Difference magnitude is not measured by discrimination steps for order of point patterns

Emmanouil D. Protonotarios
Centre for Mathematics, Physics and Engineering in the Life Sciences and Experimental Biology and Department of Computer Science, University College London, London, UK

Alan Johnston
School of Psychology, University of Nottingham, Nottingham, UK

Lewis D. Griffin
Centre for Mathematics, Physics and Engineering in the Life Sciences and Experimental Biology and Department of Computer Science, University College London, London, UK

We have shown in previous work that the perception of order in point patterns is consistent with an interval scale structure (Protonotarios, Baum, Johnston, Hunter, & Griffin, 2014). The psychophysical scaling method used relies on the confusion between stimuli with similar levels of order, and the resulting discrimination scale is expressed in just-noticeable differences (jnds). As with other perceptual dimensions, an interesting question is whether suprathreshold (perceptual) differences are consistent with distances between stimuli on the discrimination scale. To test that, we collected discrimination data, and data based on comparison of perceptual differences. The stimuli were jittered square lattices of dots, covering the range from total disorder (Poisson) to perfect order (square lattice), roughly equally spaced on the discrimination scale. Observers picked the most ordered pattern from a pair, and the pair of patterns with the greatest difference in order from two pairs. Although the judgments of perceptual difference were found to be consistent with an interval scale, like the discrimination judgments, no common interval scale that could predict both sets of data was possible. In particular, the midpattern of the perceptual scale is 11 jnds away from the ordered end, and 5 jnds from the disordered end of the discrimination scale.

Introduction

The notion of order appears in Gestalt psychology (Koffka, 1935), and is related to arrangements of objects where their relative placement has significance. As such, order may consist of specific regularities (laws), and these may interact synergistically (Wagemans, Wichmann, & de Beeck, 2005). It was suggested that the visual system seeks order (e.g., symmetry) so as to make sense of sensory signals. The ability of the visual system to detect regularities has been proposed as a method of compressing information to reduce redundancy (Attneave, 1954). Gestalt principles considered under this view are functions of the perceptual machinery that group information together and thus provide an economical description of visual reality. Order in the form of repetitive and symmetrical patterns is also aesthetically preferred (e.g., Newell, Murtagh, & Hutzler, 2013) and has been utilized historically for artistic and architectural purposes. In nature, perfect order rarely exists, and the visual system has to cope with intermediate states of order. Subjective assessments of the degree of order are common in everyday life but also in scientific research when visual patterns are examined (Cohen, Baum, & Miodownik, 2011; Cook, 2004; Marinari et al., 2012). Partial order is particularly relevant to biology and biomedicine, because living systems tend to be well-, but not perfectly, ordered. For example, in the mammalian eye, the spacing of the parafoveal receptors is less than perfectly regular, preventing Moiré-like aliasing (Wässle & Boycott, 1991). On the other hand, generation of disorder is associated with aging and disease (Guillaud et al., 2004; Hu, Li, Wang, Gou, & Fu, 2012; Sudb, Marcelpoil, & Reith, 2000). As a

Citation: Protonotarios, E. D., Johnston, A., & Griffin, L. D. (2016). Difference magnitude is not measured by discrimination steps for order of point patterns. *Journal of Vision*, *16*(9):2, 1–17, doi:10.1167/16.9.2.

doi: 10.1167/16.9.2 Received June 8, 2015; published July 5, 2016 ISSN 1534-7362

texture attribute, the degree of order is essential for texture discrimination and segmentation (Bonneh, Reisfeld, & Yeshurun, 1994; Ouhnana, Bell, Solomon, & Kingdom, 2013; Vancleef et al., 2013). Order also interacts with other perceptual dimensions, and its precise control in stimuli configurations is crucial for psychophysical experiments. For example, the degree of positional order of elements is known to affect perceived numerosity (Allik & Tuulmets, 1991; Dakin, Tibber, Greenwood, Kingdom, & Morgan, 2011; Ginsburg & Goldstein, 1987). For contour integration tasks, patterns of intermediate positional order should be carefully generated in ways to avoid density cues (Demeyer & Machilsen, 2012; Machilsen, Wagemans, & Demeyer, 2015).

Not much is known about the visual mechanisms for perceiving partial order. Recent work on perception of regularity in point patterns has investigated the effect of sensory noise on discrimination thresholds (Morgan, Mareschal, Chubb, & Solomon, 2012), while Ouhnana et al. (2013) demonstrated that regularity is an adaptable visual dimension. Using a filter-rectify-filter model (Graham, 2011), they argued that regularity is encoded via the peakedness of the distribution of the energy responses across receptive field size.

A fundamental question in the analysis of a perceptual attribute (e.g., lightness, glossiness) is whether we can represent quantitatively the intensity of the percept as a value on an interval scale. This is essential, as then almost all statistical measures are applicable (Stevens, 1946). While for some established perceptual dimensions and for limited ranges, the ordinal, interval, or even the ratio structure of the dimension is considered obvious, this should not be taken for granted for all attributes. In particular, for attributes like order where an observer may employ varying criteria, and/or the intensity of the percept may depend on the interaction of multiple parameters, even the property of transitivity is not guaranteed, and consequently even a simple ranking of the stimuli based on the attribute may not be possible.

Therefore, for the collection of data, indirect scaling methods are preferred to direct methods since they allow validation of the empirical relationships between the stimuli, and so it can be tested whether the collected judgments are consistent with a specific type of scale (Hand, 2004).

Two methods of indirect scaling can be distinguished. The first one—magnitude comparison—is based on pairwise comparisons between stimuli and originates with Thurstone (1927). Judgments are collected from a series of presented pairs of stimuli where the observer selects the one that contains the attribute in question at the greatest degree. Analysis of such judgment data yields a discrimination-based scale. The second method—magnitude difference comparison—has been considered by Maloney and Yang (Knoblauch & Maloney, 2008; Maloney & Yang, 2003) and requires observers to consider quadruples of stimuli that define two intervals and to select the pair of stimuli that shows the greater perceptual difference. This method has been applied to a range of perceptual dimensions: color (Brown, Lindsey, & Guckes, 2011; Lindsey et al., 2010; Maloney & Yang, 2003; Yang, Szeverenyi, & Ts'o, 2008), quality of compressed images or video (Charrier, Knoblauch, Maloney, & Bovik, 2011; Charrier, Knoblauch, Maloney, Bovik, & Moorthy, 2012; Charrier, Maloney, Cherifi, & Knoblauch, 2007; Menkovski & Liotta, 2012), surface texture (Emrith, Chantler, Green, Maloney, & Clarke, 2010), gloss (Obein, Knoblauch, & Viénot, 2004), transparency (Fleming, Jaekel, & Maloney, 2011), strength of the watercolor effect (Devinck, Gerardin, Dojat, & Knoblauch, 2014; Devinck & Knoblauch, 2012), similarity between pairs of faces (Rhodes, Maloney, Turner, & Ewing, 2007), correlation in scatterplots (Knoblauch & Maloney, 2008), auditory stimulus duration (Yang et al., 2008), and emotional intensity (Junge & Reisenzein, 2013). In this article we investigate whether there is a single internal interval scale of order for point patterns that underlies the perception of order both for discrimination and perceptual difference tasks.

For the construction of perceptual interval scales from indirect scaling data, one supposes the existence of a perceptual continuum where the true values of the attribute in question reside. Each time a judgment concerning a configuration of stimuli is performed, noisy realizations of the true values of the attributes are perceived and then compared, either directly or after a differencing step. Assuming the realization noise is stationary along the scale, and each perceptual realization is independent, a link function with a sigmoid form converts scale value differences, or differences of differences, into response probabilities. The form of noise distribution determines the exact shape of the link function. With a data set of responses to a set of scaling tasks, which has been well chosen with respect to stimuli spacing and is sufficiently numerous, the noisy realization model can be critically tested using likelihood analysis. If the model is found to account adequately for the data, scale values for the experimental stimuli can be estimated by maximum likelihood model fitting.

Due to their simplicity, point patterns constitute a convenient stimulus class for investigating perception of order. They are commonly displayed using circularly symmetric elements (e.g., dots, Gaussian blobs). Figure 1 shows example point patterns of varying order.

Point patterns are used in various scientific fields (e.g., ecology, developmental biology, material science) to represent systems where the focus is on the position

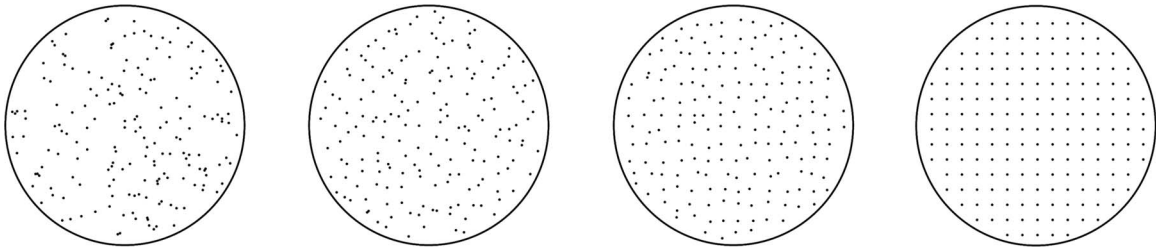

Figure 1. Point patterns exhibiting varying degree of order. All patterns of the Figure are based on the same square lattice of points, and different levels of order have been attained by varying the amount of positional jitter of the points.

of the elements. While subjective assessments of the degree of order are frequently employed, objective quantification is required for systematic analysis. However, apart from piecemeal approaches (e.g., Cliffe & Goodwin, 2013; Dunleavy, Wiesner, & Royall, 2012; Sausset & Levine, 2011; Steinhardt, Nelson, & Ronchetti, 1983; Truskett, Torquato, & Debenedetti, 2000), no generally accepted mathematical theory exists for intermediate order. Providing a mathematical scale of order as an objective surrogate of human perception seems a promising alternative. In previous work (Protonotarios, Baum, Johnston, Hunter, & Griffin, 2014) we have shown, using pairwise comparisons of point patterns from a diverse set, that observers are highly consistent in their judgments of order, and that these are compatible with an interval scale structure. This means that the preference frequencies with respect to order for pairs of point patterns of the whole set can be predicted based on the distances between the attribute values on this scale.

