Eurographics Conference on Visualization (EuroVis) 2016
K-L. Ma, G. Santucci, and J. J. van Wijk
(Guest Editors)

Volume 35 (2016), Number 3

How Ordered Is It? On the Perceptual Orderability of Visual Channels

David H. S. Chung[1], Daniel Archambault[1], Rita Borgo[1], Darren J. Edwards[1], Robert S. Laramee[1], and Min Chen[2]
[1]Swansea University, UK [2]University of Oxford, UK

Abstract

The design of effective glyphs for visualisation involves a number of different visual encodings. Since spatial position is usually already specified in advance, we must rely on other visual channels to convey additional relationships for multivariate analysis. One such relationship is the apparent order present in the data. This paper presents two crowdsourcing empirical studies that focus on the perceptual evaluation of orderability for visual channels, namely Bertin's retinal variables. The first study investigates the perception of order in a sequence of elements encoded with different visual channels. We found evidence that certain visual channels are perceived as more ordered (for example, value) while others are perceived as less ordered (for example, hue) than the measured order present in the data. As a result, certain visual channels are more/less sensitive to disorder. The second study evaluates how visual orderability affects min and max judgements of elements in the sequence. We found that visual channels that tend to be perceived as ordered, improve the accuracy of identifying these values.

1. Introduction

Given a set of visual objects in a spatial layout, determining how well they are ordered is an elementary visual task. Such tasks may be featured in more complex tasks, for instance, visual search (e.g., minima and maxima), anomaly detection (e.g., an ordered subset among mostly unordered objects), change detection (e.g., swapping order), correlation identification (e.g., among different visual channels of glyphs), and others. Chung *et al.* [CLP*15] presented a real-world example where analysts used a visualisation system to sort multivariate glyphs that encoded a number of event attributes according to one or two of them, while observing the other attributes. The authors observed that some visual channels *appeared* to give a more ordered impression than others. This became the main motivation of our work.

Bertin made a connection between orderability of a variable and the orderability of a visual channel used to encode the variable [Ber83]. While we all agree that there may be binary grouping of visual channels according to their orderability, there has not been any quantitative measure to grade, or indeed to "order", the orderability of visual channels. We anticipate that the orderability of visual channels may likely be a multi-criteria problem. This work is the first step towards a comprehensive answer to this challenging problem. In this paper, we present two empirical studies to investigate two of such criteria. We focus on two research questions in conjunction with a set of commonly used visual channels:

1. *How does red disorder affect the perception of orderedness with different visual channels?*
2. *How do visual channels affect the judgement of min and max values for ordered and unordered sequences?*

The first research question is concerned with detecting structural patterns in a visualisation. Orderedness is one of such patterns, and we investigate this in a sequence of elements encoded using different visual channels. The second research question explores how such sequences impact the judgement of minimal and maximal values. This is a typical task in many visualisation processes, for example, in discovering preferable options or filtering out undesirables. We conducted two empirical studies formulated around these two research questions. Our results show that the choice of visual channels affects the performance of both tasks.

2. Related Work

Perceptual Studies in Visualisation. A great deal of research has studied the perception of visual channels such as position, length, and colour, and their impact on effective data visualisation [CM84, War08b]. Healey *et al.* [HS12] investigate the perceptible boundary of visual elements based on pixel resolution and viewing distance. MacKinlay [Mac86] provides a technical framework to automatically design effective visualisation of data encodings of quantitative, ordinal, and categorical information. House *et*

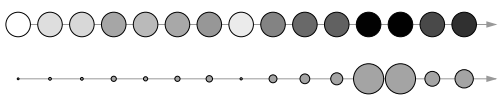


Figure 1: *The same sequence using value and size.*

al. [HBW06] describe an experimental framework to optimise visualisation designs using human-in-the-loop interaction. Their approach is adopted to select optimal textures for layered surfaces [BHW05]. Our work focuses on quantifying the orderability of visual channels when encoding a sequence of data values, an aspect not covered by the literature.

Crowdsourcing Graphical Perception. Online crowdsourcing experiments present an attractive option for evaluating the perceptual challenges in visualisation design due to its low cost and increased scalability [HB10]. Studies deployed on these platforms include semantic colouring [LFK*13], treemap design [LHH*12], and human computation algorithms [GSCO12]. The perception of correlation when displayed using spatial position has also been extensively studied [HYFC14, KH16] in this way. Our study also uses a crowdsourcing platform to collect this data to test the perception of order in a sequence.

Glyph-based Visualisation. Glyphs are visual objects that map multiple data values to visual features such as size, shape and position [War08a]. They are an effective approach to visualise multi-dimensional data [BKC*13] in many disciplines such as biological experiments [MPRSDC12], sports [LCP*12], and flow visualisation [HLNW11]. The positioning of glyphs is typically used to encode spatial information [dLvW93] or emphasise an ordering according to key attributes. As a result, glyph-based visual designs must rely on other visual mappings, such as size, in order to convey additional orderings. Our work closely follows the research by Chung *et al.* [CLP*15], who describe a framework to design visually sortable glyphs for interactive visualisation. We extend this contribution by evaluating the orderability of visual channels through a formal empirical study.

3. Orderability of Visual Channels

Figure 1 illustrates two unordered sequences using value and size, both perceptually orderable channels [Ber83]. Both impose a universal ordering where we can estimate the magnitude of each element to estimate the order of the sequence. However, it is not clear that both sequence encode the exact same values, leading us to believe that different visual channels have different levels of perceived orderedness.

3.1. Data Generation

To produce our visual stimuli, we first create synthetic data sets that model ordered and unordered sequences. We use the Body Mass Index (BMI) as a base because of its known linear correlation between the weight and height of a person. To generate our data sets, we take points along the BMI curve and map them to the visual channels shown in Figure 2. Next, we create unordered sequences using a noise-based approach adopted from signal processing theory [JJS93] to model sequences with varying levels of *disorder*.

3.2. Defintion of Disorder

A sequence can be characterised by different degrees of order, ranging from very unordered to completely ordered. We define unordered sequences as a sequence with a non-zero amount of *disorder*. Disorder can be introduced in many forms. One intuitive method is to swap elements in an ordered sequence.

