Eurographics Conference on Visualization (EuroVis) 2017
J. Heer, T. Ropinski and J. van Wijk
(Guest Editors)

Volume 36 (2017), Number 3

An Empirical Study on the Reliability of Perceiving Correlation Indices using Scatterplots

Varshita Sher[1], Karen G. Bemis[2], Ilaria Liccardi[3], and Min Chen[1]

[1] University of Oxford, UK, [2] Rutgers, The State University of New Jersey, USA, and [3] Massachusetts Institute of Technology, USA

Abstract

Scatterplots have been in use for about two centuries, primarily for observing the relationship between two variables and commonly for supporting correlation analysis. In this paper, we report an empirical study that examines how humans' perception of correlation using scatterplots relates to the Pearson's product-moment correlation coefficient (PPMCC) – a commonly used statistical measure of correlation. In particular, we study human participants' estimation of correlation under different conditions, e.g., different PPMCC values, different densities of data points, different levels of symmetry of data enclosures, and different patterns of data distribution. As the participants were instructed to estimate the PPMCC of each stimulus scatterplot, the difference between the estimated and actual PPMCC is referred to as an offset. The results of the study show that varying PPMCC values, symmetry of data enclosure, or data distribution does have an impact on the average offsets, while only large variations in density cause an impact that is statistically significant. This study indicates that humans' perception of correlation using scatterplots does not correlate with computed PPMCC in a consistent manner. The magnitude of offsets may be affected not only by the difference between individuals, but also by geometric features of data enclosures. It suggests that visualizing scatterplots does not provide adequate support to the task of retrieving their corresponding PPMCC indicators, while the underlying model of humans' perception of correlation using scatterplots ought to feature other variables in addition to PPMCC. The paper also includes a theoretical discussion on the cost-benefit of using scatterplots.

1. Introduction

A variety of visualizations, such as scatterplots and parallel coordinates plots, have been used extensively for observing correlation in data. Human perception of correlation from visualizations is likely to feature a number of cognitive influences. These may include gestalt laws of grouping, shape interpretation, learned knowledge about statistical indicators such as *Pearson's product-moment correlation coefficient* (PPMCC). In recent years, there has been growing interest in studying the underlying models of human perception of correlation (e.g., [RB10, HYFC14, KH16]).

Because an empirical study can typically examine only a few variables and a small number of variations per variable, the previous studies, such as [LMVW10,RB10,HYFC14], focused on visual stimuli featuring data points with normal distributions. This work is an extension of the previous experiments by examining human perception of correlation under different conditions, e.g., different densities of data points, different levels of symmetry of data enclosure, and different patterns of data distribution. We hope to establish how the variations of these conditions impact human perception in relation to the statistical indicator PPMCC. If the impact were insignificant, it would suggest that human perception of correlation might be modelled by a relatively simple function, such as the ones proposed in [SH78,MS92,BK79,CDM82,DAAK07], and more recently in [RB10,HYFC14,KH16]. If the impact were significant, an appropriate underlying model would have to encode the functional dependency of humans' perceived correlation on many visual features of datasets in addition to their PPMCC indicators. Developing functional models for human perception and cognition in visualization is an ultimate goal [CM84,Cum13]. This study aims to provide necessary observations about what variables a model for correlation estimation may depend on.

Although we used PPMCC as the reference indicator in this work, it is necessary to note that there are many variances of PPMCC, such as weighted, reflective, and scaled coefficients. There are also other mathematical functions for measuring correlation, e.g., mutual information, rank correlation coefficients, and distance correlation. It is also necessary to assert that values of a mathematically computed correlation function should not be regarded as a ground truth of correlation. In other words, when human perception differs from a computed indicator, it does not always mean that the human perception is wrong. Nevertheless, one major advantage of using a mathematical function, such as PPMCC, is its determinism and objectivity. We thus used PPMCC as a reference function in our study. In our discussion, we will use the word "offset" (instead of "error") to denote the amount of deviation from a human estimation of correlation based on visualization to a value computed using PPMCC.

In this work, we consider several types of controlled variations, including different *PPMCC values*, *numbers of data points* (ndp), levels of *reflective symmetry* of data enclosures, levels of *progressive symmetry* of data enclosures, and patterns of *spatial distribution*. We measure the offsets of human perception of correlation when changing 1 or 2 control variables. In addition, we also estimate the *just notifiable difference* (JND) using different reference points to define the sensory ranges. Before the study, we speculate on the likelihood of effects due to these controlled variations. Nevertheless, following the protocol of statistical inference, we define the following *null hypotheses* as the unbiased default positions to be evaluated.

1. Varying a PPMCC value does not affect offsets.
2. Varying the density of data points does not affect offsets.
3. Varying reflective symmetry does not affect offsets.
4. Varying progressive symmetry does not affect offsets.
5. Varying spatial distribution does not affect offsets.
6. Varying a PPMCC reference point does not affect JND.

2. Related Work

It is commonly believed that the scatterplot, which has also been called scatter diagram, scattergram, and scattergraph, was first introduced in the 18th century, during the boom of statistics graphics [FD05]. There are several variations and extensions, including animated scatterplot, scatterplot matrix [EDF08] and glyph-based scatterplot [CLP*15]. It can be used for observing correlation, clusters, and outliers (e.g., [LS89, KARC15]).

Psychologists and statisticians have been interested in the perception of correlation since the 1960s. Using two law cases, Bobko and Karren made a compelling argument of the risk of perceptual inconsistency [BK79]. A variety of stimuli, apparatuses and study methods were used, including audio signals (e.g., [Pol60]), data tables (e.g., [Erl66]), paper-based plots (e.g., [MS92]), projector (e.g., [CDM82]), computer screen (e.g., [LP89]), mails (e.g., [BK79]), individual interviews (e.g., [DAAK07]), and so on. The numbers of participants varied from 6 [Pol60] to over 100 [CDM82], with 20∼50 being the most common range.

The most widely studied phenomenon is human participants' *under and overestimation* in relation to different PPMCC values C_R. Pollack first observed the varying levels of sensitivity between low and high C_R values in a study with 6 participants [Pol60]. The underestimation when $0 < C_R < 1$ was subsequently confirmed by many others (e.g., [Erl66, SH78, BK79, DAAK07]). Some have also observed overestimation when $-1 < C_R < 0$ [Erl66, BK79]. Several functional models were proposed to approximate the sampled human estimation of PPMCC values, ranging from simply C_R^2 [SH78, MS92] to $1-(1-C_R^2)^{0.5}$ [BK79, CDM82], to more general polynomials [DAAK07], and to a logarithmic function [Ren13].

Several studied effects of different stimuli variations. Pollack [Pol60], Erlick [Erl66], Cleveland et al. [CDM82], Lane et al. [LAK85], and Lauer and Post [LP89] reported the effects caused by changing the *size of point clouds*. Bobko and Karren [BK79] reported no significant effect in varying *slope*, but this was contradicted by Lane et al. [LAK85], Lauer and Post [LP89], and Meyer and Shinar [MS92]. Lauer and Post [LP89] reported the effect caused by varying *number of data points*, but Doherty and Anderson [DAAK07] reported no significant effect. Some observed the effects of *knowledge* and *training* by studying participants with different education backgrounds (e.g., [LAK85, MS92]), while others showed that incorrect estimations are common among professionals with good statistics knowledge (e.g., [SH78, BK79]).

In addition, Bobko and Karren [BK79] observed the effects caused by *outliers* and *non-elliptical shapes*. Meyer and Shinar [MS92] reported some effects caused by varying among three *shapes of point clouds*. Konarski [Kon05] found effects caused by *contamination* (points in a different distribution). Bobko and Karren [BK79] observed the effect caused by *removing the middle section* of a point cloud. Beach and Scopp [BS66] studied the participants' *confidence of estimation* in relation to C_R. Several researchers studied a phenomenon called *illusory correlation* (e.g., [Cha67, HG76]). Using scatterplots as an example, Wickham et al. presented a discourse on the mutual complement of statistics and visualization [WCHB10]. More recently, Pandey et al. examined the perception of scatterplot similarity [PKF*16], while Correll and Heer studied the perception of regression in scatterplots [CH17].