The analysis that yielded the scale made use of a logistic link function (Bradley–Terry model; Bradley & Terry, 1952). The derived scale is thus expressed in just-noticeable differences (jnds), and its construction relies on the confusion between adjacent stimuli. The interval structure of the responses is important as it allows the use of such judgments as a basis for quantification of order for scientific purposes. By examining a list of preexisting and designed geometrical measures of order, we identified one that correlates very well with human perception. This measure assesses the variability of the spaces between the points, taking into account the distributions of the sizes and shapes of the triangles defined by their Delaunay triangulation (Delaunay, 1934). Rescaling an interval scale X with a linear transformation $aX + b$, where a, b are real constants with $a > 0$, does not distort the interval character, since the sign and the relative size of differences are preserved. We therefore transformed the output of the geometric measure so that certain significant patterns are anchored to memorable values. We called the resulting scale an absolute interval scale (a-scale) for the measurement of order. On this scale, the anchored values 0 and 10 correspond to total randomness (Poisson point patterns) and perfect Bravais lattice, respectively, and each unit corresponds roughly to a jnd. We demonstrated its applicability by identifying two distinct processes in the pattern formation of the *Drosophila* bristle cells during development.

Although comparisons of all pairs of patterns were used in the construction of this scale, the process relies particularly on judgments of subthreshold differences (discrimination), while suprathreshold differences do not affect the scaling apart from validating its ordinal structure. A natural question, therefore, is whether large intervals on this scale correspond to direct perceptual differences (Luce & Krumhansl, 1988). There is no fundamental reason why these two types of judgment should depend on the same sensory and/or cognitive mechanism and therefore be predicted by a common perceptual scale. However, if these depend on the same visual mechanism and there is a common underlying perceptual scale, then the derived scales from the two tasks would be in agreement only if internal noise remains constant across the perceptual dimension (Kingdom, 2009). Contrary to discrimination-based scales, it has been shown that the scales derived from suprathreshold judgments of perceptual differences are robust with respect to assumptions of constant or varying internal noise (Kingdom & Prins, 2009).

In the following sections we present two experiments we conducted for the collection of magnitude comparison and magnitude difference comparison data for the perception of order in point patterns and the construction of the corresponding scales. We then investigate whether a common interval scale can account for the data of both experiments, and therefore whether appearance and discrimination form a consistent basis for quantification of order.

Methods

Stimuli

The space of approximately ordered point patterns is vast. In our previous work (Protonotarios et al., 2014) we used a multistep process to synthesize diverse

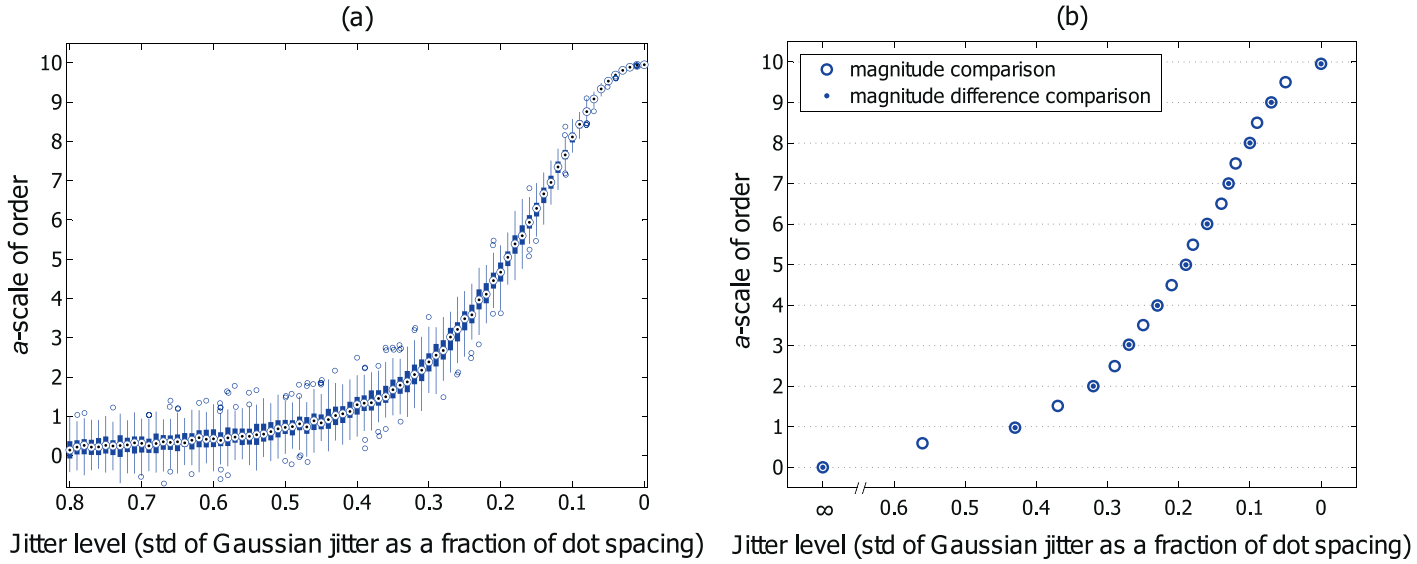

Figure 2. (a) Boxplot of the *a*-scale values of order for the generated point patterns for varying jitter level. The latter is expressed as the standard deviation of the Gaussian perturbation of the points as a fraction of the spacing of the perfect square grid. (b) *a*-scale values for the selected point patterns for the magnitude and the difference comparison task against the corresponding jitter level. The horizontal axis has been inverted in both graphs so that high-order values are on the right-hand side.

examples from this space. Starting from perfect lattices of points, which could be triangular, rectangular, or hexagonal, we independently perturbed the position of the points, randomly deleted and added varying proportions of them, and finally applied a smooth nonlinear positional warp. In this article, however, we are not concerned with a thorough investigation of the interaction of these factors in the final percept of order but rather aim to compare the two methods of scaling. It would be sufficient to demonstrate a disagreement between the two scaling methods in a subset of the space of point patterns. Additionally, we noticed in pilot experiments that in magnitude difference comparison, the effect of secondary parameters of the stimuli (e.g., type of lattice) can be significant. In particular, a quadruple of patterns may appear where one pair exhibits a large difference in the degree of order but having both patterns based on the same type of lattice (e.g., triangular), while the other pair exhibits a smaller difference in order but having its patterns based on different types of lattice (e.g., triangular and square). In this case, it is possible that the participant may be influenced by the similarity of the first pair with respect to the secondary parameter and judge the second pair as the one showing the larger difference.

Given these considerations we chose a simple method for the generation of stimuli with varying degree of order and minimum presence of secondary perceptual dimensions. We started with a point pattern based on a perfect square lattice and introduced independent Gaussian positional jitter on the points. The physical parameter that defined the strength of the jitter was the standard deviation of the Gaussian distribution (on each coordinate) expressed as a fraction of the spacing of the perfect square lattice. The final step was a random selection of a circular window containing exactly 180 points.

For our experiment to be effective, it is ideal if we use a set of stimuli that are regularly spaced in perceptual order. Even with the simple jittered square lattice we employ, this is not trivial to achieve for two reasons. The first is that patterns can vary in how ordered they appear even at a fixed level of jitter. The second is that it is unjustified to expect that average perceptual order depends linearly on jitter magnitude. We have dealt with those problems by using the geometric algorithm developed in our previous work. The *a*-scale algorithm predicts the perceptual order of patterns on a discrimination-based interval scale.

By generating many patterns at different jitter levels and computing their *a*-scale values, we have determined 21 jitter levels that on average are uniformly spaced on the *a*-scale. We then selected 21 patterns that have close to the mean *a*-scale value for their level of jitter. Figure 2a shows the boxplot of the *a*-scale order values for the generated point patterns for varying jitter levels. A set of 100 patterns was generated at each jitter level. The spread in *a*-scale values is larger for higher jitter levels. Figure 2b shows the *a*-scale values and jitter levels of the selected patterns. In all figures with jitter level on the abscissa, the axis has been inverted so that scale values corresponding to high order values are on the right-hand side. The patterns at the highest jitter level (disordered end) were generated from a Poisson process, which is equivalent to applying an infinite amount of jitter to a square lattice. We additionally excluded patterns that contained points that were so

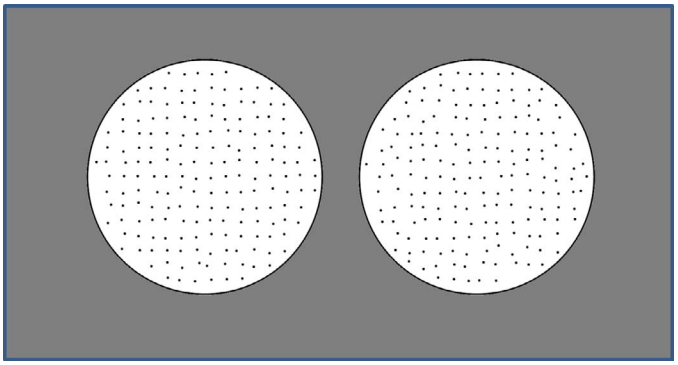

Figure 3. Magnitude comparison task. Participants were asked to select the pattern (left or right) that appears more ordered to them. For this example, the left pattern gained 17 responses and the right pattern gained seven.

close that they would overlap when displayed as dots; this happened with increasing frequency for larger amounts of jitter.

