qualitatively different pattern of log likelihoods emerges on those trials. Spatial substitution plus letter confusion remains the best fitting model on those trials (total $\log_{10}$ likelihood: −4935) and is slightly, but very significantly (based on a likelihood ratio test), better than the multiplicative model (total $\log_{10}$ likelihood: −4957; $X^2(1) = 55.48$, $p < 0.001$). Note that no new parameters were fit to this subset of data. We used the parameters from the fit to all the trials for a given subject condition but considered the likelihood of only a subset of trials.

We estimated the probability with which each model predicts similar and dissimilar flanker intru-

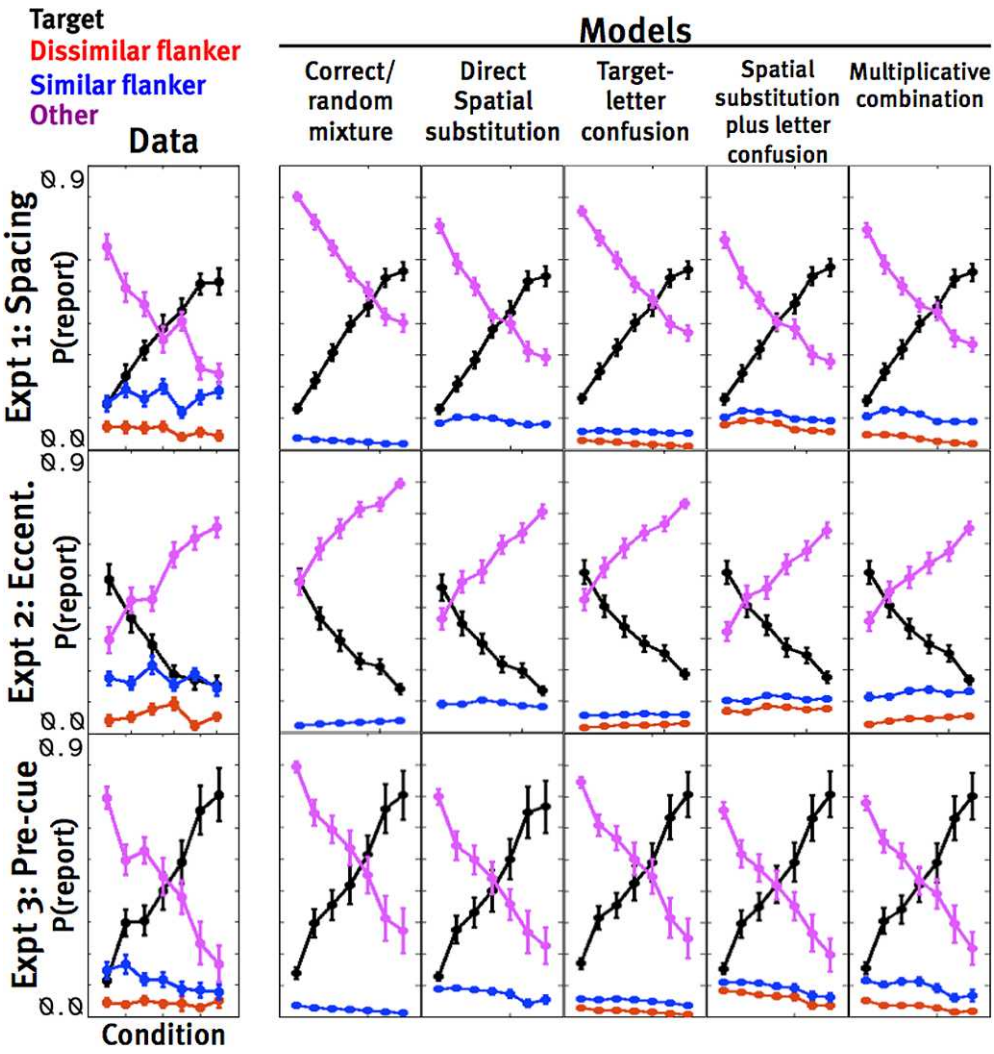

Figure 8. Comparison of report frequencies of the target (black), similar (blue), and dissimilar (red) flankers, as well as other intrusions (magenta). (left) As Freeman et al. (2012) showed, similar flankers are reported more often than dissimilar flankers. (right) This qualitative prediction is made by all models that use a confusion matrix to predict errors but not a simple correct/random mixture or direct spatial substitution (for these models the red and blue lines are exactly superimposed). The rates of the two types of intrusions under the multiplicative combination model are numerically closer to the human data. Nevertheless, the total likelihood of the multiplicative combination model is lower than the spatial substitution plus letter confusion model, presumably because it cannot account for the bulk of other kinds of errors that subjects make.

sions on this subset of trials by obtaining predictions from each model for each trial and averaging these predictions over subjects (Figure 8). We find that all models with a confusion matrix predict the qualitative effect that similar flankers are reported more often. However, the difference in predicted rates of similar and dissimilar flankers for the multiplicative combination model is much closer numerically to that seen in our subjects. The multiplicative model predicts a stronger discrepancy between reports of similar and dissimilar flankers, because it favors the reports of letters similar to two presented letters superadditively, while the other models predict only additive gains in probability.

It is intriguing that the multiplicative combination model has a lower overall likelihood on this subset of trials but can better predict the rates at which similar and dissimilar flankers are reported. We suspect this is because it does not adequately predict all of the other intrusions which dominate the error distributions even in this subset of trials. Overall, we argue that the intriguing results of Bernard and Chung (2011), Freeman et al. (2012), Huckauf and Heller (2002), and Krumhansl and Thomas (1977) do not rule out an error process that considers letters independently; however, these results are a compelling argument for the role of letter confusion in human response errors.

One dimension of difficulty

Because the statistical error models we use have a small number of interpretable parameters, we can ask how they change as we manipulate crowding. Specifically, we might imagine that the different manipulations in our three experiments—spacing, eccentricity, or precueing—would yield different kinds of errors. For instance, increasing precueing duration might increase the precision of spatial attention, thus decreasing the breadth of the spatial weighting function without increasing the confusability of isolated letters; or perhaps increasing eccentricity will make letters more confusable without changing the spatial weighting function. In this section we ask: Do the changes in error distributions across conditions in the three experiments suggest that the different manipulations of task difficulty operate in different ways?

For each of our two-parameter models (direct spatial substitution, spatial substitution plus letter confusion, and multiplicative combination), we test whether the conditions in our three experiments manipulate the error distributions in different ways or if every difficulty manipulation changes error distributions in the same way. If every difficulty manipulation has the same effect on error distributions, that would suggest that there may be only one difficulty tuning parameter, and different manipulations influence performance only via this single dimension of difficulty. We find this second alternative to be the case.