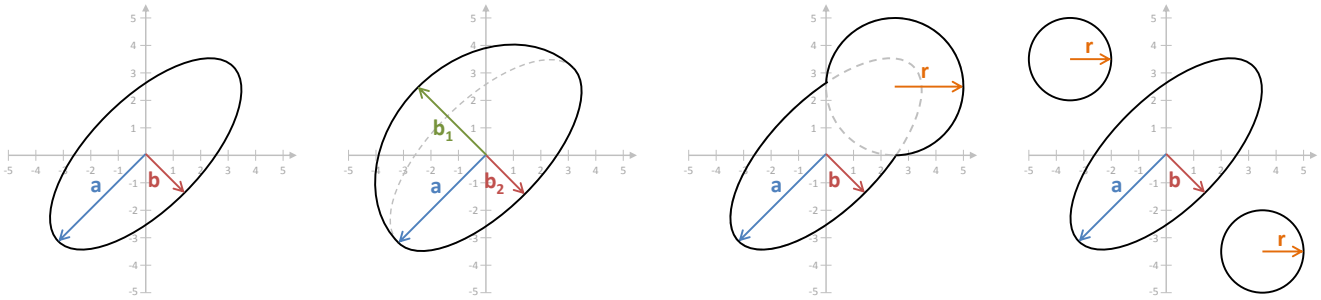
Doherty and Anderson [DAAK07] first studied the *JND* in the perception of correlation. Rensink and Baldridge [RB10] reported a study, suggesting JND as a linear function of C_R. Harrison et al. [HYFC14] reported a new set of results, suggesting a linear relation between the change of perception and that of C_R. Kay and Heer [KH16] proposed a log-linear relation based on an analysis of the results of [HYFC14].

For understandable reasons, some previous studies on the effects of *number of data points*, *shapes*, *outliers*, and *contaminations* were very small. For example, Lauer and Post [LP89] studied 200 vs. 400 data points (27 participants, reporting effect without significance analysis), while Doherty and Anderson [DAAK07] studied 9 vs. 100 data points (20 participants, reporting no significant effect). In Bobko and Karren's study [BK79], there was only 1 stimulus featuring outliers, and 1 stimulus featuring a non-elliptical shape, and they compared such stimuli with stimuli that have different C_R values, though C_R was known to have an effect on estimation. In Meyer and Shinar's study [MS92], among three shapes, there were significant differences in two pairwise comparisons. Our hypotheses **H2**∼**H5** were designed to provide a more comprehensive coverage in these aspects. Meanwhile, we use **H1** and **H6** to synchronize with the previous findings on *under and overestimation* and *JND*.

3. Experiment Overview

This study was conducted as a controlled study in a laboratory setup. The design of the study will be detailed in Section 4, its enactment in Section 5, and the analysis of its results in Section 6. Here we outline the major variables involved in this experiment.

Given a set of bivariate data points, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, we use *Pearson's product-moment correlation coefficient* (PPMCC) [Pea95] as the reference value of computer-estimated correlation. Meanwhile, human participants may visualize the corresponding scatterplot, and derive an estimation of correlation within the same range $[-1, 1]$ as PPMCC. Referring to the six null



(a) PPMCC (b) reflective symmetry (c) progressive symmetry (d) distribution

Figure 1: *Geometric specifications of four types of variations.*

hypotheses in Section 1, the dependent variable measured for the first five hypotheses is the offset $\delta = e - C_R$ between the reference PPMCC value C_R and a participant's estimation e. For the 6th hypothesis, the dependent variable γ, which has three valid values, records the notifiable difference in comparing a participant's estimations based on two juxtaposed scatterplots, S_u and S_v. With the *direct rating* method for JND [Sjö85], given two estimations e_u and e_v in relation to $C_{R,u}$ and $C_{R,v}$ where $C_{R,u} \neq C_{R,v}$,

$$\gamma = \begin{cases} \text{correct} & \text{if } (C_{R,u} - C_{R,v})(e_u - e_v) > 0 \\ \text{same} & \text{if } (e_u - e_v) = 0 \\ \text{wrong} & \text{if } (C_{R,u} - C_{R,v})(e_u - e_v) < 0 \end{cases} \quad (1)$$

The mapping from a set of bivariate data points D to its PPMCC value C_R is a many-to-one mapping. There are numerous ways of varying D while maintaining the same C_R. In this work, we use attributes of a geometric enclosure of D as the control variables. We generate data points within such an enclosure randomly with a uniform distribution. To obtain statistically meaningful observations, each null hypothesis is associated with 1 or 2 control variables.

H1: Varying PPMCC. For this hypothesis, we use an elliptical enclosure that is symmetric in relation to both of its axes. Since it was never a pre-condition that PPMCC should only be applied to data with a bivariate normal distribution, using an elliptical enclosure brings about three benefits. (i) It does not require the sampling in an unbounded space as a bivariate normal distribution would do. (ii) Controlling the two elliptical axes is more intuitive than controlling mean and standard deviation. (iii) It enables more even distribution of data points in visual stimuli without cluttering (e.g., in the case of a low standard deviation). Figure 1(a) illustrates an elliptical enclosure. We vary the two axes a and b, in conjunction with random generation of 80 data points, to obtain stimuli with $C_R = -1, -0.9, \ldots, 0, \ldots, 0.9, 1$. For a positive C_R, a is the major axis, and is fixed as $a = 7$. For a negative C_R, b is the major axis, and is fixed as $b = 7$. For $C_R = 0$, $a = b = 7$.

H2: Varying Density. For this hypothesis, there are two control variables, the number of data point *ndp*, and minor axis b of an elliptical enclosure. The former takes values $ndp = 40, 60, 80, 100, 120$, and the latter takes values $b = 4, 3, 2.5$, which correspond to $C_R = 0.3, 0.5, 0.7$. Similar to **H1**, the major axis is set to $a = 7$, and data points are randomly generated within the elliptical enclosure.

H3: Varying Reflective Symmetry. This hypothesis explores one type of geometric variation that is commonly exhibited in a data set. The underlying reason for such a variation is due to the divergence of the two statistical distributions governing the bivariate data points. When the two normal (or near-normal) distributions have different centers, means, or standard deviations, the symmetry along at least one elliptical axis will be compromised. Correlation analysis (such as PPMCC) is routinely applied to such datasets. As illustrated in Figure 1(b), we use two half-ellipses to model the asymmetry along the minor axis. These two half-ellipses have the same major axis $a = 7$, but different minor axes b_1 and b_2, which are defined as the control variables for this hypothesis. We refer to the level of such symmetry as *reflective symmetry* in reference to the constant major axis. We chose nine different pairs of (b_1, b_2), which fall into three equally-sized groups corresponding to $C_R = 0.5, 0.7, 0.9$.

H4: Varying Progressive Symmetry. This hypothesis explores one type of geometric variations. Such a variation may be caused by the divergence of the two underlying statistical distributions. It may also be caused by a hidden cluster existing within the dataset because of some unknown factors. As illustrated in Figure 1(c), the cluster is difficult to identify as it overlaps with other data points in the dataset. Nevertheless, correlation analysis is commonly applied to such datasets. We introduce a circular enclosure overlapping with the elliptical enclosure used for **H1** and **H2**, and fix the circle close to the top of the ellipse. The level of such symmetry is referred to as *progressive symmetry* in reference to increment along the major axis. We controlled the variations using two variables, the radius of the circle r, and the minor axis of the ellipse b. We maintain the major axis $a = 7$, the number of data points $ndp = 80$ with 40 in each enclosure, and $C_R = 0.7$.