Let x_i, x_j be two randomly selected data points. The swap function is defined as:

$$f(x_i, x_j, d) = \begin{cases} swap(x_i, x_j) & \text{if } \; ||i-j|| \leq d \\ \text{null} & \text{else} \end{cases} \quad (1)$$

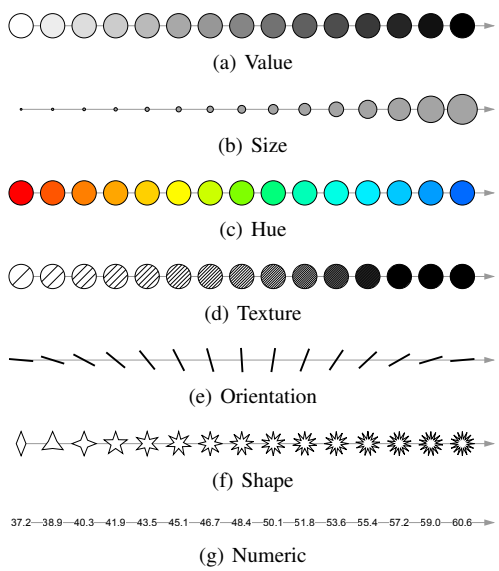
where $d \in \mathbb{R}$ is the distance between a pair of data points.

Swapping can be translated to *noise* in a sequence. For example, consider an ordered sequence of $S = [1,3,5,7,9]$. When we swap 3 and 7, we create an unordered sequence $S' = [1,7,5,3,9]$. We can rewrite the 2nd and 4th number as an expected value plus noise, i.e., $7 = 3+4$ and $3 = 7-4$. S' is thereby the result after introducing two noise components $+4$ and -4 into S.

We control disorder using two parameters: (1) the number of swaps, and (2) the swap distance between two points. The *level of disorder* is measured by applying Pearson's correlation coefficient $\eta \in [-1,1]$ to S and S'. In this work, we use only the non-negative range of η. To create a data sequence with a specific η, we first determine the number of swaps based on a predefined mapping. We then adjust the swapping distances to obtain the desired η. Through observation and pilot experiments, we noticed that the perceived orderedness was sometimes indistinct. Therefore, we selected disorder levels with a measured perceptible difference: $\eta_1 = 1.0$, $\eta_2 = 0.97$, $\eta_3 = 0.95$, $\eta_4 = 0.90$, $\eta_5 = 0.78$, $\eta_6 = 0.71$, $\eta_7 = 0.54$, and $\eta_8 = 0.12$.

3.3. Visual Mapping of Elements

Figure 2 shows the visual stimuli used in our experiments. Given that position is the most effective representation for conveying an ordering [CM84, War02], our goal is to investigate the visual channel which is next most orderable. We analyse this using 1D plot visualisations. Points (or elements) along the 1D plot are mapped using the visual channels described by Bertin [Ber83]. In addition to this, we compare their performance against raw numerical values representations, which is common practice when reading data within a table or spreadsheet view.



(a) Value (b) Size (c) Hue (d) Texture (e) Orientation (f) Shape (g) Numeric

Figure 2: *The visual stimuli used in our experiments. Each 1D plot is mapped using the visual channels taken from semiology of graphics [Ber83].*

Since perceptual differences along a visual spectrum can be non-uniform [War08b], we carefully divide each visual mapping into bins equal to the number of sample points (e.g., $n = 15$) shown in the plot. We choose the maximum number of samples to be displayed such that the just noticeable difference (JND) [Ber83] in each visual channel is maintained. In the following section, we present our visual mapping methods and JND considerations to ensure that each element within a sequence is visually discriminable.

3.4. Just Noticeable Difference

For each of our encodings, we describe how they are generated and demonstrate how consecutive elements are above their respective JND.

Value is mapped to greyscale intensity of each element (see Figure 2(a)). In order to realise increments of value with a perceptible difference, we follow previous work on visual discrimination of intensity [Hei24, War08b]. They show that human perception can distinguish a 0.5% change in luminance. The difference in luminance measured in our mapping is 7%, which is significantly larger than the JND.

Size is mapped to circle radius (see Figure 2(b)). Previous work in psychophysics demonstrates that the perception of size (e.g., area) is logarithmic, and can be effectively modelled using Weber-Fechner's Law [AR08]:

$$p = k \ln \frac{r_i}{r_j} \tag{2}$$

where p is the perception between two radii r_i and r_j. To control for the JND, we estimate the parameter $k = 0.23$ by incrementing the radius until a significantly large difference is perceived, giving a value of $p = 0.048$. Vision research shows that a Weber fraction for circle radius discrimination is $p = 0.025$ [WWH98] and therefore above the JND.

Hue is mapped to the colour (from red to blue) of each element (see Figure 2(c)). To control for hue, we measure the colour difference between two steps of hue values in *CIE-L*uv* colour space. The distance between two colours (L_1, u_1^*, v_1^*) and (L_2, u_2^*, v_2^*) is measured by:

$$\Delta E^*_{Luv} = \sqrt{(L_2 - L_1)^2 + (u_2^* - u_1^*)^2 + (v_2^* - v_1^*)^2} \tag{3}$$

where $\Delta E^*_{Luv} = 1$ is an approximation to a JND [War08b]. We set a distance of $\Delta E^*_{Luv} = 10$ in our encoding which is conservatively larger than the JND.

Texture is mapped to grain (or frequency) of the texture (see Figure 2(d)). Ware and Knight [WK92] refer to this as *contrast*, and is one of the dimensions of texture along with size and orientation. While there has been previous work on generating visually discriminable textures (e.g., [BHW05, CR68, HBW06]), they do not explicitly measure a perceptible difference. Since texture is perceived as the ratio of white (e.g., the gap separating two marks) and black, we can use these features to determine a texture difference. First, we use connected component analysis [HZ04] to detect the amount of black $b \in [0, -1]$, and the amount of white $w \in [0, +1]$ between each gap. We then calculate the difference between two textures $t_i = (b_i, w_i)$ and $t_j = (b_j, w_j)$ as:

$$|t_i - t_j| = |(b_j + w_j) - (b_i + w_i)| \tag{4}$$

To test for this JND, we performed a pilot test with five participants. We showed each participant 54 randomly selected texture pairs side-by-side on a fifteen inch laptop display, and asked whether they were equal, or not equal. Participants could only respond using the keyboard to eliminate movement delays, and we recorded their error rate and response time. During pilot testing, we found a distance $|t_i - t_j| > 0.01$ as an approximate to a JND.

