

# The nature of correlation perception in scatterplots

Ronald A. Rensink<sup>1</sup>

Published online: 26 October 2016

© The Author(s) 2016. This article is published with open access at Springerlink.com

**Abstract** For scatterplots with gaussian distributions of dots, the perception of Pearson correlation  $r$  can be described by two simple laws: a linear one for discrimination, and a logarithmic one for perceived magnitude (Rensink & Baldrige, 2010). The underlying perceptual mechanisms, however, remain poorly understood. To cast light on these, four different distributions of datapoints were examined. The first had 100 points with equal variance in both dimensions. Consistent with earlier results, just noticeable difference (JND) was a linear function of the distance away from  $r = 1$ , and the magnitude of perceived correlation a logarithmic function of this quantity. In addition, these laws were linked, with the intercept of the JND line being the inverse of the bias in perceived magnitude. Three other conditions were also examined: a dot cloud with 25 points, a horizontal compression of the cloud, and a cloud with a uniform distribution of dots. Performance was found to be similar in all conditions. The generality and form of these laws suggest that what underlies correlation perception is not a geometric structure such as the shape of the dot cloud, but the shape of the *probability distribution* of the dots, likely inferred via a form of ensemble coding. It is suggested that this reflects the ability of observers to perceive the *information entropy* in an image, with this quantity used as a proxy for Pearson correlation.

**Keywords** Perceptual organization · Visual perception · Perceptual categorization and identification · Information visualization

---

✉ Ronald A. Rensink  
rensink@psych.ubc.ca

<sup>1</sup> Departments of Computer Science and Psychology, University of British Columbia, Vancouver, BC, Canada

## Introduction

The analysis of data is important in many aspects of life. An important part of such analysis is the use of graphical representations, which can be highly effective when datasets are large, messy, and complex (see e.g., Card, Mackinlay, & Shneiderman, 1999; Thomas & Cook, 2005). If a graphical representation is designed well, analysis can be rapid, accurate, and precise; in such situations the visual system of the analyst perceives structure in a dataset in much the same way as it perceives structure in the physical world. The perception of such graphical representations therefore has considerable potential to help us investigate various aspects of our visual intelligence (Rensink, 2014; see also Cleveland & McGill, 1987; Meyer, Taieb, & Flascher, 1997).

It has been argued (Rensink & Baldrige, 2010) that a good test-bed for this approach is the estimation of Pearson correlation  $r$  in scatterplots. In part, this is because much of the estimation of  $r$  appears to be a perceptual process, one to which existing techniques of vision science can be readily applied (e.g., Doherty, Anderson, Angott, & Klopfer, 2007; Meyer & Shinar, 1992; Meyer et al., 1997). Another reason is that this domain is simple enough to explore systematically, while still being rich enough to raise interesting questions about the mechanisms involved.

Historically, the perception of correlation has been investigated in several ways (for reviews, see Boynton, 2000; Doherty et al., 2007). Most were based on *numerical estimation*—asking observers for a number that describes the magnitude of the correlation perceived. Results showed that perceived correlation  $g(r)$  tends to underestimate physical correlation  $r$  (especially at intermediate levels), with little correlation perceived when  $|r| < 0.2$  (Bobko & Karren, 1979; Boynton, 2000; Cleveland, Diaconis, & McGill, 1982; Strahan & Hansen, 1978). They also showed that much of this

process is carried out rapidly, with results largely independent of the statistical expertise of the observer (Lane, Anderson, & Kellam, 1985; Meyer & Shinar, 1992; Meyer et al., 1997; Strahan & Hansen, 1978); indeed, particular neural systems appear to be involved (Best, Hunter, & Stewart, 2006). Thus, although knowledge and expertise can influence the more sophisticated aspects of this process (Freedman & Smith, 1996; Lewandowsky & Spence, 1989), there nevertheless seems to exist a distinct *basic stage* of correlation perception—a rapidly-acting initial phase that can be considered purely perceptual, with similar characteristics for most observers.

However, although these studies were important, they had limitations. First, they paid relatively little attention to the *precision* of the process—the extent to which the same estimate results when the same stimulus is presented. Next, the *central assumption* of the estimation techniques—that numbers can be assigned to perceived magnitudes in a consistent way—is problematic, leading to the possibility of unstable or context-dependent estimates (Ellermeier & Faulhammer, 2000) (this may help explain why inconsistent results are sometimes encountered—see e.g., Doherty et al., 2007.) Third, there was little investigation of *systematicity*—the extent to which connections exist between precision and accuracy, not to mention with the rest of visual perception. Finally, little distinction was usually made between the perception of *population* properties (e.g., the correlation of a set of scatterplots) and *particular* ones (e.g., the distance of an outlier from the dot cloud of an individual scatterplot).

Rensink and Baldrige (2010) developed an approach that took many of these considerations into account. First, precision was measured via the just noticeable difference (JND—also referred to as the *difference threshold*), the difference needed to correctly discriminate two correlations 75 % of the time. Second, accuracy was measured via *bisection*: adjusting a test plot to have its perceived correlation be half-way between those of two references. Because the visual system is more concerned with relative than absolute quantities, bisection estimates are potentially more reliable and less affected by context (Tommasi, 2000; Zimmer & Ellermeier, 2006). Next, results were analyzed for possible relationships between discrimination and perceived magnitude. Finally, to ensure that the properties of populations were being perceived, dozens of scatterplots were shown for each measurement, with population parameters held constant.

Applying this approach to scatterplots with gaussian distributions, Rensink and Baldrige (2010) found that discrimination could be described by

$$\text{JND}(r) = k \left( 1 / b_{\text{disc}} - r \right) \quad (1)$$

where the *variability parameter*  $k$  and *bias* (or *offset*)  $b_{\text{disc}}$  are such that  $0 < b_{\text{disc}}, k < 1$ . Multiplying Eq. (1) by  $b_{\text{disc}}$  and

letting  $u = 1 - b_{\text{disc}} r$ , this becomes  $\text{JND}(u) = ku$ , an instance of *Weber's Law*, which characterizes discriminability for many simple properties, such as length and brightness (Billock & Tsou, 2011; Ross, 1997). Meanwhile, the estimation of perceived correlation  $g(r)$  could be described by

$$g(r) = \ln(1 - b_{\text{est}} r) / \ln(1 - b_{\text{est}}) \quad (2)$$

where  $b_{\text{est}}$  is a bias parameter describing the degree of underestimation encountered. Again letting  $u = 1 - b_{\text{est}} r$ , this becomes  $g = \ln(u) / (\ln(1 - b_{\text{est}}))$ . This is an instance of *Fechner's Law*, which has been proposed for the relation between the perceived and physical magnitudes of various properties (see Ross, 1997). In addition, Rensink and Baldrige (2010) found that  $b_{\text{disc}} = b_{\text{est}}$ , connecting discrimination and estimation in a systematic way.

Although these results led to a better understanding of behavior, they did not lead to a better understanding of the mechanisms responsible. To cast light on these, this study examined the generality of these results for different distributions of data points. Rensink and Baldrige (2010) used dot clouds with 100 points, with gaussian distributions of equal variance in both dimensions. To determine if any of these factors affect performance, four conditions were examined. The first replicated Rensink and Baldrige (2010): scatterplots had 100 points, with gaussian distributions of equal variance in both dimensions. A second condition tested the effects of dot cloud density, using only 25 dots. A third tested sensitivity to the shape of the dot cloud: this once again had 100 points, but was now compressed horizontally by a factor of 2. Finally, the fourth condition was similar to the first, but with a uniform instead of a gaussian distribution.

Results show that the laws found in Rensink and Baldrige (2010) are much the same for all these distributions. To account for the shape of these laws, it is suggested that correlation perception is based on the width of the probability distribution inferred from the points in the dot cloud. And to explain why perception of this structure might be useful, it is suggested that it reflects the perception of the information entropy in an image. Among other things, these results lead to a straightforward way to evaluate the effectiveness of a scatterplot design, as well as several predictions about the perception of correlation under various conditions. And at the most general level, they support the proposal that the study of the graphical representations used in information visualization can provide considerable insight into various aspects of the human visual system.

## General methods

The experimental design here was similar to that of Rensink and Baldrige (2010). Each observer was shown a set of

scatterplots containing data from a set of pseudo-random numbers with a fixed mean and standard deviation in each dimension. For discrimination, observers were asked to determine which of two side-by-side scatterplots was more correlated; for perceived magnitude, a test plot was adjusted until its perceived correlation was halfway between those of two reference plots. All observers carried out both tasks, the order of which was counterbalanced. Observers were told that accuracy was important, and that they could take as much time as needed. To familiarize each observer with the discrimination task, a set of eight practice trials—easy versions of the main task—were given (with feedback) beforehand, each set continuing until the observer reached 75 % accuracy, or 32 trials had been run. For the bisection task, observers were given seven practice trials; owing to the nature of the task, no feedback could be given for these.

### Observers

Each condition had 20 observers. All were undergraduates at the University of British Columbia, and were paid \$10 for a single one-hour session. All had normal or corrected-to-normal vision. Although not a requirement, all had at least some experience with scatterplots. Observers were replaced if their results (either  $k$  or  $b$ ) were more than 2.5 standard deviations beyond the average of the others; based on this criterion, 2–3 observers were replaced in each experiment.<sup>1</sup>

### Stimuli

In all tests, observers were seated 45 cm from a screen  $32^\circ \times 22^\circ$  in extent. Vertical and horizontal axes of each scatterplot extended  $6.3^\circ$ ; no tick marks or labels were used. For all conditions (except those of Experiment 3), dot clouds extended  $6.3^\circ \times 6.3^\circ$  and were centered on the midpoints of the axes; standard deviation was 0.2 of the extent of the cloud. And in all conditions (except those of Experiment 2), they contained 100 dots, each with a diameter of 4 min of arc ( $0.067^\circ$ ), ten times the visibility limit of 0.4 min of arc (Li, van Wijk, & Martens 2009).

For gaussian distributions, the  $x$ -coordinates of each dot were selected first, with this set scaled to match the given mean and standard deviation. A set of  $y$ -values was similarly

created. Each point  $(x, y)$  was then transformed to yield the correlated pair  $(x, y')$  via

$$y' = \frac{\lambda x + (1-\lambda)y}{\sqrt{\lambda^2 + (1-\lambda)^2}}, \text{ where } \lambda = \frac{r^2 - \sqrt{r^2 - r^4}}{2r^2 - 1}, \quad (3)$$

where  $r$  is the target correlation. To prevent values from exceeding the range of the scatterplot, any point greater than 2.5 standard deviations from the mean was eliminated, and a new point generated to take its place. Points were adjusted so that correlation was within 0.005 of the target. The resulting set was then rescaled again to have the designated mean and standard deviation.

### Procedure – discrimination

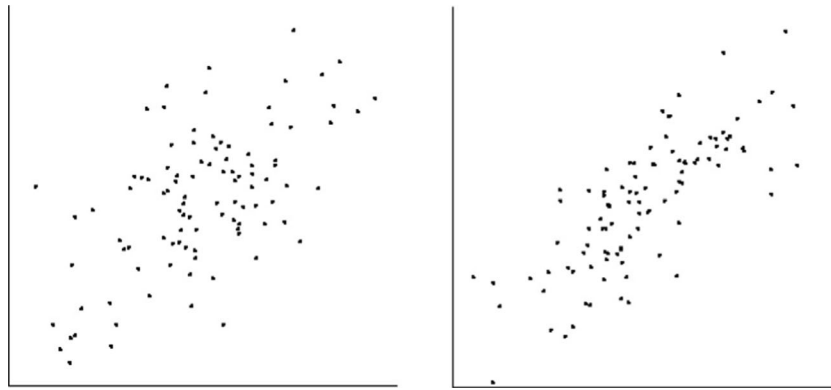
An important aspect of performance is *precision*—the scatter in the estimates made by an observer repeatedly given the same data. Following Rensink and Baldrige (2010), this was assessed via the just noticeable difference (JND), the value of  $\Delta$  for which scatterplots containing correlations  $r$  and  $r \pm \Delta$  can on average be discriminated 75 % of the time. Precision and JND are directly related: the greater the JND, the greater the separation needed to see that two scatterplots have different correlations, and consequently, the worse the precision of the perceptual estimates.

The procedure to assess JND was based on that of Rensink and Baldrige (2010). The set of correlations tested (the *base correlations*) ranged from 0.0 to 0.9, in increments of 0.1. For each, JND was obtained via a series (or *run*) of trials. In each trial, observers were shown two side-by-side scatterplots—one with the base correlation and the other a test correlation above or below it—and asked to select the more highly correlated one (Fig. 1). For each run, the initial correlation of the test plot was 0.1 above the base (when testing from above) or 0.1 below (when testing from below).<sup>2</sup> If the answer was correct, the size of the difference was decreased by 0.01, making the task more difficult; if incorrect, it was increased by 0.03, making the task easier. (The three steps for each correct answer matched a single step for each incorrect answer, resulting in a steady-state accuracy of 75 %.) Following Rensink and Baldrige (2010), performance was measured via a moving window of 24 consecutive trials, divided into three sub-windows of eight trials each. After an initial 24 trials, the ratio of the variance between sub-windows to the average variance within the sub-windows (somewhat akin to an F-test)<sup>3</sup> was continually calculated. Testing halted when

<sup>1</sup> All but one of these had a near-zero value of  $b$  in the bisection task; these observers, however, had the same value of  $b$  as others in the discrimination task. This suggests that these observers did not use the same quantity for discrimination and magnitude estimation (or use the same quantity for magnitude estimation as the other observers). It is unclear what this property might have been. However, a trimming of 10–20 % of the observers has long been done in studies of correlation perception in scatterplots (e.g., Cleveland et al., 1982), supporting the possibility that alternate strategies are used by subsets of the general populace.