H5: Varying Spatial Distribution. This hypothesis explores an extreme scenario of variations, where a dataset contains visible clusters. Many would advise against correlation analysis without first studying the clusters with care. Nevertheless, in automated data processing pipelines, PPMCC values are often obtained for such datasets without humans in the loop. It is thus interesting to examine this extreme scenario as part of this study. As asymmetry, which is studied in **H3** and **H4**, can potentially be a confounding effect, we introduce two clusters at each side of the main elliptical enclosure ($a = 7, b = 3$). As illustrated in Figure 1(d), both additional clusters have the same circular enclosure with a fixed radius $r = 1$. The control variable is the relative density

H1: PPMCC (ndp=80)

C_R	a	b	C_R	a	b
0	7	7			
0.1	7	6	-0.1	6	7
0.2	7	5	-0.2	5	7
0.3	7	4	-0.3	4	7
0.4	7	3.5	-0.4	3.5	7
0.5	7	3	-0.5	3	7
0.6	7	2.7	-0.6	2.7	7
0.7	7	2.5	-0.7	2.5	7
0.8	7	2.25	-0.8	2.25	7
0.9	7	2	-0.9	2	7
1.0	7	0	-1.0	0	7

H2: Density (a=7)

C_R	ndp	b
0.3	40	4
0.5	40	3
0.7	40	2.5
0.3	60	4
0.5	60	3
0.7	60	2.5
0.3	80	4
0.5	80	3
0.7	80	2.5
0.3	100	4
0.5	100	3
0.7	100	2.5
0.3	120	4
0.5	120	3
0.7	120	2.5

H3: Refl. Sym. (ndp=80, a=7)

C_R	b_1	b_2
0.5	3	3
0.5	5	2
0.5	7	1
0.7	2.5	2.5
0.7	4	1.5
0.7	5	1
0.9	2	2
0.9	2.5	1.5
0.9	3	1

H4: Prog. Sym. (ndp=80, a=7)

C_R	b	r
0.7	0.5	5
0.7	1	4
0.7	2	3
0.7	2.5	0
0.7	3	2
0.7	3.5	1

H5: Spatial D. (a=7, b=3, r=1)

C_R	ndp:C-E-C
0.5	0:80:0
0.25	5:70:5
-0.1	10:60:10
-0.3	15:50:15
-0.5	20:40:20
-0.7	25:30:25
-0.9	30:20:30

H6: PPMCC Reference Point in JND (Fine Set) (ndp=80)

C_R	a	b	C_R	a	b	C_R	a	b	C_R	a	b
-0.7	2.5	7	-0.3	4	7	0.3	7	4	0.7	7	2.5
-0.65	2.6	7	-0.35	3.75	7	0.35	7	3.75	0.65	7	2.6
-0.6	2.7	7	-0.4	3.5	7	0.4	7	3.5	0.6	7	2.7
-0.55	2.85	7	-0.45	3.25	7	0.45	7	3.25	0.55	7	2.85
-0.5	3	7	-0.5	3	7	0.5	7	3	0.5	7	3
-0.45	3.25	7	-0.55	2.85	7	0.55	7	2.85	0.45	7	3.25
-0.4	3.5	7	-0.6	2.7	7	0.6	7	2.7	0.4	7	3.5
-0.35	3.75	7	-0.65	2.6	7	0.65	7	2.6	0.35	7	3.75

H6: Coarse Set (ndp=80, a=7)

C_R	b	C_R	b
0.1	6	0.3	4
0.3	4	0.5	3
0.7	2.5	0.9	2
0.1	6	0.5	3
0.3	4	0.7	2.5
0.5	3	0.9	2
0.1	6	0.7	2.5
0.2	5	0.8	2.25
0.3	4	0.9	2
0.05	6.5	0.85	2.15
0.1	6	0.9	2
0.15	5.5	0.95	1

C_R: the Pearson's product-moment correlation coefficient
a: the major axis of an elliptical enclosure
b, b_1, b_2: the minor axis of an elliptical enclosure

r: the radius of a circle
ndp: number of data points

Figure 2: *The set values for different control variables associated with the six null hypotheses. For **H6** (fine set), the 3rd row consists of the four PPMCC reference points, each of which is to be paired with the six sets of control values below.*

of data points in the three clusters. In terms of the numbers of points in the three clusters (circle-ellipse-circle), seven different ratios are used, ranging from 0:80:0 to 30:20:30, corresponding to $C_R = 0.5, 0.25, -0.1, -0.3, -0.5, -0.7, -0.9$.

H6: Varying JND References. The evaluation of just notifiable difference (JND) typically requires a series of tests that compare one fixed reference-estimation pair $(C_{R,0}, e_0)$ with k other pairs $(C_{R,1}, e_1)$, $(C_{R,2}, e_2)$, ..., $(C_{R,k}, e_k)$, where the distance from $C_{R,0}$ to $C_{R,1}, C_{R,2}, \ldots, C_{R,k}$ increases. Using Eq. 1, we can obtain a ternary series $\gamma_1, \gamma_2, \ldots, \gamma_k$. By integrating such ternary series obtained from different participants, we can compute an *accuracy* curve or an *NND* (*no-notifiable-difference*) curve, from which a JND value is usually estimated [Kru89, Mor78, Sjö85].

This hypothesis is concerned with whether or not a change of the PPMCC reference point $C_{R,0}$ could affect the JND in humans' perception of correlation. The main control variable is thus the reference point $C_{R,0}$, which are set to $-0.7, -0.3, 0.3, 0.7$. For each instance of $C_{R,0}$, we define a set of six increasing distances $0.05, 0.1, 0.15, 0.25, 0.3, 0.35$. To obtain $C_{R,1}, C_{R,2}, \ldots, C_{R,6}$, we add these distances to $C_{R,0}$ in the cases of $C_{R,0} = -0.7, 0.3$ (i.e., "above") and subtract them from $C_{R,0}$ in the cases of $C_{R,0} = -0.3, 0.7$ (i.e., "below"). All stimuli for this hypothesis use the basic elliptical enclosure as shown in Figure 1(a). The major axis is fixed at 7 as in **H1**. The minor axis is used to control the changes of $C_{R,0}, C_{R,1}, \ldots, C_{R,6}$. All stimuli have 80 data points.

The tables in Figure 2 summarize the set values for different control variables discussed above. Because we restricted each hypothesis to be evaluated with only 1 or 2 control variables, we were able to prevent the explosion of stimuli due to multivariate combinations. These set values represented different samples were taken discretely along the continuous axes of C_R, a, b, and r, or sparsely along the discrete axis *ndp*. This approach to sampling was also necessary for a manageable empirical study. Inevitably, the between-group variations would likely depend on the resolution of intervals. If the intervals were too coarse, we could miss some important variations. If the intervals were too fine, many pairs of samples (especially with a small distance) would exhibit insignificant variations.