For the magnitude comparison task, the whole set of 21 patterns was used (Set A). For the magnitude difference comparison task the 11 odd-numbered patterns of A were used (Set B) (Figure 2b). The number of patterns used was chosen so that the resulting number of pairwise comparisons for a balanced task is not experimentally prohibitive while giving a relatively dense set. This is necessary since the scaling method based on the discrimination task relies on overlap of the estimated values of order.

Observers

Twelve observers (five females) with normal or corrected-to-normal vision participated in both of the experiments. Participants' age ranged from 18 to 40 years ($M = 25.4$, $SD = 6.0$ years). Our research adhered to the tenets of the Declaration of Helsinki for the protection of human subjects.

Procedure

The patterns were presented in pairs or quadruples depending on the task at a comfortable viewing distance of approximately 50 cm on a 40-cm diagonal laptop screen of resolution 1920×1080. They were presented as solid black dots of 0.5-mm diameter, on white circular disks of radius $r = 4.0$ cm, on a gray background. Participants viewed the patterns under comfortable room illumination. In figures in this article, the size of the dots in the patterns has been increased for visibility in reproduction. Presentation of stimuli and recording of responses were controlled using the MATLAB Psychtoolbox (Brainard, 1997). In both tasks participants were given unlimited time to respond for each judgment.

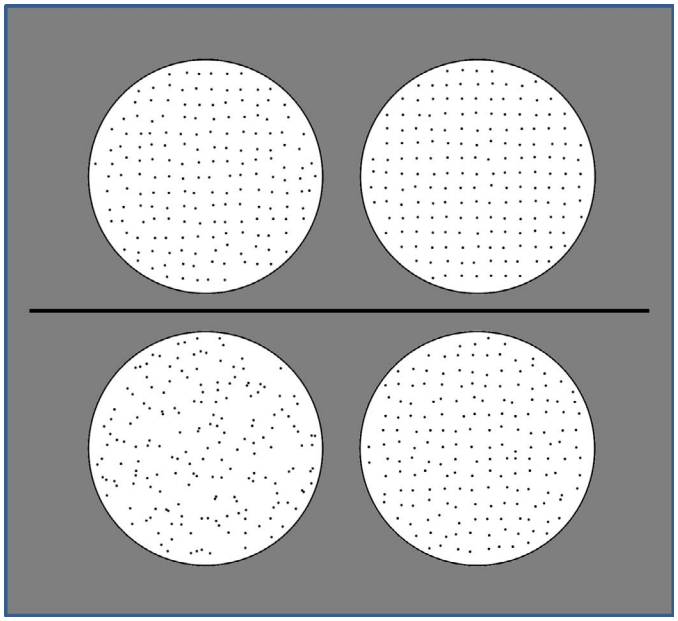

Figure 4. Magnitude difference comparison task. Participants were asked to select the pair of patterns (upper or lower) that shows the greatest perceptual difference in order. For this example, the upper pair gained three responses and the lower pair gained 21.

Magnitude comparison experiment

Observers first completed the magnitude comparison task (discrimination). Each of the $21 \times 20/2 = 210$ pairs was presented in random order in two blocks (Figure 3), resulting in 420 comparisons in total per participant. In each trial the patterns were randomly allocated to left or right in the first block and then in the opposite way in the second block. Each pattern was randomly oriented at integral multiples of 90°. Randomization aimed to minimize learning of the patterns and to reduce bias and effects of adaptation.

Participants were given written instructions to use the keyboard (left or right arrow keys) to indicate which of the patterns appeared more ordered to them. They were able to return to the previous trials if they wished to correct a keystroke error. They were free to control the pace of the experiment, but all took 12–20 min.

Magnitude difference comparison experiment

The magnitude difference comparison task followed the magnitude comparison task after a short break. Observers were presented with quadruples of patterns arranged as two horizontal pairs separated by a thick black horizontal line (Figure 4). Each of the quadruples was presented in random order in two blocks. The two

pairs were randomly allocated to the upper or lower part of the screen in the first block and the opposite way in the second block (Knoblauch & Maloney, 2012). Each pattern was again randomly oriented at integral multiples of 90°. Randomization, as in the first experiment, aimed to minimize learning of the patterns and to reduce bias and effects of adaptation. With each pair defining an interval on the perceptual continuum, all quadruples with nonoverlapping pairs of intervals were presented. This gives 330 distinct quadruples, and since all were presented twice, the total number of comparisons for the magnitude difference comparison task per participant was 660.

Participants were given written instructions to use the keyboard (up or down arrow keys) to indicate which of the pairs of patterns showed the largest difference in the degree of order. They were able to return to the previous trials if they wished to correct a keystroke error. They were free to control the pace of the experiment, but all took 30–50 min.

The method of quadruples for the magnitude difference scaling is the original suggested by Maloney and Yang (2003). However, the variant of triads has also been used for the construction of difference scales (Devinck & Knoblauch, 2012). In this case the observer still judges intervals directly but a single stimulus serves at the same time as both the high endpoint of the one interval and the low endpoint of the second interval. In the triads method, the number of possible comparisons is lower, but this does not provide any benefit since the standard deviation of the estimates depends on the number of actual trials (Maloney & Yang, 2003). Furthermore, in the triad method direct comparison is conducted only for adjacent intervals and not for intervals that are further apart.

Results

Agreement rates

We computed two measures of response variability, the intra- and interobserver agreement rates, shown in Table 1. The intra-agreement rate expresses the probability that a random participant would repeat the same judgment when faced twice with the same random trial. Observers do not always respond the same to a pair or quadruple of patterns. The rate of 91% for magnitude comparison with Set A demonstrates that there is high consistency. If we restrict the estimation of the intra-agreement rate by subsampling on the Set B, we observe a slight increase to 93%. This increase is expected since the fewer patterns of Set B are more widely spaced in order. The interagreement rate is the probability that two observers will agree on the same random trial. For the magnitude comparison task, the 91% for Set A, and the 93% for Set B, show that there is no variation between participants over and above their personal variability. For the magnitude difference comparison task, the agreement rates are lower. The intra rate is 76%, indicating lower consistency of observers in comparison to the discrimination task and the inter rate is 70%, indicating an additional 6% of interpersonal difference over and above personal variability.

	Magnitude comparison (%)	Magnitude difference comparison (%)
Set A (21 patterns)		
Intra	91	—
Inter	91	—
Set B (11 patterns)		
Intra	93	76
Inter	93	70

Table 1. Agreement rates for the two experiments.

Magnitude comparison scaling

Model fitting

For the magnitude comparison task, each stimulus S_i is assumed to have a true value M_i on a scale, and each separate perception ψ_i of it is a noisy realization of the true value. When a magnitude comparison between two stimuli S_i, S_j takes place, the two noisy realizations are compared and the observer reports which one is larger (or smaller). Assuming the noise for each perception is identically distributed and independent, then for magnitude comparison, there exists a monotonic preference function P: $\mathbb{R} \rightarrow [0, 1]$, which maps the signed difference between the true values, $\Delta M = M_i - M_j$, to the probability that the one will be preferred to the other. This preference function is the cumulative distribution function of the realization noise distribution convolved with itself. When the noise distributions of the set have sufficient overlap, then the preference rates will not all be 0 or 1 and fitting a model to the preference rate data allows the interval, not just ordinal, structure of the true values of the set to be estimated. If the noise is assumed normally distributed (Thurstone Model Case V) the preference function has the form of cumulative Gaussian distribution function (Thurstone, 1927), whereas if, for example, the noise is assumed Gumbel-distributed (Bradley–Terry Model), then the preference function has a logistic form (Bradley & Terry, 1952; David, 1988). In this article we will only consider Gaussian noise. However, this method of scaling is relatively robust to noise distribution assumptions (Stern, 1992).

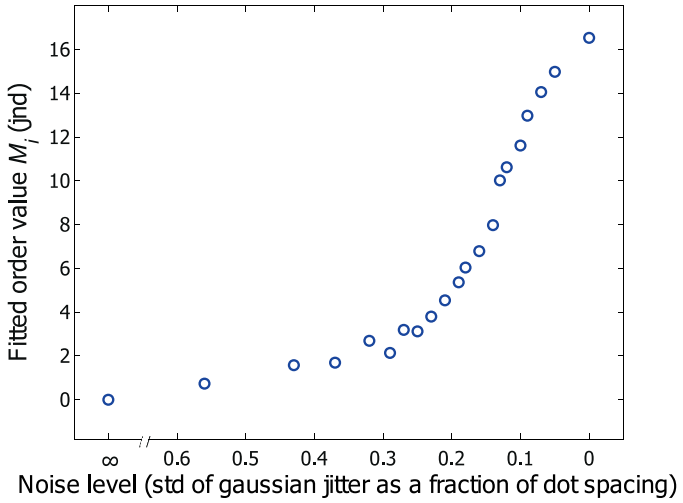

Figure 5. Fitted values of order based on the discrimination task plotted against the amount of jitter. The latter is expressed as the standard deviation of the Gaussian perturbation of the dots as a fraction of the dot spacing of the perfect square grid. For the Poisson pattern (most disordered) the equivalent amount of jitter is infinite.