<sup>2</sup> A within-observer pilot study compared four initial differences (0.0, 0.1, 0.2, and 0.3), and found no significant effect on JND. The value of 0.1 was chosen because it tended to require a minimum number of trials on average.

<sup>3</sup> Variances were calculated using a denominator of 7 for the variance within each window, and 2 for the variance between windows.



**Fig. 1** Example of discrimination task. Two side-by-side scatterplots were shown to each observer. Observers are asked to choose the one that appeared more correlated. Plots were  $6.3^\circ \times 6.3^\circ$  in extent, with

axes of length  $6.3^\circ$ ; dot cloud centers were separated by  $15^\circ$ . In this example, base correlation is 0.6 (left), and JND is from above (i.e., correlation of test plot is 0.8, higher than base correlation)

this ratio became sufficiently low ( $\leq 0.25$ ), or—consistent with the recommendations of Treutwein (1995)—52 trials had been run; the average of the sub-windows was then taken as the JND. This procedure proved reasonably effective, yielding results in 40 trials on average, and failing to converge on only 27 % of the runs.<sup>4</sup>

To assist convergence, test correlations were limited to values between the base and 1.0 (for above), or the base and 0.0 (for below). Feedback was provided immediately after each response via a 1-s sign: "+" for correct, and "-" for incorrect. New plots were generated every time a response was made, to encourage observers to respond to average properties (e.g., correlation) and avoid features of particular scatterplots (e.g., outliers).

Order of testing was determined via a latin square design (Kirk, 1995) that counterbalanced base correlations and direction of JND (above vs. below) across the 20 observers; the location of the scatterplot with the base correlation (left vs. right) was randomly assigned in each trial. To avoid floor effects at low correlations, no tests from below were run for base correlations of  $r < 0.3$ .

### Procedure – magnitude estimation

Another important aspect of performance is *accuracy*—the extent to which an observer can on average correctly determine the correlation of a scatterplot. Being a measure of central tendency, accuracy is, in principle, unrelated to precision.

Following Rensink and Baldrige (2010), accuracy was measured via the bisection of perceived correlation. Here, observers were shown two horizontally-separated *reference plots* (one with a high level of correlation, one with a low) and a *test plot* placed between them (Fig. 2). The correlation of the test plot was initially 0.1 away from that of the upper or the lower reference plot (each being equally likely); the observer then adjusted it until its correlation appeared to be halfway between

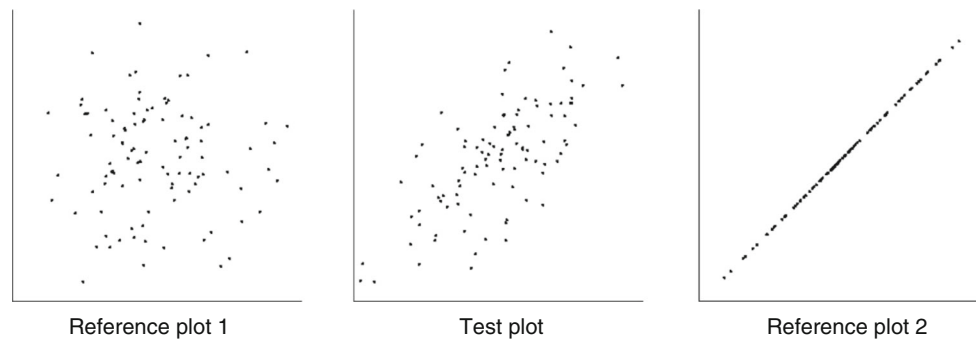
those of the references. This was done via keyboard control, with observers free to adjust the correlation of the test plot however they wished. To remove the possibility that observers could somehow use the number of steps, the size of each step was randomly assigned a value between 0 and 1/10 the difference between the reference correlations. Differing from Rensink and Baldrige (2010), each scatterplot—both reference and test—was replaced by a new instance every time an adjustment was made, or 1 s had passed. As in the case of discrimination, this encouraged observers to base their judgments on average properties rather than on some feature of a particular scatterplot.

The initial round of each test began with observers determining the point subjectively halfway between  $r = 0$  and  $r = 1$  (these corresponded to  $g = 0$  and  $g = 1$ , respectively). This was done four times, with the mean value taken as the physical correlation  $r$  corresponding to perceived correlation  $g = 1/2$ . A second round applied this recursively on two different subconditions: in the first, observers estimated the point between  $g = 0$  and  $1/2$ , and in the second, the point between  $g = 1/2$  and 1; the order of these was counterbalanced. These measurements were again made four times for each subcondition, with their averages taken as the values corresponding to  $g = 1/4$  and  $g = 3/4$ , respectively. A third round then measured the values of  $r$  corresponding to perceived magnitudes  $g = 1/8, 3/8, 5/8$ , and  $7/8$ ; subconditions here were presented in random order.

### Analysis

Following Kay and Heer (2015), JNDs were log-transformed before calculating average values across observers. JND curves were obtained by plotting these averages against adjusted correlation  $r_A$ , a symmetric measure that equates results from above and below base correlation; this is defined as  $r_A = r + 0.5 \text{ JND}(r)$ , the average of the two scatterplot correlations (Rensink & Baldrige, 2010). Variability  $k$  was the negative of the slope of this line, and bias  $b_{\text{disc}}$ , the reciprocal of its intercept with the  $r$ -axis ( $= k$  times the reciprocal of the intercept).

<sup>4</sup> For these, the value used was the average of the last 24 trials.



**Fig. 2** Example of magnitude estimation task. Observers adjusted the correlation of the test plot until its correlation appeared to be halfway between those of the reference plots. Plots were  $6.3^\circ \times 6.3^\circ$ , axes  $6.3^\circ$

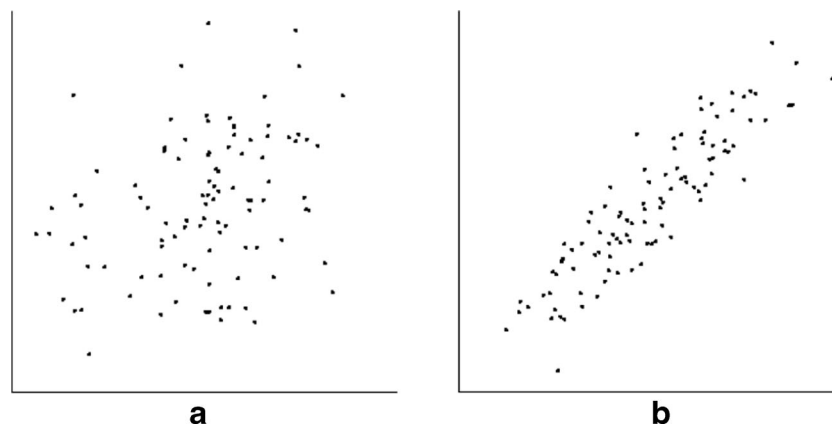
long, with dot cloud centers separated by  $7.6^\circ$ . This example is taken from the initial round of bisection; the reference plots have correlations  $r = 1$  and  $r = 0$ , while the test plot has the correlation  $r = 0.7$

Individual variabilities and biases were similarly calculated from individual JND curves. Bias  $b$  had considerable skew and kurtosis, which was reduced by use of a probit transform, with  $b$  limited to  $\leq .99$ . Also, to avoid the possibility that the limits of  $r = 0$  and  $r = 1$  were affecting JND estimates (Harrison, Yang, Franconeri, & Chang, 2014; Kay & Heer, 2015), a *range constraint* was imposed: a JND estimate was dropped if its average  $\pm 2.5$  standard deviations exceeded either of these limits.

In accord with the recommendations of Cumming (2012), effect sizes are emphasized in all analyses; 95 % confidence intervals (CIs; shown in square brackets) are given for all quantities of interest. Unless specified otherwise, any comparison is based on paired two-tailed  $t$ -tests.

### Experiment 1 – basic condition

The goal of this experiment was to replicate Rensink and Baldrige (2010), and serve as a "basic" condition against which the others could be compared. Dot clouds had 100 points, in a gaussian distribution with a mean of 0.5 and standard deviation of 0.2 in both dimensions (Fig. 3).



**Fig. 3** Examples of scatterplots for the basic condition. Here, scatterplots had 100 points with the same gaussian distribution in both dimensions. (a) Typical scatterplot for  $r = 0.3$ , (b) typical scatterplot for  $r = 0.9$

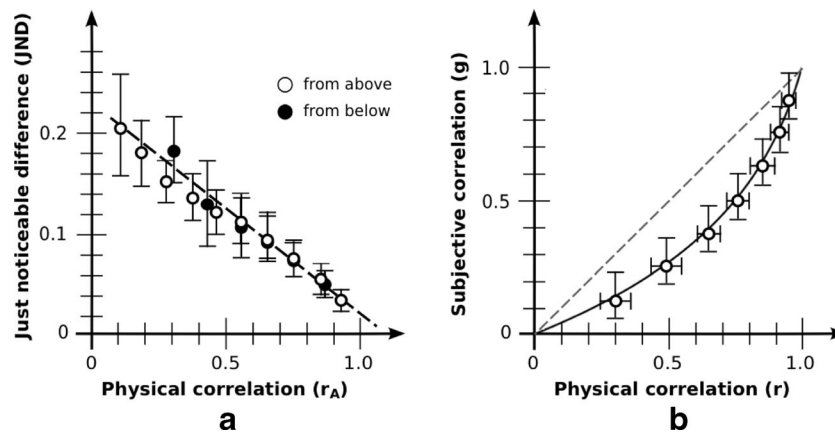
### Results: Discrimination

JNDs based on aggregate data are shown in Fig. 4a. Due to the range constraint, JNDs for base correlation  $r_B = 0.3$  (from below) were omitted from the analysis. Similar to what was found in Rensink and Baldrige (2010), JND was a strongly linear function of adjusted correlation both for JNDs from above ( $R^2 = .987$ ) and below ( $R^2 = .941$ ). For each observer, slopes and intercepts with the  $y$ -axis were then obtained by fitting a least-squares line through their data points. Consistent with Rensink and Baldrige (2010), there was no effect of JND direction on either slope ( $-0.20$  [ $-0.15, -0.24$ ] for above;  $-0.23$  [ $-0.18, -0.28$ ] for below;  $t(19) = 1.12$ ;  $p = .28$ ) or intercept ( $0.23$  [ $0.19, 0.26$ ] for above;  $0.25$  [ $0.21, 0.29$ ] for below;  $t(19) = 1.14$ ;  $p = .27$ ). Combining the data for both directions, JND retained a strong linearity ( $R^2 = .970$ ). Performance is therefore well described by Eq. (1), viz.,

$$\text{JND}(r) = k \left( 1 / b_{\text{disc}} - r_A \right)$$

with  $k$  the negative of the slope, and  $b_{\text{disc}}$  the reciprocal of the intercept with the  $r$ -axis.





**Fig. 4** Aggregate results: basic condition. **(a)** Discrimination as measured via JND. Error bars denote 95 % CIs. White dots indicate that comparison is made against a test correlation from above; black dots indicate test correlation from below. As is evident, these give rise to much the same line. Note that the maximum error in  $r$  in the generated plots (of 0.005) is much less than the JNDs found, even those for high

correlations, making it unlikely that this affects estimates in any significant way. **(b)** Magnitude estimation as measured via bisection. The curve for perceived correlation is  $g(r) = \ln(1 - b_{est}r) / \ln(1 - b_{est})$ ; best fit is for  $b_{est} = 0.90$ . Vertical error bars show  $g(r \pm 1 \text{ JND})$ ; horizontal error bars the 95 % CIs. The reference line  $g(r) = r$  makes explicit the degree of underestimation from physical correlation

Based on the variability and bias for each observer, average results were  $k = 0.21$  [0.17, 0.24] and  $b_{disc} = 0.90$  [0.84, 0.94]. These are broadly similar to the values  $k = 0.24$  and  $b_{disc} = 0.91$  of Rensink and Baldrige (2010).

**Results: Magnitude estimation**

Bisection results based on aggregate data are shown in Fig. 4b. Consistent with other reports (e.g., Cleveland et al., 1982), an underestimation of correlation appeared, especially in the range  $0.2 < r < 0.6$ . And consistent with Rensink and Baldrige (2010), data show a good fit to Eq. (2)

$$g(r) = \ln(1 - b_{est}r) / \ln(1 - b_{est})$$

where  $b_{est}$  is the bias obtained via magnitude estimation.

The best fit with the aggregate data is for  $b_{est} = 0.90$ . The resulting fit is excellent: root mean square error (RMSE) from the set of observed values is only 0.018. The value of  $b_{est}$  obtained via individual estimates is similar: 0.91 [0.85, 0.95]. These results are also not far from the value  $b_{est} = 0.87$  found in Rensink and Baldrige (2010).

**Results: Systematicity**

If the *Fechner assumption* holds for this situation—i.e., that each JND corresponds to the same difference in subjective experience—Eq. (1) implies Eq. (2), with the bias in the two equations being identical (Rensink & Baldrige, 2010). The value of bias obtained via discrimination ( $b_{disc} = 0.90$  [0.84, 0.94]) was indeed much the same as that obtained via magnitude estimation (0.91 [0.85, 0.95]); the slight difference of 0.01 was not statistically significant ( $t(19) = 0.40$ ;  $p = .69$ ).

As such, the Fechner assumption appears to hold in this condition.

These results therefore replicate the findings of Rensink and Baldrige (2010): for this condition, average precision and accuracy of correlation perception can be described via a pair of simple performance laws (linear and logarithmic, respectively) that are closely related, with  $b_{disc}$  and  $b_{est}$  essentially measuring the same quantity  $b$ .

