Probability and the changing shape of response distributions for orientation

Britt Anderson

Dept. of Psychology and Centre for Theoretical Neuroscience,
University of Waterloo,
Waterloo, ON, Canada

Spatial attention and feature-based attention are regarded as two independent mechanisms for biasing the processing of sensory stimuli. Feature attention is held to be a spatially invariant mechanism that advantages a single feature per sensory dimension. In contrast to the prediction of location independence, I found that participants were able to report the orientation of a briefly presented visual grating better for targets defined by high probability conjunctions of features and locations even when orientations and locations were individually uniform. The advantage for high-probability conjunctions was accompanied by changes in the shape of the response distributions. High-probability conjunctions had error distributions that were not normally distributed but demonstrated increased kurtosis. The increase in kurtosis could be explained as a change in the variances of the component tuning functions that comprise a population mixture. By changing the mixture distribution of orientation-tuned neurons, it is possible to change the shape of the discrimination function. This prompts the suggestion that attention may not "increase" the quality of perceptual processing in an absolute sense but rather prioritizes some stimuli over others. This results in an increased number of highly accurate responses to probable targets and, simultaneously, an increase in the number of very inaccurate responses.

Introduction

We are faster (Posner, 1980) and we are better (Anderson & Druker, 2013; Prinzmetal et al., 1998; Prinzmetal et al., 1997) at identifying visual targets when we are cued to locations or pertinent features (Eckstein, 2011; Kastner & Ungerleider, 2000). Such cues can be external, such as pointing arrows, or they may be internal and take the form of statistical information or expectations (Chun, Golomb, & Turk-Browne, 2011; Gottlieb, Oudeyer, Lopes, & Baranes, 2013). It is common to view location information and feature information as independent, each source of information utilized by a distinct, separate attentional process (Maunsell & Treue, 2006). In particular, feature attention is asserted to be location-independent (Cohen & Maunsell, 2011; Sàenz, Buracas, & Boynton, 2002, 2003; Yantis, 2000). It is further asserted that we can attend to only one particular feature of any particular visual dimension (e.g., blue/color) at a time and that our feature priority is applied across the whole visual field, and it can be conjoined in an additive fashion with spatial preferences in something like a Boolean "AND" operation (Hayden & Gallant, 2009; Huang, 2010; Huang & Pashler, 2007). One consequence of such a partitioning would be that we could not attend simultaneously and with spatial precision to orthogonal features that are both relevant for target identification (Andersen, Hillyard, & Müller, 2013).

Accepting this conclusion seems premature. First, the definition of a visual feature has been elided, and the operationalizing of distinct features is coarse (red or blue colors, vertical or horizontal orientations). A full evaluation of the independent nature of feature prioritization would benefit from experiments using more subtle feature gradations and tasks in which the graded distinctions may be pertinent. Second, most tasks demonstrating the independence of spatial and feature attentional effects require only that the target be detected and reported. Although this is an understandable laboratory simplification, it results in omitting many typical uses of visual information in natural environments. Third, accepting such a conclusion seems to mean accepting extreme limits on human performance that are not well matched to behavioral requirements. When one of my distant forebears tried to pluck a ripe berry from a bush or select a target for her spear, she was likely looking for features that exist on continua and are defined by precise conjunctions of shade, hue, orientation, size, motion, *and* location. It may be that only particular conjunctions of these elements are relevant and that the fitness of a response may be best quantified with a continuous measure and not collapsed to a binary classification (right/wrong or present/absent).

Citation: Anderson, B. (2014). Probability and the changing shape of response distributions for orientation. *Journal of Vision*, *14*(13):15, 1–11, http://www.journalofvision.org/content/14/13/15, doi:10.1167/14.13.15.

doi: 10.1167/14.13.15 Received April 15, 2014; published November 18, 2014 ISSN 1534-7362

Testing these ideas, I report data from three experiments. All three experiments used the same procedures. In each experiment, participants made orientation judgments. Across all locations, all orientations were equally probable, but there where particular conjunctions of location and orientation that were more likely. If one assumes spatial invariance for feature attention, no position or orientation could be favored over any other. The results, in contrast, reveal consistent improvements in orientation judgments for probable conjunctions of orientation and position; median absolute error is less for high-probability conjunctions. In trying to understand the nature of this improvement in more detail, I was led to the recognition that the shapes of the error distributions were different for the high- and low-probability conjunctions. This observation led me to the recognition that the differences could be well described as changes in kurtosis.

Kurtosis is related to variance and skew. Where variance is the second central moment and skew the third central moment, kurtosis is a function of the fourth moment. Where variance gives a sense of the spread of a distribution and skew its symmetry, kurtosis provides an estimate of the "shoulders" of a distribution (DeCarlo, 1997). Compared to the normal, or Gaussian, distribution, a leptokurtotic distribution (a relative increase in kurtosis) will tend to have more mass around the mode and more mass in the tails. A platykurtotic distribution (a lower kurtosis than the Gaussian) tends to be flatter with greater mass at the shoulders and less mass around the mode and tails. An interesting fact about the Gaussian distribution is that when normalized by the variance all Gaussians have exactly the same kurtosis. Gaussians can be broader or narrower, but relative to their variance, they cannot shift the probability mass to the tails or to the mode.

One connection between kurtosis and Gaussian distributions is the notion of a mixture. In a mixture distribution, one mixes simpler component distributions in some weighted combination (McLachlan & Peel, 2004). Although individual Gaussian distributions all have the same kurtosis, mixtures of Gaussians may have the same, less, or greater kurtosis. Models of attentional effects that propose shifts in the shape of tuning functions (a mixing of tuning functions with varying "standard deviations"; Reynolds, Pasternak, & Desimone, 2000) or differential weighting of tuning functions that either are maximal at a displayed orientation or at nearby orientations (a mixing of tuning functions with different "means"; Navalpakkam & Itti, 2005) are essentially theories of attention that propose different mixture distributions. We can adjudicate between them by comparing the theoretical consequences of each for changes in kurtosis and compare those predicted changes with the ones we experimentally observed.