We fit models by maximization of the likelihood (ML) of the data. The likelihood is computed using binomial probabilities for each stimulus pair. It has been shown that ML fitting is extremely sensitive to probabilities near 0 or 1 (Harvey, 1986; Wichmann & Hill, 2001a). It is possible that during the psychophysical experiment errors that are independent of the attribute differences occur (lapses), even when the difference in the attributes between two stimuli is large. The effect of lapsing appears at the data near the edges of the preference function and can result in strong biases in the estimates. Incorporating the lapse rate in the model corrects this problem. This is easily done by rescaling the preference function as $\lambda + (1 - 2\lambda)P(\Delta M)$, where λ, positive but small, is the lapse rate parameter (Wichmann & Hill, 2001a). The model is parameterized by (a) the unknown true values of each stimulus, (b) the lapse rate parameter, and (c) any additional parameters used to vary the form of P. We used gradient descent for model fitting, with multiple random starts to check for stability. The equal variance Gaussian noise model we used required no additional parameters.

Since the interval scale is invariant under linear transformation, we can choose the unit distance on the scale to correspond to a jnd. By convention, the magnitude of a jnd on a discrimination scale is such that an observer will have a 75% chance of correctly ordering two stimuli whose attributes are different by this amount (Torgerson, 1958). On the Thurstone scale, the jnd width does not refer to a change of a physical stimulus intensity that yields a particular discrimination performance, but to a specific spacing of two stimuli that corresponds to the fixed preference rate. In this model the noise intensity (standard deviation of the Gaussian noise distribution) and the jnd width are related and remain constant across the scale. The fitted values for Set A with respect to the physical parameter (amount of positional jitter of the points) are shown in Figure 5. This figure shows the estimated degree of perceived order on the discrimination scale in jnd units; the degree of perceived order increases as the amount of jitter decreases. Since only differences in fitted values, not absolute values, are used to predict preference rates, ML fitting only estimates values up to an additive constant for the entire set. In particular, the value zero has been selected for the Poisson pattern, which is judged as the most disordered. The estimated value for the most ordered pattern is 16.5 jnds, meaning that the whole range of order for Set A is 16.5 jnd units. The ML estimated lapse rate parameter was 0.3%. We note that 16.5 jnds is larger than the 10 jnds we found between order and disorder in our previous work (Protonotarios et al., 2014). We attribute this difference to the greater variation in pattern types in the previous experiment.

Our ML estimates for the scale values of our stimuli are uncertain because we have collected only a finite number of responses to each stimulus pair. To estimate this uncertainty we used a parametric bootstrap method (Efron, 1979; Wichmann & Hill, 2001b). The method uses the ML-fitted equal variance Gaussian model (with lapsing) to generate sets of synthetic experimental data. A new model is fitted to each synthetic data set. To remove the effect of arbitrary additive constants, we align different bootstrap estimates by subtracting the mean of each set. We thus obtain a set of estimates for each stimulus, the standard deviation of which provides an estimate of uncertainty. Across the stimuli of Set A the standard deviations ranged from 0.46 to 1.04 jnds, with 0.67 jnds being the root mean square (RMS) average. The size of the range of the fitted values in Set A is 16.5 ± 1.2 jnds. The fitted values for the patterns and the resulting 95% confidence intervals are shown in Figure 6. Patterns are numbered according to their predicted a-scale values with 1 corresponding to the least ordered and 21 to the most ordered.

Both Figures 5 and 6 show the estimated order values for the 21 patterns we used in the experiment. Figure 5 shows how these vary with respect to the physical parameter, while Figure 6 shows the fitted values with respect to the pattern number. Latter figure offers a better view of the pattern spacing on the discrimination scale. We observe that graph is approximately linear confirming the effectiveness of our a-scale computation of order for jittered square lattice patterns, which is a different population than used to establish the a-scale. Patterns 5 and 7 in both graphs appear to violate the ordinal scale. This is not

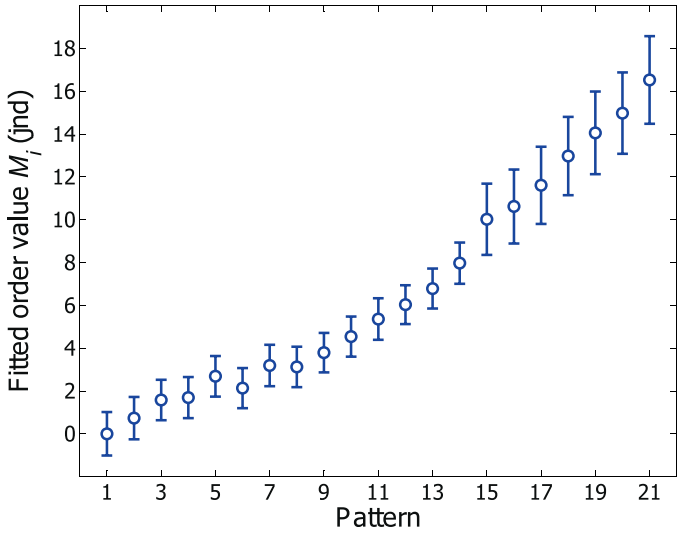

Figure 6. Fitted values of order based on the discrimination task. Bars show 95% confidence intervals as estimated from bootstrap analysis (200 repetitions). Patterns are numbered according to their predicted *a*-scale values with 1 corresponding to the least ordered and 21 to the most ordered.

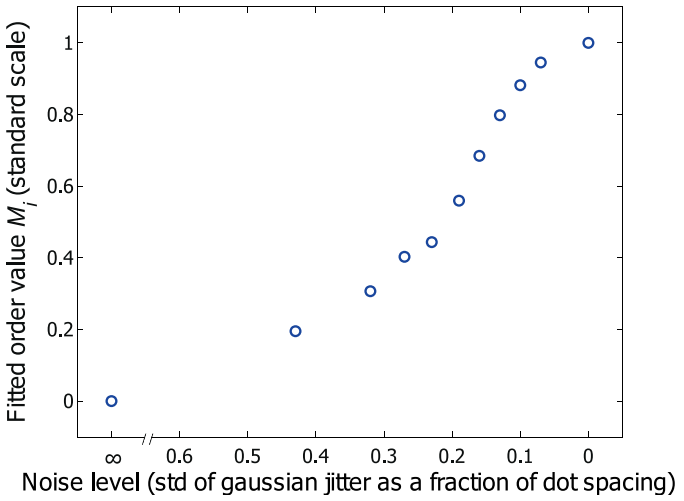

Figure 7. Fitted values of order based on the magnitude difference comparison task plotted against the amount of jitter. The latter is expressed as the standard deviation of the Gaussian perturbation of the dots as a fraction of the dot spacing of the perfect square grid. The fitted values have been scaled so that patterns of extreme order/disorder get the values of 1/0 respectively (standard scale).

surprising for two reasons. First, our *a*-scale has limited accuracy in predicting order, which is comparable to the patterns dense spacing. Second, we have collected a finite number of responses and thus order estimates are uncertain. Bar size in Figure 6 does not allow concluding whether the observed violations are real (i.e., whether these correspond to significant perceptual differences). These violations do not affect the analysis since we do not rely at any point on the initial ranking of the patterns of Set A. The *a*-scale has been employed only to achieve approximate equal spacing of the patterns on the discrimination scale.

Goodness of fit

Having determined a maximum likelihood model we assess its goodness of fit (GoF) by comparing its empirical deviance to the distribution of deviances that result by Monte Carlo generation of random datasets from the ML model. If the empirical deviance lies within the range of simulated deviances that encompass the 95% most common, then we accept the model. Deviance of a dataset (original or simulated) is twice the difference between the log-likelihood of the dataset given the ML model and the log-likelihood of the dataset given a saturated model. The saturated model in this case specifies a separate ML probability for each trial.

For the equal variance Gaussian model, the empirical deviance is 145 and the 95% interval of acceptable deviances for 10,000 repetitions is [107, 170], hence the model is accepted. Therefore, the consensus of discrimination-based order for Set A has the structure of an interval scale. We refer to the resulting scale as the discrimination scale.

Examining the responses, we notice that preference rate was exactly 1 (or 0) in 129 out of the 210 presented pairs. This means that a large number of trials simply validated the ordinal structure of the scale. Although we have established in previous work (Protonotarios et al., 2014), the interval structure of the responses for diverse point patterns, for the sake of completeness, we preferred again a balanced design to check for that for the new set. We could not exclude beforehand that patterns of high jitter may appear more ordered than patterns of intermediate jitter (e.g., because of random pairs or clusters of points). The reader is reminded that observers were not asked to judge the amount of positional jitter, but the degree of order and these two may not coincide. Independently of the monotonicity with respect to jitter amount, judgments only between similar order levels (within a few jnds) could not detect possible violations of transitivity that would render the existence of a one-dimensional scale impossible. Furthermore, a validated obvious ordering of the stimuli set is a requirement for the analysis of the difference comparisons that follows. A balanced design allows the estimation of the probability of consistent rankings for the subset of stimuli of the magnitude comparison that are being used in the difference comparison (Set B). The traditional increment threshold paradigm for the construction of the discrimination scale was not preferred for an additional reason. For our set of patterns, as Figure 2 indicates, the intensity of the attribute, unlike, for example, brightness, is not

uniquely defined by the physical parameter; patterns generated at a specific jitter amount vary stochastically in apparent order. Generating patterns only based on jitter amount would increase the variability of responses. We avoided this problem by using common stimuli in our two experiments. This way we can construct and compare the two scales without relying on any physical parameter.