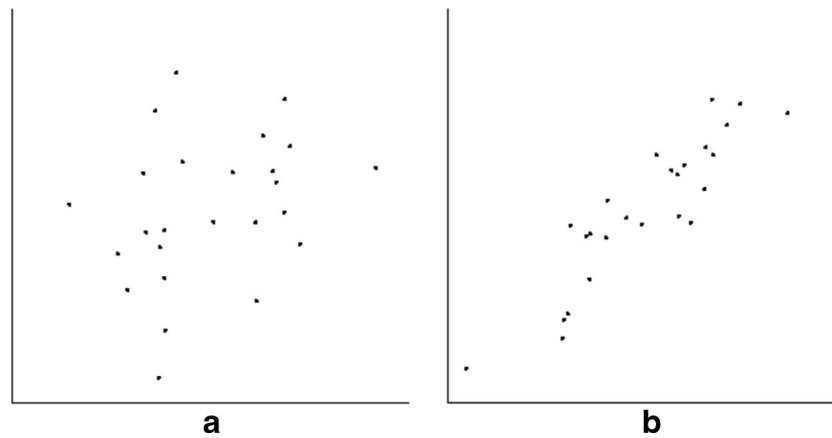
**Experiment 2 – low density**

This was the same as the basic condition, but with only 25 points. The goal here was to determine whether performance would change if the number of points—and thus, the density of the dot cloud—were markedly lower (Fig. 5).

**Results: Discrimination**

JNDs based on the aggregate data are shown in Fig. 6a. Due to the range constraint, results were omitted from base correlations 0.3 and 0.4 for JNDs for the below condition. The remaining JNDs were again a linear function of adjusted correlation  $r_A$  both from above ( $R^2 = .973$ ) and from below ( $R^2 = .961$ ). No effect of JND direction was found on either slope (0.30 [0.25, 0.34] for above; 0.30 [0.23, 0.36] for below;  $t(19) = 0.02$ ;  $p = .98$ ) or intercept (0.34 [0.31, 0.38] for above; 0.33 [0.28, 0.37] for below;  $t(19) = 0.78$ ;  $p = .45$ ). Combining both sets, behavior again remained quite linear ( $R^2 = .963$ ).

The slopes of individual fits yielded an average variability  $k = 0.30$  [0.26, 0.35], noticeably higher than for the basic condition ( $k = 0.20$ ). Bias  $b_{disc} = 0.91$  [0.85, 0.95] remained about the same (cf.  $b_{disc} = 0.90$  [0.84, 0.94]).



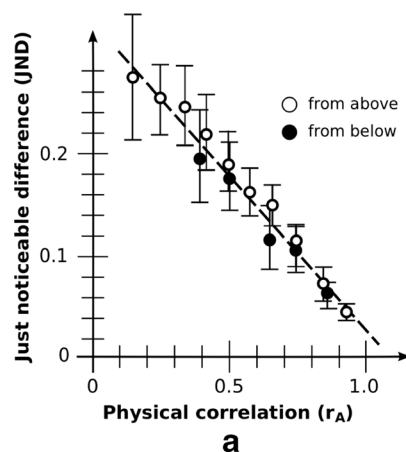
**Fig. 5** Examples of low-density scatterplots. Here, scatterplots had 25 points with the same gaussian distribution in both dimensions. (a) Typical scatterplot for  $r = 0.3$ , (b) typical scatterplot for  $r = 0.9$

### Results: Magnitude estimation

Estimates based on the aggregate data are shown in Fig. 6b. Data again show a good fit to the logarithmic function of Eq. (2). The best fit was for  $b_{\text{est}} = 0.89$ , close to the value of 0.90 for the basic condition; the fit for this against the set of observed estimates was again excellent, with an RMSE of only 0.013. The estimate of  $b_{\text{est}}$  obtained via individual observers was also similar:  $b_{\text{est}} = 0.90$  [0.84, 0.94], and much the same as that of the basic condition ( $b_{\text{est}} = 0.91$  [0.85, 0.95]).

### Results: Systematicity

The bias obtained via discrimination ( $b_{\text{disc}} = 0.91$  [0.85, 0.95]) was much the same as that obtained via magnitude estimation ( $b_{\text{est}} = 0.90$  [0.84, 0.94]). The difference of 0.01 was not statistically significant ( $t(19) = 0.14$ ;  $p = .89$ ), showing that the Fechner assumption again holds, with  $b_{\text{disc}}$  and  $b_{\text{est}}$  simply being measures of the same quantity.



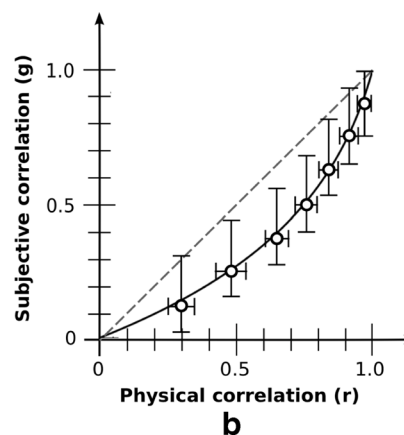
**Fig. 6** Aggregate results: low-density condition. (a) Discrimination as measured via JND. White dots indicate test correlation from above; black dots indicate test correlation from below. As is evident, these give much the same estimates. Error bars denote 95 % CIs. (b) Magnitude estimation

### Experiment 3 – high aspect ratio

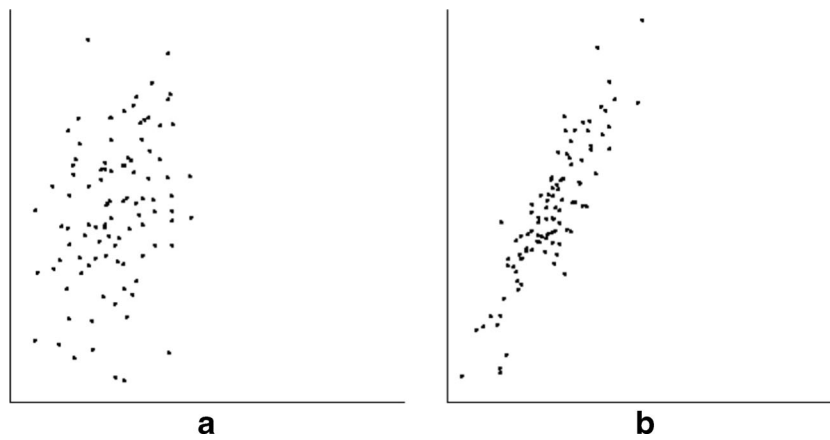
An important property of any scatterplot is its *aspect ratio*—the ratio of vertical to horizontal extent. This ratio may affect the ability of observers to detect trends in data, including the perception of clusters (Fink, Hauer, Spoerhase, & Wolff, 2013), possibly by reducing the distances between the dots. To test whether perceived correlation is affected by this factor, the dot cloud of the basic condition was horizontally compressed so as to have an aspect ratio of 2:1. The size of the axes and dots themselves were left unaltered (Fig. 7).

### Results: Discrimination

JNDs based on the aggregate data are shown in Fig. 8a. Results were omitted from base correlation  $r_B = 0.3$  (JNDs from below) due to the range constraint. Dependence of JND on adjusted correlation  $r_A$  was again quite linear, both for JNDs from above ( $R^2 = .942$ ) and below ( $R^2 = .942$ ). There was no significant effect



as measured via bisection. The curve for perceived correlation is  $g(r) = \ln(1 - b_{\text{est}}r) / \ln(1 - b_{\text{est}})$ ; best fit is for  $b_{\text{est}} = 0.89$ . Vertical error bars show  $g(r \pm 1 \text{ JND})$ ; horizontal error bars the 95 % CIs



**Fig. 7** Examples of scatterplots with high aspect ratios. Here, scatterplots had 100 points with a gaussian distribution in both dimensions, but with the horizontal component compressed by a factor of 2. (a) Typical scatterplot for  $r = 0.3$ , (b) typical scatterplot for  $r = 0.9$

of JND direction on either slope (0.22 [0.18, 0.26] for above; 0.24 [0.18, 0.29] for below;  $t(19) = 0.57$ ;  $p = .57$ ) or intercept (0.25 [0.22, 0.29] for above; 0.26 [0.22, 0.31] for below;  $t(19) = 0.49$ ;  $p = .63$ ). When both sets of data were combined, JND remained a linear function of  $r_A$  ( $R^2 = .941$ ).

Analysis of individual slopes yielded an average variability  $k = 0.22$  [0.19, 0.26], similar to that for the basic condition ( $k = 0.21$ ). Average bias  $b_{disc} = 0.89$  [0.84, 0.92] was also much the same as the corresponding basic value ( $b_{disc} = 0.90$ ). Thus, even when standard deviations in the two dimensions differ by a factor of 2, performance appears largely unaffected.

**Results: Magnitude estimation**

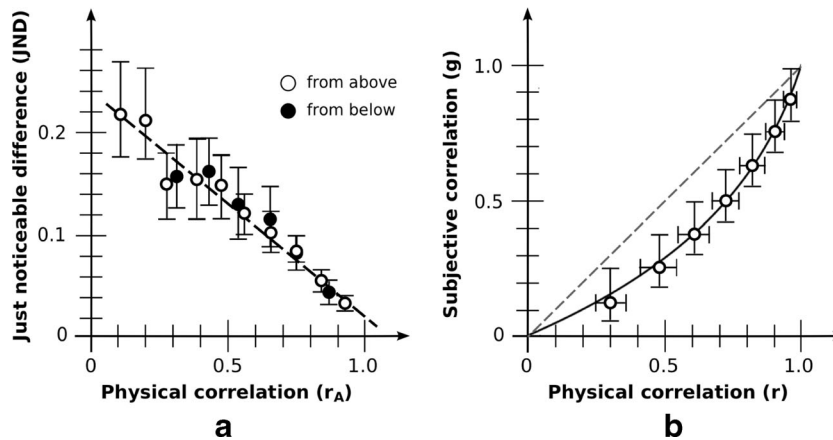
Average estimates are shown in Fig. 8b. Data again show a good fit with Eq. (2). The best fit was for  $b_{est} = 0.85$ ; the resulting fit is excellent, with RMSE less than 0.013. The average value of  $b_{est}$  obtained via individual observers was similar:  $b_{est} = 0.83$  [0.71, 0.92].

**Results: Systematicity**

For this condition, the bias obtained via discrimination ( $b_{disc} = 0.89$  [0.84, 0.92]) was not far from that obtained via magnitude estimation ( $b_{est} = 0.83$  [0.71, 0.92]). This difference was not statistically significant ( $t(19) = 1.00$ ;  $p = .33$ ), suggesting that the link between precision and accuracy exists for this condition as well. Thus, for high aspect ratios, the shape of both performance curves remains much the same as for the other conditions, as does the link between them.

**Experiment 4 – uniform distribution**

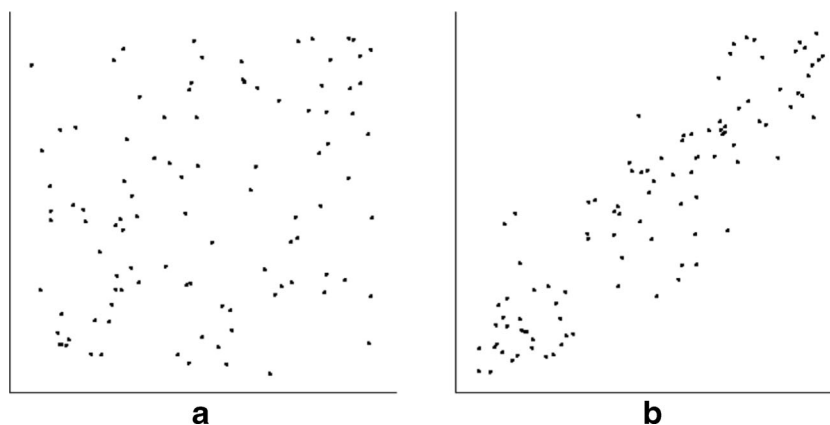
The results of Experiments 1–3 suggest that correlation perception is robust to variations in the parameters of a scatterplot dot cloud, at least for gaussian distributions. To examine what happens when distributions move away from being gaussian, Experiment 4 used scatterplots with uniform distributions.



**Fig. 8** Aggregate results: high aspect ratio. (a) Discrimination as measured via JND. White dots indicate test correlation from above, black from below; these give much the same estimates. Error bars denote 95 % CIs. (b) Magnitude estimation as measured via bisection.

The curve for perceived correlation is  $g(r) = \ln(1 - b_{est}r) / \ln(1 - b_{est})$ ; best fit is for  $b_{est} = 0.85$ . Vertical error bars show  $g(r \pm 1 \text{ JND})$ ; horizontal error bars the 95 % CIs





**Fig. 9** Examples of scatterplots with uniform distributions. Here, scatterplots had 100 points with a uniform distribution in both dimensions. **(a)** Typical scatterplot for  $r = 0.3$ , **(b)** typical scatterplot for

$r = 0.9$ . These tend to be slightly blockier in shape, and have more sharply-defined borders than their gaussian counterparts (cf. Fig. 3)

Such distributions have often been used to study visual perception (e.g., Chong & Treisman, 2003; Cohen, Singh, & Maloney, 2008). Moreover, they are not entirely unnatural: for example, uniform distributions appear to be the basis of internal representations of visuomotor error in speeded reaching tasks (Zhang, Daw, & Maloney, 2015).