Figure 9 (top) shows the best fitting parameters (breadth of spatial weighting, rate of random guessing) for the direct spatial substitution model for individual subjects in each condition as well as the standard error ellipse (across subjects, taking into account the across-subject covariance of parameter estimates) for each specific condition in each experiment. The rightmost panel shows the across-subject standard error ellipses for all conditions in all three experiments. Across all experiments as the difficulty increases (redder colors), the spatial weighting increases in breadth and the probability of random guessing increases. Moreover, pooling results across all experiments suggest (without much squinting required) that all conditions from all experiments seem to fall on a single straight line in parameter space. Regardless of how we manipulate difficulty, the relationship between the rate of random guessing and the breadth of the spatial weighting function remains fixed, suggesting that all difficulty manipulations effectively have the same effect.

We can quantify the apparent one dimensionality of the changes in error distributions via principal component analysis. A detailed description of this analysis appears in Appendix B, but here we convey the intuition. If there is only one latent dimension to the changes in error distributions, the pairs of parameter estimates across different experiments and conditions should all fall along a single line in parameter space and should not deviate significantly from that line. In other words, there should be significant across-condition variation in parameter estimates only along their first principal component, and there should be no significant variation along the second (orthogonal) principal component. Thus, to test for deviations from unidimensionality, we must assess whether variation along the second principal component is greater than would be expected by chance. We can test for above-chance variation along the second principal component by asking whether the loadings onto the second component divided by the appropriate standard error term have variance greater than one. Under the null hypothesis of no variance along the second component except for error, these studentized residuals should have a variance of one. We assess whether the variance along the second principal component differs from one in three ways: (a) obtaining a 95% posterior credible interval on the variance along the second principal component (see Appendix B for details)—this interval should include one if this variance does not differ from chance; (b) evaluating the posterior probability that this second-component variance is greater than one (this should be low if variance is not higher than chance: 0.95 or higher would correspond to a traditional alpha value of 0.05); (c) performing a chi-squared test on the sum of the squared studentized residuals and obtaining a p value (which should be high if variance does not differ from chance). If we can be reasonably confident that the variance of the residuals along the second principal component is indistinguishable from one, then we have evidence (albeit evidence by failing to reject the null) that all the variation in accuracy across conditions boils down to a single dimension.

For the two parameters of the direct spatial substitution model, we find that (a) 95% posterior confidence intervals on the variance along the second principal component span one [0.63, 2.3], indicating that the null hypothesis is within our 95% confidence interval, (b) the posterior probability that this second-component variance is greater than one is relatively low (0.65), and (c) the p value for a chi-squared test to see if this variance is greater than one is not significant, $\chi^2(20) = 23$, $p = 0.28$. In contrast, the 95% posterior confidence interval on the variance of studentized residuals along the first principal component is [9.02, 32.93], significantly greater than one by any measure. Thus, when error distributions are summarized in terms of the breadth of the spatial weighting function and the rate of random guessing in a direct spatial substitution model, changes in these parameters, regardless of how we manipulate difficulty, fall along a single line.