Tuning function mixtures also appear in theoretical models of population coding, and here again there are corresponding changes in kurtosis. One particularly relevant model is that of Ringach (2010). Ringach derived an equation for capturing the discrimination performance of a population of tuning functions representing V1 neurons responding to orientation. His starting point was to consider the population tuning function as, essentially, a mixture of localized tuning functions. From this formulation, Ringach was able to theoretically derive expressions for an equation describing the discrimination function. His conclusion was that different tuning function shapes, shapes that vary in kurtosis, may advantage certain ranges of stimuli while achieving the same aggregate discrimination function. This finding has definite implications for "why" kurtosis may be a signature of error distributions for varying experimental conditions.

To summarize, the paper has several overlapping goals. First, the paper evaluates the claim that feature prioritization is spatially agnostic. Experimental results are presented showing that, contrary to this prediction, it is possible to prioritize conjunctions of features and locations. A second, and possibly greater, goal of this paper is to share the observation that the way this benefit is achieved seems to be due to changes in the shape of response distributions—changes that can be summarized by the kurtosis. Because kurtosis is not used commonly in psychology or neuroscience research, a third goal of the paper is to review the definition of kurtosis and show how kurtosis varies depending on the type of Gaussian distributions that are mixed together. Fourth, as a proof of the value of measuring kurtosis, we compare the implications of mixing Gaussians of different types with the presented experimental observations to compare the predictions of two popular accounts of attention at the neural level. Last, to show how such observations might be leveraged to move from descriptive to causal accounts, we combine the preceding with Ringach's (2010) theoretical model of population coding to suggest that kurtosis changes reflect a constraint whereby shape changes are used to favor important subdomains of perceptual space while keeping aggregate discrimination ability at its maximum.

Methods

The principal method was an adaptation of the procedures used by Anderson and Druker (2013).

Participants

Participants were recruited from the undergraduate Research Experience Group at the University of Waterloo and were compensated with course credit. All

experiments were approved by the University of Waterloo's Office of Research Ethics and adhered to the Declaration of Helsinki, and all participants signed an informed consent prior to beginning their research participation. Experiment 1 had 27 participants (two eliminated), Experiment 2 had 22 participants (two eliminated), and Experiment 3 had 21 participants (four eliminated). Reasons for elimination were either not understanding the instructions or failing to follow them. This was diagnosed based on participants pushing only one of the choice buttons for almost all trials or having a global performance indistinguishable from chance. In addition, only trials with response times (RTs) between 0.1 and 5 s were included in order to eliminate anticipation and distraction errors.

General methods

Participants sat approximately 60 cm from a 33 cm × 26.5 cm CRT that refreshed at 85 Hz. Responses were made with a computer keyboard. The experiments were programmed in Python using the PsychoPy library (Peirce, 2009). Participants were instructed to look at the center of the screen throughout the experiments. Eye movements were not recorded, and the stimuli were equally likely to be to the left and right of fixation (6°).

Experimental sessions consisted of 100 practice trials followed by blocks of 150 trials. For Experiment 1, there were five blocks. For Experiments 2 and 3, there were four blocks, but only the first two blocks were identical to Experiment 1. Those are the blocks that are included in the present analyses.

The CRT background was gray, and trials began with a centered white fixation cross (1.2°) that remained on through the presentation of the stimulus. The stimulus and the fixation spot had a simultaneous offset. Targets were "Gabors," sine-wave textures of 4 c/° visual angle along a single dimension filtered with a Gaussian mask (the standard deviation [sd] or space constant was 2/3 deg). The patches were randomly rotated. Targets remained on screen 83 ms. After a one-frame blank screen (12 ms), the response meter was presented at the same location as the target. This was a black line 4° in length and oriented vertically. Participants used a set of three keys to rotate the line clockwise, counterclockwise, and to "lock in" their response. There was no time pressure to make this response, and participants were encouraged to emphasize accuracy; they were told that it would be easier to respond when the stimulus was fresh in their mind.

For calculating response errors, the horizontal orientation was designated as 0° and the vertical direction as +90°. RT was calculated from when the response meter appeared to when the participants terminated their response.

For practice trials, there was visual feedback with a line indicating the correct orientation overlaid on their response. For test trials, participants were provided with auditory feedback after each test trial to encourage accuracy. Participants were not told the threshold for accuracy nor the direction of the error. A higher pitched "ding" sound (http://www.freesound.org/people/HardPCM/sounds/32950/) indicated they were below threshold (10°), or a negative lower pitched "donk" sound (http://www.freesound.org/people/tombola/sounds/49219/) indicated they were above the threshold. After the computer task, participants filled out a brief questionnaire.

Specific methods

The principal manipulation of all experiments was a manipulation of the probability of the Gabor being tilted to the right or left as a function of whether the stimulus appeared on the participant's right or left. For each participant, Gabor angles between 0° and 90° appeared 80% of the time if the Gabor was on the participant's right and were between 90° and 180° of tilt 20% of the time. The distribution was reversed (90° to 180°: 80%) when the target was on the left. This was counterbalanced across participants. More simply, this can be described as a right tilt on the right and a left tilt on the left (or a left tilt bias on the right and a right tilt bias for stimuli on the left) being 80% likely. The aggregate result of the manipulation was that for each individual participant stimuli were equally likely to be on their left and right, and all angles between 0° and 180° were equally likely to appear.