For the condition here, the same mean and range were used as for the basic condition; this entailed a somewhat larger standard deviation ( $1/\sqrt{12} = 0.29$ , rather than 0.2). In contrast to the other conditions, points here were created by first obtaining a list of  $x$ -coordinates from a uniform distribution (subject to a particular mean and standard deviation). The corresponding  $y$ -coordinates were formed by creating a scrambled copy of this list and then ordering the values using a comb sort (Box & Lacey, 1991; Harrison, 1995) until the required correlation had been achieved. A comb sort initially compares—and if need be, swaps—items separated by large distances, with this distance being reduced over time. The result is a dot cloud with a shape somewhat similar to that of the gaussian distribution with the same correlation (Fig. 9).

## Results: Discrimination

JNDs based on the aggregate data are shown in Fig. 10a. Due to the range constraint, results were omitted for base correlation  $r_B = 0.3$  and 0.4 (JNDs from below). As in previous conditions, dependence of JND on adjusted correlation  $r_A$  was highly linear, both for JNDs from above ( $R^2 = .975$ ) and below ( $R^2 = .988$ ). There was a tendency for JND direction to affect slope (0.28 [0.23, 0.32] for above; 0.22 [0.16, 0.28] for below ( $t(19) = 2.09$ ;  $p = .05$ ).<sup>5</sup> There

<sup>5</sup> When data from the two lowest base correlations ( $r = 0.0$  and 0.1) were removed from the analysis, slopes became much more alike ( $t(19) = 0.75$ ;  $p = .46$ ), as did intercepts ( $t(19) = 1.07$ ;  $p = .30$ ), with average  $k = 0.24$  [0.20, 0.28]. The divergence may therefore have been caused by the squarish shape of the dot clouds at low levels of correlation. These two conditions are therefore omitted in the subsequent analysis.

was also an effect of JND direction on intercept (0.29 [0.25, 0.32] for above; 0.24 [0.19, 0.28] for below;  $t(19) = 2.48$ ;  $p = .02$ ). When these data were combined, behavior remained highly linear ( $R^2 = .983$ ).

Analysis of individual slopes yielded a variability  $k = 0.24$  [0.20, 0.28], not far from that of the basic condition ( $k = 0.20$ ). Bias  $b_{\text{disc}} = 0.94$  [0.89, 0.97] appeared to be slightly higher than the basic value  $b_{\text{disc}} = 0.90$  [0.84, 0.94], possibly because of the greater standard deviation (Cleveland et al., 1982; Lauer & Post, 1989).

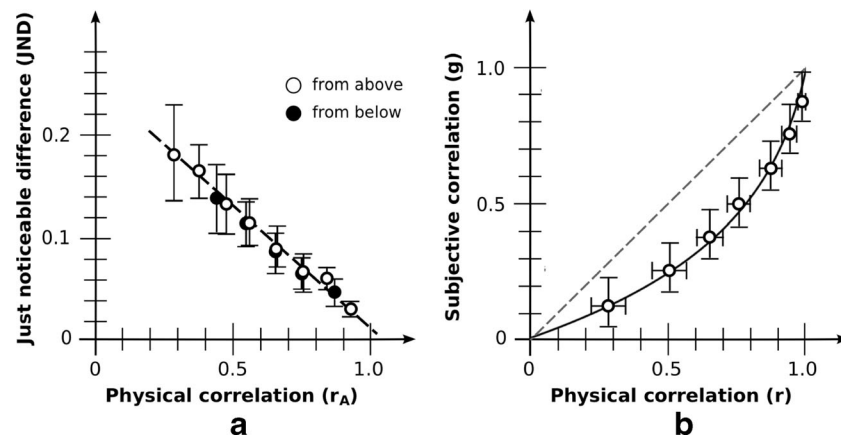
## Results: Magnitude estimation

Average estimates for this condition are shown in Fig. 10b. The best fit of the aggregate data with Eq. (2) was for  $b_{\text{est}} = 0.93$ . As was the case for the other conditions, the fit is a good one, with an RMSE of 0.032. This value of  $b_{\text{est}}$  is similar to that obtained via the fits of individual observers: 0.94 [0.91, 0.97]. Similar to the case of discrimination-based estimates, this latter value appears slightly higher than that for the basic condition ( $b_{\text{est}} = 0.91$  [0.85, 0.95]).

## Results: Systematicity

As in the case of the gaussian distributions, the bias obtained via discrimination ( $b_{\text{disc}} = 0.94$  [0.89, 0.97]) was much the same as that obtained via magnitude estimation ( $b_{\text{est}} = 0.94$  [0.91, 0.97]). This difference was not significant ( $t(19) = 0.54$ ;  $p = .59$ ), indicating that the link between discrimination and magnitude estimation is reasonably good here as well.

In summary, then, performance for uniform distributions was largely the same as for gaussian ones: performance for discrimination remained highly linear, perceived magnitude remained logarithmic, and the two curves remained closely linked.



**Fig. 10** Aggregate results: uniform distribution. (a) Discrimination via JND. White dots indicate test correlation from above; black dots from below. As is evident, estimates are much the same. Error bars denote 95 % CIs. (b) Magnitude estimation as measured via bisection. The curve

for perceived correlation is  $g(r) = \ln(1 - b_{\text{est}}r) / \ln(1 - b_{\text{est}})$ ; best fit is for  $b_{\text{est}} = 0.93$ . Vertical error bars show  $g(r \pm 1 \text{ JND})$ ; horizontal error bars the 95 % CIs

## General discussion

The experiments here show that under a fairly wide range of conditions the perception of correlation in scatterplots obeys two linked laws: a linear (Weber) law for discrimination, and a logarithmic (Fechner) law for perceived magnitude. Fit to observed values was good for all conditions tested. As such, these laws will likely hold reasonably well for many distributions, including those that are non-gaussian to some extent.

## Mechanism

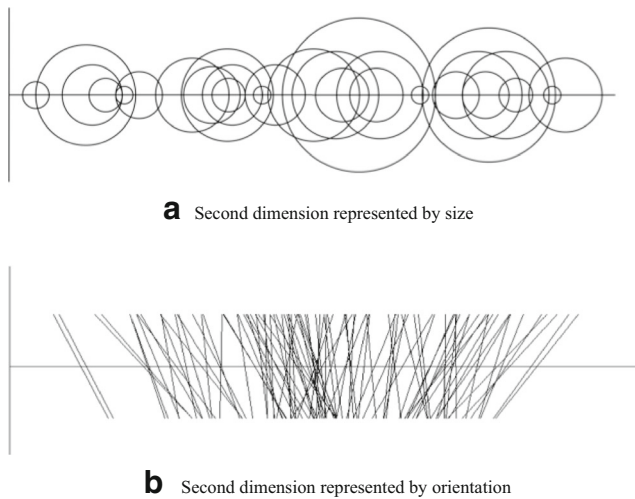
Why do the laws describing correlation perception have such generality? And why do they have the form that they do? In what follows, it is suggested that (i) perceived correlation is based on the probability distribution of data points in an abstract parameter space, (ii) performance depends on the shape—and in particular, the width—of this distribution, and (iii) this reflects the ability of human observers to perceive the information entropy (Shannon entropy) in the image. Each of these suggestions will now be discussed in turn.

- i) *Probability distribution.* When considering what might underlie the perception of correlation, it is worth noting that discrimination and magnitude estimation are both functions of  $u = 1 - br$ , not  $r$  alone. The quantity  $u$  is akin to the average perpendicular distance  $X$  from the regression line proposed by Meyer et al. (1997) in that it has a small value at  $r = 1$  and increases as  $r$  decreases. The possibility that a quantity of this kind is involved is further supported by the finding that the areas of the brain involved in correlation perception increase in activity as  $r$  moves away from 1 (Best et al., 2006).

One candidate consistent with this behavior is the area of the region encompassing the set of dots in the cloud—e.g., their bounding box or convex hull. This seems unlikely, however, given that outliers (e.g., from contaminating distributions) have relatively little effect on performance, which is based instead on the bulk of the points (Bobko & Karren, 1979; Konarski, 2005; Meyer & Shinar, 1992). Moreover, Weber's Law applies best to properties that—like lightness, color, or density—are *intensive*, i.e., when components with the same property are combined, the whole has that property too (see Ross, 1997).<sup>6</sup> Intensive properties that are geometric—such as orientation—are possible candidates. But those that are possible in theory do not seem likely in practice: perceived correlation corresponds neither to the orientation of the regression line (Experiment 3; Lane et al., 1985) nor to the ratio of the major and minor axes of the dot cloud (Cleveland et al., 1982; although see Boynton, 2000). Furthermore, the perceived magnitude of most physical properties (including area or distance) is generally described best by a power function of its physical magnitude (see e.g., Billock & Tsou, 2011; Ross, 1997), not a logarithmic function of the kind found here.

An even more important consideration perhaps is the invariance of correlation perception to different kinds of graphical representation. Large dots in a scatterplot, for example, make the dot cloud blobby and give it a larger outer boundary. They do not, however, affect performance (Rensink, 2012, 2014). Indeed, estimation and discrimination of correlation follow similar laws

<sup>6</sup> In contrast, properties such as area or length are *extensive*, behaving in a rather different way (see, e.g., Carnap, 1966). For example, the area of a given region is the *sum* of the areas of its component sub-regions.



**Fig. 11** Examples of different representations of data. In these *augmented stripplots*, the first data dimension of each data element is represented spatially (as in a scatterplot), but the value of the second dimension is mapped to a value on a different physical property. **(a)** Second dimension represented by size (area). **(b)** Second dimension represented by orientation. Despite their different visual appearance, both result in correlation perception that is much the same as that for scatterplots: perceived magnitude is still described by  $g(r) = \ln(1 - br) / \ln(1 - b)$ , and discrimination by  $JND(r) = k(1/b - r)$ . From Rensink (2014, 2015)

even when the graphical representations involved are entirely different in appearance (Fig. 11)—for example, when the second data dimension of a data element is represented by size or color rather than vertical position (Rensink, 2012, 2014, 2015), or when line graphs or bar charts are used (Harrison et al., 2014).

This indifference suggests that performance is based not on a geometric structure inferred from the dot cloud, but on something more abstract. One possibility is the *probability distribution*  $p(x_i, y_i)$ , which states that  $np(x_i, y_i)$  dots—or more generally, data points—are expected at each position  $i$  in a two-dimensional array. Such distributions are the basis of various scene statistics, which play an important role in scene perception (see e.g., Geisler, 2008; Olshausen & Field, 1996). Distributions involving two perceptual dimensions would not be problematic, especially if they pertain not to different properties (e.g., one for space and one for orientation), but to a common, more abstract parameter space.

The mechanisms responsible are less clear, but may well be those that underlie ensemble coding, which enable various statistics of a set of items to be determined rapidly and with relatively little attention (see e.g., Alvarez, 2011; Haberman & Whitney, 2012). It has been suggested that the shape of one-dimensional distributions of various properties can be perceived this way (Utochkin, 2015). An extension to two dimensions would appear to be fairly natural. And given that summary statistics can be obtained in as little as 100 ms (Chong & Treisman, 2003; Robitaille

& Harris, 2011), it could also explain why correlation perception can be achieved within a similarly brief amount of time (Rensink, 2014).

- ii) *Distribution shape*. If the density of points is sufficiently high,  $p(x_i, y_i)$  can be approximated by a continuous *probability density function*  $f(x, y)$ , such that  $nf(x, y)\Delta x\Delta y$  dots are expected in the area  $\Delta x\Delta y$  centered on  $(x, y)$ . The function  $nf(x, y)$ —the *dot density function*—has a shape largely unaffected by outliers. Moreover, the shape of  $f(x, y)$  does not depend on the number of dots present; the greater variability in  $k$  when relatively few dots are used (cf. Experiment 2; Rensink, 2014) is likely due to  $f(x, y)$  being sampled insufficiently finely. If the center of mass of each dot were used as the basis of  $f(x, y)$  (instead of raw pixels, say), it would also explain why performance is largely indifferent to their size, shape, and color (Rensink, 2014).

For a bivariate gaussian distribution,  $f(x, y)$  has the form:

$$f(x, y) = \frac{e^{-q(x, y)}}{2\pi\sigma_x\sigma_y(1-r^2)^{1/2}} \quad (4)$$

where

$$q(x, y) = \frac{\frac{(x-\mu_x)^2}{\sigma_x^2} - 2r\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2}}{2(1-r^2)}, \quad (5)$$

where  $\mu_i$  and  $\sigma_i$  are respectively the mean and standard deviation of dimension  $i$ , and  $r$  is the correlation<sup>7</sup> (see e.g., Timm, 2002). One way of determining its shape via a set of *isofraction* points—points whose value is a fixed fraction of the maximum (i.e., the value at the center of the dot density function). Writing this fraction as  $e^{-K_1^2}$ , where  $K_1$  is some fixed constant, these points form an ellipse  $q(x, y) = K_1^2$  (Fig. 12).<sup>8</sup>

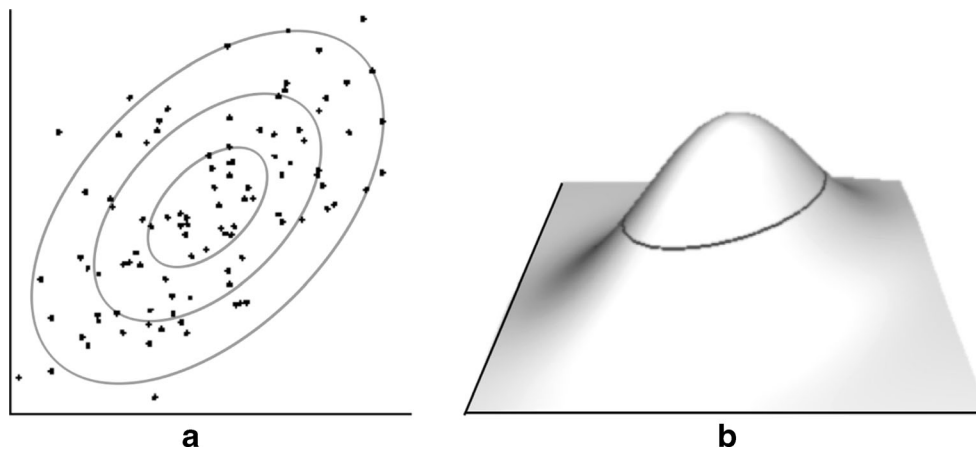
Applying the formula for the semi-minor axis  $L_{min}$  of an ellipse (e.g., Johnson, 2015), the width  $w$  of this *isofraction ellipse* is

$$w(r) = 2 \times L_{min} = \frac{4K_1\sigma_x\sqrt{1-r^2}}{\left[ \left(1 + 1/\kappa^2\right) + \sqrt{\left(1 - 1/\kappa^2\right)^2 + 4r^2/\kappa^2} \right]^{1/2}}, \quad (6)$$

where  $\kappa = \sigma_y/\sigma_x$ . Note that  $w(r)$  is independent of  $n$ . Given Eq. (2), it is natural to assume that perceived correlation  $g(r)$

<sup>7</sup> The correlation  $r$  is technically the population parameter  $\rho$ . In the work here, however,  $r$  is set to the same value in all instances, and so is used here to minimize notational switching.