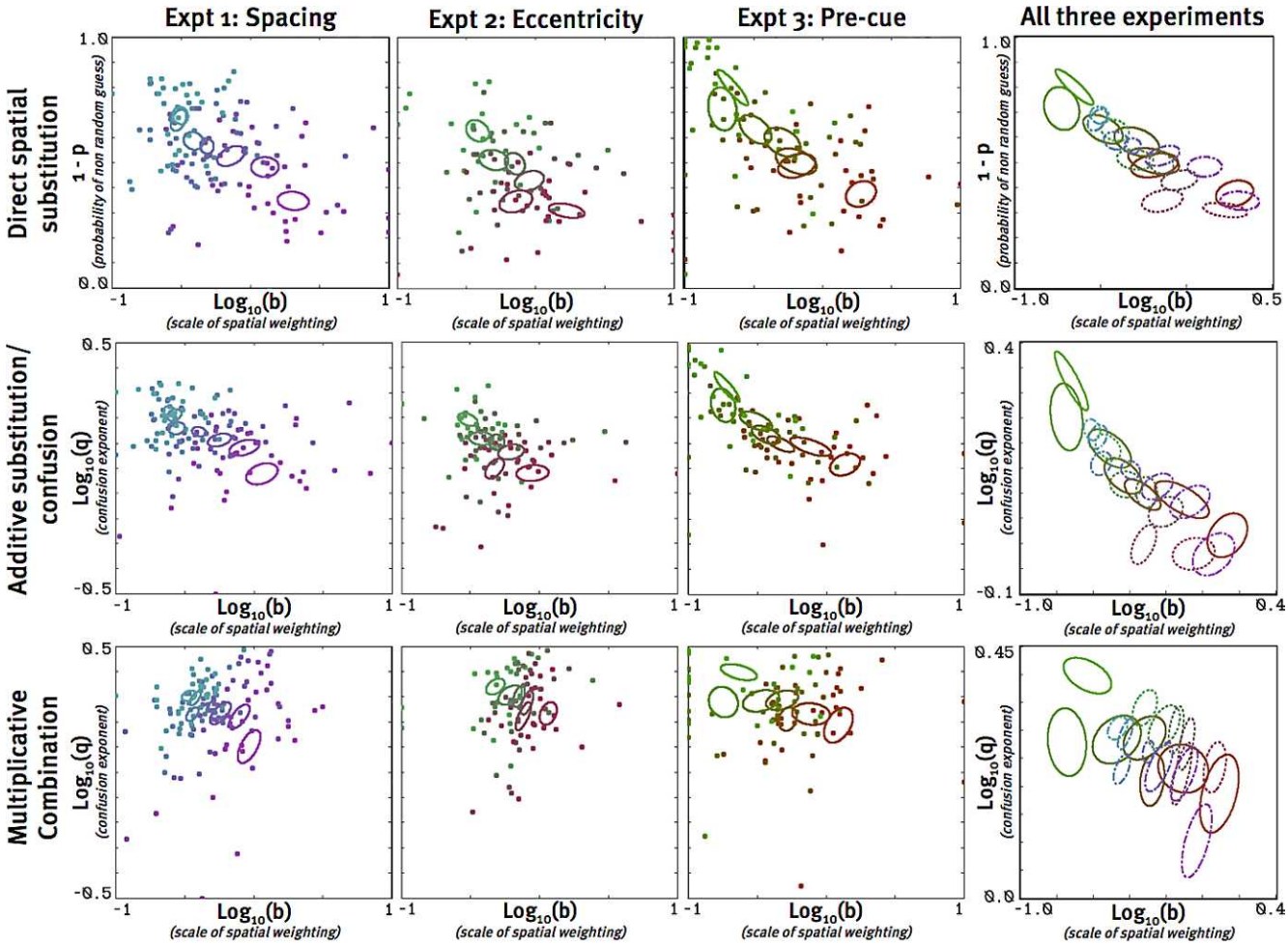

Figure 9. Changes in parameters for three two-parameter error models as a function of experiment and condition. Each panel shows the maximum likelihood parameter estimates for individual subjects within a given condition (points) and the across-subject standard error ellipses for the average parameter values within a given condition. Because the covariance of the pairs of parameter estimates across subjects is not zero, these standard error ellipses have some rotation in parameter space. Within each experiment, different conditions are color coded (in the same manner as Figure 6) so that more difficult conditions within each experiment have redder hues. The rightmost panels show the average parameter (standard error ellipses) for all conditions and all experiments. Experiment- and condition-specific color coding is preserved; different experiments are further distinguished by line property: Experiment 1a is dot-dashed, Experiment 2 is dashed, and Experiment 3 is solid. We argue that the across-experiment/condition variation in parameters falls along a straight line. (top row) The direct spatial substitution model (Equation 10) has two parameters: the probability of random guessing (**p**; *y*-axis) and the breadth of the spatial weighting function (**b**; *x*-axis). (middle row) The spatial substitution plus letter confusion model (Equation 13) has the spatial weighting function breadth on the *x*-axis (**b**), and the confusion matrix exponent (**q**) on the *y*-axis. (bottom row) The multiplicative combination model (e1. 15) also has the confusion matrix exponent on the *y*-axis and the spatial breadth on the *x*-axis, but this model is qualitatively different from the independent spatial substitution plus letter confusion model.

We can undertake the same analysis for the other two-parameter models we might consider: spatial substitution plus letter confusion and multiplicative combinations.

Figure 9 (middle) shows parameter results for the spatial substitution plus letter confusion model; thus, instead of a random guessing parameter, we consider **q**, the confusion matrix exponent. In all experiments, conditions that increase difficulty of the trial also increase the breadth of the spatial weighting function and decrease the confusion matrix exponent (bringing the confusion matrix closer to uniformity). We find that collapsing over experiments and conditions, parameters again appear to fall on one dimension: (a) 95% intervals on the variance along the second principal component spans one [0.61, 2.25], (b) the probability that this variance is greater than one is low (0.62), and (c) the chi-squared test comparing this

variance to one is not significant, $\chi^2(20) = 22.64$, $p = 0.31$. As should be clear from the graph, variance along the primary principal component is much greater than one (95% confidence interval: [7.4, 27.3]). Thus, if we use the spatial substitution plus letter confusion model, we again find that all manipulations of difficulty appear to have effects that can be described by one dimension.

Figure 9 (bottom) follows the same figure structure but shows parameters for the multiplicative combination model. Although the two parameters are (as in the spatial substitution plus letter confusion model) the breadth of spatial weighting and the confusion matrix exponent, those parameters interact in qualitatively different ways in the multiplicative model. We again find that conditions of increasing difficulty tend to yield broader spatial weightings and lower confusion matrix exponents; however, the multiplicative error model seems to result in considerably greater dispersion in fits for individual subjects, particularly along the axis orthogonal to difficulty, causing standard errors to be quite large along this orthogonal axis. We believe this greater across-subject dispersion is symptomatic of a poor model, but we have no quantitative argument to support this claim. Nevertheless, despite the qualitatively different model and across-subject dispersion, we again find little evidence that the variation of parameters across experiments and conditions is more than one dimensional: (a) 95% intervals on the variance along the second principal component span one [0.84, 3.13], (b) the probability that this variance is greater than one is not sufficient to reject the null (although higher than the two other models; 0.91), and (c) the chi-squared test comparing this variance to one is marginally significant, $\chi^2(20) = 31.31$, $df = 20$, $p = 0.051$. As should be clear from the graph, variance along the primary principal component is much greater than one (95% confidence interval: [6.9, 26]). Although the results under the multiplicative combination model are less persuasive (presumably because of the correlation in parameter estimates orthogonal to the principal component), we again find little evidence that different manipulations of difficulty affect error distributions in different ways.

From these results, we argue that changes in error distributions as we manipulate perceptual difficulty all fall along one dimension. This seems to be the case regardless of which of the three two-parameter models we consider to best summarize subjects' error distributions. In other words, whether we manipulate the spacing or eccentricity of the array, or even the precue duration, we find that all conditions across all experiments appear to make different settings along one parameter of difficulty. This surprising result comes with certain caveats of which the reader should be cautious; we discuss these next.

Discussion

We characterized crowding errors by specifying a set of statistical error models that summarized how error distributions vary across three different difficulty manipulations: letter spacing (Experiment 1), eccentricity (Experiment 2), and precueing (Experiment 3). We have three major conclusions: First, human error distributions reflect both spatial weighting (responses depend on letters adjacent to the target) and orthographic letter confusion (observers tend to report some specific letters in place of others), and models that best account for human error distributions must take both factors into account. Second, once both factors are taken into account, it is difficult to differentiate models that consider the presented letters independently from models that operate over combinations of letters, suggesting that distinguishing between spatial substitution and spatial pooling models will be difficult. Third, regardless of how we manipulate perceptual difficulty (spacing, eccentricity, or precueing) and which model we use to measure error distributions, all conditions from all experiments seem to manipulate difficulty in the same manner, suggesting that all manipulations influence one latent difficulty parameter.

Our first conclusion—that both the distance-dependent influence of flankers and letter-letter confusions must be taken into account to capture human errors in crowding—strikes us as uncontroversial and retrospectively obvious. However, most studies of intrusions and substitutions consider letters to be homogenously confusable stimuli and consider only direct spatial substitutions. Indeed, that was also our first inclination until reading Freeman et al. (2012). While retrospectively obvious, this point proves important: If one considers only the possibility of spatial intrusions without taking into account the individual letter confusability, experimental findings that show preferential reporting of similar flankers seem to argue for a spatial pooling model rather than a model in which letters are considered independently (Freeman et al., 2012; Huckauf & Heller, 2002; Krumhansl & Thomas, 1977). Instead, we find that as long as letter-letter confusion and spatial weighting are taken into account, independent and nonindependent error models are hard to distinguish, and both favor intrusions of similar flankers.