Statistical analyses

Statistical analyses used the R programming language (R Development Core Team, 2011). Kurtosis was calculated using the type II method of the kurtosis function in the E1071 package (Meyer, Dimitriadou, Hornik, Weingessel, & Leisch, 2012).

To measure bias in participants' responses, we used signed errors (range −90° to 90°). Signed errors were coded with respect to vertical such that errors that undershot were negative and overshooting errors were positive. For example, consider two trials in which the Gabor was tilted as if it were pointing to the two (or 10) of a clock face. Responses pointing toward one (or 11) would be negative signed errors, and responses pointing toward three (or nine) would be positively signed errors. For measuring the precision of the participants' responses, we used the absolute value of the error (measured in degrees of angular deviation; range 0° to

Data set (no. of subjects)	Mean signed error (degrees)	Median absolute error (degrees)	Median log RT
1 (25)	−3.14	8.48	0.15
2 (20)	−3.73	9.30	0.32
3 (17)	−5.52	10.94	0.32
F (2, 59)	1.0 ($p = 0.36$)	2.11 ($p = 0.13$)	4.45 ($p = 0.02$)

Table 1. Descriptive data for the three experimental data sets. *Notes*: The last line of the table presents the F values for separate analyses of variance conducted for each variable. Negative signed errors result from vertical biases. Accuracy measures are comparable across experiments.

90°; chance 45°). The absolute error was computed *after* subtracting the mean deviation. This was done separately for each participant and for each combination of presentation side and probability condition. This measure was selected because the unsigned error yields a measure that is less sensitive to outliers and is similar to the measure of precision used by Prinzmetal, Nwachuku, Bodanski, Blumenfeld, and Shimizu (1997) and Prinzmetal, Amiri, Allen, and Edwards (1998). In addition, the absolute error value is a superior measure for non-Gaussian data (Gorard, 2005). As a summary statistic for each participant's RTs, we used the median value of log-transformed RT data because the distribution of RT is right-skewed.

Results

Accuracy measures are consistent across experiments

Table 1 presents the descriptive data for all three experiments. The number of participants from each replication are shown as well as the results of an analysis of variance with "data set" as the only factor. The primary end point of accuracy was nonsignificantly different across experiments. Therefore, the data from the participants in all three experiments are analyzed as a single group. There is a difference for RT, possibly representing practice effects. The participants in Experiment 1 performed almost twice as many trials.

High-probability tilts are judged more quickly and accurately and with less vertical bias

Even though participants had not been instructed to speed their responses, high-probability targets were judged more quickly. Low-probability trials had a mean RT of 1.29 s. That was 34 ms slower than high-probability trials, $F(1, 61) = 13.21$, $p = 0.001$.[1] It is worth noting that this speed advantage for high-probability trials comes about despite a greater accuracy. Because the orientation of the response line began vertically and because there was a tendency to undershoot the correct orientation, a decrease in bias means a further movement of the response line (further details follow).

Trials with high-probability tilts were judged with less bias than low-probability tilts. Responses to low-probability tilts showed an average bias across participants of 4.46° and were tilted more vertically than they should have been. Trials of high-probability tilts were judged more accurately, with 1.61° less bias, $F(1, 61) = 12.48$, $p = 0.001$.

The absolute error magnitude for high-probability trials was smaller. The average error for low-probability tilts across participants was 9.55°, and when trials were of high-probability tilts this error was reduced by 0.5°, $F(1, 61) = 5.01$, $p = 0.03$. The different probability classes did not differ for the variance of the mean corrected angular difference, $F(1, 61) = 0.1$, $p = 0.76$.

On a first consideration, the significant results for absolute angular deviation and the lack of a significant result for variance might seem to be in conflict, but in fact, it merely reflects the fact that the error distributions for the two probability conditions differ in shape. The response distributions have different kurtoses. High-probability location–target conjunctions have error distributions that are both sharply peaked near zero *and* that have larger tails. As the variance gives added weight (by virtue of squaring the deviation) to outliers, this combination of high peaks, fat tails, and small shoulders leads to no appreciable change in variance despite more highly accurate responses. The error distributions for trials with high-probability tilts show a consistent increase in kurtosis compared to the trials with low-probability tilts (mean kurtosis high-probability tilts = 4.51; mean kurtosis low-probability tilts = 2.8), $F(1, 61) = 16.53$, $p = 0.0001$. This result has nominally the largest statistical effect and is generally consistent across participants (Figure 1).

Probability effects emerge early

In order to observe for effects of learning on accuracy, speed, and kurtosis, the data were reanalyzed after subdividing them into pseudoblocks of length 50 and using only the first two experimental blocks from each cohort.[2] In addition, since pseudoblocks of this