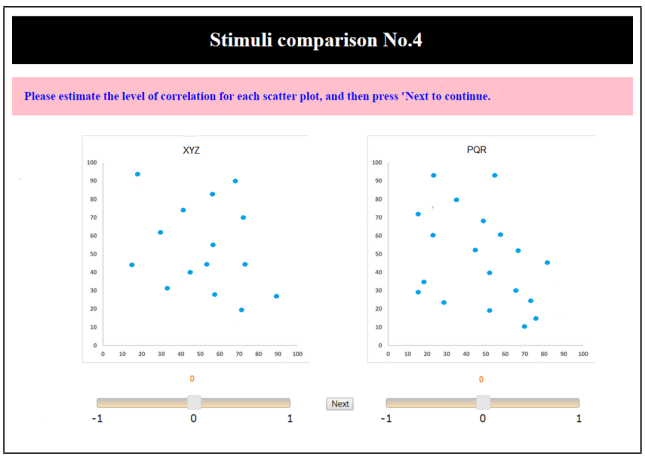
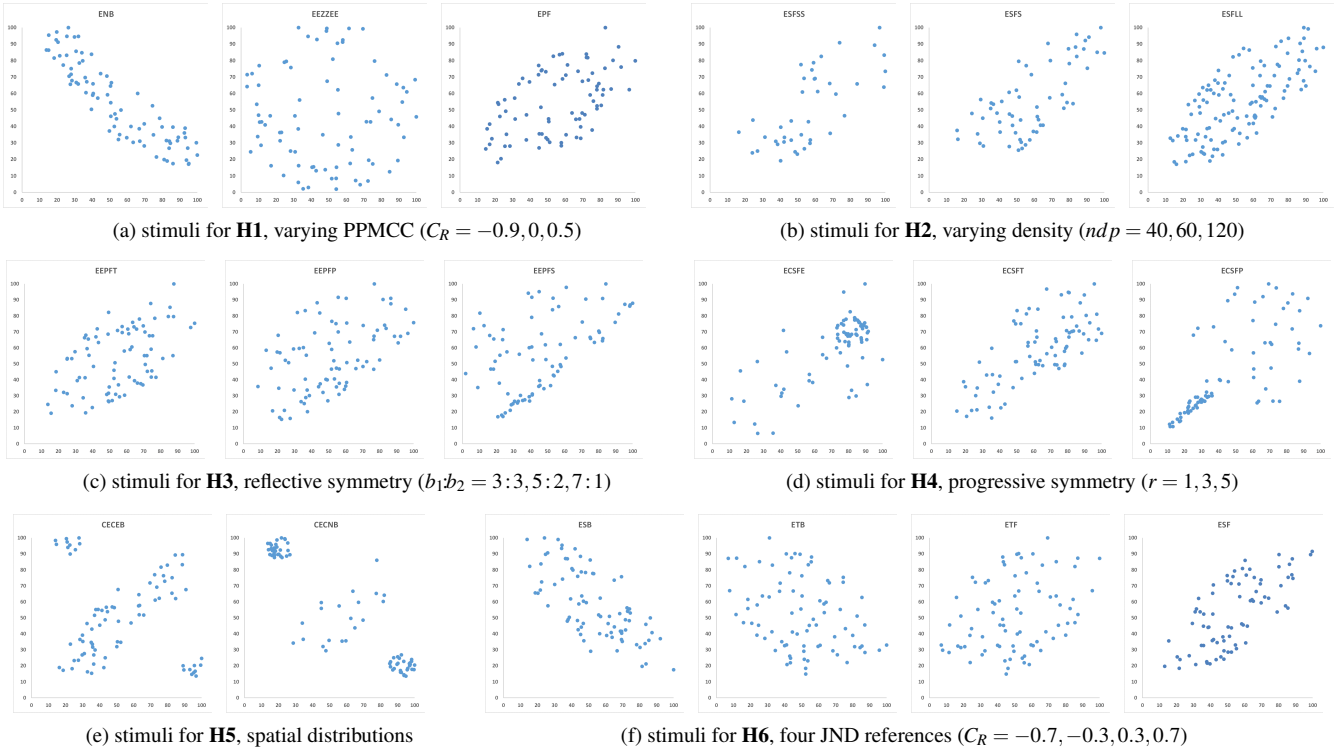


Figure 3: *Each trial consists of two stimuli. This generic design enables trials for JND (Just Noticeable Difference) evaluation to be integrated with other trials seamlessly.*

4. Study Design

User Interface and Task Design. Each stimulus for the first five hypothesis, **H1**∼**H5**, can be evaluated independently, while **H6** requires the consideration of two stimuli at the same time. In order to provide a coherent user interface to accommodate all six hypotheses, we defined a common task, that is, to estimate the PPMCC value of a given scatterplot stimulus. A common user interface, as shown in Figure 3, presented two stimuli to participants in each trial. The participants were asked to estimate the PPMCC values of the two stimuli independently of one another. Only for **H6**, comparative results were recorded. Two slider bars were provided to adjust the estimation ranging from $[-1, 1]$ with 0.05 step size. The selected value is shown in red above the slider.

Stimuli Generation. For all hypotheses, we first defined a set of PPMCC values, C_R, as shown in cyan color in Figure 2. For each specific C_R in a hypothesis, we estimated the relevant geometric variables, a, b, b_1, b_2, or r, by randomly generating a number of bivariate datasets. Note that a pre-defined geometric enclosure does not guarantee a unique C_R. For each given geometric enclosure, therefore, we randomly generated 20 datasets and then selected one



(a) stimuli for **H1**, varying PPMCC ($C_R = -0.9, 0, 0.5$)

(b) stimuli for **H2**, varying density ($ndp = 40, 60, 120$)

(c) stimuli for **H3**, reflective symmetry ($b_1 b_2 = 3:3, 5:2, 7:1$)

(d) stimuli for **H4**, progressive symmetry ($r = 1, 3, 5$)

(e) stimuli for **H5**, spatial distributions

(f) stimuli for **H6**, four JND references ($C_R = -0.7, -0.3, 0.3, 0.7$)

Figure 4: *A selection of stimuli that are designed for different hypotheses.*

with a PPMCC value C'_R such that $\Delta = |C_R - C'_R|$ is the smallest. In some cases where $\Delta \geq 0.001$, we manually modified a data point in the dataset to ensure $\Delta < 0.001$.

Stimuli with the Same Geometry Characteristics. It is necessary to generate stimuli from the same geometric enclosure for **H6** and training stimuli. For example, **H6** required at least two stimuli to be computed from the geometric enclosure $[C_R = 0.4, a = 7, b = 3.5, ndp = 80]$. For all multiple stimuli with the same geometric enclosure, we ensured that they were rendered from different datasets generated using the above procedure. From Figure 2, one can also observe some other duplications. For example, the basic elliptical enclosure $[C_R = 0.7, a = 7, b = 3, ndp = 80]$ can serve **H1**, **H2**, **H3**, **H5**, and **H6**. In this latter case, the response of such a stimulus can be reused for evaluating a difference hypothesis.

Stimuli Organization. This study consisted of 100 trials, 5 in the training session and 95 in the main session. In each trial, participants were presented with two stimuli. The 10 stimuli shown in the training session were designed to enable participants to familiarize themselves with the user interface and typical visual patterns in the stimuli. They also allowed them to synchronize their estimation with the actual PPMCC values, while provided the study organizers with any signs of poor understanding or random clicks. Among the 190 stimuli shown in the main session, 5 were used as *checkpoints* for poor attention or random clicks, providing further validation of the experimental data captured in the study. There were thus 185 stimuli, which contributed directly to the evaluation of the six hypotheses.

Using a C++ program, we generated 72 unique stimuli datasets. Figure 4 shows 18 example stimuli used for evaluating hypotheses **H1**-**H6**. All stimuli were drawn with the same visual mapping onto a region of $[10, 100] \times [10, 100]$. The canvas region $x, y \in [0, 10)$ was intentionally avoided to prevent perceptual errors that may be caused by the two axes. Each stimulus had an encrypted label, which was used by experimenters to identify the numerical attributes of the stimulus and the corresponding hypothesis in a confidential manner.

The pairing of stimuli in each trial was manually assigned to ensure the correct coverage for **H6**, while mixing stimuli for different hypotheses in non-**H6** trials. The left or right positions of the two stimuli in each trial were determined randomly. The 95 trials in the main sessions were ordered pseudo-randomly. The reason for manual adjustment was to ensure that no stimulus would be repeated in nearby trials. A pixel-based visualization is given in the supplementary materials, depicting the distribution of the 72 stimuli in relation to different hypotheses and trials.

5. Study Implementation

Apparatus. The study took place in a laboratory, where each computer had 24 inch LCD display with 1920×1200 pixels. The software was implemented using HTML, JavaScript and PHP, and was run using Chrome web browser in the full screen mode. About 6∼10 computers were used during each experimental session and they were connected to a PHP server.