This leads to our second conclusion—at least with letter stimuli as used here, error models that treat flankers independently are difficult to distinguish from those that nonlinearly combine adjacent letters, as long as both models include a spatial weighting function and the possibility of orthographic confusion. We believe that this result implies that despite the fact that spatial-uncertainty and texture-synthesis models offer qualitatively different mechanistic/computational accounts

of crowding, they will also be quite difficult to tease apart with letter-based response error data such as ours. Moreover, although the error distributions in our data are more likely under the independent error model (spatial substitution plus letter confusion) than the naïve pooling model (multiplicative combination), there are two senses in which our findings cannot separate pooling and spatial substitution accounts of crowding in general. First, while the independent spatial substitution plus letter confusion model has significantly higher likelihood aggregated over all conditions and experiments, those differences are quite small and unstable across experiments and conditions; thus, we do not believe that this result is particularly diagnostic. Moreover, this conclusion is limited to the specific naïve pooling model we consider. Many pooling models could exist, and some of them are likely to capture patterns in our data that independent spatial substitution plus letter confusion cannot. Indeed, by using texture synthesis models that specify the space of features and how those features are used to infer letters, there may be no need to empirically estimate a letter-letter confusion matrix; it may instead be derived from the texture synthesis models themselves. The second and more serious reason is that spatial-uncertainty and texture-synthesis models can both be arranged to yield error distributions that correspond to our independent, substitution/confusion, and multiplicative combination models. For instance, texture synthesis models could yield error distributions like our multiplicative combination models (given the right basis set of textures and an appropriate decision model to choose letters based on the texture representation), but they might also produce errors like our spatial substitution/confusion model (if the basis set of textures emphasized complex orthographic features). In turn, spatial uncertainty models would yield our substitution/confusion model if subjects sample spatial positions and then sample letters conditioned on the sampled letter position, but a spatial-uncertainty model would yield multiplicative combinations of letters if each letter were interpreted as evidence about a single common cause. Of course, some of these model structures make more sense than others; however, the crucial point is that neither overarching class of models yields substantive constraints about the errors that might arise during crowding unless minute details are specified.

Our conclusions are further limited outside of the specific experimental paradigm we employed: using arrays of nine letters. Greenwood, Bex, and Dakin (2009) used intersecting lines to test pooling and substitution models and found a slight advantage for pooling (although, like us, they note the difficulty of discerning such models). We can think of at least two reasons why this discrepancy might arise. First, Greenwood et al. (2009), like most crowding studies, used fewer flankers (typically, arrays of just three or five letters). Perhaps the larger nine-item array we used introduces a greater degree of spatial uncertainty and encourages substitution. The second, and perhaps more exciting, possibility, is that specific stimuli may influence whether they are pooled or substituted. Perhaps when the visual system summarizes a portion of the visual field, it must solve a problem analogous to the distinction between prototype and exemplar categorization models (Griffiths, Sanborn, Canini, & Navarro, 2008). Are the contents of this portion of the visual field best represented as a few unique entities or as samples from one distribution that can be summarized as a coherent statistical ensemble (Alvarez, 2011) or a texture (Freeman & Simoncelli, 2011)? If the visual system aims to figure out whether to agglomerate or individuate elements in the scene, then substitution and pooling are the two limiting cases of the range of solutions the visual system may entertain. The visual system is then likely to reach different answers for different sorts of stimuli; perhaps stimuli that differ along basic visual dimensions (like the height of a bisecting horizontal line in Greenwood et al., 2009) may be more likely to be represented as an ensemble/texture than those that are not simple parametric variations of one another. While it is interesting to speculate that such a flexible inference process underlies the discrepancies across stimuli between substitution and pooling, in the meantime, we are left with just the caveat that our conclusions are limited to the specific experimental paradigm we employ.

Our final conclusion is that different manipulations of perceptual difficulty yield indistinguishable changes in behavior; all manipulations of difficulty that we considered—spacing, eccentricity, and precueing—yield error distributions that can be summarized along one axis of accuracy. Although the error distributions across experiments and conditions fall along a single axis in our error-model parameters, they reflect changes in both spatial uncertainty and letter confusion. It would appear that although both spatial uncertainty and letter confusion are requisite elements of a crowding-error model; they themselves appear to change together, as though they arose from a common mechanism. We find this result quite compelling but also very surprising. Indeed, we chose the three specific manipulations here precisely because we expected them to yield qualitatively different effects on human error distributions. Thus, we eagerly list the caveats that come along with this result. First, this conclusion is robust to the specific error models we considered (direct spatial substitution, independent substitution/confusion, or multiplicative combinations), but there may well be better, or at least different, error models that can differentiate among the qualitatively different

perceptual difficulty manipulations. Because we do not know what these other error models might be, we cannot rule out the possibility that a different way of characterizing error distributions would find qualitatively different consequences of the three difficulty manipulations. In short, this conclusion is predicated on the set of models we considered. Second, we considered only three manipulations of difficulty: spacing, eccentricity and precueing. We chose these manipulations because they struck us as being common, robust, and qualitatively different (apparently, as far as we can tell, we were wrong about them being qualitatively different). However, other crowding manipulations may well manipulate difficulty and error distributions in qualitatively different ways. In the most extreme example, we suspect that manipulating difficulty by blindfolding subjects will yield error distributions that do not fall along the curves we observed. Similarly, when characters are near contrast threshold, a spatial cue can offset part of a crowding-induced deficit by increasing contrast sensitivity without influencing flanker confusion (Strasburger & Malania, 2013). In sum, although our results argue for a single tuning parameter governing difficulty and error distributions of crowding tasks, other manipulations may well uncover different dimensions of error distributions.

Conclusion

We find that human error distributions in crowding tasks reflect both spatial imprecision and confusion of individual items, and it is hard to distinguish qualitatively different models of errors that include both of these factors. Furthermore, increasing the difficulty of crowding tasks by manipulating spacing, eccentricity, or precueing appears to have the same effect: One dimension describes the complicated ways in which error distributions change as the task becomes harder. While we argue that our results, and results of most experiments of the form we considered, cannot separate texture-synthesis and spatial-uncertainty models, we believe that setting as a target the prediction of the across-trial variation in human errors in crowding tasks will provide useful advances for both types of crowding models.

Keywords: Crowding, texture, spatial pooling, spatial uncertainty, orthographic confusion

Acknowledgments

This work was supported by the Office of Naval Research MURI N00014-07-1-0937. We thank Hans Strasburger and an anonymous reviewer for uncommonly insightful and helpful comments.

Commercial relationships: none.
Corresponding author: Edward Vul.
Email: evul@ucsd.edu.
Address: Psychology, University of California, San Diego, San Diego, CA, USA.

Footnotes

[1] Although it is possible to define distance with respect to the weighting function in terms of angle of arc or degrees visual angle, it is not necessary for our analysis, and it would require a few additional assumptions. Moreover, these physical measures of distance would be useful if the breadth of a spatial weighting function defined over them remained constant across conditions, but that is not the case.