Orientation is mapped to a line rotated between $[5°, 175°]$ to avoid ambiguous orientations (Figure 2(e)). A number of experiments indicate that line orientation discrimination ranges between $[1.06°, 6.44°]$ as a factor of length [VVO86]. Our vision is particularly sensitive to changes in orientation from lines that are vertical and horizontal where a JND is as small as 0.71° [OVV84, VO85]. The orientation between consecutive elements used in our mapping changes by 11.3°, which is significantly greater than this JND.

Shape is mapped to the number of *spikes* of a star-shaped glyph (see Figure 2(f)). For our studies, it is important that

participants can extract the underlying values from our chosen shape encoding. From a restricted shape set, we considered two designs: (1) elementary shapes (number of sides), and (2) star shape glyphs (number of spikes).

To measure shape difference, we use image moment statistics [Hu62] and found that star-shaped glyphs, according to this metric, converge to a greater JND. We then performed a second pilot study, which did not find a clear difference in performance. Participants were slightly more accurate with star-shaped glyphs (Error = 11.11%) when compared to elementary shapes (Error = 12.04%). Therefore, we chose star-shaped glyphs for our encoding.

4. Experimental Overview

In order to compare the perceptual orderability of visual channels, we performed a within subject experiment design to analyse orderability based on two criteria:

- **How ordered is it?** — we measure the perceived orderedness in a sequence of elements from 1 (unordered) to 5 (ordered).
- **Which is smallest? Which is largest?** — we assess the ability to identify whether a target element has the smallest value, largest value, or neither. Both accuracy and response time are measured.

We therefore developed two experiments, one for each question, which we conducted using Amazon's Mechanical Turk [HB10, KZ10, LFK*13] online crowdsourcing platform. Each study followed a similar experimental procedure. Participants were first introduced to the experiment interface through a video tutorial embedded on the experiment web page. These demonstration allowed participants to gain familiarity with the interface prior to the experiment, and to understand the required task. For each experiment, we recorded the following details: gender, age group, whether they were colour blind, and the device they were using. After the study, each participant completed a qualitative survey:

1. *When the sequence of elements are **ordered**, how difficult did you find the task? (Easy) 1 - 5 (Hard)*
2. *When the sequence of elements are **unordered**, how difficult did you find the task? (Easy) 1 - 5 (Hard)*
3. *When the sequence of elements are **unordered**, how difficult did you find the task? (Easy) 1 - 5 (Hard)*

Participants could provide free form comments as well.

4.1. Interface

We developed the web-based interface using HTML, Javascript and PHP. The experiment interface consists of a single view of the visual stimuli, and the question at the top of the screen as shown in Figure 3. A series of radio buttons appears at the bottom of the interface indicating the possible answers. Once a radio button is selected, participants then confirm their response by clicking the large submit button at the bottom of the screen, and the next stimuli is presented in the view. A progress bar is also provided on the top right corner of the screen. Due to the repetitive nature of the task, such a visual cue helps participants monitor progress, and reduce the risk of a participant rushing to complete the study.

4.2. Pilot Studies

Lab-Based Pilot We conducted a pilot study using five participants with the following focus: (1) evaluate interface usability, (2) estimate completion time, and (3) evaluate the experimental design. For each experiment, there were 168 trials for each condition: 7 visual channels $\times$ 8 disorder levels $\times$ 3 repetitions. Repetitions in Experiment 2 corresponded to the three possible answers of: smallest, largest, neither. Each trial was randomised for every participant. After each experiment, the participants gave feedback individually.

Most participants found the interface intuitive to use. However, some mentioned that the radio buttons were too small, comment which we addressed by enlarging the buttons in the final interface (see Figure 3 for an example). Since we cannot control the environment of online contributors (e.g., preventing remote participants, also referred to as "workers", from taking long breaks), we designed our experiments with a strong emphasis on objective (2) to minimise fatigue, and encourage workers continuity towards completion of the study. During pilot testing, participants felt Experiment 1 was too long ($\sim$45 minutes). On the other hand, participants felt comfortable with Experiment 2 ($\sim$15 minutes). We therefore modified the conditions in Experiment 1 by reducing the levels of disorder from 8 $(\eta_1,\ldots,\eta_8) \rightarrow 5$ $(N_1,\ldots,N_5)$. To avoid confusion between different disorder sets we use η_i to refer to the set of 8 disorder levels and N_i to refer to the set of 5 disorder levels. The new N_i disorder set includes only those η_i for which the variation was found significant in our pilot study, meaning: $\eta_1 = 1.0$, $\eta_2 = 0.97$, $\eta_4 = 0.90$, $\eta_6 = 0.71$, and $\eta_8 = 0.12$.

Mechanical Turk Pilot In order to verify the experimental updates derived from our initial pilot, as well as gauge the behaviour of online contributors, we ran a second pilot of 20 participants using Mechanical Turk. This data is not included in our overall analyses.

Many online platforms such as Mechanical Turk provide an option to target jobs to skilled contributors only as an approach to improve the quality of the crowdsourcing data. We apply this feature in our studies. Each participant can only perform the study once. The maximum time allowed for each session is one hour. We found that all participants finished within time and no further issues are reported.