The software was used to structure the study, and record participants' inputs. It runs the pre-experiment data collection, training

session, and main study session. After each trial, it displays a noise-based masking screen for 2 seconds to neutralize the effect of the two stimuli just shown. It computes various measurements, including the offset between a participant's estimation, and the original PPMCC value for each stimuli, and the JND classification. It sends the captured records back to the server in a .txt file, while keeping a backup copy (a downloadable .json file) on the client in case of technical emergency.

Procedures. Before the study, participants were given an *Information Session*. The experimenters explained the aim of the research, the structure of the study, and the remuneration for their time. The experimenters presented various examples of scatterplots, and showed how they correspond to PPMCC correlation values. (None of these scatterplots were stimuli used in the study, except ones with $C_R = \pm 1$.) During this session, participants could read the consent form provided and ask questions about the study and the use of their data. The information session lasted about 10∼15 minutes, depending on participants' questions.

The experiment started with *Pre-experiment Data Collection* for basic demographic information (e.g., gender, age group and education level). This was followed by a *Training Session*, which comprised of 5 trials (each with 2 stimuli). As mentioned in Section 4, these trials were means for quality assurance, including familiarization of visual display and interaction, synchronization of participants' estimation, and validation of usability of the captured data.

The *Main Study Session* comprised of four sections, with 20, 25, 25, and 25 trials respectively. At the end of each section, participants were required to have a break for at least 2 minutes. Any longer break was at participants' own discretion.

At the end of the study, participants were asked to complete a paper-based *Post-Experiment Survey* with three multiple choice questions. It was designed to collect participants' subjective assessment as to how easy it is to infer correlation values from scatterplots showing different visual features (e.g., positive vs. negative correlation, symmetric vs. asymmetric enclosure, and sparse vs. dense data points).

Data Validation. Stimuli exhibiting full positive and negative correlation (i.e., $C_R = \pm 1$) were expected to be answered correctly throughout the study. We used five such checkpoints in the main study session to detect poor attention or random clicks.

Participants. A total of 37 participants took part in the experiment in return for a £10 Amazon gift voucher. There were 27 males and 10 females. Among them, 25, 10, and 2 were in the 20∼29, 30∼39, 40∼49 age groups respectively. They include 27 university students, 9 university staff, and 1 other (unspecified). Two participants' responses were excluded from analysis. One (male, 20∼29) failed to estimate easy correlation (i.e., $C_R = \pm 1$) in the training session, possibly due to poor understanding. One (female, 30∼39) passed the training but failed the validation during the main session with 3 incorrect estimations for the 5 check points, suggesting random clicks possibly due to boredom. Therefore, the responses of 35 participants were included in the results analysis in Section 6.

6. Results and Analysis

In this section, we present the results for each hypothesis. We will discuss our main findings in Section 7.

Following the study, the captured data records were first divided into groups according to hypotheses. For hypotheses **H1**∼**H5**, each stimulus had 3 or more repeated measures per participant. Let $e_{i,j,k}$ be the k^{th} repeated measure for the j^{th} stimulus from the i^{th} participant. A signed offset is thus $\delta_{i,j,k} = e_{i,j,k} - C_{R,j}$. For each (i,j), we compute three types of averages over m repeated measures. They are:

$$\Theta_{i,j} = \frac{\sum_k \delta_{i,j,k}}{m}, \quad \Phi_{i,j} = \frac{|\sum_k \delta_{i,j,k}|}{m}, \quad \Psi_{i,j} = \frac{\sum_k |\delta_{i,j,k}|}{m}.$$

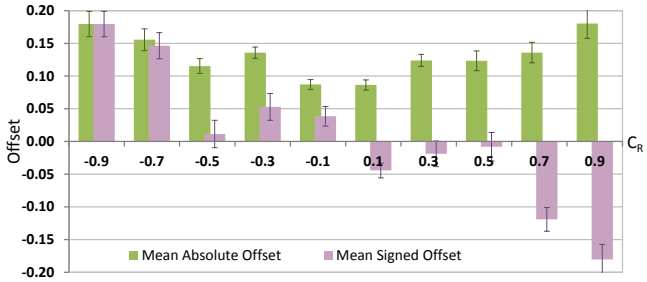
For n participants, the *Mean Signed Offset* (MSO) for the j^{th} stimulus is thus $\sum_i \Theta_{i,j}/n$. The *Absolute Mean Offset* (AMO) is $\sum_i \Phi_{i,j}/n$. The *Mean Absolute Offset* (MAO) is $\sum_i \Psi_{i,j}/n$.

H1: Varying PPMCC. Figure 5(a) shows the MSOs (lilac), and MAOs (green) for C_R between $[-0.9, 0.9]$ with a 0.2 interval. Positive MSOs suggest overestimation, and negative MSOs suggest underestimation, while MAOs are always ≥ 0. We can thus observe the symmetric pattern of overestimation when $C_R < 0$, and underestimation when $C_R > 0$. This result synchronizes well with most of the previous findings in the literature (e.g., [Erl66, BK79]).

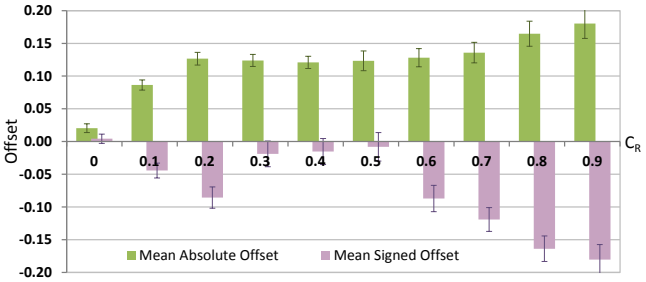
When we examine in detail the range $[0.0, 0.9]$ with a 0.1 interval as shown in Figure 5(b), we can observe the pattern of increasing underestimation when $C_R \geq 0.6$, and some underestimation around $C_R = 0.2$. Further examining the numbers of participants who produced positive, zero, and negative offsets for $C_R \geq 0$ in Figure 5(c), we can find that there were strong biases towards underestimation, when $C_R = 0.1, 0.2$ and $C_R = 0.7, 0.8, 0.9$. Only when $C_R = 0$, most participants estimated correctly. Symmetrically, for $C_R < 0$ as shown in Figure 5(d), there were strong biases towards overestimation, when $C_R = -0.9, -0.7$. When $C_R = \pm 0.5$, a similar number of positive and negative offsets were recorded, while its MAO is similar to that of $C_R = \pm 0.3$.

For the MSOs (lilac) shown in Figure 5(a), the ANOVA analysis indicates $F(9, 340) = 32$, $p < 0.01$, and for the MSOs in Figure 5(b), $F(9, 340) = 13$, $p < 0.01$. Both suggest that there are significant between-group variations. Figure 5(e) shows the p-values after applying t-Test (unpaired, two-tailed, Bonferroni correction) to each pair of the MSOs (lilac) in Figure 5(b). The last three columns in Figure 5(e) provide overwhelming evidences for supporting differences between the set $\{C_R = 0.7, 0.8, 0.9\}$ and the other samples, while the variations among the three samples within the set are lessened because of the closer C_R distances.