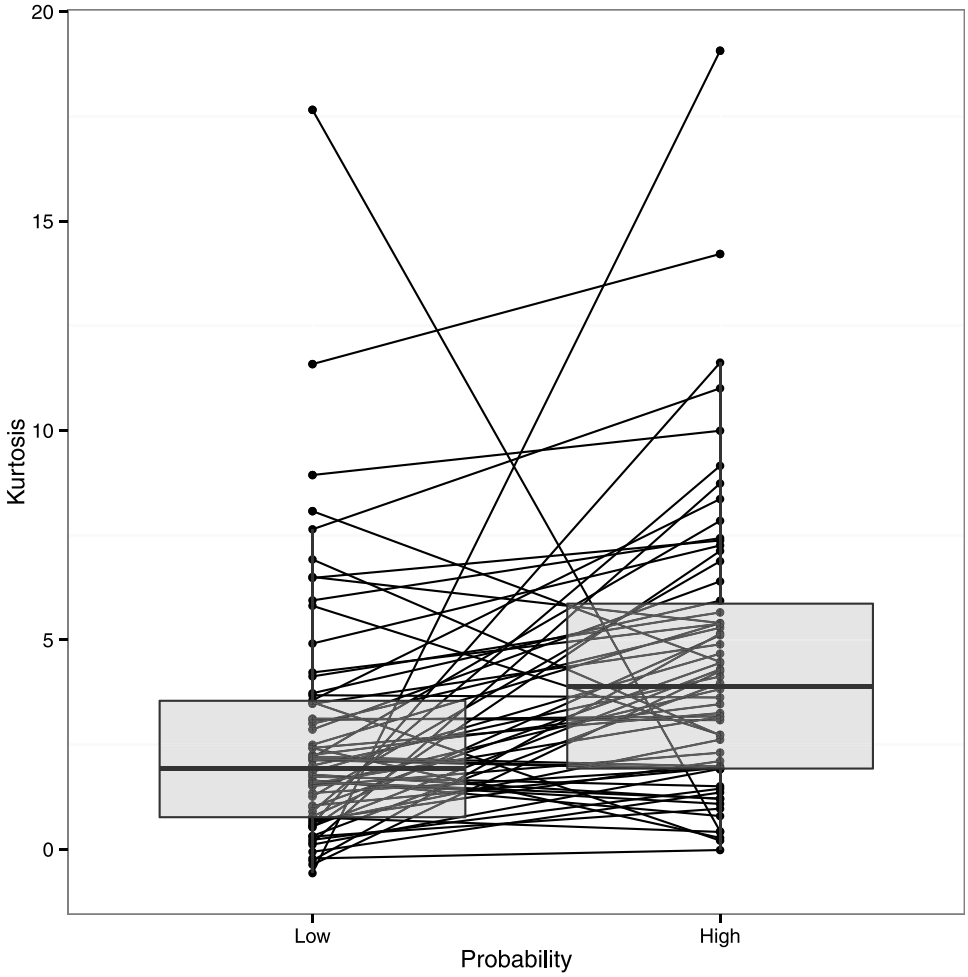

Figure 1. The excess kurtosis for each participant is plotted for each of their high-probability and low-probability conjunction trials. The values for each individual are connected by lines. The gray boxes show the median plus and minus one quartile. The kurtosis differences are highly significant by repeated measures ANOVA (see text). For 44 of the 62 participants, the kurtosis was higher for high-probability conjunctions ($p = 0.001$; binomial test).

size sometimes contained only a few trials for the low-probability conjunctions, I randomly sampled with replacement to create equal-sized samples of 20 for each of the high- and low-probability conditions.

None of the measures for accuracy, speed, or kurtosis showed an interaction between the amount of time spent doing the task and the effect of whether a trial was a high-probability conjunction. Accuracy and speed did improve with experience. For each of the comparisons, a repeated-measures ANOVA was computed with factors of pseudoblock number (1–6) and high (or low) probability conjunction. For the median log RT, there was a significant effect of pseudoblock, $F(5, 305) = 19.95$, $p < 0.0001$; probability of the conjunction, $F(1, 61) = 16.19$, $p = 0.0002$; and no interaction, $F(5, 305) = 1.4$, $p = 0.214$. For the accuracy measure of the median of the absolute angular deviation, the effects were significant for time on task, $F(5, 305) = 3.93$, $p = 0.002$, whether it was a high-probability conjunction, $F(1, 61) = 5.38$, $p = 0.024$, but no interaction, $F(5, 305) = 0.737$, $p = 0.6$. The kurtosis measure was only significant for the factor of probability, $F(1, 61) = 36.7$, $p < 0.0001$, but not for time on task, $F(5, 305) = 0.79$, $p = 0.558$, or the interaction, $F(5, 305) = 1.02$, $p = 0.41$.

Beneficial effects of high-probability conjunctions are implicit

After completing the task, all participants completed a short questionnaire with three open-ended questions. The questions asked the participants whether anything in the task stood out, whether they used any particular strategies, and whether any stimuli seemed different from the others. None of the participants reported that any stimuli were more common than any others or that stimuli appearing on the two sides were in any way different. Many participants reported feeling that they improved with experience, and many participants

reported that stimuli near the vertical and horizontal were different, but individual participants could find them either easier or harder. No participants made any mention of patterns formed by sequential stimuli across the sides such as a "v" or chevron. Some strategies reported were to look for patterns, to try and guess on which side the stimulus was to appear, or to keep strictly focused on the fixation cross.

Discussion

There are two principal findings of this research. The first is that high-probability conjunctions are prioritized and, second, that error distributions change their shape; when conjunctions are probable, kurtosis is larger.

Prioritizing probable conjunctions of features and positions

We are faster and more accurate in detecting targets when we are cued to particular locations. This is true even when the cues are implicit (Chun, 2000) or statistical (Druker & Anderson, 2010; Jiang, Swallow, & Rosenbaum, 2013). We are also better when cued to particular features (Maunsell & Treue, 2006) although in this case the claim is that the feature prioritization is invariant to spatial location. Although much of the evidence for this position comes from electrophysiological studies (e.g., Hayden & Gallant, 2005), it has also been demonstrated behaviorally (White & Carrasco, 2011). However, evidence showing that search times or neuron firing rates are altered by spatially distant cued features is not the same as showing that conjunctions of features and locations cannot be prioritized.

The data reported here confirm that high-probability conjunctions of location and tilt are judged more accurately. With the marginal probabilities of locations and orientations uniform, there is no information that would allow prioritizing any particular location or tilt.