4.3. Crowdsourcing Consistency

When running crowdsourcing studies, we need to ensure the worker is engaged with the task. In our pilot, we observed

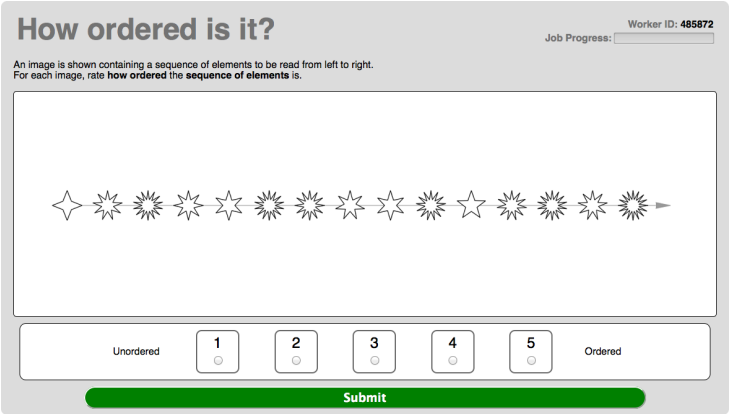


Figure 3: *Interface for Experiment 1: How ordered is it? Shape is tested in this example.*

that this was not always the case and some participants selected the same answer for all stimuli. Previous crowdsourcing studies [CSD*09] propose the use of consistency checks to filter out non-invested participants and we propose a set of checks based on our Mechanical Turk pilot:

- **C1.** Errors are close to chance (e.g., $\mu > 0.66$).
- **C2.** The distribution of answers in an experimental task is close to zero ($\sigma < 0.2$). Participants are choosing the same answer for most questions.
- **C3.** We check how reliable a user's answer is for a known data condition (e.g., an ordered sequence for each visual channel) by looking at answer variance.

We use checks **C2** and **C3** for *How ordered is it?*. As we have a range of error values (difference between entered and correct answer), **C1** is not appropriate. For *Which is smallest? Which is largest?*, we use **C1** and **C2** since there is a right/wrong answer to the task. Approximately 95% of the data collected passes all our criteria and is included.

5. Experiment 1: How ordered is it?

The goal of our first experiment is to investigate whether different visual channels affect the perceptual order of a sequence of elements. In addition, we hypothesise that the perceived order of different visual channels are more/less sensitive to disorder than others. Such a property can be informative in improving the performance of various analytics tasks. We test this hypothesis through an experiment conducted on Amazon Mechanical Turk. Participants were asked to rate how ordered a sequence of elements is using 1D plot graphs. Each 1D plot showed 15 data samples which we mapped using each visual channel (Section 3.3).

Participants 115 contributors on Mechanical Turk (paid $1.00) participated in the experiment. Two participants reported they were colour-blind, and their data was discarded. Another three participants failed our consistency checks, and were also removed. Therefore, 110 participants (62 male and 48 female) were included in our final analysis. The devices used were: 49 desktop, 54 laptop, 6 tablet, and 1 phone.

Experimental Design The experiment followed a within subject design and consisted of seven visual channels, five disorder levels, and three repetitions (105 trials). At the start of the experiment, participants completed a training block of 14 sample questions showing an ordered and unordered sequence for each visual channel, making a total of 119 trials. To limit any confounding effects of fatigue and learning, the two blocks of trials were randomised per participant. To counteract memory, the data set used for each repetition of a disorder level was generated independently with equal correlation coefficients to two significant figures. The correlation coefficient measured at each disorder is: $N_1 = 1.0$, $N_2 = 0.97$, $N_3 = 0.90$, $N_4 = 0.71$, and $N_5 = 0.12$.

For a single trial, we showed participants a 1D plot and asked them to rate *how ordered* the elements are using a 5-point Likert scale that corresponds to 1 (unordered) through to 5 (ordered) respectively (see Figure 3). After observing the sequence of elements, participants provided their rating by clicking on one of the radio buttons below the image followed by the submit button. We measured both answer and response time after each trial, before the next stimuli was automatically displayed. In order to overcome change blindness, we use an Inter-Stimulus-Interval (ISI) [BBK09] in between each stimuli. This served for two purposes: 1) to indicate that the stimulus has actually changed, and 2) to 'reset' the visual system and remove any possibilities of a previous stimuli influencing the following one.

5.1. Results

We perform our analysis in two stages. First, we consider the effect of perceptual orderability based on visual channel overall, as this is our primary research question. To check for normality, we ran a Shapiro-Wilk test on each distribution. We find that the data is not always normally distributed, and therefore use a non-parametric Friedman's test with standard

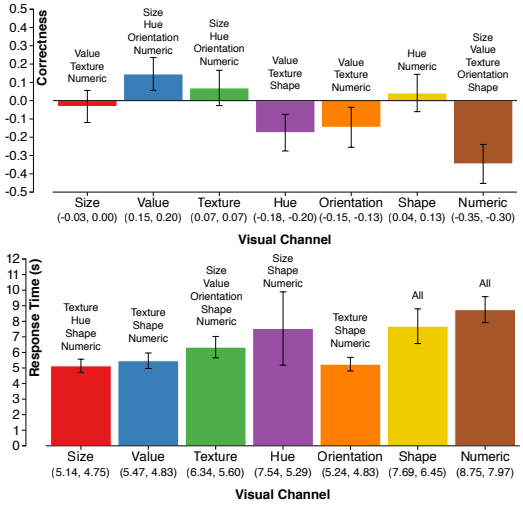


Figure 4: *Correctness (top) and response time (bottom) of each visual channel in Experiment 1. Significant differences are listed above each bar, with (mean, median) values indicated below. Error bars show 95% confidence intervals.*

statistical level $\alpha = 0.05$ to determine the statistical significance between conditions. Post-hoc analysis was conducted using Nemenyi-Damico-Wolfe-Dunn test.

We then analyse how visual channel affects perceptual orderability under the condition of disorder. When dividing the data by disorder, we apply a Bonferroni correction, reducing the significance level to $\alpha = 0.01$ for this second second-level analysis. Post-hoc analysis was conducted as above.

5.1.1. Orderability of Visual Channel Overall

Figure 4 shows the mean correctness and response time results with significant pairwise differences indicated over each bar. We measure correctness by taking the difference between the expected orderability rating for a sequence with disorder N_i and a user's actual response. Given a disorder level N_i with $i = 1 \ldots 5$, visual channel v, and user's corresponding response $u_{i,v}$, we derive correctness as $\delta_p = (5 - i + 1) - u_{i,v}$ according to our 5-point Likert scale. We find a significant difference in both correctness ($\chi^2(6) = 149.05$, $p \ll 0.05$) and response time ($\chi^2(6) = 289.61$, $p \ll 0.05$).