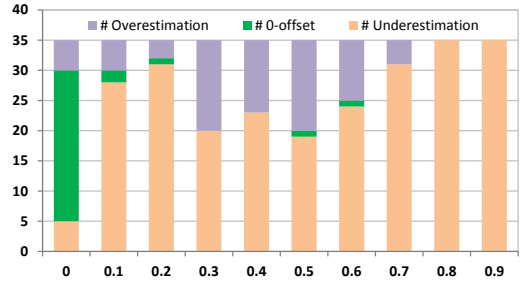
Meanwhile, the variation between the sample $C_R = 0.2$ and those nearby (i.e., $C_R = 0.0, 0.1, 0.3, 0.4, 0.5$) also suggest that the variation at $C_R = 0.2$ shown in Figure 5(b) is significant. The sine wave pattern of the MSOs (lilac) between -0.5 and 0.5 in Figure 5(a) also suggests that the variations around $C_R = \pm 0.2$ will be interesting to study in the future. Out of the 45 pairs of t-Tests for the MSOs in Figure 5(a), 29 tests show $p < 0.05/45$. The majority counts confirm that the variations observed are not accidental. Out of the 45 pairs of t-Tests for the MSOs in Figure 5(b), 18 tests show $p < 0.05/45$. As shown in Figure 5(e), $\{0.7, 0.8, 0.9\}$



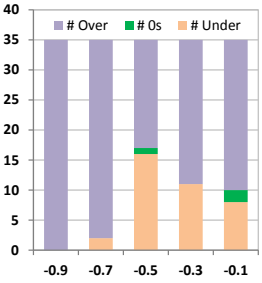
(a) $C_R \in [-0.9, 0.9]$ with a 0.2 interval



(b) $C_R \in [0, 0.9]$ with a 0.1 interval



(c) $C_R \in [0, 0.9]$ with a 0.1 interval

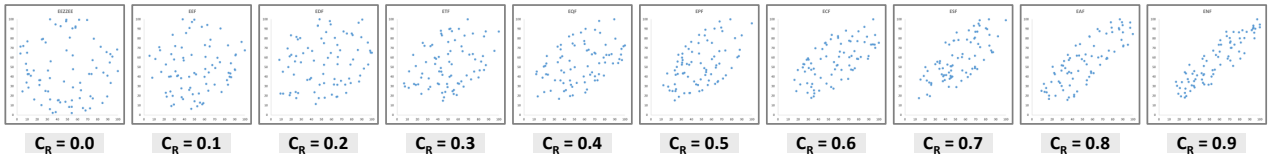


(d) $C_R \in [-0.9, -0.1]$

Bonferroni correction: 0.05/45 = 0.00111

p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	
0.0	6.E-04	7.E-06	0.27	0.36	0.59	1.E-04	1.E-07	3.E-10	2.E-09	0.0
0.1		0.04	0.27	0.21	0.15	0.07	9.E-04	2.E-06	2.E-06	0.1
0.2			0.01	0.01	0.01	0.96	0.17	3.E-03	1.E-03	0.2
0.3				0.90	0.71	0.02	4.E-04	2.E-06	1.E-06	0.3
0.4					0.81	0.01	3.E-04	1.E-06	8.E-07	0.4
0.5						0.01	2.E-04	1.E-06	7.E-07	0.5
0.6							0.24	0.01	3.E-03	0.6
0.7								0.10	0.04	0.7
0.8									0.58	0.8
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	

(e) p-values resulting from pairwise t-Tests



(f) 10 examples of the stimuli used for testing **H1**, $C_R \in [0.0, 0.9]$, corresponding to (b), (c), and (e)

Figure 5: *Summary statistics for **H1: Varying PPMCC**. (a) MSOs and MAOs for $C_R \in [-0.9, 0.9]$, indicating a symmetric pattern between positive and negative C_R values; (b) MSOs and MAOs for $C_R \in [0.0, 0.9]$, showing a non-linear pattern of underestimation. (c) the numbers of participants with positive, zero, and negative offsets while estimating non-negative correlation, (d) while estimating negative correlation, and (e) p-values for pairwise t-tests (in terms of MSOs, $C_R \in [0.0, 0.9]$) with insignificant pairs marked in red.*

vs. $\{0.0, 0.1, 0.3, 0.4, 0.5\}$ is significant. As **H1** postulates a similar mean across different C_R, the result clearly contravenes this null hypothesis.

Meanwhile, when we consider MAOs (green) in Figure 5(a), we can observe an increasing trend from ± 0.1 to ± 0.9. The t-Tests show that the differences of all five pairs are insignificant, i.e., $-0.1 \boxtimes 0.1, \ldots, -0.9 \boxtimes 0.9$. Here $a \boxtimes b$ denotes that the difference of a vs. b is insignificant. Similarly $\boxtimes$ denotes significant.

We also compared MAOs and AMOs for $C_R = 0, 0.1, \ldots, 0.9$. We observed that the differences appeared to depend on the values of C_R. It means that when AMOs reduce noises in repeated measures, they may also remove a small amount of the effect due to varying C_R. We thus consider only MAOs below.

H2: Varying Density. Figures 6(a,b) show the participants' responses to 15 stimuli, which represent combinations of 5 density levels (40, 60, 80, 100, 120 data points) and 3 PPMCC values ($C_R = 0.3, 0.5, 0.7$). Between three PPMCC values, we can observe the increasing underestimation, which is consistent with the finding for $C_R > 0$ in **H1**. We can also notice that the responses to stimuli with 40 data points appear to be more different.

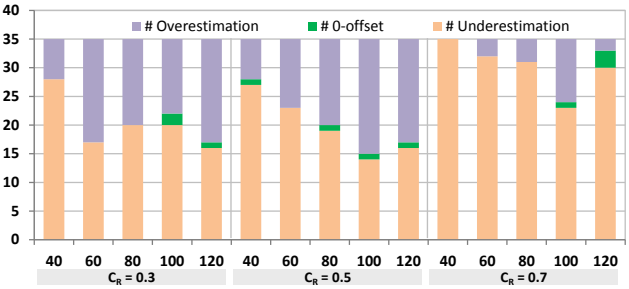
Since the analysis of **H1** showed that C_R has an impact on participants' estimation, we thereby applied ANOVA analysis to each condition of C_R separately. For $C_R = 0.3$, the ANOVA analysis yields $F(4, 170) = 4.4$, $p < 0.01$. When we exclude the sample of $ndp = 40$, the analysis indicates insignificant variations $F(3, 136) = 0.9$, $p = 0.45$. This suggests that the between group variations are mainly contributed by the sample of $ndp = 40$. This can be confirmed by observing the pairwise t-Test results in Figure 6(c).

For $C_R = 0.5$, the ANOVA analysis yields $F(4, 170) = 7.4$, $p < 0.01$. When the sample of $ndp = 40$ is excluded, we have $F(3, 136) = 3.7$, $p = 0.01$. A further examination of Figure 6(c) shows that both samples of $ndp = 40, 60$ contribute to the detected variations.

For $C_R = 0.7$, the ANOVA analysis yields $F(4, 170) = 13.6$, $p < 0.01$. When we exclude the sample of $ndp = 40$, the analysis indicates insignificant variations $F(3, 136) = 0.5$, $p = 0.65$ Figure 6(c) also shows that the sample of $ndp = 40$ is the main contributor of the variations.

No significant difference was detected in other two-sample comparisons, including all pairings of $ndp = 80, 100, 120$. This sug-

(a) mean signed offsets (MSOs) and mean absolute offsets (MAOs)

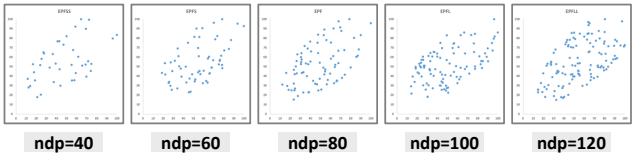


(b) numbers of participants with positive, zero, negative offsets

Bonferroni correction: 0.05/10 = 0.005

p	60	80	100	120	60	80	100	120	60	80	100	120	
40	7.E-04	3.E-03	0.03	1.E-03	0.17	5.E-04	3.E-05	8.E-05	2.E-06	8.E-07	1.E-07	2.E-06	40
60		0.48	0.16	0.92		0.04	4.E-03	0.01		0.91	0.33	0.95	60
80			0.42	0.55			0.39	0.66			0.36	0.85	80
100				0.19				0.65				0.28	100
	60	80	100	120	60	80	100	120	60	80	100	120	
	C_R = 0.3				C_R = 0.5				C_R = 0.7				

(c) p-values resulting from pairwise t-Tests



(d) 5 examples of the stimuli used for testing **H2**, $C_R = 0.5$

Figure 6: *Summary statistics for **H2: Varying Density**.*

gests that the impact of varying density only occurs when there is a large difference in numbers of data points.