One reason why these experiments found a conjunction effect, when earlier work has not, could be due to the specific character of the experimental task (Harel, Kravitz, & Baker, 2014). In the current task, the target was relatively large, was easily visible, was present in every trial, and was presented without distractors, and its general character was known in advance. Detection was not the challenge; rather, it was the discrimination of the target that was the challenge. In contrast, visual search tasks, which form the method of many prior studies, often make it hard to find a target but easy to classify it. Whether a task is a challenging detection task or a discrimination task may have an important effect on the collaboration demonstrated by attentional subsystems.

In addition to the basic demonstration that accuracy was improved for high-probability conjunctions, the data also revealed that the basis for this was at least partly due to changes in the shape of the error distributions—a change that can be quantified by kurtosis. In the next section, I discuss the implications of this observation for these particular data and for research on perception more generally.

Kurtosis: What is it, what does it mean, and why should we care?

Common approaches to data analysis focus on summary statistics that emphasize the common, modal response, such as the median absolute accuracy statistic used here. In order to determine the significance of such numbers, we determine how expected they are by using hypothesis tests or confidence intervals. In doing so, we frequently make the assumption that our errors and our data are normally distributed. We are generally aware that when our data depart from normality this affects our statistical analysis, but less often do we see these departures as useful information in and of themselves. When analyzing these data, I tripped over the changes in kurtosis. In trying to understand why kurtosis might change, I found that the changes in kurtosis were another useful source of evidence for and against various theoretical accounts of attentional effects. The general conclusion is that we should not assume normal distributions, and when we find non-normally distributed data, we should not view it merely as an inconvenience to be transformed away. A second conclusion is that experimental measurements that do not force responses into two or a few categories can yield a richer set of data for identifying such departures. The important specific conclusion is that kurtosis changes in conditions in which stimulus probability changes. Because kurtosis is a well-defined mathematical entity, we can use kurtosis to make strong claims about what processes can or cannot give rise to the changes seen in our data. The organization for the rest of the discussion will be as follows: Because kurtosis has not been used frequently in analyzing psychological data, I will first give its definition. Next, to support the claim that kurtosis can be used as evidence for comparing theoretical claims, I produce a simulated example. I then show how kurtosis can be used to provide even stronger evidence by an analytical demonstration that two common theories of attentional effects lead to necessarily opposite and incompatible effects on kurtosis. Last, I show how we can use kurtosis to link to an earlier theory of population

coding. This leads to the conclusion that changes in kurtosis may be the only basis for prioritizing some stimuli at the expense of others, all the while maintaining the same overall high level of performance.

Kurtosis defined

Kurtosis is a function of the fourth moment of a probability distribution. The first moment of a distribution is the mean, and the second central moment is the variance. The n^{th} moment for discrete data is computed by weighting the n^{th} power of each data value by its probability, $\mathrm{E}(x^n) = \sum_i x_i^n p(x_i)$. The central moments are computed similarly except that the n^{th} power of each data point has the mean subtracted [$(x_i - \mu)$ replaces x in the above formula].

For normal, also called Gaussian, distributions, the higher-order central moments are either zero (the odd powers) or invariant under a suitable standardization. The fourth standardized central moment provides a measure of how the mass of probability beneath the shoulders of a distribution relates to the central peak and the tails (DeCarlo, 1997). There is more than one formula for quantifying kurtosis, but a common one is to standardize this fourth central moment by the square of the variance and to subtract three. Equation 1 shows this form of the equation for excess kurtosis (Frühwirth-Schnatter, 2006).

By this measure, all normal distributions will have *excess kurtosis* 0, and other distributions will be characterized by either a relatively greater or lesser kurtosis. The distribution with the smallest excess kurtosis value is the binomial distribution. When $p = 0.5$, all the mass is at the shoulders, and excess kurtosis is -2. There is no bound on how large excess kurtosis can grow. Increasing the probability mass in the tails *and* near the mean both increase kurtosis and make it possible to have distributions with identical variances but very different kurtoses.

Kurtosis: Using it to compare attentional theories?

What are the implications of a change in kurtosis? First, it indicates that the data are not normally distributed; thus it is not enough to report only the mean and variance. Higher-order moments should also be evaluated, not simply because they provide a more complete quantitative description of non-normal distributions, but because the natures of changes in these higher-order moments may provide useful constraints on the nature of the generative processes. For example, the present results suggest gain enhancement (Boynton, 2005) rather than off-center tuning (Navalpakkam & Itti, 2005; Scolari, Byers, & Serences, 2012) as a basis for the attentional modulation with probability as tested in these experiments.

Figure 2 demonstrates this logic with a simple simulation. On the left, two normal distributions with the same mean but different variances represent the gain-enhancement model of attention. On the upper right, two normal distributions with the same variance bracket the target orientation to simulate off-center tuning. Simulating what happens to the kurtosis when we mix equal proportions of observations from each of the two distributions for each of the two different models are the empirical distributions shown in the lower panels. Overlaid are the normal curves that have the same mean and variance as each of the two empirical distributions. Note that, by design, the normal distributions described by the mean and variance are very similar for both models even though the actual empirical data is very different. Note further that on the left we have a distribution of observations that is highly peaked whereas on the right the shoulders of the mixture poke up above that of the normal distribution with the same mean and variance. The empirical distribution on the left is leptokurtotic and on the right platykurtotic. The shape of the distribution of the simulated data on the left is quantitatively similar to what was seen in the participants' error distributions. These simple simulations suggest that the gain-enhancement model provides a better account of the participant data. For some simple cases, this impression can be proved mathematically.