5.1.2. Disorder Sensitivity of Visual Channels

Figure 5 compares the effects of disorder (i.e., measured orderedness) to the perceived orderedness under different visual channels. Our results show that disorder has a significant effect on the perceived order ($p \ll 0.01$). The significant differences are shown in Table 1.

5.2. Discussion

Overall Our results show significant evidence that different visual channels affect the perception of order in sequences. For example, participants tend to rate a sequence as being more ordered using value, while other visual channels (e.g., hue) are often judged as being less ordered. This suggests that if a visualisation task involves detecting ordered or unordered patterns, mappings to different visual channels will lead to different judgements.

Overall, value and texture lead to higher degree of perceived orderedness. Given the encoding for texture we used, this makes sense as both channels can be viewed as a form of grayscale. Surprisingly, we find shape to be orderable. However, using shape also led to increased response time overall by 1.23s compared to value and texture. Looking at our error measure, we see that size has a mean and median very close to zero, meaning that it closely approximates the disorder present in the data.

If the goal is to detect an ordered pattern, participants respond at least three seconds faster per task with visual channels of a higher degree of perceived orderedness (e.g., value and texture) when compared to hue and numeric. However, if the task is to detect an unordered pattern, participants respond faster using orientation. Since visualisation users often engage in many trend extracting tasks both within and across charts, this may improve both performance (e.g., detecting an ordering), and reduce cognitive load depending on the choice of visual channel used.

Sensitivity to Disorder Looking at the effect sizes of disorder (see Figure 5), we observe that the perceived orderedness of different visual channels decreases at different rates. This behaviour is consistent across all visual channels tested, and that the relationship between measured orderedness (i.e., disorder level), and perceived orderedness is non-linear. There is a common trend in the middle where the perceived orderedness dips between the measured orderedness $N_2 = 0.97$ and $N_4 = 0.71$ illustrated by the slope of the curves. For example, shape decreased in perceived orderedness by 1.55 in this range. Comparing this to a visual channel such as hue, we see a greater difference of 1.74. This difference indicates that some visual channels (e.g., shape) are less sensitive to disorder such that viewers may perceive an ordered pattern that does not exist in the data. Similarly, other visual channels (e.g., hue) are more sensitive to disorder such that viewers are less likely to see an ordered pattern.

We present further comparative analyses by plotting visual channels against an average curve (dashed line) as shown in Figure 5. There are two observable clusters above and below the curve. Our results show that value, texture, and shape seem to lead to participants to estimate a higher degree of orderedness than orientation, size, hue, and numeric. We also find that this gap is significant as shown in Table 1. Notice that for N_4, the visual channels that lie above the average (e.g., value, shape and texture) are significantly different to those below (e.g., size, orientation, hue and numeric) indicated by the p-values highlighted in red.

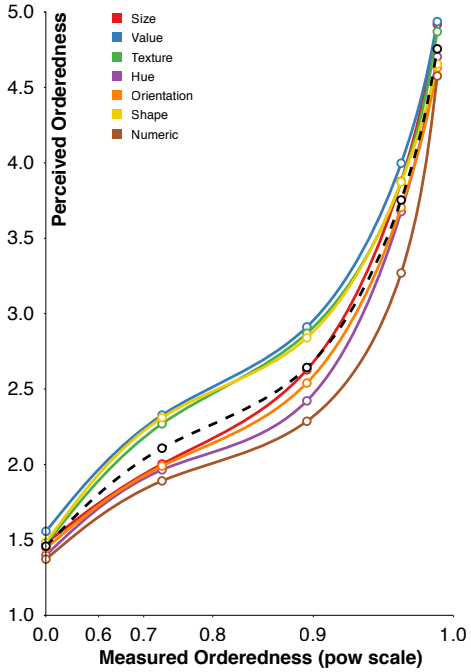


Figure 5: *Measured orderedness (correlation coefficient) versus perceived orderedness. Points are the perceived orderedness for each visual channel given its measured orderedness (power scale). Average plotted as dashed line.*

The same differences can mostly be seen in N_3. At this condition, size and orientation perform closer to the average, and thus we find this gap becomes less distinct. One interesting observation is that size starts and finishes above and below the average as measured orderedness decreases. This result shows that the orderability of size is less sensitive to low levels of disorder, but more sensitive to high levels of disorder, which makes size a good visual channel for detecting both ordered and unordered patterns.

For very ordered and very unordered sequences (e.g., N_5 and N_1), different visual channels lead to almost the same judgement. Participants mostly rate the order of such sequences to be 1 (unordered) and 5 (ordered) respectively, independent of the visual channel used. We therefore found fewer significant differences. At such levels of measured orderedness, the decision becomes a binary process (e.g., ordered, or not ordered). Hence, we expected such results to appear in our data. Based on our findings, we conclude that human's judgement of orderedness is sensitive to the choice of visual channel. The amount of difference depends on the level of orderedness. We look at how this may affect the performance of a visual task in Experiment 2.

6. Experiment 2: Which is smallest? Which is largest?

The goal of our second experiment is to investigate how different visual channels affects the judgement of min and max values in a sequence of elements. In particular, we hypothesise that visual channels that are perceptually orderable may improve a user's performance. To test our hypothesis, we adopt the method proposed by Bertin [Ber83] on ordered perception, which follows a similar design to Experiment 1.

Table 1: *Experiment 1 results. Post-hoc p-values of disorder vs perceived orderedness for different visual channels. Significant differences are highlighted in red using a Bonferroni corrected* $\alpha = 0.01$.