Our result of $40 \approx \{60, 80, 100, 120\}$ contradicts the finding in [DAAK07] ([*ndp*]: $9 \not\approx 100$). It provides more concrete evidence in supporting the observation in [LP89], which did not report any analysis of variance for evaluating their hypothesis (200 vs. 400). In addition to an improvement of the *ndp* resolution from two samples to five samples, we had a larger group of participants than that in [DAAK07] (i.e., 35 vs. 20), we had tested more C_R values than [LP89] (3 vs. 1).

From Figure 6, one may suggest an explanation for the more noticeable underestimation when $ndp = 40$, that is, a sparse distribution might affect one's perception of a geometric enclosure, or the confidence associated with the perception. Unconsciously one might "downgrade" an estimation towards less correlation. If this hypothesized explanation could be verified in future studies, it would be a positive confirmation of humans' ability to perceive and reason about uncertainty in visualization.

H3: Reflective Symmetry. Figures 7(a,b) show the responses to 9 stimuli, which represent combinations of 3 levels of symmetry and 3 PPMCC values ($C_R = 0.5, 0.7, 0.9$). In each C_R set, the first stimulus with $b_1 = b_2$ are used to synchronize with the results in **H1**. There is a common pattern that the most asymmetric stimulus (rightmost) in each set has the highest level of underestimation. The ordering of the other two sets does not show a consistent pattern.

Because of the potential impact of C_R, we applied ANOVA analysis to each condition of C_R separately. For $C_R = 0.5$, the ANOVA analysis yields $F(2,102) = 2.9$, $p = 0.06$, indicating insignificant variation. For $C_R = 0.7$, the ANOVA analysis gives $F(2,102) = 6.0$, $p < 0.01$, suggesting some significant variation. For $C_R = 0.9$, the ANOVA analysis returns $F(2,102) = 3.4$, $p = 0.04$, also suggesting some significant variation.

By having a closer look at the t-Test results in Figure 7(c), we can observe some pairwise variations, such as:

$$\text{For } C_R = 0.7, \quad [b_1 : b_2] : \ 5:1 \approx \{2.5:2.5, 4:1.5\}$$
$$\text{For } C_R = 0.9, \quad [b_1 : b_2] : \ 3:1 \approx 2.5:1.5$$

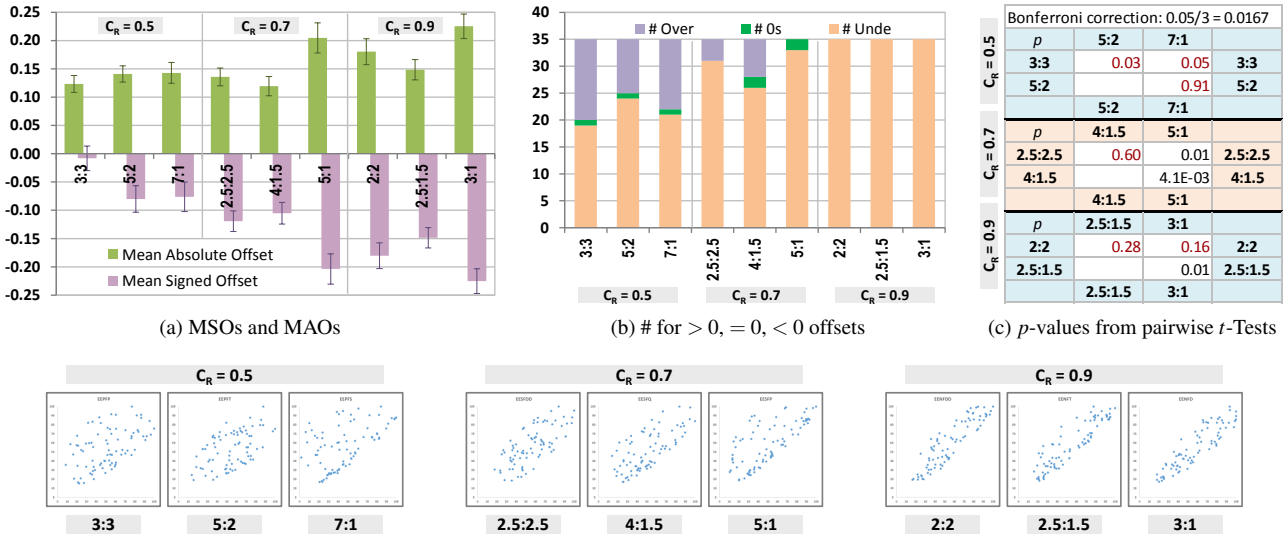
In general, the experiment shows some significant variations when $C_R = 0.7$, and $C_R = 0.9$, suggesting the possibility to negate the null hypothesis for **H3**. However, because we cannot confidentially eliminate the possibility of accidental errors in sampling, further studies with a higher sampling resolution in terms of $[b_1 : b_2]$ will be necessary for rejecting **H3** conclusively.

H4: Progressive Symmetry. Figures 8(a,b) show the responses to 6 stimuli (all with $C_R = 0.7$). Among these $r = 0$ corresponds to an elliptical enclosure in **H1** without an additional cluster, while $r = 5, 4, 3, 2, 1$ correspond to different sizes of clusters that were added to the main elliptical enclosure. While the numbers of underestimation are similar, there are noticeable changes in magnitudes. The nearly symmetric pattern in Figure 8 centered around $r = 0$ (i.e., the normal elliptical enclosure) also suggests the unlikelihood of any major accidental error. Meanwhile, Figure 8(c) shows a consistent pattern of underestimation by a noticeable amount.

The ANOVA analysis yield $F(5,204) = 17.5$, $p < 0.01$, indicating significant between-group variations. When we examine the 15 two-sample comparisons using t-Test, 6 pairs have insignificant differences. These are all occurred at either sampling pairs next to each other, or at the opposite sides of the symmetric pattern in Figure 8. Hence the null hypothesis for **H4** can be rejected.

H5: Spatial Distribution. Figure 9(a) shows the responses to 7 stimuli. Among these $ndp = 0{:}80{:}0$ corresponds to an elliptical enclosure ($C_R = 0.5$) in **H1** without any additional cluster. The other 6 stimuli all have three clusters, one main elliptical shape and two additional clusters. The sizes of three enclosures and the total numbers of data points remain the same, but the 80 data points are distributed differently.

Visually, we can observe the noticeable changes from $ndp = 0{:}80{:}0$ to other distributions in terms of both the numbers and magnitudes of overestimation. This becomes more dramatic in comparison with the changes observed in the context of **H1** as shown in



(a) MSOs and MAOs

(b) # for $> 0, = 0, < 0$ offsets

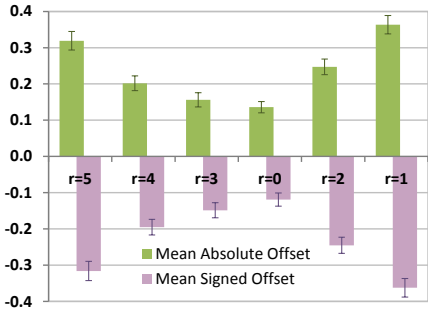
Bonferroni correction: 0.05/3 = 0.0167

C_R	p			
0.5	3:3	5:2	7:1	3:3
	5:2	0.03	0.05	5:2
		5:2	0.91	
0.7	p	4:1.5	7:1	
	2.5:2.5	0.60	5:1	2.5:2.5
	4:1.5		0.01	4:1.5
		4:1.5	4.1E-03	
0.9	p	2.5:1.5	5:1	
	2:2	0.28	3:1	2:2
	2.5:1.5		0.16	2.5:1.5
		2.5:1.5	0.01	
			3:1	

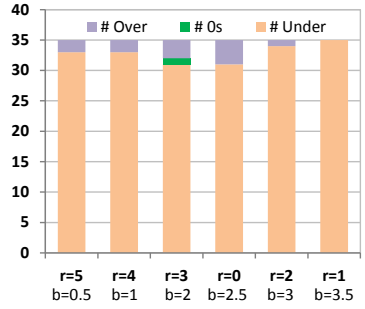
(c) p-values from pairwise t-Tests

(d) 9 examples of the stimuli used for testing **H3**, corresponding to (a) and (b) from left to right