Magnitude difference comparison scaling

Model fitting

The magnitude difference comparison scaling method is based on a stochastic model of difference measurement (Krantz, Luce, Suppes, & Tversky, 1971; Maloney & Yang, 2003; maximum likelihood difference scaling). Similarly to the Thurstone scaling assumptions, each stimulus is associated with a number on an interval scale expressing the real attribute contained in the stimulus. The aim is to recover these numbers for a set of stimuli by collecting direct interval comparison judgments.

We again denote the set of stimuli as $S_1, S_2, \ldots, S_N$, numbered in such a way so that the physical parameter, φ_i, related to each stimulus is ranked as $\varphi_1 < \varphi_2 < \ldots < \varphi_N$ (here $N = 11$). We assume that each stimulus, S_i, in the set evokes a perceptual response for the degree of order, which can be numerically represented as $\psi_i = M_i + N(0, \sigma)$, with M_i being the true value of the attribute. We assume that when an observer compares the perceptual difference between pairs (S_i, S_j) and (S_k, S_l), they respond on the basis of the sign of $|\psi_j - \psi_i| - |\psi_l - \psi_k|$. If the pair differences $|M_j - M_i|$ and $|M_l - M_k|$ are always sufficiently larger than the noise level, then observers will never make an error about which pattern of a pair is the more ordered, and the response probability arises from a link function depending on $(M_j - M_i) - (M_l - M_k)$. The link function is the cumulative distribution function of a Gaussian of variance 4 times the variance of the realization noise. We are justified in using this link function for our data because of the spacing of patterns in Set B used for the difference task. Within the discrimination data, pairs of patterns from Set B are correctly ordered in 96% of trials. In this article we only consider Gaussian noise when examining the magnitude difference comparison scaling. However, Maloney and Yang (2003) showed, with the use of simulations, that the resulting scale is robust with respect to the distributional assumptions for the noise.

Similarly to the discrimination scale fitting, we rescaled the link function to incorporate the lapse rate. We used the maximum likelihood method with gradient decent and multiple random starts for stability checking to estimate the parameters $M_1, M_2, \ldots, M_N$ and σ.

Goodness of fit

For the GoF assessment we compared the empirical deviance with the histogram of deviances based on Monte Carlo simulations using 10,000 repetitions. Pooling data for all observers did not result in an acceptable model for the equal-variance Gaussian model (empirical deviance = 494 with 95% interval of acceptable deviances: [303, 402]). A model of a common scale for all observers was accepted when we allowed different sensitivities ($1/\sigma$) for each of the two sessions completed by each observer (a noticeable variation in the sensitivity per observer has been found also in other studies where the same difference scaling method was applied; e.g., Devinck & Knoblauch, 2012). For this model the empirical deviance was 7,051, inside the 95% interval of acceptable deviances [6,898, 7,213]. Therefore, also for the case of the magnitude difference comparison task, a common scale for the stimuli can fit the collected data for all participants, when independent sensitivities are allowed per session.

The fitted attribute values for the common scale with respect to the physical parameter (amount of jitter) are shown in Figure 7. We will call this scale the difference scale. In the figure, the fitted values of the difference scale have been normalized so that patterns of extreme order/disorder get the values of 1/0 respectively (standard scale; Knoblauch & Maloney, 2012).

Independently of the varying sensitivities across the observers, we would expect that the pooled preference rates should be around 1 when the perceptual difference between patterns of one pair is considerably larger than the corresponding difference of the other pair. The extreme cases of this type appear when the smallest interval on either side of the spectrum is compared with the remaining largest nonoverlapping interval of the opposite side. Checking this for both cases validates that the largest difference has been preferred in 24 or 23 out of the total 24 trials. This was in general true for large differences; 75 out of the 330 quadruples corresponded to 24 or 23 preferences.

To exclude the possibility of a change in sensitivity because of learning, we examined the distribution of sensitivities for the first and second session per observer. For comparison we normalized the sensitivities so that the highest sensitivity of all sessions is set to 1. Learning is expected to cause an increase in sensitivity. When, for each observer, the sensitivity of the second session is plotted against the sensitivity of the first session, it is clear that there is no overall tendency for increase (Figure 8). The number of data points positioned above the identity line is five, while the other seven are positioned below the line. In the same graph, it is also noted that the average sensitivities for the two sessions are highly correlated meaning that the observers exhibit consistently low, medium, or high sensitivity.

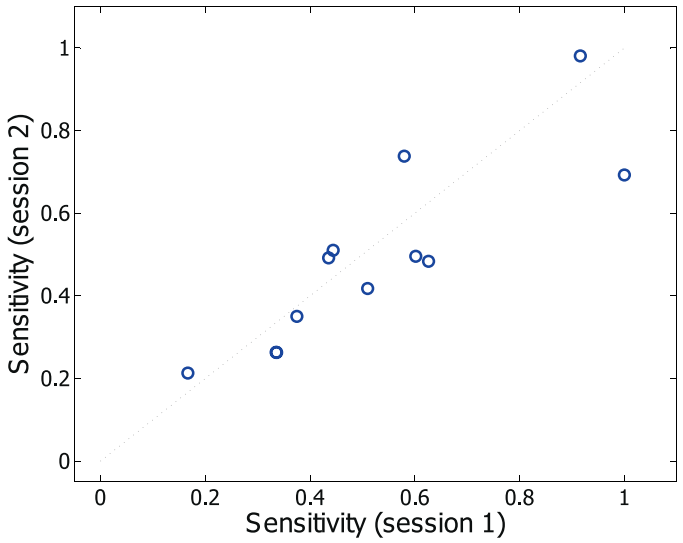

Figure 8. Scatter plot of the normalized sensitivity per observer for the first and second session. The identity line is shown for reference. Two data points (below the identity line) overlap.

Discrimination and difference scales

One has to be careful when interpreting Thurstone-type discrimination-based scales. If the discrimination data are consistent with such an equal-variance Gaussian model, this does not necessarily imply that the derived scale is the actual perceptual scale associated with the attribute or that internal noise remains constant along the dimension. The difficulty to determine the perceptual scale based only on discrimination judgments has been emphasized by Kingdom (2009) who highlighted the contribution of Whittle (1986, 1992) on the analysis of internal noise for discrimination scales. In his work on perception of luminance for a series of disks on a uniform background, Whittle was able to superimpose (with appropriate scaling) the functions relating jnds and equal-perceived differences to contrast magnitude. As Kingdom argued, this good overlap implied that contrast transduction noise is additive (constant). In a more detailed analysis García-Pérez and Alcalá-Quintana (2009) explained that with only two alternative forced choice discrimination data, it is not possible to disentangle the three underlying functions that affect discrimination, namely, (a) the transducer function, (b) how the variance of the noise varies with stimulus level, and (c) the distribution of the noise. They showed that models of constant or varying internal noise can fit equally well the same discrimination data. Whether the Thurstonian-type interval scale is mainly a theoretical construct or it has a more substantial existence corresponding to something happening in the nervous system had puzzled psychophysicists in the past (Luce, 1994). In any case, it is of particular practical importance. First, it allows testing of the collected responses for their consistency with an interval scale and thus verifies whether the perceptual dimension can be properly quantified. Second, it offers a simple model that allows direct comparison of intervals and their conversion to preference rates. Further, and most importantly, if the perceptual scale happens to be of constant noise, then the Thurstone Case V model will recover the real spacing of the stimuli, since for constant noise the discrimination scale coincides with the perceptual (Kingdom & Prins, 2009).

Common interval scale model for both experiments

Having established that both the discrimination and the difference comparison data can be described by interval scale models with Gaussian noise, we attempted models of equal variance noise that would account for both sets of data simultaneously. This is important as it would allow a simple interval scale for both supra- and subthreshold judgments and would imply a common underlying mechanism. We do not expect that such a model would be accepted though for the combination of tasks for all observers, since even for the magnitude difference comparison task alone such a model was not possible unless the per session sensitivities were adjusted. Indeed, by constraining the noise level to be the same for discrimination and difference data, we achieved a deviance of 800, which is outside the 95% interval of acceptable deviances [444, 566] computed using 10,000 repetitions. For a fixed noise level for the discrimination scale, if we allowed different noise levels for each of the difference comparison task sessions, the empirical deviance was 7,361, just inside the 95% acceptable interval of [7,067, 7,389], for 10,000 repetitions. Thus we cannot reject this model based solely on the total deviance analysis. However, by focusing on the deviance residuals distributions for the two experiments, we notice that for the difference experiment the empirical deviance of 7,059 was inside the 95% interval of acceptable deviances [6,901, 7,216], while for the discrimination experiment the empirical deviance value was 302, far outside the 95% interval of acceptable deviances [141, 203]. This is a strong indication that this model does not describe the data for both tasks accurately.

The different but constant noise level for the two tasks is equivalent to a linear transformation between the two scales. Since the empirical deviance value for the total model does not reject it, and since a nonlinear relationship linking the two scales could be preferred, we examined a model with a simple nonlinearity; we implemented a quadratic transformation linking the two scales' values, which required the addition of one

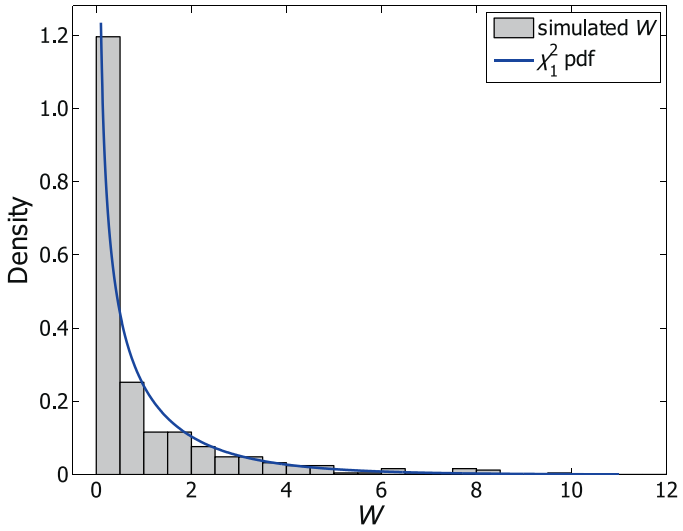

Figure 9. Simulation of the W statistic distribution for 500 repetitions. The curve corresponds to the probability density function of the χ^2_1.