<sup>8</sup> Isofraction lines are a special type of isopleth (lines of equal probability). By definition, they have the property that the same fraction (determined by  $K_1$ ) can specify an ellipse independent of any particular value of  $r$  or  $n$ .



**Fig. 12** Shape of dot density function. **(a)** Two-dimensional view of a scatterplot ( $r = 0.5$ ), with several possible isopleths (iso-probability contours). **(b)** Three-dimensional view of the corresponding dot density

function  $n_f(x,y)$ , with iso-fraction ellipse for values 1/2 that of the maximum (i.e., the value at its center)

might be proportional to the logarithm of this quantity. This assumption can be expressed as

$$g(r) = G[\ln(w(r)) + h]. \tag{7}$$

where  $G$  and  $h$  are constants, to be chosen such that  $g(0) = 0$  and  $g(1) = 1$ . Note that this implies a calibration step to map perceived quantities to the appropriate values of correlation; skipping this step may explain some of the incoherence in magnitude estimates occasionally encountered (see e.g., Doherty et al., 2007).

When both dimensions have the same standard deviation  $\sigma$ , Eq. (6) reduces to

$$w(r) = 2\sqrt{2}K_1\sigma\sqrt{1-r}, \tag{8}$$

and Eq. (7) takes the form

$$g(r) = \frac{G}{2} \ln(K_2^2(1-r)) + Gh \tag{9}$$

where  $K_2 = 2\sqrt{2}K_1\sigma$ . Note that  $g(r)$  diverges as  $r$  approaches 1, since the width of the ellipse approaches zero. However, owing to perceptual noise and blurring, a residual width  $w_{res}$  still exists in this situation. Placing this into Eq. (9) yields

$$g(r) = \frac{G}{2} \ln(w_{res}^2 + K_2^2(1-r)) + Gh, \tag{10}$$

which can be rewritten

$$g(r) = \frac{G}{2} [\ln(1-br) + \ln(w_{res}^2 + K_2^2)] + Gh, \tag{11}$$

where

$$b = \frac{K_2^2}{w_{res}^2 + K_2^2}. \tag{12}$$

Choosing  $h$  such that  $g(0) = 0$ , Eq. (11) becomes

$$g(r) = \frac{G}{2} \ln(1-br),$$

which becomes Eq. (2) when  $G$  is chosen to let  $g(1) = 1$ .

In this view, the parameter  $b$  reflects the relative contribution of residual and "primary" factors. This can be made more explicit by setting

$$w_{res} = cK_2,$$

where  $c > 0$  describes the relative contribution of  $w_{res}$ ; Eq. (12) then becomes

$$b = \frac{1}{c^2 + 1}. \tag{13}$$

A relatively small residual  $c$  (and thus,  $c^2$ ) would explain why values of  $b$  are often close to 1. Furthermore, since  $w_{res}$  is fixed and  $K_2$  is proportional to  $\sigma$  (cf. Eq. (9)),  $c$  is inversely proportional to  $\sigma$ , which may explain why accuracy of correlation perception improves when dot clouds have smaller standard deviations (Cleveland et al., 1982).<sup>9</sup> The larger standard deviations for uniform distributions might likewise account for the (nonsignificantly) larger biases found in Experiment 4: the ratio of .29/.2 for the relative sizes of the standard deviation would predict  $b = .95$ , close to the observed value of .94.

Note that these developments do not rely on the way that information is represented; instead of corresponding to positions in the image, dimensions  $x$  and  $y$  might correspond to values in a more abstract parameter space. This could explain

<sup>9</sup> Fitting the data in Fig. 2 of Cleveland et al. to eq. (2) yields  $b$  values of .91, .89, .94, and .96 for increasing sizes of the dot cloud. Taking the reference value to be  $b = .89$  (size 2 in Cleveland et al), predicted  $b$  values are .84, (.89), .94, and .97 respectively, a good match to most of the data. Note that size 2 in Cleveland et al. is roughly that of the basic condition here; the similar values of  $b$  indicate a degree of consistency across studies.



the existence of similar laws when other graphical representations are used (Harrison et al., 2014; Rensink, 2014, 2015).

iii) *Information entropy.* The assumption that perceived correlation depends on the width of the probability distribution can account for the empirical results found here, as well as those of several other studies. But why would human vision be concerned with this particular quantity (cf. Marr, 1982)? And given that the square of the width (cf. Eq. (8)) is an accurate estimator of correlation, why would the visual system instead use its logarithm? As a possible answer, an *entropy theory* is proposed here: observers can perceive the information entropy in a scene, with this quantity then used as a proxy for correlation.

Information entropy (Shannon entropy) can be defined as  $H = -\sum p(x_i, y_i) \ln p(x_i, y_i)$  (see, e.g., Lemons, 2013). This is an inherently statistical quantity, reflecting the number of possible configurations of a given probability distribution. Physical entropy—the physical instantiation of  $H$ —is useful for describing many natural structures (see e.g., Ben-Naim, 2008; Lemons 2013; Weber, Depew, & Smith, 1988). Information entropy could likewise be a useful descriptor of visual structure. In computer vision systems measures based on entropy have been used to assess statistical structure in images of real-world scenes and textures (e.g., Chang, Du, Wang, Guo, & Thouin, 2006; Zhu, Wu, & Mumford, 1998). In human vision, it has been suggested that eye movements are guided by the perception of contrast entropy, enabling the greatest amount of information gained at each fixation (Raj, Geisler, Frazor, & Bovik, 2005; Renninger, Verghese, & Coughlan, 2007). A variant of this quantity—the number of bits needed to encode an image using a subset of wavelets—has also been considered as a possible measure of the visual clutter in a scene (Rosenholz, Li, & Nakano, 2007).

In general, entropy is difficult to estimate (Archer, Park, & Pillow, 2014). For a bivariate gaussian distribution, however, the situation can be simplified. The differential entropy<sup>10</sup> of the corresponding probability density function (Eq. 4) is

$$H(r) = \ln(2\pi e\sigma_x\sigma_y\sqrt{1-r^2}) \quad (14)$$

(see e.g., Gokhale, Ahmed, & Res, 1989); for a distribution

of  $n$  dots, this simply becomes

$$H(r) = n \ln(2\pi e\sigma_x\sigma_y\sqrt{1-r^2}) \quad (15)$$

Meanwhile, the isofraction ellipse  $q(x,y) = K_1^2$  for the dot density function  $n f(x,y)$  has an area  $A$  given by (e.g., Johnson, 2015)

$$A(r) = 2\pi\sigma_x\sigma_y K_1^2 \sqrt{1-r^2} \quad (16)$$

allowing Eq. (15) to take the form

$$H(r) = n \ln(Ae/K_1^2) = n \ln(A) + n \ln(e/K_1^2) \quad (17)$$

Given that the methods used here involve only differences in  $g(r)$ , entropy theory implies

$$g(r_1) - g(r_2) = G'(H(r_1) - H(r_2)), \quad (18)$$

or equivalently,

$$g(r) = G'[H(r) + h'], \quad (19)$$

where  $G'$  and  $h'$  are real-valued constants. Substituting Eq. (17) into Eq. (19) leads to

$$g(r) = G[\ln(A(r)) + \ln(e/K_1^2) + h']. \quad (20)$$

where  $G = nG'$ .

The issue now is to determine the area  $A$  of the isofraction ellipse. For geometric structures, the perceived area of an ellipse is the product of separate one-dimensional measurements (Morgan, 2005); the existence of capacity limitations for ensemble coding (Attarha & Moore, 2015) would suggest a similar situation here. If only one ensemble descriptor can be determined at a time,  $A(r)$  might be approximated by the product of the width  $w(r)$  of the ellipse and some fixed value  $D$  representing its length; Eq. (20) then becomes

$$g(r) \approx G[\ln(w(r)) + \ln(D) + \ln(e/K_1^2) + h']. \quad (21)$$

Setting  $h = h' + \ln(D) + \ln(e/K_1^2)$  yields Eq. (7), which—as shown in the previous section—becomes Eq. (2) when  $\sigma_x$  and  $\sigma_y$  are equal. This approximates the exact formula for entropy (Eq. (20)) fairly well (Fig. 13).

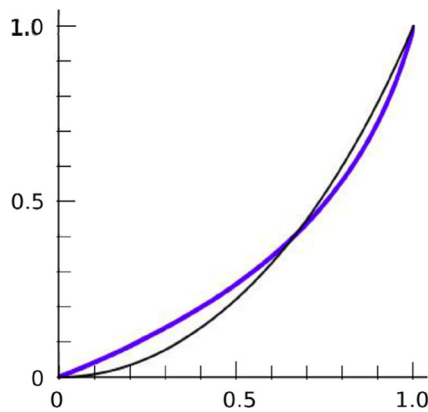
To show the linkage between discrimination and estimation, the derivative of Eq. (2) yields

$$\Delta g(r) / \Delta r \approx dg/dr = -b / [(1-br)\ln(1-b)]. \quad (22)$$

Because of the close approximation of Eq. (2) to the actual entropy, any change  $\Delta g$  is reasonably proportional to the associated change  $\Delta H$ . And because  $\Delta H$  is a measure of information gain (see e.g., Lemons, 2013), all  $\Delta g$ 's of equal size involve about the same number of bits. In this situation, then,

<sup>10</sup> Although much the same as entropy for the discrete case, differential entropy—based on a probability density function rather than an array of discrete probabilities—can be negative in some situations (e.g., as  $r$  goes to 1 in eq. (9)). However, it is a good approximation here, with the fixed residual width  $w_{\text{res}}$  preventing the argument of the logarithm from becoming arbitrarily small. Moreover, the approach here relies on *differences* in differential entropy, which does not have such problems (see e.g., Norwich, 1987).





**Fig. 13** Approximation of entropy. Fechner's law (Eq. (2); thicker line) vs. exact formula based on entropy (Eq. (20); thinner line). Both have been adjusted to have best fits to the data of Experiment 1. As is evident, Fechner's law is slightly higher at low correlations and slightly lower at high ones. But the two curves match each other reasonably well: average RMSE over the range  $0 \leq r \leq 1$  is only 0.04

the Fechner assumption that each JND corresponds to the same difference in subjective experience  $\Delta g$  takes the weaker form that each JND corresponds to the same number of perceived bits  $I_{75}$  (where the subscript denotes 75 % discrimination accuracy<sup>11</sup>). Letting  $JND(r)$  denote the value of  $\Delta r$  that corresponds to this number, Eq. (22) becomes

$$\begin{aligned}
 JND(r) &\approx -I_{75}(1-br)\ln(1-b) / (-b) \\
 &= -I_{75}\ln(1-b) \left(1/b-r\right), \tag{23}
 \end{aligned}$$

which then becomes Eq. (1) by setting  $k = -I_{75} \ln(1 - b)$ . (Note that since  $I_{75}$  depends on the particular representation and number  $n$  of data points used, if these are constant,  $k$  will be proportional to  $-\ln(1-b)$ .<sup>12</sup>) The finding that discrimination and estimation lead to the same estimates of  $b$  shows that this weaker form of the Fechner assumption holds fairly well. It is worth noting that the resulting logarithmic relationship between physical and perceived quantities is relatively rare, since most such relationships involve power laws rather than logarithms (Billock & Tsou, 2011).

Interestingly, the above treatment also goes through if  $g(r)$  were proportional to *entropy density*, an intensive property

<sup>11</sup> The value of a threshold depends somewhat on the method used to measure it (see e.g., Chong & Treisman, 2005; Treutwein, 1995). Thus,  $I_{75}$ —and therefore,  $k$ —may differ slightly for different methods. But it will be constant when the same method is used throughout.