$$\text{excess kurtosis} = \frac{\sum_{k=1}^{K} p_k((\mu_k - \mu)^4 + 6(\mu_k - \mu)^2\sigma_k^2 + 3\sigma_k^4)}{\left(\sum_{k=1}^{K}(\mu_k^2 + \sigma_k^2)p_k - \mu^2\right)2} - 3 \quad (1)$$

If all components of a mixture have the same mean, as is assumed in the version of the gain-enhancement model depicted on the left side of Figure 2, then the overall mean of the mixture will be equal to the component means, and the only terms of the numerator of Equation 1 we need to concern ourselves with are those involving variance. If we further assume that the mean is zero, which merely involves translating the units of our scale and not changing the distribution itself, the denominator also simplifies, and proving that kurtosis increases for such a mixture is demonstrated if we can show that $\sum_i p_i\sigma_i^4/(\sum_i p_i\sigma_i^2)^2 \geq 1$ or, equivalently, that the numerator is greater than the denominator. This can be demonstrated by starting with the formula for the squared deviation of the individual variance components from the overall variance as follows and simply relies on some rearrangements and

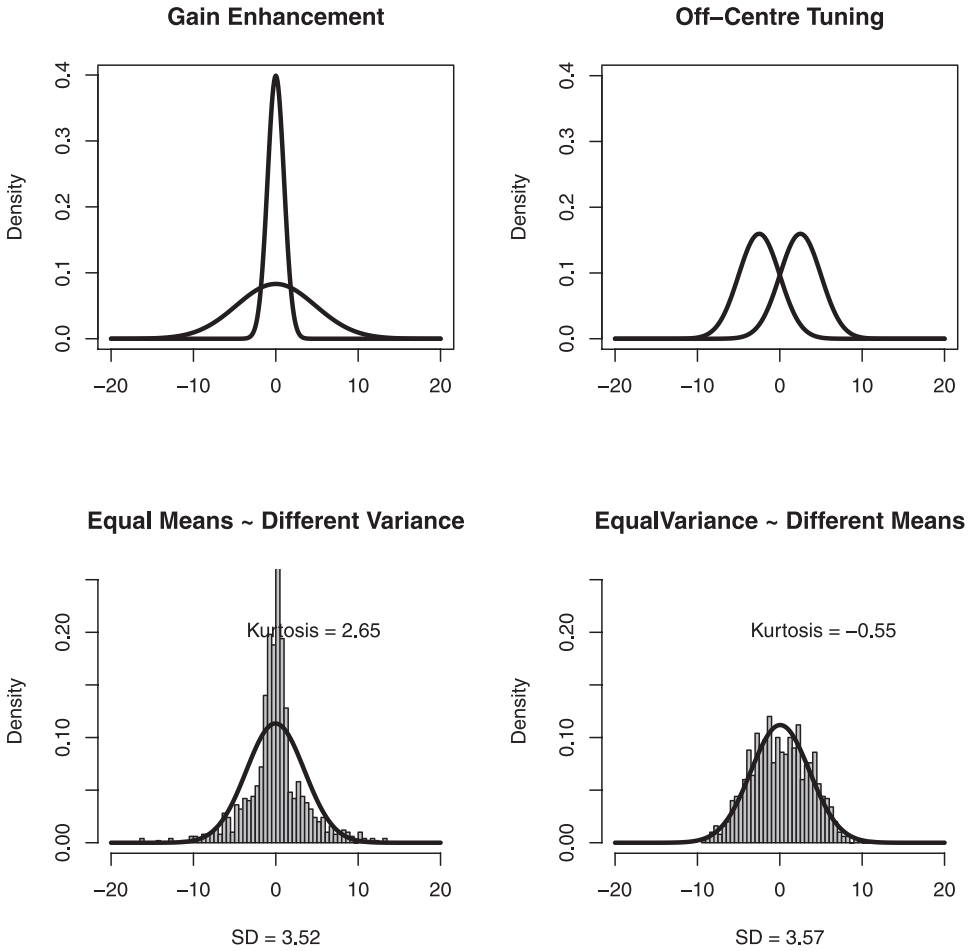

Figure 2. Two mixture models. This figure shows how different combinations of normal distributions can yield mixture distributions with either increased or decreased kurtosis. The upper panels show two different combinations of two normal distributions that are meant as idealizations of two common accounts of how attention might operate at the level of neuronal-tuning functions. The normal distributions are idealizations of neuronal-tuning functions. On the left, the two normal distributions have the same mean but have different variances. This is meant to capture the idea that attention works by sharpening or broadening the response functions of subsets of the responsive neuronal population. In the right upper panel are shown two normal distributions that have the same variance but have slightly different means. This is meant to capture the idea that when making difficult discriminations attention has the effect of giving the most weight to neuronal populations that are tuned slightly off from target orientation. In the lower panels are shown the mixture distributions that result from mixing these two pairs of distributions with equal probability. The mixture distributions have similar means and variances but very different kurtoses. The black curves in the bottom panels show the normal distributions that have the same mean and variance as the mixture distributions. Because normal distributions are determined by their mean and variance, the two normal fits in the bottom panels are nearly identical even though the actual mixture distributions, as delineated by the histograms of random samples, show substantial departures from normality and a distinct shape for each of the two combinations.

use of the formula for computing the variance of a zero-mean heteroscedastic mixture:

$$\sum_i p_i(\sigma_i^2 - \sigma^2)^2 \geq 0$$

$$\sum_i p_i((\sigma_i^2)^2 - 2\sigma_i^2\sigma^2 + \sigma^2)$$

$$\sum_i p_i(\sigma_i^2)^2 - 2\sigma^2\sum_i p_i\sigma_i^2 + (\sigma^2)^2\sum_i p_i$$

$$\sum_i p_i\sigma_i^4 - (\sigma^2)^2 \Longrightarrow \sum_i p_i\sigma_i^4 \geq \left(\sum_i p_i\sigma_i^2\right)^2$$

Whenever we mix normal distributions of distinct variances, even if we mix distributions with identical means, the observed distribution will *not* be normally distributed, and it will have a greater kurtosis than a normal distribution of identical mean and variance (McLachlan & Peel, 2004).