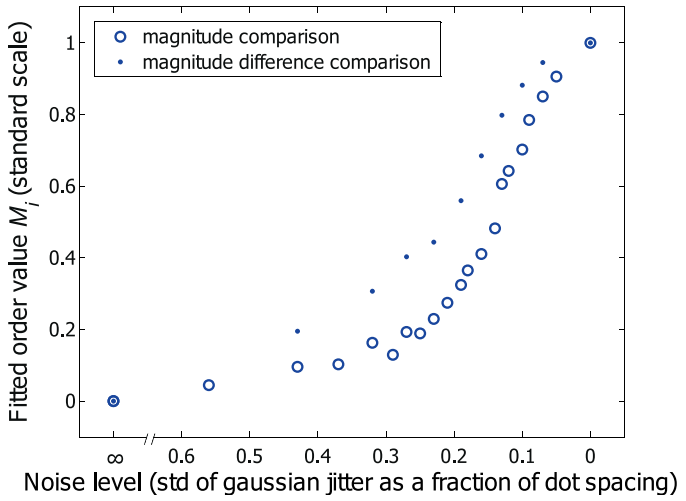

Figure 10. Magnitude comparison and magnitude difference comparison scales superimposed versus noise level. Both scales are rescaled to [0, 1] (standard scale).

parameter. This model achieved a deviance of 6,974, which fell inside the 95% interval of acceptable deviances [6,780, 7,115]. Examining the deviances per experiment, for the difference comparison task the empirical deviance was 6,797, inside the 95% interval of acceptable deviances [6,640, 6,970], and for the discrimination task, we achieved a better result than in the linear case, as the empirical deviance value of 177 fell just inside the 95% interval of acceptable deviances [114, 178].

Although it is not expected that the simple quadratic transformation captures the exact relationship between the two scales, the nonlinear model can be used for the rejection of a common interval scale for both the discrimination and the difference comparison tasks. The linear and the simple nonlinear models form a nested pair, the linear one being a special case of the nonlinear, and therefore the linear relationship (null hypothesis) between the two scales can be tested with the use of the likelihood ratio test (Kingdom & Prins, 2009). For the application of the likelihood ratio test we need to assume that the statistic W, which is twice the difference of the log-likelihoods of the two models in comparison, follows a χ^2_{df} distribution, where df is the difference in the number of parameters between the two models (Wilks, 1938; in our case the two models differ by 1 degree of freedom). Although this is asymptotically true, it is not easy to estimate the number of observations that are necessary to provide a good such approximation (Wichmann & Hill, 2001a). We can, however, simulate the distribution of the statistic W. By accepting the linear model that has been estimated with the ML method, we generated a large number of simulated data and refitted both the linear and the nonlinear models. For 500 repetitions we computed the W statistic and examined how well the simulated distribution is approximated by the χ^2_1 theoretical density. In Figure 9 we can see that this approximation is satisfactory in shape and extent. The log-likelihood for the estimated linear model was −3,789.31, and the log-likelihood for the estimated nonlinear model was −3,595.97. This resulted in a value for $W_1 = 386.68$. Getting a value of W, which is much greater than the 95% cut off of the χ^2_1 distribution, 3.84, means that we have to reject the linear model in favor of the nonlinear one. Since the nonlinear transformation distorts the spacing between the scale values in one of the two scales, we conclude that no common interval scale model with Gaussian noise is able to describe both the discrimination and the difference comparison data.

Discussion

When the discrimination- and difference-based scales are plotted against the jitter level (Figures 5 and 7), they present very similar shapes. Similar to Whittle's approach, we can rescale one in an attempt to superimpose them and observe their overlap (Figure 10). If the agreement is sufficient, and we assume a common transducer function, then the internal noise can be considered constant across the perceptual dimension, and a common interval scale could predict both discrimination and difference data. However, due to the shape of the curves in Figure 10, the degree of agreement between the two scales is hard to assess. Therefore, we prefer to plot one scale against the other (Figure 11). It is then evident that the plotted curve deviates smoothly and systematically from a linear relationship. This provides a visual explanation of the

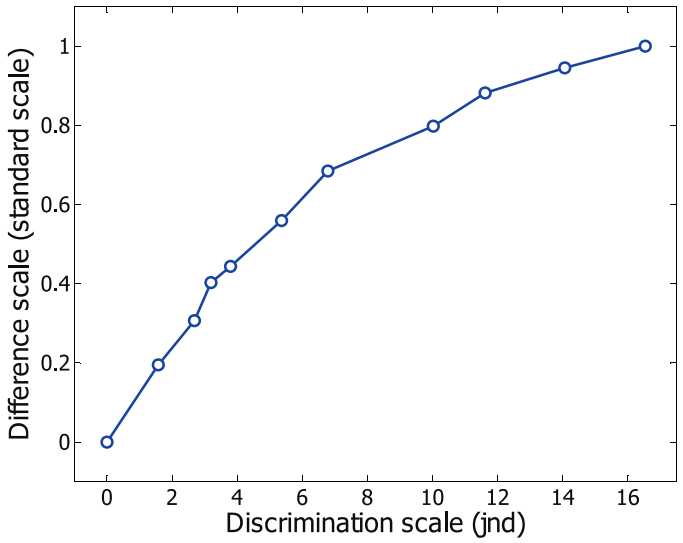

Figure 11. Scatter plot of the estimated values of order for the point patterns (Set B) derived from the discrimination task and from the magnitude difference comparison task.

failure of a common interval scale to describe both tasks. The slope of the curve is higher near the disordered end and lower near the ordered end. This indicates that, when matched for discrimination jnd steps, suprathreshold differences near disorder appear larger than near order. Conversely, considering the difference scale as reference, a perceptual difference on the ordered end corresponds to more jnd steps on the discrimination scale than the same perceptual difference on the disordered end. As a consequence, there is no way of transforming one scale to match the other without affecting its interval structure.

Allowing varying noise across a common perceptual scale for the two tasks is expected to affect the spacing of the stimuli on the two scales to a different degree. The discrimination scale being sensitive in the local extent of the noise will be affected more; as mentioned in the introduction, the difference scales are relatively robust to whether noise is constant or varying (Kingdom & Prins, 2009). On such a common scale, equal differences at different positions on the scale will not correspond to equal discrimination performances. However, if the particular function that defines the extent of the noise distribution at each point is known, then the jnd width at a particular location can be computed. The simplest implementation of such a model assumes a constant rate of change in noise intensity. Such a model, together with the linear model, which was examined in the previous section, form a nested pair, and thus we can repeat the same type of likelihood ratio analysis we followed for the quadratic transformation. The varying noise model has one additional parameter in comparison to the linear model—the rate of change. Fitting with ML we get a log-likelihood value of −3,713.11, which results in a value $W_2 = 152.40$, a number much greater than the 95% cut off of the χ^2_1 distribution, which is 3.84. This means that the linear model should be rejected in favor of the varying noise model.

Although we can reject a common equal variance model in both types of analysis—(a) quadratic transformation of the difference scale, and (b) common perceptual scale with varying noise—it is not possible to decide between the two options by comparing the values of W_1 and W_2. A major reason is that the proposed functional forms are not necessarily the best among all possible. The two models offer two different interpretations: According to (a), we can assume that observers use different visual mechanisms or employ different criteria when judging sub- and suprathreshold differences. Both scales are consistent with an interval scale structure but are different from each other. Alternatively, assuming a common mechanism and a common transducer function, the internal noise has to vary across the perceptual dimension. For the constantly increasing noise model we examined, (b), there is approximately a 2-fold increase (1.85) in the standard deviation of the noise distribution in the direction towards disorder. Both interpretations agree qualitatively on their consequences about how jnd width varies with respect to the difference scale.

Our analysis shows that our previous approach for the construction of the *a*-scale of order based on discrimination judgments is not compatible with perceptual judgments of large differences. The two views cannot be reconciled in one interval scale for the measurement of order in point patterns. Although there is no ordinal disagreement, for practical purposes in the analysis of evolving systems, differences, and therefore rates of change, are not consistent across the whole range for both types of judgment. In our experiment there is a smooth relationship between the two scales. However, the actual form may depend on the particular class of patterns.

The apparent deviation of the curve in Figure 11 from a straight line, although significant, could not necessarily have practical or testable implications. However, as Figure 12 illustrates, when focusing on the midpoints of each scale, this curve offers a simple and easily testable prediction. By comparing the midpattern of the discrimination scale with the midpattern of the difference scale, we see that the first one is noticeably more ordered than the second. We perceive the midpattern of the discrimination scale as being more toward the ordered end, and conversely, the pattern that we directly perceive as equidistant from the ends is not placed an equal number of jnds from the ends.