<sup>12</sup> This may explain why variability in the uniform distribution (Experiment 4) tends to be higher than in the basic condition (Experiment 1). If  $b$  is greater in the uniform distribution ( $b = 0.94$  [0.89, 0.97] vs.  $b = 0.91$  [0.85, 0.95]), then  $k = 0.21 * \ln(1-0.94)/\ln(1-0.91) = 0.245$ , quite close to the  $k = 0.24$  observed. Given that these trends were not found to be significant in the present study, another study with more power would be needed before conclusively accepting or rejecting this proposal.

that describes the amount of entropy per unit area (with this area being, e.g., that subtended by the scatterplot axes). Because this differs formally from information entropy only by a multiplicative constant, the calibration done to set  $g(0) = 0$  and  $g(1) = 1$  would result in no formal difference in  $g(r)$ , and thus, no difference in performance.

iv) *Unequal standard deviations.* Much of the development above assumes that the standard deviations of both dimensions are equal (i.e.,  $\kappa = 1$ ). When this is not the case, two limiting cases can be singled out. First, when  $\sigma_y \ll \sigma_x$  ( $\kappa \rightarrow 0$ ), Eq. (6) becomes

$$w(r) = 2\sqrt{2}K_1\sigma_y\sqrt{1-r^2}. \tag{24}$$

Second, when  $\sigma_y \gg \sigma_x$  ( $\kappa \rightarrow \infty$ ), Eq. (6) becomes

$$w(r) = 2\sqrt{2}K_1\sigma_x\sqrt{1-r^2}. \tag{25}$$

In both cases, a similar development as above yields

$$g(r) = \frac{\ln(1-br^2)}{\ln(1-b)}, \tag{26}$$

where  $b$  is as before, but with  $K_2 = 2\sqrt{2}K_1\sigma_y$ , when  $\sigma_y \ll \sigma_x$ , and  $K_2 = 2\sqrt{2}K_1\sigma_x$  when  $\sigma_y \gg \sigma_x$ . A development parallel to that for Eq. (23) then yields the JND curve

$$JND(r) = k' \left(1/b-r\right). \tag{27}$$

Since  $A(r)$  is always proportional to  $\sqrt{1-r^2}$  (Eq. 16), Eqs. (24) and (25) have an exact match with  $A(r)$  under these conditions. Consequently, Eq. (21) should be a reasonable approximation for most choices of  $\sigma_x$  and  $\sigma_y$ .

### Comparison with previous work

The basic stage of correlation perception in scatterplots has been the focus of many studies over the years. Relatively few, however, examined discrimination. Pollack (1960) and Doherty et al. (2007) characterized this in terms of  $d'$  (signal to noise ratio), finding performance to be better at high correlations; similar results were obtained by Li, Martens, and van Wijk (2010). These are all consistent with the results found here. Indeed, transforming the  $d'$  measures of Experiment 2 of Doherty et al. (2007) into JNDs (these quantities being inversely related) yields a highly linear behavior ( $R^2 = 0.972$ ) that obeys Eq. (1), with variability  $k \approx 0.17$  and bias  $b \approx 0.90$ , values not far from those found here. Harrison et al. (2014) found that Eq. (1) held for several kinds of graphical representation (e.g., line plots), again with parameters broadly similar to those found here.

The more intensively-studied aspect of correlation perception, however, is magnitude estimation. (For reviews, see Boynton, 2000; Doherty et al., 2007; Konarski, 2005). Several studies have proposed particular equations for the relationship of perceived to physical correlation. Table 1 shows the most popular ones (and a few variants) and their fits to the estimation data obtained here. These proposals are grouped according to the number of free parameters they contain:

- i) *Free parameters = 0.* These are simply functions of correlation  $r$ . As is evident from Table 1, average RMSE is about 0.08. Note that  $g=r^2$  (row A) is not the best model for most conditions, indicating that the underlying quantity is unlikely to be simple variance, as was sometimes suggested (e.g., Pollack, 1960; Strahan & Hansen, 1978). More generally, the relatively poor fit for these equations supports the proposal that—even for simple bivariate gaussian distributions—perceived correlation depends upon more than  $r$  alone (Boynton, 2000; Lane et al., 1985).
- ii) *Free parameters = 1.* When a single free parameter is allowed, the best fit is with Eq. (2)—row (J); RMSE here is less than half that of the other models with one free parameter, at least for gaussian distributions. The fit remains good for uniform conditions as well, although the somewhat higher RMSE here suggests that gaussian distributions may in some sense be the more natural ones.<sup>13</sup> Note that the fit is better than for the average perpendicular distance  $X$  (row G); since  $X$  is essentially proportional to the width of the distribution, the somewhat better fit with Eq. (2) would seem to be due to the use of the logarithm and the presence of a residual. The fit also tends to be better than for a power of the distance from  $r = 1$  (row H), or of the ratio of the width of the isofraction ellipse to its length (row I). Note that power laws—which describe the perceived magnitude of most perceptual properties (see e.g., Billock & Tsou, 2011; Ross, 1997)—do not generally fit much better than equations based only on  $r$ . Importantly, the fit with Eq. (2) is also better than for the logarithm of  $1-br^2$ , (row K), the accurate equation for entropy.
- iii) *Free parameters = 2.* Table 1 shows that the double-power law of Cleveland et al. (row M) has the best fit of any two-parameter proposal. This is not entirely a surprise: all things being equal, fit should be better for equations with more free parameters. However, Eq. (2) still provides the best global fit for the basic and low-density conditions, as well as the second-best fit overall, despite having only one free parameter (bias).

<sup>13</sup> This pattern can also be seen in terms of CIs: For the three gaussian distributions, eq. (2) falls within the 95 % CIs for all values tested. For the uniform distribution, this did not occur for one value (the highest); distance beyond the 95 % CI here was 0.01.

All things considered, then, Fechner's Law (Eq. (2)) fits the data at least as well as any other proposal to date. Moreover, it also shows a systematic link between discrimination and estimation, and points to a mechanism that connects it with the rest of visual perception. Note that in this formulation, perceived correlation—especially for gaussian distributions—essentially involves just one parameter, bias  $b$ , which summarizes the effects of all the factors that influence correlation perception, essentially acting as a modulator.

Although care must be taken when comparing the results of different experiments, a few tendencies are apparent. For the low-density condition (Experiment 2),  $k$  is considerably higher (0.30 [0.26, 0.35], vs. 0.21 [0.17, 0.24];  $t(38) = 3.29$ ;  $p = .002$ ), a phenomenon also found by Doherty et al. (2007) and Rensink (2014). This likely reflects the greater sampling noise due to the smaller number of dots. Bias was not noticeably affected, consistent with the pilot studies reported by Bobko and Karren (1979). Although it is sometimes stated that perceived magnitude—and therefore bias—is affected by density (e.g., Boynton, 2000), those studies manipulated density by changing the standard deviation of the dot cloud; when density is manipulated by changing the number of dots present, effects are much weaker (Lauer & Post, 1989; Rensink, 2014). Finally, in the uniform condition (Experiment 4), bias tended to be somewhat higher; if so, it may be either because the uniform distributions did not match the gaussian structure assumed by the perceptual systems involved (Utochkin, 2015; Utochkin & Tiurnia, 2014), or simply because of the greater standard deviation.

### Applications to visualization

The view of correlation perception put forward here has several implications for information visualization. To begin with, the relatively simple nature of Eqs. (1) and (2) suggests an interesting application to the evaluation of designs. Visualization designs are typically evaluated using user studies, which can often be quite time-consuming (Carpendale, 2008). The results here, however, suggest that some aspects of this process—at least for the visualization of correlation using scatterplots—could be done in a considerably faster way. For example, if the datasets being visualized have near-gaussian distributions, and if each level of correlation is equally likely to be encountered, precision can be characterized by the scatter  $S$ , defined as the average JND over the range  $0 \leq r \leq 1$ ; this corresponds to the area under the curve of Eq. (1), or equivalently for this situation, its value at the midpoint  $r = 0.5$ :

$$S = k \left( \frac{1}{b} - \frac{1}{2} \right), \quad 0 < k, b < 1. \quad (28)$$

**Table 1** Fits of perceived correlation

Par	Src	Equation	Basic	Low Density	High Aspect	Uniform
0	A	$g = r^2$	0.059	0.053	0.036	0.088
	B	$g = 1 - \sqrt{1 - r^2}$	0.147	0.150	0.167	0.111
	C	$g = 1 - \sqrt{1 - r}$	0.051	0.047	0.049	0.025
	D	$g = 1 - \sqrt{(1 - r)/(1 + r)}$	0.121	0.116	0.102	0.125
	E	$g = 1 - \sqrt{(1 - r^2)/(1 + r^2)}$	0.051	0.052	0.064	0.023
	F	$g = 1 - (1 - r)^{0.71}(1 + r)^{0.66}$	0.043	0.045	0.064	0.042
1	G	$g = 1 - bX$	0.042	0.039	0.043	0.025
	H	$g = 1 - (1 - r)^a$	0.049	0.044	0.037	0.018
	I	$g = 1 - ((1 - r)/(1 + r))^a$	0.078	0.073	0.064	0.047
	J	$g = \ln(1 - br)/\ln(1 - b)$	<b>0.018</b>	<b>0.013</b>	0.013	0.033
	K	$g = \ln(1 - br^2)/\ln(1 - b)$	0.034	0.033	0.033	0.062
2	L	$g = 1 - bX^a$	0.037	0.025	0.034	0.015
	M	$g = 1 - b(1 - r)^a$	0.031	0.026	0.012	0.011
	N	$g = 1 - (1 - r)^a(1 + r)^b$	0.023	0.016	<b>0.005</b>	<b>0.011</b>
	O	$g = 1 - (1 - br)^a$	0.049	0.044	0.037	0.018
	P	$g = 1 - b((1 - r)/(1 + r))^a$	0.043	0.039	0.025	0.019

RMSE of various proposals for perceived correlation  $g(r)$ ; fits are to the seven levels of perceived correlation measured here. Gray squares indicate best overall fit for each number of free parameters; numbers in bold are the best fits for each condition

Parameters (Par): The number of free parameters in the equation. Source (Src): (A) Variance (Pollack, 1960). (B) Coefficient of alienation  $w(r)$  of Jennings, Amabile, and Ross (1982); based on area of isopleths. (C) Modified version of  $w(r)$ , based on width rather than area. (D)  $g(r)$  of Cleveland et al., 1982; based on the ratio of minor to major axes. (E) Version of  $g(r)$  modified by Meyer et al. (1997). (F) Double-power law of note 15 of Cleveland et al. (1982), with suggested constants  $a = 0.71$  and  $b = 0.66$ . (G) Linear function of average absolute perpendicular distance  $X$  from the regression line (Meyer et al., 1997). (H) Power law for distance from  $r = 1$ , with exponent  $a$  (best fit  $a \approx 0.5$  for most conditions). (I) Power law for ratio of minor to major axes (Boynton, 2000); best fit is  $a \approx 0.38$  for most conditions. (J) Fechner's law (Eq. (2)). (K) Logarithmic function of  $r^2$  instead of  $r$ —the entropy based on the true area of the isopleth ellipse. (L) Power law for average absolute perpendicular distance  $X$ , with two parameters (Meyer et al., 1997). (M) Power law for distance from  $r = 1$ , with two free parameters. (N) Double-power law of Cleveland et al. (1982; note 15). (O) Power law for  $u = 1 - br$ , with exponent  $a$  (best fit  $a \approx 0.5$ ,  $b = 1$  for all conditions). (P) Modified power law for ratio of minor to major axes; variant of the measure proposed in Boynton (2000)

Accuracy can likewise be characterized as the average (under)estimation error  $E$ , corresponding to the average difference between  $g(r)$  and  $r$ . Via Eq. (2), this takes the form:

$$E = \frac{1}{b} - \frac{1}{2} + \frac{1}{\ln(1-b)}, \quad 0 < b < 1. \tag{29}$$

Using these formulae, any scatterplot design (e.g., one with a particular size or color of dots) can be rated in terms of precision and accuracy, at least for the visualization of correlation in near-gaussian distributions: all that is needed are the

parameters  $k$  and  $b$ .<sup>14</sup> Different dot colors, sizes, etc. could likewise be compared. And even if Eqs. (28) and (29) turn out to be valid only for near-gaussian distributions, any results found using these (for different colors, say) might still be applicable for other kinds of distributions.

An important issue is to what extent such evaluation could be sped up. In the experiments here, parameters  $k$  and  $b$  were measured using a fairly large set of base correlations. This was

<sup>14</sup> Clearly, such measures may not be relevant for other uses of scatterplots, such as the visualization of clusters or outliers. They also do not capture aspects such as the aesthetics of the display.

**Table 2** Estimates using different numbers of sampling points ( $n$ ). Upper row for each parameter (in gray) corresponds to the estimates given in the main text

	$n$	Basic	Low Density	High Aspect	Uniform
$k$	-	.21 [.17, 0.24]	.30 [.26, .35]	.22 [.19, .26]	.24 [.20, .28]
	4	.19 [0.15, 0.23]	.31 [.26, .37]	.25 [.18, .32]	.25 [.21, .30]
	3	.20 [0.15, 0.24]	.33 [.27, .39]	.25 [.18, .33]	.26 [.21, .30]
	2	.19 [0.15, 0.23]	.34 [.28, .40]	.26 [.19, .33]	.26 [.21, .31]
$b_{\text{disc}}$	-	.90 [.84, .94]	.91 [.85, .95]	.89 [.84, .92]	.94 [.89, .97]
	4	.89 [.80, .95]	.91 [.85, .95]	.91 [.84, .95]	.95 [.91, .97]
	3	.90 [.84, .95]	.92 [.87, .95]	.92 [.88, .95]	.95 [.90, .97]
	2	.89 [.81, .95]	<b>.95 [.88, .97]</b>	<b>.94 [.91, .96]</b>	.95 [.92, .97]
$b_{\text{est}}$	-	.91 [.85, .95]	.90 [.85, .95]	.83 [.71, .92]	.94 [.91, .97]
	3	.91 [.84, .95]	.90 [.85, .94]	.86 [.76, .93]	.93 [.88, .96]
	1	.90 [.85, .94]	.90 [.85, .94]	.81 [.65, .92]	.90 [.84, .94]

As is apparent, reducing the number of points does not cause estimates to deteriorate greatly. The two significant variations in measurements are in bold; these were for 2-point estimates of  $b_{\text{disc}}$  for low densities and high aspect ratios, which were significantly higher than those for large numbers of sampling points. Square brackets indicate 95 % CIs. To reduce clutter, leading zeroes have been omitted

done to enable the shape of the performance curves to be mapped out in detail. But if conditions are similar to those tested here,  $k$  and  $b$  might be measured using far fewer tests. To examine the feasibility of this, the discrimination data in Experiments 1–4 were reanalyzed using a smaller number of base correlations: either {0.3, 0.5, 0.7, 0.9}, {0.3, 0.6, 0.9}, or {0.3, 0.9}, with JNDs from above. In addition, magnitude estimates were made using 3 subdivisions (corresponding to the first two stages of the method used here), and a single (initial) subdivision.<sup>15</sup> As seen from Table 2, estimates were quite robust, remaining largely the same. The only exceptions were significantly higher values for  $b_{\text{disc}}$  when two base correlations were used on distributions with low densities or high aspect ratios, and a (nonsignificant) trend toward higher values for  $b_{\text{est}}$  when just one subdivision was used on uniform distributions.