[2] http://www.edvul.com/crowding-models/

[3] Because proportional changes in **q** between zero and one are just as meaningful as proportional changes above one, we consider the logarithm of **q** as our measure because that linearizes the relationship between our parameter estimate and accuracy reporting the target.

[4] Note that the error bars in Figure 7 correspond to across-subject standard errors of the mean, while the comparisons are done within subjects; hence the error bars are larger than the relevant standard errors of the difference.

[5] For the likelihood ratio test, the natural logarithm is relevant, elsewhere we report the base 10 logarithm as we find it more intuitively interpretable.

References

Alvarez, G. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences*, *15*(3), 122–131.

Andriesen, J., & Bouma, H. (1976). Eccentric vision: Adverse interactions between line-segments. *Vision Research, 16*(1), 71–78.

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, http://www.journalofvision.org/content/9/12/13, doi:10.1167/9.12.13. [PubMed] [Article]

Bernard, J., & Chung, S. (2011). The dependence of crowding on flanker complexity and target–flanker

similarity. *Journal of Vision, 11*(8):1, 1–16, http://www.journalofvision.org/content/11/8/1, doi:10.1167/11.8.1. [PubMed] [Article]

Bouma, H. (1970). Interaction effects in parafoveal letter recognition. *Nature*, *226*(4), 177–178.

Bouma, H. (1973). Visual interference in the parafoveal recognition of initial and final letters of words. *Vision Research, 13*(4), 767–782.

Brainard, D. (1997). The psychophysics toolbox. *Spatial Vision*, *10*(4), 433–436.

Carrasco, M., & Frieder, K. (1997). Cortical magnification neutralizes the eccentricity effect in visual search. *Vision Research*, *37*(1), 63–82.

Cavanagh, P., & Holcombe, A. (2007). Non-retinotopic crowding. *Journal of Vision, 7*(9):338, http://www.journalofvision.org/content/7/9/338, doi:10.1167/7.9.338. [Abstract]

Chastain, G. (1982). Feature mislocalizations and misjudgments of intercharacter distance. *Psychological Research*, *44*(1), 51–65.

Ehlers, H. (1953). Clinical testing of visual acuity. *Archives of Ophthalmology*, *49*(4), 431.

Ehlers, H. (1936). The movements of the eyes during reading. *Acta Opthalmologica*, *14*(4), 56–63.

Eriksen, C., & Collins, J. (1969). Temporal course of selective attention. *Journal of Experimental Psychology*, *80*(2), 254–261.

Eriksen, C., & Eriksen, B. (1972). Visual backward masking as measured by voice reaction time. *Attention, Perception, & Psychophysics, 12*(1), 5–8.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task *Attention, Perception, & Psychophysics, 16*(1), 143–149.

Eriksen, C. W., & Rohrbaugh, J. (1970). Some factors determining efficiency of selective attention. *The American Journal of Psychology, 83*(3), 330–342.

Estes, W., Allmeyer, D., & Reder, S. (1976). Serial position functions for letter identification at brief and extended exposure durations. *Attention, Perception, & Psychophysics, 19*(1), 1–15.

Estes, W. K., & Wolford, G. L. (1971). Effects of spaces on report from tachistoscopically presented letter strings. *Psychonomic Science, 25*(2), 77–80.

Flom, M., Weymouth, F., & Kahneman, D. (1963). Visual resolution and contour interaction. *Journal of the Optical Society of America*, *53*(9), 1026–1032.

Freeman, J., Chakravarthi, R., & Pelli, D. G. (2012). Substitution and pooling in crowding. *Attention, Perception, & Psychophysics, 74*(2), 379–396.

Freeman, J., & Simoncelli, E. (2011). Metamers of the ventral stream. *Nature Neuroscience*, *14*(9), 1195–1201.

Geiger, G., & Lettvin, J. (1986). Enhancing the perception of form in peripheral vision. *Perception*, *15*(2), 119.

Gelman, A., & Rubin, D. (1995). Avoiding model selection in Bayesian social research. *Sociological Methodology, 25,* 165–174.

Gervais, M., Harvey, L., & Roberts, J. (1984). Identification confusions among letters of the alphabet. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(5), 655.

Greenwood, J., Bex, P., & Dakin, S. (2009). Positional averaging explains crowding with letter-like stimuli. *Proceedings of the National Academy of Sciences*, *106*(31), 13130–13135.

Griffiths, T. L., Sanborn, A. N., Canini, K. R., & Navarro, D. J. (2008). Categorization as nonparametric Bayesian density estimation. In A. B. Lastname (Ed.), *The probabilistic mind: Prospects for Bayesian cognitive science* (pp. 303–328). Oxford, UK: Oxford University Press.

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

Huckauf, A., & Heller, D. (2002). What various kinds of errors tell us about lateral masking effects. *Visual Cognition, 9*(7), 889–910.

Korte, W. (1923). Über Die Gestaltauffassung im Indirecten Sehen. *Zeitschrift für Psychologie, 93*(3), 17–82.

Krumhansl, C., & Thomas, E. (1977). Effect of level of confusability on reporting letters from briefly presented visual displays. *Attention, Perception, & Psychophysics, 21*(3), 269–279.

Levi, D. (2008). Crowding—An essential bottleneck for object recognition: A mini-review. *Vision Research, 48*(5), 635.

Levi, D., McGraw, P., & Klein, S. (2000). Vernier and contrast discrimination in central and peripheral vision. *Vision Research, 40*(8), 973–988.

Luce, D. (1959). *Individual choice behavior a theoretical analysis*. New York, NY: John Wiley and Sons.

Monti, P. (1973). Lateral masking of end elements by inner elements in tachistoscopic pattern perception. *Perceptual & Motor Skills, 36*(3), 777–778.

Mueller, S. T., & Weidemann, C. T. (2012). Alphabetic letter identification: Effects of perceivability, similarity, and bias. *Acta Psychologica, 139*, 19–37, doi:10.1016/j.actpsy.2011.09.014.

Parkes, L., Lund, J., Angelucci, A., Solomon, J., Morgan, M. (2001). Compulsory averaging of

crowded orientation signals in human vision. *Nature Neuroscience*, *4*(7), 739–744.

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

Pelli, D., & Tillman, K. (2008). The uncrowded window of object recognition. *Nature Neuroscience*, *11*(10), 1129–1135.

Pelli, D., Tillman, K., Freeman, J., Su, M., Berger, T., & Majaj, N. (2007). Crowding and eccentricity determine reading rate. *Journal of Vision, 7*(2):20, 1–36, http://www.journalofvision.org/content/7/2/20, doi:10.1167/7.2.20. [PubMed] [Article]

Portilla, J., & Simoncelli, E. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, *40*(1), 49–70.