Devinck and Knoblauch (2012) previously made a comparison of scales based on magnitude and magnitude difference comparisons and, contrary to our

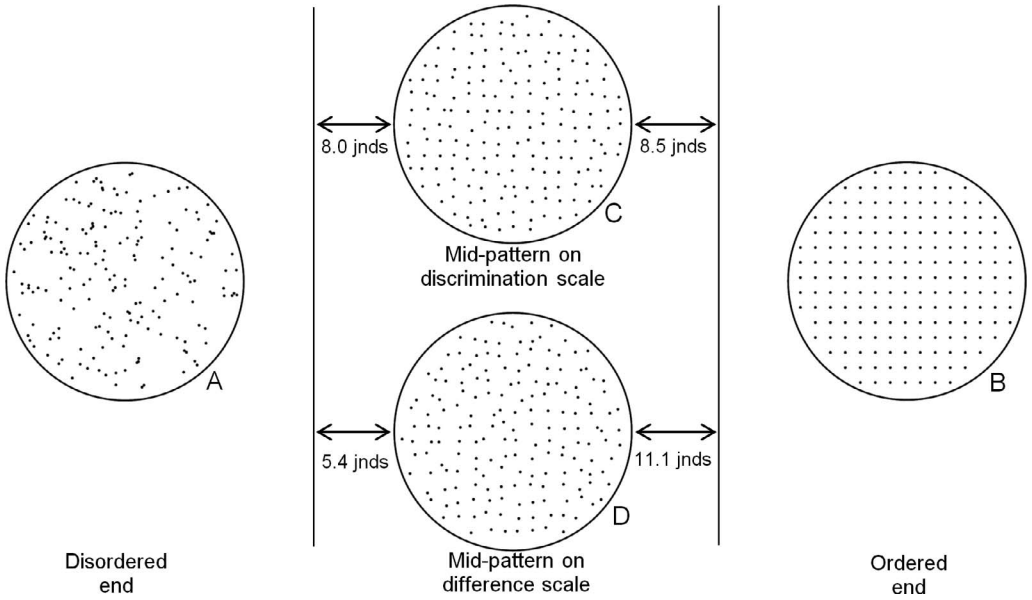

Figure 12. Illustrates the mismatch between discrimination- and difference-based scales by focusing on end- and midpoints of the scales. We expect that the reader will consider the difference AC as substantially greater than CB, even though they are almost equal in jnd steps. Also, the difference AD will be considered as similar to the difference DB even though they are different in jnd steps.

results, found them compatible. They collected comparison judgments for the strength of the fill-in percept of the watercolor effect (Pinna, Brelstaff, & Spillmann, 2001) as they varied the luminance ratio between the two components of the generating contour. For magnitude difference comparison data they found it necessary to construct separate scales for each observer. They compared the discrimination and difference scales and claimed that they were compatible, in contrast to our findings. However, they tested the discrimination performance of each observer only at a narrow range of the physical parameter. It is our view that to exclude the possibility of only a local agreement between the two scales, verification over a wide range of the scale is necessary, as the nonlinear relation between the two scales is only apparent when a larger extent is considered.

The method of maximum likelihood difference scaling as suggested by Maloney and Yang (2003) provides a convenient tool for the construction of perceptual scales based on suprathreshold judgments. This method in combination with the traditional Thurstone-type scaling can address the important psychophysical questions regarding the existence of a unique perceptual scale for both discrimination and appearance, and the form of internal noise. These two methods do not require the existence of a physical parameter, and therefore are particularly useful when such a parameter does not exist or the stimuli are generated with multiple parameters. The disagreement between the two scales can be relevant in cases where the experimental method involves both sub- and suprathreshold comparisons in the same task. For example, recently Jogan and Stocker (2014) suggested a new two-alternative forced choice method for the characterization of the perceptual bias caused by stimuli secondary parameters, where a common equal-variance signal detection model is assumed for both short and long stimulus distances. Not verifying the validity of such a model can confound the analysis as depending on the perceptual dimension under study the effect may be significant.

Conclusion

In this work we have analyzed judgments of order for a specific class of point patterns (jittered square lattice of points). Our proposed method makes combined use of the Thurstone-type scaling for subthreshold judgments and the maximum likelihood difference scaling for suprathreshold judgments. The method does not rely on the existence of a physical parameter for the control of the intensity of the perceptual attribute.

Analysis showed that, within the limits of our collected data, it is possible to construct separate interval scales for the description of magnitude comparison and magnitude difference comparison data. However, a common interval scale with Gaussian noise model describing both is not possible. Since the mismatch between the scales is most apparent in comparisons involving the endpoints and midpoints, and the scaling methods are relatively robust to noise distribution assumptions, it seems unlikely that a

common scale can be achieved by any more complex model that properly treats the interval structure of the scales, or the use of Gaussian noise. This signifies that it is not possible to construct an interval scale for the quantification of order in agreement with human judgments for both small and large differences. Our findings are consistent with either two separate scales for sub- and suprathreshold judgments of order implying two distinct perceptual mechanisms, or a common scale with varying internal noise that increases in the direction of disorder. Therefore, with regard to the degree of order in point patterns, the magnitude of large differences is not determined by the number of jnd steps.

Keywords: psychophysics, scaling, discrimination scale, difference scale, signal detection theory, order, regularity, point pattern

Acknowledgments

CoMPLEX is an Engineering & Physical Sciences Research Council (UK) (EPSRC) funded Centre for Doctoral Training. Emmanouil Protonotarios is supported by the Greek State Scholarship Foundation (IKY). The authors would like to thank Keith May for reading the manuscript and for providing useful comments.

Commercial relationships: none.
Corresponding author: Emmanouil D. Protonotarios.
Email: emmanouil.protonotarios.10@ucl.ac.uk.
Address: Centre for Mathematics, Physics and Engineering in the Life Sciences and Experimental Biology and Department of Computer Science, University College London, London, UK.

References

Allik, J., & Tuulmets, T. (1991). Occupancy model of perceived numerosity. *Perception & Psychophysics*, *49*(4), 303–314.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, *61*(3), 183–193.

Bonneh, Y., Reisfeld, D., & Yeshurun, Y. (1994). Quantification of local symmetry: Application to texture discrimination. *Spatial Vision*, *8*(4), 515–530.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, *39*(3/4), 324–345, doi:10.2307/2334029.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436, doi:10.1163/156856897X00357.

Brown, A. M., Lindsey, D. T., & Guckes, K. M. (2011). Color names, color categories, and color-cued visual search: Sometimes, color perception is not categorical. *Journal of Vision*, *11*(12):2, 1–21, doi:10.1167/11.12.2. [PubMed] [Article]

Charrier, C., Knoblauch, K., Maloney, L. T., & Bovik, A. C. (2011). Calibrating MS-SSIM for compression distortions using MLDS. In *2011 18th IEEE International Conference on Image Processing (ICIP)*. New York: IEEE.

Charrier, C., Knoblauch, K., Maloney, L. T., Bovik, A. C., & Moorthy, A. K. (2012). Optimizing multiscale SSIM for compression via MLDS. *IEEE Transactions on Image Processing*, *21*(12), 4682–4694, doi:10.1109/TIP.2012.2210723.

Charrier, C., Maloney, L. T., Cherifi, H., & Knoblauch, K. (2007). Maximum likelihood difference scaling of image quality in compression-degraded images. *Journal of the Optical Society of America, A: Optics, Image Science, and Vision*, *24*(11), 3418–3426, doi:10.1364/JOSAA.24.003418.

Cliffe, M. J., & Goodwin, A. L. (2013). Quantification of local geometry and local symmetry in models of disordered materials. *Physica Status Solidi (B)*, *250*(5), 949–956, doi:10.1002/pssb.201248553.

Cohen, M., Baum, B., & Miodownik, M. (2011). The importance of structured noise in the generation of self-organizing tissue patterns through contact-mediated cell-cell signalling. *Journal of the Royal Society Interface*, *8*(59), 787–798, doi:10.1098/rsif.2010.0488.

Cook, J. E. (2004). Spatial regularity among retinal neurons. In *The visual neurosciences* (pp. 463–477). Cambridge, MA: MIT Press.

Dakin, S. C., Tibber, M. S., Greenwood, J. A., Kingdom, F. A. A., & Morgan, M. J. (2011). A common visual metric for approximate number and density. *Proceedings of the National Academy of Sciences*, *108*(49), 19552–19557, doi:10.1073/pnas.1113195108.

David, H. A. (1988). *The method of paired comparisons*. London: Oxford University Press.

Delaunay, B. (1934). Sur la sphère vide [Translation: On the empty sphere]. *Izvestia Akademii Nauk SSSR, Otdelenie Matematicheskikh I Estestvennykh Nauk*, *7*, 793–800.

Demeyer, M., & Machilsen, B. (2012). The construc-

tion of perceptual grouping displays using GERT. *Behavior Research Methods*, *44*(2), 439–446, doi:10.3758/s13428-011-0167-8.

Devinck, F., Gerardin, P., Dojat, M., & Knoblauch, K. (2014). Spatial selectivity of the watercolor effect. *Journal of the Optical Society of America, A: Optics, Image Science, and Vision*, *31*(4), A1–A6, doi:10.1364/JOSAA.31.0000A1.

Devinck, F., & Knoblauch, K. (2012). A common signal detection model accounts for both perception and discrimination of the watercolor effect. *Journal of Vision*, *12*(3):19, 1–14, doi:10.1167/12.3.19. [PubMed] [Article]

Dunleavy, A. J., Wiesner, K., & Royall, C. P. (2012). Using mutual information to measure order in model glass-formers. *Physical Review E*, *86*(4-1), 041505.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, *7*(1), 1–26, doi:10.1214/aos/1176344552.

Emrith, K., Chantler, M. J., Green, P. R., Maloney, L. T., & Clarke, A. D. F. (2010). Measuring perceived differences in surface texture due to changes in higher order statistics. *Journal of the Optical Society of America, A*, *27*(5), 1232–1244, doi:10.1364/JOSAA.27.001232.