Based on these results, a simple method of assessing of  $k$  and  $b$  can be suggested: measure JND from above for three base correlations (e.g., 0.3, 0.6, and 0.9), and fit the results to Eq. (1). Owing to the link between discrimination and perceived magnitude, no bisection is needed, although if the

Fechner assumption is to be tested, a small number of subdivisions (3, or possibly even 1) should suffice. For maximal sensitivity, a within-observer design could be used, with each observer tested on the same set of (counterbalanced) designs.

The developments in this paper also have applications at a more general level. For example, the proposal that correlation is based on probability distributions over an abstract parameter space implies that the values of data points need not be conveyed by spatial position—they could instead be represented by other properties, such as color or orientation (cf. Fig. 11). Given that the perception of correlation in such representations is similar to that found in scatterplots (Rensink, 2014, 2015), and given that such visualizations could take up less space (Fig. 11), there may be practical advantages to their use.

More generally yet, the developments here show that our understanding of visualization can be improved not only via knowledge of the *mechanisms* underlying human perception and cognition (e.g., Card et al., 1999; Ware, 2012), but also via the *methodologies* used to obtain that knowledge. Indeed, developing this approach in a more thorough and systematic way may even result in a science of visualization for some aspects of this domain, an area of research that could connect with several parts of psychology—in particular, vision science (Rensink, 2014),

<sup>15</sup> Consistent with the approach used in the main analysis, values of  $b$  or  $k$  were dropped if they were outside the range of 2.5 standard deviations. However, in this case the associated observers were not replaced.

## Applications to vision science

This study has shown that for gaussian and uniform distributions, the perception of correlation in scatterplots can be described by a pair of simple laws: a linear one for discrimination and a logarithmic one for magnitude. These laws appear to derive from the width of inferred probability distributions, which in turn may reflect the perception of entropy in the image. If so, there would be several important implications for our understanding of the mechanisms underlying visual perception.

The proposal of inferred probability distributions suggests that the mechanisms involved in correlation perception may be related to those underlying ensemble coding (see Alvarez, 2011; Haberman & Whitney, 2012). Early studies of ensemble coding focused on simple scalar properties (estimators), such as mean size (Ariely, 2001), orientation (Dakin & Watt, 1997), and center of mass (Alvarez & Oliva, 2008; Drew, Chubb, & Sperling, 2010). However, it has become increasingly clear that these mechanisms can also respond to the *shape* of the underlying probability distributions, at least over one-dimensional spaces (Chetverikov, Campana, & Kristjansson, 2016; Utochkin, 2015). The results here suggest that the shape of *two-dimensional* distributions can also be determined this way, with this shape serving as the basis of entropy estimation. (Note that this shape need not be used exclusively for this purpose; it might also serve as the basis for other things, such as categorization.) The results also suggest that—similar to the limits on the number of ensemble properties can be determined at a time (Brand, Oriet, & Tottenham, 2012)—limits exist on the number of properties that can be concurrently determined about the shape of such distributions.

Another connection involves the proposal that the goal of the mechanism outlined here is the perception of the entropy in an image. Statistical structure has long been thought to play a critical role in the visual perception of scenes (see e.g., Geisler, 2008; Haberman & Whitney, 2012; Olshausen & Field, 1996; Rensink, 2000), with several statistical quantities apparently perceived quite rapidly (e.g., Fei-Fei, Iyer, Koch, & Perona, 2007; Oliva & Torralba, 2006).<sup>16</sup> Given that entropy is an important statistical structure, it is reasonable to suppose that it too might be such a quantity. Among other things, entropy has been suggested as an important quantity in the guidance of eye movements (Raj et al., 2005; Renninger et al., 2007) as well as a possible measure of clutter in a scene (Rosenholz et al., 2007). The proposal here supports suggestions of this kind; indeed, similar—or even the same—mechanisms may be

<sup>16</sup> Statistical representations in vision can use two different kinds of referencing: one based on sets of *positions*, the other on sets of *items*. Position-based referencing is the basis of representations such as texture (e.g., Bergen & Adelson, 1988); item-based referencing is the basis of ensemble coding. For a scatterplot, correlation stems from the positions *derived* from a set of items (dots); the process does not necessarily use the position of each dot to *reference* it.

involved. And if correlation perception does indeed reflect a form of entropy perception, it would indicate that entropy is not only *used* in perceptual processing, but can be *accessed* by higher-level mechanisms. The use of scatterplots to visually convey structure would then be an interesting example of this ability being harnessed for practical purposes.

In any event, the view proposed here consolidates much of the work on the perception of correlation in scatterplots over the past several decades. In particular, it can account for several key aspects of this process:

- i. The logarithmic form of perceived correlation  $g(r)$ .
- ii. The linear form of  $JND(r)$ , possibly via an intensive quantity (entropy density).
- iii. The linkage between these two quantities (Fechner assumption).
- iv. The relative indifference of these to the presence of individual outliers.
- v. The greater accuracy (lower bias  $b$ ) when the standard deviation of the cloud is smaller.
- vi. The considerable indifference of  $b$  to the density of dots in a scatterplot.
- vii. The invariance of the above in regards to different ways of representing information (e.g., via position or color).

Several predictions also follow:

- i. The two-dimensional shape of probability distributions can be determined in ensemble coding (although only a single aspect of it may be accessible at any one time).
- ii. Perceived correlation  $g(r)$  is a logarithmic function of  $v=I-br^2$  in situations where the two dimensions have markedly unequal standard deviations.
- iii. If the Fechner assumption holds,  $JND(r)$  will be proportional to  $(I/br - r)$  under such conditions.
- iv. For a given property to represent information, a given number of data elements, and a given way to measure  $JND$ , variability  $k$  will be proportional to  $-\ln(I-b)$ .
- v. The above aspects and predictions will hold to the extent that the property used to represent information obeys an *isometric constraint*—equal perceptual distances map to equal distances in numerical (value) space. The properties for which this is possible will likely be related to the basic features of visual perception (e.g., Treisman, 1988).
- vi. Different ways of representing correlation can be compared without any great loss in performance—e.g., the correlation in an augmented stripplot of the type shown in Fig. 11 could be accurately matched against that in a scatterplot, or JNDs would be much the same when measured using scatterplots with different densities or aspect ratios.<sup>17</sup>

<sup>17</sup> Preliminary experiments appear to support this.



- vii. It should be possible to adapt to correlations conveyed by different graphical representations.

### Future directions

Although the view put forward here can explain much of correlation perception, many issues still remain to be investigated. For example, what happens at transitions between positive and negative correlations? When base correlations less than 0.2–0.3 are removed, *post-hoc* analysis shows that behavior for uniform distributions is similar to that of the basic condition (Experiment 4). This suggests that a transition of some kind may exist there, in which the proxy for correlation at high values (entropy) is replaced at low ones by a different one (e.g., density), possibly due to the latter quantity supporting a stronger signal. If so, distinct "zones" for high and low correlations may exist, with interesting effects at their transition points.<sup>18</sup>

Another important set of issues concerns the nature of the data distributions themselves. Although the laws here apply fairly well to at least some non-gaussian distributions, it is not clear how far this goes. Distributions with the same means, standard deviations, and correlations can vary considerably in their structure (Anscombe, 1973). It would be useful to know how far the approach developed here would apply. Another issue is the extent to which a second, irrelevant distribution can affect performance (cf. Konarski, 2005; Lewandowsky & Spence, 1989; Wainer & Thissen, 1979). More generally, it may be worth looking at the extent to which multiple distributions can be separated out, based on the two-dimensional shape of the probability distribution; this might be investigated by an adaptation of current approaches to segmenting probability distributions (e.g., Cohen et al., 2008; Feldman, Singh, & Froyen, 2013; Utochkin, 2015). A related issue is whether the gaussian is a natural distribution for the processes involved, as appears to be the case for ensemble coding (Alvarez, 2011; Utochkin & Tiurnia, 2014). Other issues in this vein include the extent to which nonlinear correlations can be perceived, and whether information entropy could also account for the perception of correlation in higher-dimensional datasets.

A somewhat different set of questions concerns the effectiveness of display factors such as the size and shape of the dot cloud, or the size and shape of its dots to convey correlation (cf. Cleveland & McGill, 1984a). If gaussian distributions are used to test these, the evaluation procedure suggested above could be readily applied and the values of  $k$  and  $b$  measured; once this has been done, Eqs. (28) and (29) could provide quantitative measures of precision and accuracy for each design parameter. Results of this kind would not only be of

practical importance, but might also cast further light on the nature of the perceptual mechanisms involved. Indeed, investigation into the kinds of properties that give rise to laws similar to those found here could provide a new, independent source of insight into the nature of the visual features believed to support the early stages of perception.

In this context it is worth mentioning that the aspect of correlation perception investigated here is its basic stage—i.e., the part that is carried out rapidly and intuitively by most observers. The finding that this is largely complete within 100–150 ms (Rensink, 2014) suggests a similarity to the initial stage of scene perception, where processes are spatially parallel and act rapidly, typically within a few hundred milliseconds (see e.g., Rensink, 2000). If these stages turn out to be identical, some interesting implications follow. For example, the estimates used in the probability distributions could be properties of *proto-objects*—localized structures believed to be created early in visual processing (Rensink & Enns, 1995). If so, correlation perception would have interesting connections to visual search and clutter perception, both of which appear to be based on measurements derived from proto-objects rather than raw pixels in the image (e.g., Rensink & Enns, 1995; Yu, Samaras, & Zelinsky, 2014).

And just as scene perception has an attentional stage that depends on the knowledge of the observer, so does correlation perception have a subsequent stage that supports more sophisticated operations, such as the selection of particular data points (Freedman & Smith). Although the extent to which this stage involves attention is not yet clear, it does appear to require deliberation and is aided by expertise (e.g., Lewandowsky & Spence, 1989). As such, many of the same processes may be involved, further supporting the proposal of a deep connection between vision and visualization (Rensink, 2014).

Clearly, more can be perceived in a scatterplot than just correlation: many other kinds of visual structure are possible. Possible candidates include not only clusters and outliers, but also such things as the convexity, skinniness, or clumpiness of the dot cloud itself (Wilkinson, Ananad, & Grossman, 2005; Wilkinson & Wills, 2008). Techniques analogous to those described here might be developed to explore such possibilities.

Finally, it may be worth emphasizing that research issues in vision science and information visualization are often interlinked: the design of a visualization can often be aided by knowledge of the underlying perceptual mechanisms, while careful investigation into its operation can shed new light on the nature of these mechanisms (Rensink, 2014). This study is one example of how the latter could be done. But examples also exist for other aspects of visualization, such as the perception of average value (e.g., Gleicher, Correll, Nothelfer, & Franconeri, 2013; Legge, Gu, & Lubker, 1989) and the perception of structure in graphs (Cleveland & McGill, 1984b). More generally, the graphical representations used to display data can form a useful class of stimuli for

<sup>18</sup> Such a phenomenon has also been found in pilot experiments on parallel co-ordinate plots (Lane Harrison, personal communication; Kyle Melnick, personal communication).

research into human perception and cognition. It is sometimes believed that artifacts have arbitrary structure, and as such are irrelevant for the study of human perception and cognition. But although humans did not evolve to work with artifacts, artifacts in common use essentially evolved to work with us. Representations such as scatterplots are survivors of considerable competition; there are likely good reasons why they remain in use. Finding those reasons may therefore not only help us better understand the kinds of visualizations that have been or could be developed, but may also help give us new insights into the nature of our perceptual and cognitive systems.