Posner, M. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, *32*(1), 3–25.

Rovamo, J., & Virsu, V. (1979). An estimation and application of the human cortical magnification factor. *Experimental Brain Research*, *37*(3), 495–510.

Rovamo, J., Virsu, V., & Näsänen, R. (1978). Cortical magnification factor predicts the photopic contrast sensitivity of peripheral vision. *Nature*, *271*(5640), 54–56.

Strasburger, H. (2005). Unfocussed spatial attention underlies the crowding effect in indirect form vision. *Journal of Vision, 5*(11):8, 1024–1037, http://www.journalofvision.org/content/5/11/8, doi:10.1167/5.11.8. [PubMed] [Article]

Strasburger, H., Harvey, L., & Rentschler, I. (1991). Contrast thresholds for identification of numeric characters in direct and eccentric view. *Attention, Perception, & Psychophysics, 49*(6), 495–508.

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

Stuart, J., & Burian, H. (1962). A study of separation difficulty: Its relationship to visual acuity in the normal amblyopic eye. *American Journal of Ophthalmology, 53*(471), 163–169.

Taylor, S., & Brown, D. (1972). Lateral visual masking: Supraretinal effects when viewing linear arrays with unlimited viewing time. *Perception & Psychophysics, 12*(1), 97–99.

Townsend, D. (1971). Theoretical analysis of an alphabetic confusion matrix. *Perception & Psychophysics, 9*(1A), 40–50.

Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.

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

Vlaskamp, B., & Hooge, I. (2006). Crowding degrades saccadic search performance. *Vision Research*, *46*(3), 417–425.

Vul, E., Hanus, D., & Kanwisher, N. (2009). Attention as inference: Selection is probabilistic; responses are all-or-none samples. *Journal of Experimental Psychology: General, 138*(4), 546.

Vul, E., & Rich, A. (2010). Independent sampling of features enables conscious perception of bound objects. *Psychological Science*, *21*(8), 1168–1175.

Westheimer, G., & Hauske, G. (1975). Temporal and spatial interference with vernier acuity. *Vision Research*, *15*(10), 1137–1141.

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

Wilkinson, F., Wilson, H., & Ellemberg, D. (1997). Lateral interactions in peripherally viewed texture arrays. *Journal of the Optical Society of America*, *14*(9), 2057–2068.

Wolford, G. (1975). Perturbation model for letter identification. *Psychological Review*, *82*(3), 184.

Wolford, G., & Chambers, L. (1983). Lateral masking as a function of spacing. *Attention, Perception, & Psychophysics, 33*(2), 129–138.

Woodrow, H. (1938). The effect of pattern upon simultaneous letter-span. *The American Journal of Psychology*, *51*(1), 83–96.

Woodworth, R. (1938). *Experimental psychology*. New York, NY: Holt.

Appendix A: Example of calculating trial likelihood

Each model specifies a likelihood function for the data. For instance, if the letter array on a single trial were “A”, “B”, “C”, “D”, “E”, “F”, “G”, “H”, “I”, (the target being the central letter, “E”) and the subject

reported "E", the likelihood of this trial given under the spatial substitution/confusion model (Equation 12) with parameters $\mathbf{p} = 0.2$, $\mathbf{b} = 0.5$, $\mathbf{q} = 1.2$ would be 0.2535:

$$P(\text{'E'}|\,\mathcal{L}) = \mathbf{p}\frac{1}{26} + (1-\mathbf{p})\sum_{x=-4}^{4} P_L(x|\mathbf{b})\frac{Q\left(\mathcal{L}(x),\text{'E'}\right)^{\mathbf{q}}}{\sum_{j'=1}^{26} O\left(\mathcal{L}(x),j'\right)^{\mathbf{q}}}$$

$$P(\text{'E'}|\,\mathcal{L}) = \frac{0.2}{26} + 0.8\sum_{x=-4}^{4} P_L(x|0.5)\frac{Q\left(\mathcal{L}(x),\text{'E'}\right)^{1.2}}{\sum_{j'=1}^{26} O\left(\mathcal{L}(x),j'\right)^{1.2}}$$

$$\begin{aligned}P(\text{'E'}|\mathcal{L}) = {} & \frac{0.2}{26} + 0.8*\Big(P_L(-4|\mathbf{b})\frac{Q(\text{'A'},\text{'E'})^{\mathbf{q}}}{\sum_{j'=1}^{26} Q(\text{'A'},j')^{\mathbf{q}}} \\ & + 0.8*P_L(-3|0.5)\frac{Q(\text{'B'},\text{'E'})^{1.2}}{\sum_{j'=1}^{26} Q(\text{'B'},j')^{1.2}} \\ & + 0.8*P_L(-2|0.5)\frac{Q(\text{'C'},\text{'E'})^{1.2}}{\sum_{j'=1}^{26} Q(\text{'C'},j')^{1.2}} \\ & + 0.8*P_L(-1|0.5)\frac{Q(\text{'D'},\text{'E'})^{1.2}}{\sum_{j'=1}^{26} Q(\text{'D'},j')^{1.2}} \\ & + 0.8*P_L(0|0.5)\frac{Q(\text{'E'},\text{'E'})^{1.2}}{\sum_{j'=1}^{26} Q(\text{'E'},j')^{1.2}} \\ & + 0.8*P_L(1|0.5)\frac{Q(\text{'F'},\text{'E'})^{1.2}}{\sum_{j'=1}^{26} Q(\text{'F'},j')^{1.2}} \\ & + 0.8*P_L(2|0.5)\frac{Q(\text{'G'},\text{'E'})^{1.2}}{\sum_{j'=1}^{26} Q(\text{'G'},j')^{1.2}} \\ & + 0.8*P_L(3|0.5)\frac{Q(\text{'H'},\text{'E'})^{1.2}}{\sum_{j'=1}^{26} Q(\text{'H'},j')^{1.2}} \\ & + 0.8*P_L(4|0.5)\frac{Q(\text{'I'},\text{'E'})^{1.2}}{\sum_{j'=1}^{26} Q(\text{'I'},j')^{1.2}}. \end{aligned} \quad (17)$$

Using the appropriate "A", "B", . . . , "I" rows of our confusion matrix Q and the Laplacian spatial weighting function described in Equation 2, we obtain the following:

$$\begin{aligned}P(\text{'E'}|\mathcal{L}) = {} & \frac{0.2}{26} + 0.8*0.0004*\frac{0.0097}{0.7011} \\ & + 0.8*0.0029*\frac{0.0516}{0.6043} \\ & + 0.8*0.0215*\frac{0.0096}{0.6278} \\ & + 0.8*0.159*\frac{0.0111}{0.6744} \\ & + 0.8*0.6321*\frac{0.2758}{0.6059} \\ & + 0.8*0.159*\frac{0.0578}{0.5826} \\ & + 0.8*0.0215*\frac{0.0086}{0.6184} \\ & + 0.8*0.0029*\frac{0.0295}{0.5777} \\ & + 0.8*0.0004*\frac{0.0127}{0.5692} \end{aligned}$$

$$P(\text{'E'}|\mathcal{L}) = 0.2535 \quad (18)$$

Thus, the likelihood of the subject reporting an "E", on this trial, with parameters $\mathbf{p} = 0.2$, $\mathbf{b} = 0.5$, $\mathbf{q} = 1.2$, is 0.2535.