Fleming, R. W., Jaekel, F., & Maloney, L. T. (2011). Visual perception of thick transparent materials. *Psychological Science*, *22*(6), 812–820, doi:10.1177/0956797611408734.

García-Pérez, M. A., & Alcalá-Quintana, R. (2009). Fixed vs. variable noise in 2AFC contrast discrimination: Lessons from psychometric functions. *Spatial Vision*, *22*(4), 273–300, doi:10.1163/156856809788746309.

Ginsburg, N., & Goldstein, S. R. (1987). Measurement of visual cluster. *The American Journal of Psychology*, *100*(2), 193–203.

Graham, N. V. (2011). Beyond multiple pattern analyzers modeled as linear filters (as classical V1 simple cells): Useful additions of the last 25 years. *Vision Research*, *51*(13), 1397–1430, doi:10.1016/j.visres.2011.02.007.

Guillaud, M., Cox, D., Adler-Storthz, K., Malpica, A., Staerkel, G., Matisic, J., & MacAulay, C. (2004). Exploratory analysis of quantitative histopathology of cervical intraepithelial neoplasia: Objectivity, reproducibility, malignancy-associated changes, and human papillomavirus. *Cytometry*, *Part A: The Journal of the International Society for Analytical Cytology*, *60*(1), 81–89, doi:10.1002/cyto.a.20034.

Hand, D. J. (2004). *Measurement theory and practice: The world through quantification*. London: Oxford University Press.

Harvey, L. O. (1986). Efficient estimation of sensory thresholds. *Behavior Research Methods, Instruments, & Computers*, *18*(6), 623–632, doi:10.3758/BF03201438.

Hu, W., Li, H., Wang, C., Gou, S., & Fu, L. (2012). Characterization of collagen fibers by means of texture analysis of second harmonic generation images using orientation-dependent gray level co-occurrence matrix method. *Journal of Biomedical Optics*, *17*(2), 26007, doi:10.1117/1.JBO.17.2.026007.

Jogan, M., & Stocker, A. A. (2014). A new two-alternative forced choice method for the unbiased characterization of perceptual bias and discriminability. *Journal of Vision*, *14*(3):20, 1–18, doi:10.1167/14.3.20. [PubMed] [Article]

Junge, M., & Reisenzein, R. (2013). Indirect scaling methods for testing quantitative emotion theories. *Cognition & Emotion*, *27*(7), 1247–1275, doi:10.1080/02699931.2013.782267.

Kingdom, F. (2009). The significance of Whittle's experiments on luminance discrimination and brightness scaling for the multiplicative-versus-additive contrast-noise question. *Journal of Vision*, *9*(8):362, doi:10.1167/9.8.362. [Abstract]

Kingdom, F., & Prins, N. (2009). *Psychophysics: A practical introduction*. New York: Academic Press.

Knoblauch, K., & Maloney, L. T. (2008). MLDS: Maximum likelihood difference scaling in R. *Journal of Statistical Software*, *25*(2), 1–26.

Knoblauch, K., & Maloney, L. T. (2012). *Modeling psychophysical data in R*. New York: Springer. Retrieved from http://www.springer.com/gp/book/9781461444749

Koffka, K. (1935). *Principles of Gestalt psychology*. London: Lund Humphries.

Krantz, D., Luce, D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement, Vol. I: Additive and polynomial representations*. New York: Academic Press.

Lindsey, D. T., Brown, A. M., Reijnen, E., Rich, A. N., Kuzmova, Y. I., & Wolfe, J. M. (2010). Color channels, not color appearance or color categories, guide visual search for desaturated color targets. *Psychological Science*, *21*(9), 1208–1214, doi:10.1177/0956797610379861.

Luce, R. D. (1994). Thurstone and sensory scaling: Then and now. *Psychological Review*, *101*(2), 271–277, doi:10.1037/0033-295X.101.2.271.

Luce, R. D., & Krumhansl, C. L. (1988). Measurement,

scaling, and psychophysics. *Stevens' Handbook of Experimental Psychology*, *1*, 3–74.

Machilsen, B., Wagemans, J., & Demeyer, M. (2015). Quantifying density cues in grouping displays. *Vision Research*, in press, doi:10.1016/j.visres.2015.06.004

Maloney, L. T., & Yang, J. N. (2003). Maximum likelihood difference scaling. *Journal of Vision*, *3*(8):5, 573–585, doi:10.1167/3.8.5. [PubMed] [Article]

Marinari, E., Mehonic, A., Curran, S., Gale, J., Duke, T., & Baum, B. (2012). Live-cell delamination counterbalances epithelial growth to limit tissue overcrowding. *Nature*, *484*(7395), 542–545, doi:10.1038/nature10984.

Menkovski, V., & Liotta, A. (2012). Adaptive psychometric scaling for video quality assessment. *Signal Processing: Image Communication*, *27*(8), 788–799, doi:10.1016/j.image.2012.01.004.

Morgan, M. J., Mareschal, I., Chubb, C., & Solomon, J. A. (2012). Perceived pattern regularity computed as a summary statistic: Implications for camouflage. *Proceedings of the Royal Society, B: Biological Sciences*, *279*(1739), 2754–2760, doi:10.1098/rspb.2011.2645.

Newell, F., Murtagh, R., & Hutzler, S. (2013). A role for Gestalt principles of organisation in shaping preferences for non-natural spatial and dynamic patterns. In *ECVP 2013* (Vol. 42). Bremen, Germany: Perception.

Obein, G., Knoblauch, K., & Viénot, F. (2004). Difference scaling of gloss: Nonlinearity, binocularity, and constancy. *Journal of Vision*, *4*(9):4, 711–720, doi:10.1167/4.9.4. [PubMed] [Article]

Ouhnana, M., Bell, J., Solomon, J. A., & Kingdom, F. A. A. (2013). Aftereffect of perceived regularity. *Journal of Vision*, *13*(8):18, 1–13, doi:10.1167/13.8.18. [PubMed] [Article]

Pinna, B., Brelstaff, G., & Spillmann, L. (2001). Surface color from boundaries: A new "watercolor" illusion. *Vision Research*, *41*(20), 2669–2676, doi:10.1016/S0042-6989(01)00105-5.

Protonotarios, E. D., Baum, B., Johnston, A., Hunter, G. L., & Griffin, L. D. (2014). An absolute interval scale of order for point patterns. *Journal of the Royal Society: Interface*, *11*(99), 20140342, doi:10.1098/rsif.2014.0342.

Rhodes, G., Maloney, L. T., Turner, J., & Ewing, L. (2007). Adaptive face coding and discrimination around the average face. *Vision Research*, *47*(7), 974–989, doi:10.1016/j.visres.2006.12.010.

Sausset, F., & Levine, D. (2011). Characterizing order in amorphous systems. *Physical Review Letters*, *107*(4), 045501.

Steinhardt, P. J., Nelson, D. R., & Ronchetti, M. (1983). Bond-orientational order in liquids and glasses. *Physical Review B*, *28*(2), 784–805, doi:10.1103/PhysRevB.28.784.

Stern, H. (1992). Are all linear paired comparison models empirically equivalent. *Mathematical Social Sciences*, *23*(1), 103–117, doi:10.1016/0165-4896(92)90040-C.

Stevens, S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677–680, doi:10.1126/science.103.2684.677.

Sudb, J., Marcelpoil, R., & Reith, A. (2000). New algorithms based on the Voronoi Diagram applied in a pilot study on normal mucosa and carcinomas. *Analytical Cellular Pathology: The Journal of the European Society for Analytical Cellular Pathology*, *21*(2), 71–86.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*(4), 273–286, doi:10.1037/h0070288.

Torgerson, W. S. (1958). *Theory and methods of scaling*. New York: Wiley.

Truskett, T. M., Torquato, S., & Debenedetti, P. G. (2000). Towards a quantification of disorder in materials: Distinguishing equilibrium and glassy sphere packings. *Physical Review E*, *62*(1), 993–1001, doi:10.1103/PhysRevE.62.993.

Vancleef, K., Putzeys, T., Gheorghiu, E., Sassi, M., Machilsen, B., & Wagemans, J. (2013). Spatial arrangement in texture discrimination and texture segregation. *I-Perception*, *4*(1), 36–52, doi:10.1068/i0515.

Wagemans, J., Wichmann, F., & de Beeck, H. (2005). Visual perception I: Basic principles. In K. Lamberts & R. Goldstone (Ed.), *Handbook of cognition* (pp. 3–47). London: Sage.

Wässle, H., & Boycott, B. B. (1991). Functional architecture of the mammalian retina. *Physiological Reviews*, *71*(2), 447–480.

Whittle, P. (1986). Increments and decrements: Luminance discrimination. *Vision Research*, *26*(10), 1677–1691.

Whittle, P. (1992). Brightness, discriminability and the "crispening effect." *Vision Research*, *32*(8), 1493–1507.

Wichmann, F. A., & Hill, N. J. (2001a). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics*, *63*(8), 1293–1313, doi:10.3758/BF03194544.

Wichmann, F. A., & Hill, N. J. (2001b). The

psychometric function: II. Bootstrap-based confidence intervals and sampling. *Perception & Psychophysics*, *63*(8), 1314–1329, doi:10.3758/BF03194545.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, *9*(1), 60–62, doi:10.1214/aoms/1177732360.

Yang, J. N., Szeverenyi, N. M., & Ts'o, D. (2008). Neural resources associated with perceptual judgment across sensory modalities. *Cerebral*, *18*(1), 38–45, doi:10.1093/cercor/bhm029.