**Acknowledgments** Many thanks to Kyle Melnick and Ben Shear for the huge effort they put into data collection and preliminary analysis. Thanks also to Gideon Baldridge for locating and implementing the comb filter, and to Paulo Apolinar for assistance with several figures. Also thanks to Minjung Kim and the reviewers for helpful feedback on earlier versions of this paper. This work was supported by grants from the Natural Sciences and Engineering Research Council (NSERC) and The Boeing Company.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## References

- Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Science*, *15*, 122–131.
- Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science*, *19*, 392–398.
- Anscombe, F. J. (1973). Graphs in statistical analysis. *American Statistician*, *27*, 17–21.
- Archer, E., Park, I. M., & Pillow, J. W. (2014). Bayesian entropy estimation for countable discrete distributions. *Journal of Machine Learning Research*, *15*, 2833–2868.
- Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, *12*, 157–162.
- Attarha, M., & Moore, C. M. (2015). The capacity limitations of orientation summary statistics. *Attention, Perception, & Psychophysics*, *77*, 1116–1131.
- Ben-Naim, A. (2008). *A Farewell to Entropy: Statistical Thermodynamics Based on Information*. London: World Scientific.
- Bergen, J. R., & Adelson, E. H. (1988). Early vision and texture perception. *Nature*, *333*, 363–364.
- Best, L. A., Hunter, A. C., & Stewart, B. M. (2006). Perceiving relationships: A physiological examination of the perception of scatterplots. In D. Barker-Plummer et al. (eds.), *Diagrams 2006*, pp. 244–257.
- Billock, V. A., & Tsou, B. H. (2011). To honor Fechner and obey Stevens: Relationships between psychophysical and neural nonlinearities. *Psychological Bulletin*, *137*, 1–18.
- Bobko, P., & Karren, R. (1979). The perception of Pearson product moment correlations from bivariate scatterplots. *Personnel Psychology*, *32*, 313–325.
- Box, R., & Lacey, S. (1991). A fast easy sort. *Byte*, *16*, 315–320.
- Boynton, D. M. (2000). The psychophysics of informal covariation assessment: Perceiving relatedness against a background of dispersion. *JEP: HPP*, *26*, 867–876.
- Brand, J., Oriet, C., & Tottenham, L. S. (2012). Size and emotion averaging: Costs of dividing attention after all. *Canadian Journal of Experimental Psychology*, *66*, 63–69.
- Card, S. K., Mackinlay, J. D., Shneiderman, B., & Information visualization. (1999). In S. K. Card, J. D. Mackinlay, & B. Shneiderman (Eds.), *Readings in Information Visualization: Using Vision to Think*. San Francisco: Morgan Kaufman. **Chapter 1**.
- Carnap, R. (1966). In M. Gardner (Ed.), *The Philosophy of Science* (pp. 70–77). New York: Basic Books.
- Carpendale, S. (2008). Evaluating Information Visualizations. In A. Kerren, J. T. Stasko, J.-D. Fekete, & C. North (Eds.), *Information Visualization: Human-Centered Issues and Perspectives* (pp. 19–45). Berlin: Springer. doi:10.1007/978-3-540-70956-5\_2
- Chang, C.-I., Du, Y., Wang, J., Guo, S.-M., & Thouin, P. D. (2006). Survey and comparative analysis of entropy and relative entropy thresholding techniques. *IEE Proceedings - Vision, Image, and Signal Processing*, *153*, 837–850.
- Chetverikov, A., Campana, G., & Kristjánsson, Á. (2016). Building ensemble representations: How the shape of preceding distractor distributions affects visual search. *Cognition*, *153*, 196–201. doi:10.1016/j.cognition.2016.04.018
- Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research*, *43*, 393–404.
- Chong, S. C., & Treisman, A. (2005). Attentional spread in the statistical processing of visual displays. *Perception & Psychophysics*, *67*, 1–13.
- Cleveland, W. S., Diaconis, P., & McGill, R. (1982). Variables on scatterplots look more highly correlated when scales are increased. *Science*, *216*, 1138–1141.
- Cleveland, W. S., & McGill, R. (1984a). The many faces of a scatterplot. *Journal of the American Statistical Association*, *79*, 807–822.
- Cleveland, W. S., & McGill, R. (1984b). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, *79*, 531–554.
- Cleveland, W. S., & McGill, R. (1987). Graphical perception: The visual decoding of quantitative information on graphical displays of data. *Journal of the Royal Statistical Society*, *150*, 192–229.
- Cohen, E. H., Singh, M., & Maloney, L. T. (2008). Perceptual segmentation and the perceived orientation of dot clusters: The role of robust statistics. *Journal of Vision*, *8*(7):6, 1–13.
- Cumming, G. (2012). *Understanding The New Statistics: Effect Sizes, Confidence Intervals, And Meta-Analysis*. New York: Routledge.
- Dakin, S., & Watt, R. J. (1997). The computation of orientation statistics from visual texture. *Vision Research*, *37*, 3181–3192.
- Doherty, M. E., Anderson, R. B., Angott, A. M., & Klopfer, D. S. (2007). The perception of scatterplots. *Perception & Psychophysics*, *69*, 1261–1272.
- Drew, S. A., Chubb, C. F., & Sperling, G. (2010). Precise attention filters for Weber contrast derived from centroid estimations. *Journal of Vision*, *10*, 20. doi:10.1167/10.10.20
- Ellermeier, W., & Faulhammer, G. (2000). Empirical evaluation of axioms fundamental to Stevens's ratio-scaling approach: I. Loudness production. *Perception & Psychophysics*, *62*, 1505–1511.
- Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision*, *7*(10), 1–29.
- Feldman, J., Singh, M., & Froyen, V. (2013). Perceptual grouping as Bayesian mixture estimation. In L. Gepshtein, L. Maloney, & M. Singh (Eds.), *The Oxford Handbook of Computational Perceptual Organization*. Oxford: University Press. doi:10.1093/oxfordhb/9780199829347.013.5
- Fink, M., Haunert, J.-H., Spoerhase, J., & Wolff, A. (2013). Selecting the aspect ratio of a scatter plot based on its Delaunay triangulation.

- IEEE Transactions on Visualization and Computer Graphics*, 19, 2326–2335.
- Freedman, E. G., & Smith, L. D. (1996). The role of data and theory in covariation assessment: Implications for the theory-ladenness of observation. *Journal of Mind and Behavior*, 17, 321–344.
- Geisler, W. S. (2008). Visual perception and the statistical properties of natural scenes. *Annual Review of Psychology*, 59, 167–192.
- Gleicher, M., Correll, M., Nothelfer, C., & Franconeri, S. (2013). Perception of average value in multiclass scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, 19, 2316–2325.
- Gokhale, D. V., Ahmed, N. A., & Res, B. C. (1989). Entropy expressions and their estimators for multivariate distributions. *IEEE Transactions on Information Theory*, 35(3), 688–692. doi:10.1109/18.30996
- Haberman, J., & Whitney, D. (2012). Ensemble Perception: Summarizing the scene and broadening the limits of visual processing. In J. Wolfe & L. Robertson (Eds.), *From Perception to Consciousness: Searching with Anne Treisman* (pp. 339–349). Oxford: University Press.
- Harrison, J. (1995). Combsort11 algorithm. <http://www.cs.ubc.ca/~harrison/Java/CombSort11Algorithm.java.html>
- Harrison, L., Yang, F., Franconeri, S., & Chang, R. (2014). Ranking visualizations of correlation using Weber's law. *IEEE Transactions on Visualization and Computer Graphics*, 20, 1943–1952.
- Jennings, D., Amabile, T. M., & Ross, L. (1982). Informal covariation assessment: Data-based vs. theory-based judgments. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment Under Uncertainty: Heuristics And Biases* (pp. 211–230). New York: Cambridge University Press.
- Johnson, R. (2015). Calculating the length of the semi-major axis from the general equation of an ellipse. <http://math.stackexchange.com/questions/1227369/calculating-the-length-of-the-semi-major-axis-from-the-general-equation-of-an-el>
- Kay, M., & Heer, J. (2015). Beyond Weber's Law: A second look at ranking visualizations of correlation. *IEEE Transactions on Visualization and Computer Graphics*, 22(1), 469–478.
- Kirk, R. E. (1995). *Experimental Design: Procedures for the Behavioral Sciences* (3rd ed., pp. 37–40). Boston: Brooks-Cole.
- Konarski, R. (2005). Judgments of correlation from scatterplots with contaminated distributions. *Polish Psychological Bulletin*, 36, 51–61.
- Lane, D. M., Anderson, C. A., & Kellam, K. L. (1985). Judging the relatedness of variables: The psychophysics of covariation detection. *Journal of Experimental Psychology: Human Perception and Performance*, 11, 640–649.
- Lauer, T. W., & Post, G. V. (1989). Density in scatterplots and the estimation of correlation. *Behaviour & Information Technology*, 8, 235–244.
- Legge, G. E., Gu, Y., & Lubker, A. (1989). Efficiency of graphical perception. *Perception & Psychophysics*, 46, 365–374.
- Lemons, D.S. (2013). *A Student's Guide to Entropy*. Cambridge: University Press. Ch. 8. (pp. 140–158).
- Lewandowsky, S., & Spence, I. (1989). Discriminating strata in scatterplots. *Journal of the American Statistical Association*, 84, 682–688.
- Li, J., Martens, J.-B., & van Wijk, J. J. (2010). Judging correlation from scatterplots and parallel coordinate plots. *Information Visualization*, 9, 13–30.
- Li, J., van Wijk, J.J., & Martens, J.-B. (2009). Evaluation of symbol contrast in scatterplots. *IEEE Pacific Visualization Symposium*, pp. 97–104.
- Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: Freeman.
- Meyer, J., & Shinar, D. (1992). Estimating correlations from scatterplots. *Human Factors*, 34, 335–349.
- Meyer, J., Taieb, M., & Flascher, I. (1997). Correlation estimates as perceptual judgments. *Journal of Experimental Psychology: Applied*, 3, 3–20.
- Morgan, M. J. (2005). The visual computation of 2-D area by human observers. *Vision Research*, 45, 2564–2570.
- Norwich, K. H. (1987). On the theory of Weber fractions. *Perception & Psychophysics*, 42, 286–298.
- Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*, 155, 23–36.
- Olshausen, B. A., & Field, D. J. (1996). Natural image statistics and efficient coding. *Network: Computation in Neural Systems*, 7, 333–339.
- Pollack, I. (1960). Identification of visual correlational scatterplots. *Journal of Experimental Psychology*, 59, 351–360.
- Raj, R., Geisler, W. S., Frazor, R. A., & Bovik, A. (2005). Contrast statistics for foveated visual systems: Fixation selection by minimizing contrast entropy. *J. Opt. Soc. Am A*, 22, 2039–2049.
- Renninger, L. W., Verghese, P., & Coughlan, J. (2007). Where to look next? Eye movements reduce local uncertainty. *Journal of Vision*, 7, 6. doi:10.1167/7.3.6
- Rensink, R. A. (2000). The dynamic representation of scenes. *Visual Cognition*, 7, 17–42.
- Rensink, R.A. (2012). Invariance of correlation perception. *Journal of Vision*, 12: 433. <http://www.journalofvision.org/content/12/9/433.short> [Vision Sciences Society, Sarasota, FL, USA. May 2012.]
- Rensink, R. A. (2014). On the prospects for a science of visualization. In W. Huang (Ed.), *Handbook of Human Centric Visualization* (pp. 147–175). New York: Springer.
- Rensink, R.A. (2015). Visual features as carriers of information. *Journal of Vision*, 15: 890. [Vision Sciences Society, St. Petersburg, FL, USA. May 2015.] doi: 10.1167/15.12.893
- Rensink, R. A., & Baldrige, G. (2010). The perception of correlation in scatterplots. *Computer Graphics Forum*, 29, 1203–1210.
- Rensink, R. A., & Enns, J. T. (1995). Preemption effects in visual search: Evidence for low-level grouping. *Psychological Review*, 102, 101–130.
- Robitaille, N., & Harris, I. M. (2011). When more is less: Extraction of summary statistics benefits from larger sets. *Journal of Vision*, 11(18), 1–8.
- Rosenholz, R., Li, Y., & Nakano, L. (2007). Measuring visual clutter. *Journal of Vision*, 7, 17. doi:10.1167/7.2.17
- Ross, H. E. (1997). On the possible relations between discriminability and apparent magnitude. *British Journal of Mathematical and Statistical Psychology*, 50, 187–203.
- Strahan, R. F., & Hansen, C. J. (1978). Underestimating correlation from scatterplots. *Applied Psychological Measurement*, 2, 543–550.
- Thomas, J.J., & Cook, K.A. (2005). *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Center.
- Timm, N. H. (2002). *Applied Multivariate Analysis* (pp. 85–86). New York: Springer.
- Tommasi, M. (2000). Bisections on backgrounds with different luminances. *Review of Psychology*, 7, 15–18.
- Treisman, A. (1988). Features and objects: The fourteenth Bartlett memorial lecture. *Quarterly Journal of Experimental Psychology*, 40A, 201–237.
- Treutwein, B. (1995). Adaptive psychophysical procedures. *Vision Research*, 15, 2503–2522.
- Utochkin, I. S. (2015). Ensemble summary statistics as a basis for rapid visual categorization. *Journal of Vision*, 15, 8. doi:10.1167/15.4.8
- Utochkin, I. S., & Tiurnia, N. A. (2014). Parallel averaging of size is possible but range-limited: A reply to Marchant, Simons, and De Fockert. *Acta Psychologica*, 146, 7–18.
- Wainer, H., & Thissen, D. (1979). On the robustness of a class of naive estimators. *Applied Psychological Measurement*, 3, 543–551.

- Ware, C. (2012). *Information Visualization: Perception for Design* (3rd ed.). New York: Morgan Kaufmann.
- Weber, B. H., Depew, D. J., & Smith, J. D. (1988). *Entropy, Information, and Evolution: New Perspectives on Physical and Biological Evolution*. Cambridge: MIT Press.
- Wilkinson, L., Anand, A., & Grossman, R. (2005). Graph-theoretic scagnostics. In *IEEE Symposium on Information Visualization, 2005*, 157–164.
- Wilkinson, L., & Wills, G. (2008). Scagnostics distributions. *Journal of Computational and Graphical Statistics, 17*, 473–491.
- Yu, C.-P., Samaras, D., & Zelinsky, G. J. (2014). Modeling visual clutter perception using proto-object segmentation. *Journal of Vision, 14*, 1–16.
- Zhang, H., Daw, N. D., & Maloney, L. T. (2015). Human representation of visuo-motor uncertainty as mixtures of orthogonal basis distributions. *Nature Neuroscience, 18*, 1152–1158. doi:10.1038/nn.4055
- Zhu, S. C., Wu, Y., & Mumford, D. (1998). *International Journal of Computer Vision, 27*, 107–126.
- Zimmer, K., & Ellermeier, W. (2006). Axiomatic approaches to Stevens' magnitude scaling: Recent developments. *Proceedings of the Annual Meeting of the International Society for Psychophysics, 22*, 49–56.