Appendix B: One-dimensional difficulty

To determine whether the various experiments and conditions all manipulate a single dimension (i.e., tuning parameter) of difficulty, we used the following analysis (shown in Figure 10).

For each two-parameter model we have two vectors of parameters for each experimental condition $x_j^{(i)}$ and $y_j^{(i)}$, where x and y denote the two first parameters, i is the experimental condition (with conditions concatenated across experiments), and j is the index over all subjects contributing in that experimental condition.

First, we calculate the mean and covariance of x and y for each condition across subjects, thus obtaining the mean (vector) parameter estimates for the ith condition, denoted $\mu^{(i)}$, and the across-subject covariance of parameter estimates $\sum^{(i)}$. The standard error of the mean is given by dividing the covariance of parameter estimates by the number of subjects: $\sum_{\mu}^{(i)} = \sum^{(i)}/n$.

Second, we obtained the principal components of the distribution of mean vectors, that is the two orthogonal components of the variation (along experiment conditions, i) of the mean parameters for each condition ($\mu^{(i)}$). This yields a vector of component loadings $\mathbf{C}$ (describing the orientation of each principal component in the original parameter space) and the mean vector within that coordinate frame

(expressed as deviations from the grand mean along the two components), $v^{(i)}$.

Third, we projected the standard error covariance matrices into the principal component coordinate frame: $\Xi^{(i)} = \mathbf{C}^{\mathrm{T}} * \sum_{\mu}^{(i)} * \mathbf{C}$. Thus, we obtain standard errors of the mean (for each condition) within the principal component coordinate frame.

Fourth, we studentized every mean vector in the principal component space by dividing the deviation from the grand mean by the standard error along that dimension. In other words, we calculated the deviation (in units of standard error) of each condition parameter estimate mean from the grand mean across all conditions along the two principal components. These are studentized (alternatively z-scored) deviations of the means along the two principal components.

$$z_1^{(i)} = v_1^{(i)} / \sqrt{\Xi_{1,1}^{(i)}}$$

$$z_2^{(i)} = v_2^{(i)} / \sqrt{\Xi_{2,2}^{(i)}} \quad (19)$$

These studentized residuals in principal component space can be analyzed in two ways: either using the Bayesian posterior distribution over their covariance (to obtain credible intervals and estimates of the magnitude of the variance along the second component) or via frequentist chi-squared tests to assess whether variation along the second component is significantly different from one. If we can be reasonably confident that the variance of the residuals along the second principal component is indistinguishable from one, then we have evidence (albeit evidence by failing to reject the null), that all the variation in accuracy across conditions boils down to a single dimension.

Posterior covariance

We calculated the covariance of these studentized residuals in the principal component space and obtained a posterior estimate of this covariance matrix using an inverse-Wishart distribution (using a non-informative prior) (Gelman & Rubin, 1995).

By obtaining parametric bootstrap samples from this inverse-Wishart distribution (we used 10,000 samples), we can calculate:

(a) posterior credible intervals on the variance along the second principal component.
(b) the posterior probability that the variance along the second principal component is greater than one.
(c) a credible interval on the ratio of the variance along the first principal component divided by the variance along the second.

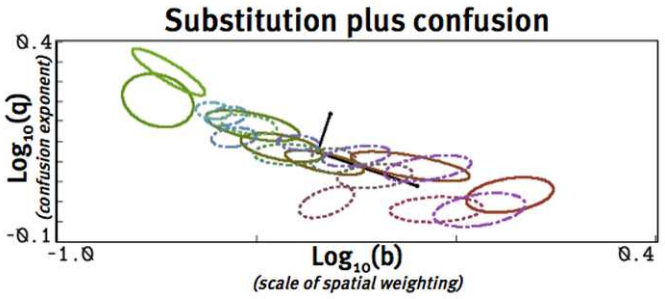

Figure 10. Illustration of the analysis via which we argue that all the experiments and conditions manipulate accuracy along only one dimension (illustrated using the parameters of the spatial substitution plus letter confusion model from Figure 9). We obtain the principal components of the variation in parameters across conditions, (black lines at right angles). Then we studentize the deviation of each condition from the grand mean along each component by dividing these deviations by the standard error of that condition mean as projected onto that principal component (see text). The crucial test is whether these studentized residuals along the second principal component have variance greater than one (where variance of one is the expected variance under the null hypothesis).

Chi-squared test for variance

As a simple classical alternative, we can run a chi-squared test on the sum of squared studentized residuals along the second principal component to test whether it is significantly greater than it would be under the null model of no variance, except for the error along the second principal component. The chi-squared distribution describes the sampling distribution of the summed squared z scores sampled from a standard normal distribution. That is what the studentized residuals along the second principal component should be under the null hypothesis of zero variation (other than error) along this component. Thus, we can calculate a chi-squared statistic as $\chi^2 = \sum_1^n (z_2^{(i)})^2$, and under the null hypothesis this should follow a chi-squared distribution with n degrees of freedom ($df = n$). We can obtain a p value by assessing the upper tail probability of this chi-squared distribution (the probability of seeing a chi-squared as large as ours or larger under the null hypothesis that variation along the second principal component arises only from sampling error).

Appendix C: Experiment 1b

Experiment 1b differed from Experiment 1a only in that the precue was 200 ms rather than 50 ms in order to attempt to create a family of conditions where cue

uncertainty played no role. Because performance was near ceiling for a number of conditions, and the goal of this experiment is to assess whether our conclusions hold when spatial uncertainty is eliminated, we excluded this experiment from the primary analyses and analyze it separately here.

The primary questions of interest in Experiment 1b are whether our model comparison results hold under conditions that try to eliminate spatial uncertainty in the location of the cue and target letter by using a long precue. Figure 11 shows the log likelihood per trial gain for each of the seven models (plot style follows that seen in Figure 7, see text of Model comparison section for details).