For the alternative case of mixing two normal distributions with similar variances but different means, we can begin with Equation 2 derived by Preston (1953), in which a is the ratio of the larger mixing coefficient to the smaller (a has a minimum value of one that occurs when the two components are weighted equally), and Δ is the distance between the means of the two distributions in standard deviation units.

$$\frac{a(a^2 - 4a + 1)\Delta^4}{(a\Delta^2 + (a+1)^2)^2} \quad (2)$$

Note that whether we see a negative excess kurtosis depends entirely on the second term in the numerator. For the special case in which the components are mixed equally, this simplifies to -2. It remains negative unless one term is weighted $2 + \sqrt{3}$ times the other. The separation between the distributions scales but does not change the sign. We can conclude that all reasonably balanced mixtures of two normal distributions will have a negative excess kurtosis; the magnitude of which depends on the separation of their means.

Kurtosis changes: Implications for neural mechanisms?

An additional implication of the observation of a change in kurtosis for high-probability conjunctions is that it provides a connection between "attention" and the theoretical results on probability coding reported by Ringach (2010). Ringach derived the form for a population-tuning function for orientation discrimination. To quantify the performance of two populations, Ringach used the information curve (Seung & Sompolinsky, 1993). Ringach demonstrated that discrimination performance was proportional to d'^2, and he represented the discrimination performance as a Fourier expansion. This transformation revealed that it was only the amplitude components and not the phase components that affected the overall discrimination performance. Thus, a number of discrimination curves could achieve, in the aggregate, similar discrimination performance despite having different shapes. As demonstrated in Ringach's figure 3b and 3c, these shape changes are the sort that alter kurtosis (shifting tail and center mass toward or away from the shoulders). From this observation, we can revisit a central question of attentional research: What is attention for?

It is popular to discuss attention as providing improved processing of sensory signals. However, this raises its own question: If we have a capacity for improved perceptual performance, why don't we use it all the time? One conclusion that can be inferred from Ringach's (2010) work, combined with the empirical results reported here, is that it well might be that we *are* always working at our optimum, at least when evaluated by a global metric such as the area under the information discrimination curve. Our global performance may be as good as it can get. The freedom we have is not to make our discrimination performance "better," per se, but to make it better in respect to the sorts of judgments we have to make now. The freedom we may have is to prioritize certain subsets of the problem space (at the expense of others). The subsets deserving prioritization are those that are common or important.

Considering the importance of stimuli may be the way to resolve a paradox that may occur with increased kurtosis. When kurtosis increases, it may not only pack more observations closer to the mean, but it may also pack more mass into the tails. This means more highly accurate judgments occur as companion to more really bad judgments. How can this trade-off be understood?

It does make sense if we are trying to minimize the cost of errors rather than simply error magnitude. Is this a plausible conjecture? Imagine the penalty for driving one's car into a parking garage with a narrow entrance. The "penalty" for missing the entrance by 1 foot or 2 feet is essentially the same: major body damage. And the benefit for shifting some of the 1-cm errors to no error is substantial: Instead of a long scrape down the side of one's car, there is no damage at all. This prioritization for reducing small errors, with less concern for a compensatory increase in gross errors, can be conjectured as a general principle of our natural environment. When shooting at prey, a miss by 1 foot or 2 feet makes no difference, but even a slight decrease in the number of near misses could mean a great decrease in the number of empty stomachs.

These are speculations, but they provide a consistent account of how increasing kurtosis for judgments about high-probability events could be optimal even if it resulted in an increase in the proportion of gross errors. However, the novelty of the presented results does not depend on whether this idea is persuasive. The present results contradict the claim that we can only attend to one stimulus feature per dimension at a time by showing that high-probability conjunctions of features and locations are judged more accurately even when the aggregate distribution of features and locations is uniform and the features favored at each location are orthogonal. Further, these data demonstrate that error distributions are non-normal. This is easier to see when tasks require continuous reports rather than limiting themselves to binary classifications. From such data, one can examine the shape of error distributions. From these higher-order moments, one can make additional assessments of cognitive and neural theories. An important manipulation in the present experiments was probabilistic. It may be that manipulations of low-level salience that do not depend

on expectation or probability may have different effects. Also, this task required that participants use the information to make a difficult discriminative judgment whereas prior procedures have emphasized challenging searches and easy classifications. The purpose for which perceptual data are to be used may also have important effects on how it is processed. Lastly, the *costs* of misperceptions and motor errors may combine to sculpt perceptual experience.

Keywords: attention, probability, kurtosis, vision, feature attention, conjunctions

Acknowledgments

This research was supported by an NSERC Discovery Grant.

I would like to thank the three students who helped with data collection: Michael Druker, Tracy Dow, and Tim Moy.

Commercial relationships: none.
Corresponding author: Britt Anderson.
E-mail: britt@uwaterloo.ca.
Address: Dept. of Psychology and Centre for Theoretical Neuroscience, University of Waterloo, Waterloo, ON, Canada.

Footnotes

[1] Statistics were conducted on the median log RTs, but the inverse transformed values are reported because they are more familiar and more easily interpreted.

[2] The only variable to differ between experiments was the RT measure (see text above). It was significantly shorter for the experiment with five experimental blocks as participants got quicker over time. This was the justification for limiting to the data for the first two blocks of each repetition.