Pair-wise test	Disorder				
	N5	N4	N3	N2	N1
Value - Shape	0.86	0.97	0.53	0.99	0.00
Value - Texture	0.76	0.83	0.96	0.49	0.77
Value - Size	0.98	0.00	0.00	0.84	0.98
Value - Orientation	0.40	0.00	0.00	0.05	0.00
Value - Hue	0.07	0.00	0.00	0.00	0.00
Value - Numeric	0.00	0.00	0.00	0.00	0.00
Shape - Texture	1.00	0.99	0.97	0.89	0.00
Shape - Size	0.99	0.00	0.30	0.99	0.00
Shape - Orientation	0.98	0.00	0.03	0.26	0.35
Shape - Hue	0.73	0.00	0.00	0.02	0.40
Shape - Numeric	0.25	0.00	0.00	0.00	1.00
Texture - Size	0.99	0.00	0.03	0.99	0.99
Texture - Orientation	0.99	0.00	0.00	0.94	0.30
Texture - Hue	0.84	0.00	0.00	0.45	0.26
Texture - Numeric	0.36	0.00	0.00	0.00	0.00
Size - Orientation	0.90	0.98	0.97	0.67	0.06
Size - Hue	0.45	0.99	0.00	0.16	0.05
Size - Numeric	0.09	0.35	0.00	0.00	0.00
Orientation - Hue	0.98	0.99	0.06	0.97	1.00
Orientation - Numeric	0.72	0.84	0.08	0.00	0.38
Hue - Numeric	0.98	0.59	1.00	0.00	0.43

Participants 88 Mechanical Turk participants (paid $1.00) took part in the experiment. One participant failed our consistency checks and was removed. Therefore, 87 participants (42 male and 45 female) were included in our final analysis. The devices used were: 40 desktop, 42 laptop and 5 tablet.

Experimental Design The experiment followed a within subject design. Participants saw a series of 1D plots containing a sequence of elements with one target element highlighted in a red bounding box as shown in Figure 6. For each trial, we asked participants to identify whether the highlighted element has: (1) the **smallest** value, (2) the **largest** value, or (3) **neither**. The experiment required participants to answer all three question types under the following conditions: seven visual channels and eight disorder levels $\eta_1, \ldots, \eta_8$ (see Section 3.2) resulting in 168 trials. Where question type (3) is tested, we choose the median value to represent the *neither* condition. Similar to Experiment 1, we first showed a training block of 16 sample questions unrelated to later trials. Each participant therefore completed a total of 184 trials which were randomised in both blocks. We used a similar interface to Experiment 1, with a list of answers presented below the stimuli. Participants respond by selecting one of these answers. The keywords of each question type is highlighted in bold text to enable easy identification. We measured both error rate and response time.

6.1. Results

Figure 7 shows mean error and response time results with significant pairwise differences indicated above each bar.

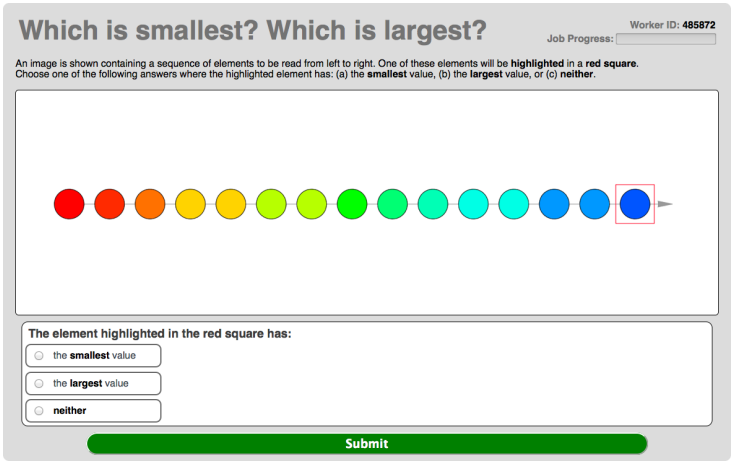


Figure 6: *Interface for Experiment 2: Which is smallest? Which is largest?*

Since the data we collected is not normally distributed, we once again apply non-parametric statistics. A Friedman's test shows significant differences in error rate ($\chi^2(6) = 276.15$, $p \ll 0.05$) and response time ($\chi^2(6) = 121.46$, $p \ll 0.05$) under the effects of visual channel. Post-hoc analysis was conducted similar to Experiment 1 with $\alpha = 0.05$.

6.2. Discussion

Overall Given a sequence of elements, we find that different visual channels have a significant effect on the error rate of min-max judgements. Participants produced fewest errors with numeric. This is what we expected, since the value is explicitly given and the number of samples shown is relatively small. Despite the numerical values observed being not very complex (e.g., 3 digits) which meant that the cognitive load on short term memory is relatively low, we find participant's spending a significant amount of time searching numbers as shown in their response data (see Figure 7(bottom)). Thus visual encodings seem to help in this task.

Judgements using size also produced fewer errors which cannot be explained in our data. To investigate potential reasons, we refer back to Bertin's classification of visual properties and find that size is the only channel which is quantitative [Ber83]. A quantitative variable means we can perceive the numerical ratio between two sizes, for example, this circle is twice the size of that circle. This may explain why size performed so well in our tests. However, further experimentation is needed to fully understand the exact causes.

Conversely, error rate significantly increased when using hue and orientation. It is easy to see that such encodings can be misleading (e.g., they do not impose a universal perceived order) and would therefore produce more errors in such a task. This is consistent with previous claims, for example, the error-prone use of rainbow colour-mapping within the visualisation community [BT07].

We find few significant effects in our response time data.

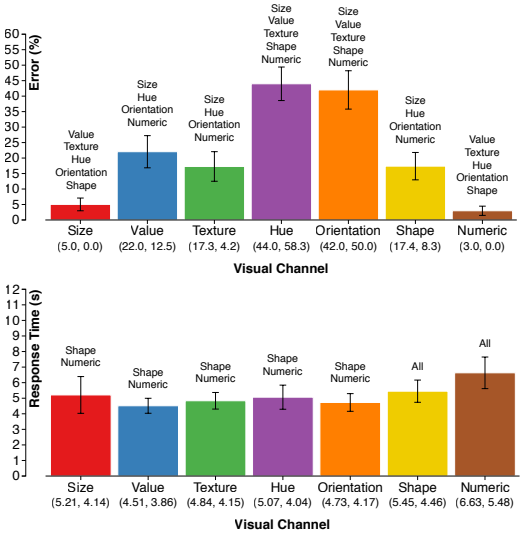


Figure 7: *Comparing the effects of visual channel against error rate (top) and response time (bottom) in Experiment 2. Significant differences are listed above each bar, with (mean,median) values indicated below. Error bars show 95% confidence intervals.*

However, shape and numeric were again the slowest as found in Experiment 1, which supports that both encodings are cognitively demanding, and thus, increases response times.