In this experiment we again find:

(1) In particularly difficult conditions, all models perform poorly. The log-likelihood gain over the random guessing baseline is much lower for closer spacings (Figure 11 top), although in this experiment, all conditions are easier due to the longer precue.

(2) Unlike Experiments 1a, 2, and 3, more sophisticated error models provide a smaller and less consistent advantage over a simple correct/random mixture model. We believe this is the case because subjects' performance in Experiment 1b is much higher, therefore there are fewer errors that can be used to inform different error models.

(3) Nonetheless, the direct spatial substitution model provides a significant benefit over the correct/random mixture model, indicating that observers tend to substitute flankers directly for the target (the comparison between the purple data points to the baseline (zero) in Figure 11, bottom). This is evident both in the $\log_{10}$ likelihood per trial gain averaged across all conditions (mean: 0.027, *SD*: 0.0077) and in across-subject t tests within 6/7 conditions. The overall advantage of direct spatial substitution over a simple correct/random mixture can be tested via a likelihood ratio test that aggregates likelihoods over all subjects and conditions and accounts for the extra parameters in the spatial substitution model. The spatial substitution model (total $\log_e$ likelihood $= -8542$ with 196 parameters) has a higher likelihood than the correct/random mixture (total $\log_e$ likelihood $= -8936$ with 98 parameters), with a very significant likelihood ratio test ($X^2(98) = 787$, $p \approx 0$).

(4) However, there is a further advantage over the direct spatial substitution model of models that take into account the confusion matrix for the spatial substitution plus letter confusion model (comparison between purple and red points; $X^2(650) = 301$, $p \approx 0$ for the spatial substitution model with no random guessing parameter). For

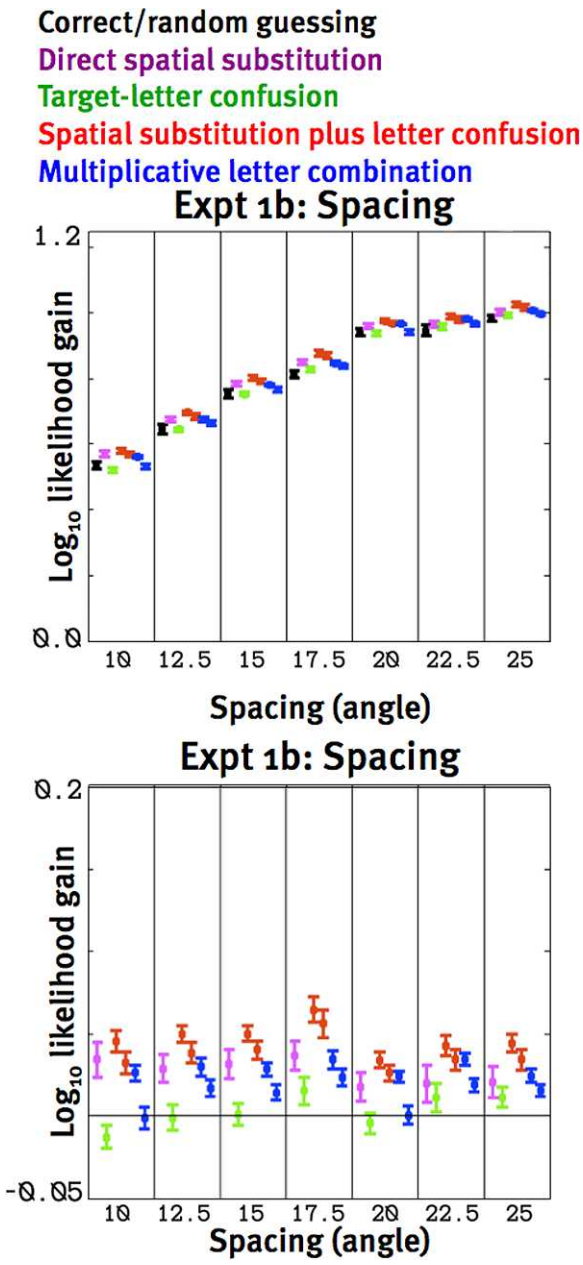

Figure 11. Improvements in $\log_{10}$ likelihood per data point for different error models compared to different baselines (higher is better) for all conditions in Experiment 1b (plot style follows that seen in Figure 7). (top) Other error models compared to uniform random guessing. Each point corresponds to the average gain in log likelihood per trial over the fully random model (more is better). Plotted points are averages over subjects ± 1 *SEM* (for reference, random guessing predicts constant $\log_{10}$ likelihood $= -1.415$ per data point). These plots show that the average likelihood per data point is largely dominated by variation across conditions: Responses on harder trials are less predictable. (bottom) Average log likelihood per trial gain compared to the correct/random mixture model (black points in top panel). This baseline takes into account variation in difficulty across conditions, thus the differences between error models are easier to discern.

the multiplicative combination model, the advantage over direct spatial substitution is less reliable (comparison between purple and blue points in Figure 11). The full multiplicative combination model slightly outperforms direct spatial substitution ($G^2 = 53$, $df = 748$, $p \approx 0$), but the multiplicative combination model without random guessing has a significantly lower net log likelihoods (−8750) than the direct substitution model (−8542).

(5) Similarly, taking the letter confusion matrix into account provides considerable advantage over a simple correct/random mixture model: The average $\log_{10}$ likelihood/trial gain: 0.0157, $SD = 0.0065$—comparison between green points and baseline in Figure 7 bottom; ($G^2 = 84$, $df = 650$, $p \approx 0$). But again, there is a considerable further benefit of adding spatial weighting to the model so that responses are not based on the target alone but incorporate influences from adjacent letters. Again, this is the case both for the spatial substitution plus letter confusion model ($G^2 = 1005$, $df = 98$, $p \approx 0$—comparison between green and red points) and multiplicative models ($G^2 = 287$, $df = 98$, $p \approx 0$—comparison between green and blue points).

The previous two findings again indicate that neither spatial substitution nor letter confusion alone are sufficient to account for the error distributions; a successful model should include both.

(6) Consistent with the results of Experiments 1a (and the net results over Experiments 1a, 2, and 3), we see a stable and significant advantage of the spatial substitution plus letter confusion model over the multiplicative combination model (red points compared to blue). The average log-likelihood/trial gain of the spatial substitution plus letter confusion model is small but reliable (across condition mean: 0.0248, *SD*: 0.0069 for the no random guessing models), all within condition t tests significant at $p \leq 0.037$, and the overall comparison is highly significant ($G^2 = 718$, $df = 1$, $p \approx 0$). This conclusion also holds for versions of these models that include the random guessing parameter.

Thus the primary results of our model comparison hold in this control experiment where we have tried to eliminate spatial uncertainty by greatly increasing precue duration.