References

Andersen, S. K., Hillyard, S. A., & Müller, M. M. (2013). Global facilitation of attended features is obligatory and restricts divided attention. *Journal of Neuroscience, 33,* 18200–18207.

Anderson, B., & Druker, M. (2013). Attention improves perceptual quality. *Psychonomic Bulletin & Review, 20,* 120–127.

Boynton, G. M. (2005). Attention and visual perception. *Current Opinion in Neurobiology, 15,* 465–469.

Chun, M. M. (2000). Contextual cueing of visual attention. *Trends in Cognitive Sciences, 4*(5), 170–178.

Chun, M. M., Golomb, J. D., & Turk-Browne, N. B. (2011). A taxonomy of external and internal attention. *Annual Review of Psychology, 62,* 73–101.

Cohen, M., & Maunsell, J. (2011). Using neuronal populations to study the mechanisms underlying spatial and feature attention. *Neuron, 70,* 1192–1204.

DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological Methods, 2,* 292–307.

Druker, M., & Anderson, B. (2010). Spatial probability aids visual stimulus discrimination. *Frontiers in Human Neuroscience, 4,* 1–10.

Eckstein, M. P. (2011). Visual search: A retrospective. *Journal of Vision*, *11*(5):14, 1–36, http://www.journalofvision.org/content/11/5/14, doi:10.1167/11.5.14. [PubMed] [Article]

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. New York: Springer.

Gorard, S. (2005). Revisiting a 90-year-old debate: The advantages of the mean deviation. *British Journal of Educational Studies, 53,* 417–430.

Gottlieb, J., Oudeyer, P. Y., Lopes, M., & Baranes, A. (2013). Information-seeking, curiosity, and attention: Computational and neural mechanisms. *Trends in Cognitive Sciences, 17,* 585–593.

Harel, A., Kravitz, D. J., & Baker, C. I. (2014). Task context impacts visual object processing differentially across the cortex. *Proceedings of the National Academy of Sciences, USA, 111,* E962–E971.

Hayden, B. Y., & Gallant, J. L. (2005). Time course of attention reveals different mechanisms for spatial and feature-based attention in area V4. *Neuron, 47*(5), 637–643.

Hayden, B. Y., & Gallant, J. L. (2009). Combined effects of spatial and feature-based attention on responses of V4 neurons. *Vision Research, 49,* 1182–1187.

Huang, L. (2010). The speed of feature-based attention: Attentional advantage is slow, but selection is fast. *Journal of Experimental Psychology: Human Perception and Performance, 36,* 1382–1390.

Huang, L., & Pashler, H. (2007). A Boolean map theory of visual attention. *Psychological Review, 114,* 599.

Jiang, Y. V., Swallow, K. M., & Rosenbaum, G. M. (2013). Guidance of spatial attention by incidental learning and endogenous cuing. *Journal of Experimental Psychology: Human Perception and Performance, 39*(1), 285–297.

Kastner, S., & Ungerleider, L. (2000). Mechanisms of visual attention in the human cortex. *Annual Review of Neuroscience, 23*(1), 315–341.

Maunsell, J. H., & Treue, S. (2006). Feature-based attention in visual cortex. *Trends in Neurosciences, 29,* 317–322.

McLachlan, G., & Peel, D. (2004). *Finite mixture models*. New York: John Wiley & Sons.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2012). e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.6-1. http://CRAN.R-project.org/package=e1071.

Navalpakkam, V., & Itti, L. (2005). Modeling the influence of task on attention. *Vision Research, 45,* 205–231.

Peirce, J. W. (2009). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics, 2,* 1–8.

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

Preston, E. J. (1953). A graphical method for the analysis of statistical distributions into two normal components. *Biometrika, 40,* 460–464.

Prinzmetal, W., Amiri, H., Allen, K., & Edwards, T. (1998). Phenomenology of attention: 1. Color, location, orientation, and spatial frequency. *Perception, 24,* 261–282.

Prinzmetal, W., Nwachuku, I., Bodanski, L., Blumenfeld, L., & Shimizu, N. (1997). The phenomenology of attention. 2. Brightness and contrast. *Consciousness and Cognition, 6,* 372–412.

R Development Core Team. (2011). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Reynolds, J. H., Pasternak, T., & Desimone, R. (2000). Attention increases sensitivity of V4 neurons. *Neuron, 26,* 703–714.

Ringach, D. L. (2010). Population coding under normalization. *Vision Research, 50,* 2223–2232.

Sàenz, M., Buracas, G. T., & Boynton, G. M. (2002). Global effects of feature-based attention in human visual cortex. *Nature Neuroscience, 5,* 631–632.

Sàenz, M., Buraĉas, G. T., & Boynton, G. M. (2003). Global feature-based attention for motion and color. *Vision Research, 43,* 629–637.

Scolari, M., Byers, A., & Serences, J. T. (2012). Optimal deployment of attentional gain during fine discriminations. *Journal of Neuroscience, 32,* 7723–7733.

Seung, H. S., & Sompolinsky, H. (1993). Simple models for reading neuronal population codes. *Proceedings of the National Academy of Sciences, USA, 90,* 10749–10753.

White, A. L., & Carrasco, M. (2011). Feature-based attention involuntarily and simultaneously improves visual performance across locations. *Journal of Vision, 11*(6):15, 1–10, http://www.journalofvision.org/content/11/6/15, doi:10.1167/11.6.15. [PubMed] [Article]

Yantis, S. (2000). Goal-directed and stimulus-driven determinants of attentional control. In S. Monsell & J. Driver (Eds.), *Attention and performance XVIII*, Vol. 18 (pp. 73–103). Cambridge, MA: MIT Press.