7. General Discussion

Perceptual Orderability and Min/Max Judgement Overall, we noticed that visual channels that are perceived as ordered in Experiment 1, perform well in Experiment 2. A summary of this combined performance is shown by the two middle axes in Figure 8(top). With the exception of numeric, the ranking is fairly consistent across both experiments. Scaling our tests to a larger number of samples may

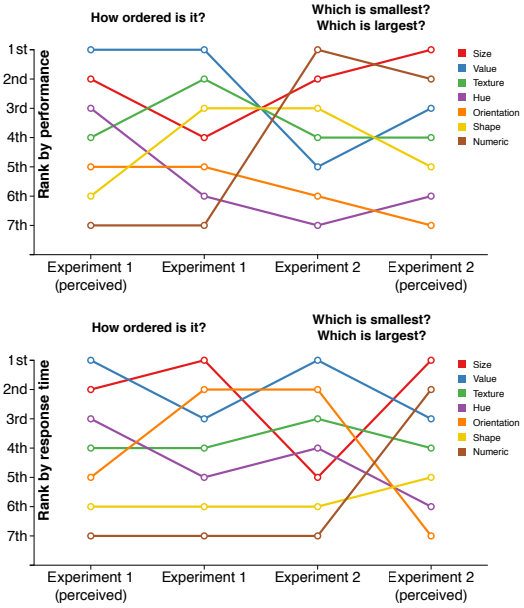


Figure 8: *Ranking of visual channels from worst (7th) to best (1st) for each experiment based on performance (top) and response time (bottom). These measurements are compared to the participants' perceived ranking.*

provide a more accurate ranking that reflects our results, since we predict that numeric will be greatly affected due to its high cognitive demand (see response time). Surprisingly, we find value to perform worse than expected for min and max judgements, considering that participants perceived this visual channel as most orderable. However, it is still significantly better than less orderable channels such as hue and orientation. Our results suggest that in practice, visual orderability can improve the accuracy of visual tasks such as the one presented in our study.

Measured vs. Perceived Difficulty At the end of each experiment, we collected survey data as outlined in Section 4 to understand what participants felt was least and most difficult about the task. Most participants found both experiments to be easier when the sequence of elements are ordered, compared to unordered. In addition, we asked participants to rank the visual channels in terms of their perceived difficulty. Figure 8 compares this feedback against their measured rankings based on average performance and response time. There were no clear overall trends between perceived and actual performance (see Figure 8(top)). However, one observation is there is a negative correlation between hue's performance, and its perceived difficulty. This tells us that participants tend to perceive hue to be less difficult than their actual results. Similarly, shape performed better than what they expected, but not in response time.

Across both studies, we find that the perceived ranking is generally well correlated to participants response times. For example, the ranking of shape, texture, and value remain relatively constant (straight line) as shown in Figure 8(bottom). It suggests that the higher the preference of a visual channel, the faster their response for that task.

Bertin's Categorisation on Ordered Perception In this paper, we investigated the perceptual orderability of different visual channels for ordered and unordered sequences. The original concept described by Bertin show that shape, hue, and orientation are not ordered. However, our crowdsourced results indicate that shape can be orderable. Of course, any arbitrary encoding of shapes will not be orderable as argued by Bertin [Ber83]. The reason behind our results is that our shapes can be considered as using two types of channels. While shape itself is not ordered, we find that counting (e.g., the number of spikes or edges) is. This raises another interesting research question: *"How does the combination of visual channels affect the perceived order?"*. For example, in Experiment 2, by combining value (fastest response time) with numeric (most accurate), do we gain the performance advantage from both in the resulting composition? Further experiments might therefore explore the trade-offs between such combinations.

8. Conclusion

We have presented two experiments to measure the perceptual orderability of visual channels and how they impact the performance of min and max judgements. Our results indicate that, depending on the visual channel selected, an encoded sequence can appear more ordered (value and texture) or more disordered (hue, orientation, and numeric) than the underlying data. Overall, we find that visual channels that appear more ordered improve the performance of min and max judgements. In order to meet the dynamic environment of online contributors, we developed visual designs that maximise the encoding each visual channel has to offer by sampling points significantly above a JND, since we did not want to disadvantage visual channels from each other. A limitation of this design is that perceptual differences between two elements across visual channels may not be equal. Normalising these JND gaps is therefore interesting future work, but will require significant research towards perceptually uniform models, of which only hue and value have confirmed studies. Other areas we would like to investigate include tasks such as categorical search, which is often performed with this type of data.

References

[AR08] AUGUSTIN T., ROSCHER T.: Empirical evaluation of the near-miss-to-weber's law: a visual discrimination experiment. *Psychology Science Quarterly 50*, 4 (2008), 469–488. 3

[BBK09] BUONOMANO D. V., BRAMEN J., KHODADADIFAR M.: Influence of the interstimulus interval on temporal processing and learning: testing the state-dependent network model.

Philosophical Trans. of The Royal Society B 365 (2009), 1865–1873. 5

[Ber83] BERTIN J.: *Semiology of graphics*. University of Wisconsin Press, 1983. 1, 2, 3, 7, 8, 9

[BHW05] BAIR A., HOUSE D., WARE C.: Perceptually optimizing textures for layered surfaces. In *Proc. of the 2nd Symposium on Applied Perception in Graphics and Visualization* (2005), APGV '05, ACM, pp. 67–74. 2, 3

[BKC*13] BORGO R., KEHRER J., CHUNG D. H., MAGUIRE E., LARAMEE R. S., HAUSER H., WARD M., CHEN M.: Glyph-based visualization: Foundations, design guidelines, techniques and applications. In *Eurographics State of the Art Reports* (2013), pp. 39–63. 2

[BT07] BORLAND D., TAYLOR R. M.: Rainbow color map (still) considered harmful. *IEEE Computer Graphics and Applications 27*, 2 (2007), 14–17. 8

[CLP*15] CHUNG D. H. S., LEGG P. A., PARRY M. L., BOWN R., GRIFFITHS I. W., LARAMEE R. S., CHEN M.: Glyph sorting: interactive visualization for multi-dimensional data. *Information Visualization 14*, 1 (2015), 76–90. 1, 2

[CM84] CLEVELAND W. S., MCGILL R.: Graphical perception: theory, experimentation and application to the development of graphical methods. *Journal of the American Statistical Association 79*, 387 (1984), 531–554. 1, 2

[CR68] CAMPBELL F. W., ROBSON J. G.: Application of fourier analysis to the visibility of gratings. *Journal of Physiol 197*, 3 (1968), 551–566. 3

[CSD*09] COLE F., SANIK K., DECARLO D., FINKELSTEIN A., FUNKHOUSER T., RUSINKIEWICZ S., SINGH M.: How well do line drawings depict shape? *ACM Trans. on Graphics 28*, 3 (2009), 28:1–28:9. 5

[dLvW93] DE LEEUW W. C., VAN WIJK J. J.: A probe for local flow field visualization. In *Proc. IEEE Visualization Conf. (Vis '93)* (1993), pp. 39–45. 2

[GSCO12] GINGOLD Y., SHAMIR A., COHEN-OR D.: Micro perceptual human computation for visual tasks. *ACM Trans. on Graphics 31*, 5 (2012), 119:1–119:12. 2

[HB10] HEER J., BOSTOCK M.: Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems* (2010), pp. 203–212. 2, 4

[HBW06] HOUSE D. H., BAIR A. S., WARE C.: An approach to the perceptual optimization of complex visualizations. *IEEE Trans. on Visualization and Computer Graphics 12*, 4 (2006), 509–521. 2, 3

[Hei24] HEICHT S.: The visual discrimination of intensity and the weber-fechner law. *The Journal of General Physiology 7*, 2 (1924), 235–267. 3

[HLNW11] HLAWATSCH M., LEUBE P., NOWAK W., WEISKOPF D.: Flow radar glyphs - static visualization of unsteady flow with uncertainty. *IEEE Trans. on Visualization and Computer Graphics 17*, 12 (2011), 1949–1958. 2

[HS12] HEALEY C. G., SAWANT A. P.: On the limits of resolution and visual angle in visualization. *ACM Trans. on Applied Perception 9*, 4 (2012), 20:1–20:21. 1

[Hu62] HU M.-K.: Visual pattern recognition by moment invariants. *IRE Trans. on Information Theory 8*, 2 (1962), 179–187. 4

[HYFC14] HARRISON L., YANG F., FRANCONERI S., CHANG R.: Ranking visualizations of correlation using weber's law. *IEEE Trans. Visualization and Computer Graphics 20*, 12 (2014), 1943–1952. 2

[HZ04] HARTLEY R. I., ZISSERMAN A.: *Multiple View Geometry in Computer Vision*, second ed. Cambridge University Press, ISBN: 0521540518, 2004. 3

[JJS93] JAYANT N., JOHNSTON J., SAFRANEK R.: Signal compression based on models of human perception. *Proc. of the IEEE 81*, 10 (1993), 1385–1422. 2

[KH16] KAY M., HEER J.: Beyond weber's law: A second look at ranking visualizations of correlation. *IEEE Trans. on Visualization and Computer Graphics 22*, 1 (2016), 469–478. 2

[KZ10] KOSARA R., ZIEMKIEWICZ C.: Do mechanical turks dream of square pie charts? In *Proc. of the 3rd BELIV'10 Workshop: BEyond time and errors: novel evaLuation methods for Information Visualization* (2010), ACM, pp. 63–70. 4

[LCP*12] LEGG P. A., CHUNG D. H. S., PARRY M. L., JONES M. W., LONG R., GRIFFITHS I. W., CHEN M.: Matchpad: Interactive glyph-based visualization for real-time sports performance analysis. *Computer Graphics Forum 31*, 3pt4 (2012), 1255–1264. 2

[LFK*13] LIN S., FORTUNA J., KULKARNI C., STONE M., HEER J.: Selecting semantically-resonant colors for data visualization. *Computer Graphics Forum (Proc EuroVis)* (2013). 2, 4

[LHH*12] LIANG J., HUA J., HUANG M. L., NGUYEN Q. V., SIMOFF S.: Rectangle orientation in area judgment task for treemap design. In *Proc. of the 24th Australian Computer-Human Interaction Conference* (2012), OzCHI '12, ACM, pp. 349–352. 2

[Mac86] MACKINLAY J.: Automating the design of graphical presentations of relational information. *ACM Trans. on Graphics 5*, 2 (1986), 110–141. 1

[MPRSDC12] MAGUIRE E., P. ROCCA-SERRA S.-A. S., DAVIES J., CHEN M.: Taxonomy-based glyph design: with a case study on visualizing workflows of biological experiments. *IEEE Trans. on Visualization and Computer Graphics* (2012). 2

[OVV84] ORBAN G. A., VANDENBUSSCHE E., VOGELS R.: Human orientation discrimination tested with long stimuli. *Vision Research 24*, 2 (1984), 121–128. 3

[VO85] VOGELS R., ORBAN G. A.: The effect of practice on the oblique effect in line orientation judegements. *Vision Research 25*, 11 (1985), 1679–1687. 3

[VVO86] VANDENBUSSCHE E., VOGELS R., ORBAN G. A.: Human orientation discrimination: Change with eccentricity in normal and amblyopic vision. *Investigative Opthalmology & Visual Science 27*, 2 (1986), 237–245. 3

[War02] WARD M. O.: A taxonomy of glyph placement strategies for multidimensional data visualization. *Information Visualization 1*, 3-4 (2002). 2

[War08a] WARD M. O.: Multivariate data glyphs: Principles and practice. In *Handbook of Data Visualization* (2008), Chen C.-H., Hardle W., Unwin A., (Eds.), Springer Handbooks Comp. Statistics. Springer, pp. 179–198. 2

[War08b] WARE C.: *Visual Thinking: for Design*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2008. 1, 3

[WK92] WARE C., KNIGHT W.: Orderable dimensions of visual texture for data display: Orientation, size and contrast. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems* (1992), CHI '92, ACM, pp. 203–209. 3

[WWH98] WILKINSON F., WILSON H. R., HABAK C.: Detection and recognition of radial frequency patterns. *Vision Research 38*, 22 (1998), 3555–3568. 3