Figure 7: *Summary statistics for **H3: Reflective Symmetry**. Samples labelled with* $3:3$*,* $2.5:2.5$*, and* $2:2$ *are symmetric samples. The variations between them and asymmetric samples are observable. However, finer sampling will be desirable for a conclusive finding.*



(a) MSOs and MAOs

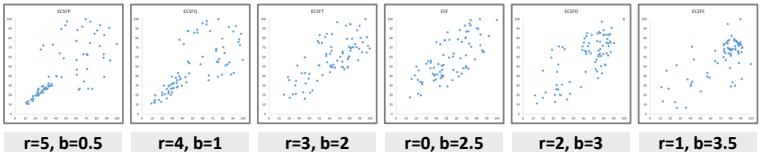


(b) # for $> 0, = 0, < 0$ offsets

Bonferroni correction: 0.05/15 = 0.00333

p	r=4	r=3	r=0	r=2	r=1	
r=5	7.9E-04	6.0E-06	8.5E-08	0.05	0.22	r=5
r=4		0.13	0.01	0.11	4.9E-06	r=4
r=3			0.29	2.3E-03	1.6E-08	r=3
r=0				4.3E-05	1.3E-10	r=0
r=2					9.8E-04	r=2
	r=4	r=3	r=0	r=2	r=1	

(c) p-values resulting from pairwise t-Tests



(d) 6 examples of the stimuli used for testing **H4**, corresponding to (a) and (b) from left to right, $C_R = 0.7$

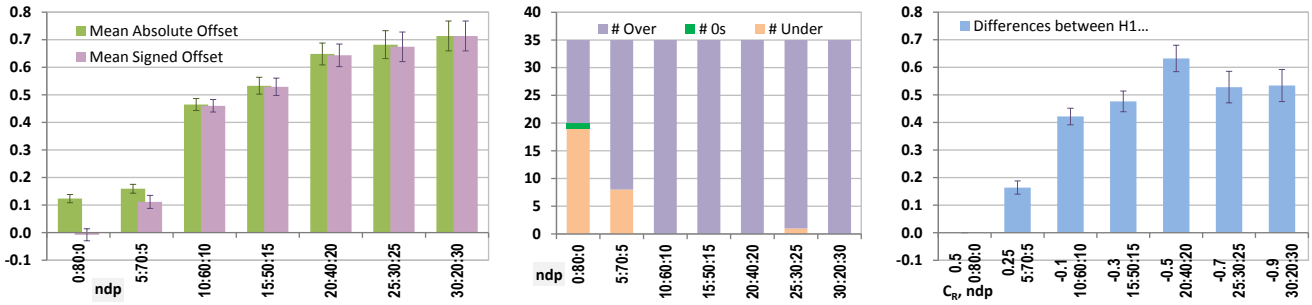
Figure 8: *Summary statistics for **H4: Progressive Symmetry**. The comparisons between* $r = 0$ *and* $r > 0$ *show significant variations.*

Figure 9(c,d). Although the PPMCC values of the 7 stimuli in **H5** are $C_R = 0.5, 0.25, -0.1, -0.3, -0.5, -0.7, -0.9$, the magnitudes of overestimation resulting from six of these stimuli (i.e., except $ndp = 0:80:0, C_R = 0.5$) are significantly higher than the corresponding stimuli in **H1** without two additional clusters. Note that there is no stimulus in **H1** featuring $C_R = 0.25$. We actually use the average of the statistical indicators of $C_R = 0.2$ and $C_R = 0.3$ here.

This noticeable variation due to different spatial distribution can also be observed by comparing Figure 9(b) with Figure 9(e). Much of the underestimation for $C_R = 0.25, -0.1, -0.3, -0.5$ in Figure 9(d) has been replaced with overestimation in Figure 9(b).

Each of the 7 stimuli in **H5** features distinct PPMCC values, as well as different spatial distributions. It is thus not appropriate to apply ANOVA directly to the data collected in the experiment without removing the impact of C_R as shown in **H1**. We thus used the individual participants' offsets for **H1** as the base line and subtract these from the offsets of the same participants for **H5**, and then apply ANOVA to the difference values. The mean differences are shown in Figure 9(c). The ANOVA analysis yields $F(6, 238) = 30.7$, $p < 0.01$, indicating significant between-group variations.

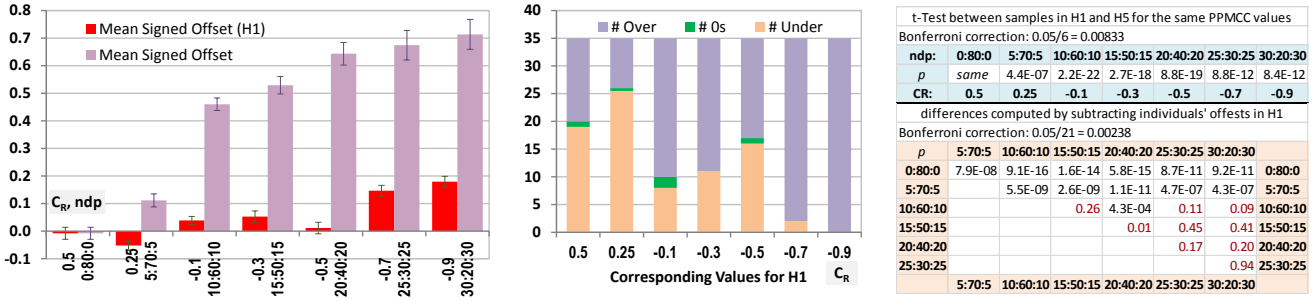
We further examine each pair of samples between **H1** and **H5** for $C_R = 0.25, -0.1, -0.3, -0.5, -0.7, -0.9$ using t-Test. As shown in Figure 9(f), all p-values are well below 0.05/6 (Bonferroni cor-



(a) MSOs and MAOs of H5 samples

(b) # different offsets (H5)

(c) Differences (**H5** offset−**H1** offset)



(d) MSOs: H1 samples vs. H5 samples

(e) # different offsets (H1)

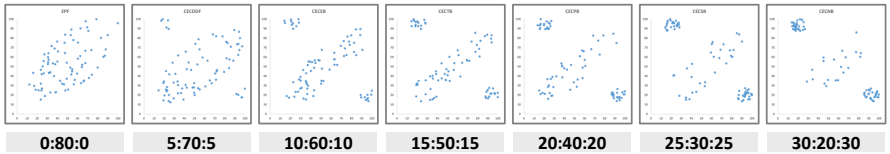
t-Test between samples in H1 and H5 for the same PPMCC values
Bonferroni correction: 0.05/6 = 0.00833

ndp:	0:80:0	5:70:5	10:60:10	15:50:15	20:40:20	25:30:25	30:20:30
p	*same*	4.4E-07	2.2E-22	2.7E-18	8.8E-19	8.8E-12	8.4E-12
CR:	0.5	0.25	-0.1	-0.3	-0.5	-0.7	-0.9

differences computed by subtracting individuals' offests in H1
Bonferroni correction: 0.05/21 = 0.00238

p	5:70:5	10:60:10	15:50:15	20:40:20	25:30:25	30:20:30	
0:80:0	7.9E-08	9.1E-16	1.6E-14	5.8E-15	8.7E-11	9.2E-11	0:80:0
5:70:5		5.5E-09	2.6E-09	1.1E-11	4.7E-07	4.3E-07	5:70:5
10:60:10			0.26	4.3E-04	0.11	0.09	10:60:10
15:50:15				0.01	0.45	0.41	15:50:15
20:40:20					0.17	0.20	20:40:20
25:30:25						0.94	25:30:25
	5:70:5	10:60:10	15:50:15	20:40:20	25:30:25	30:20:30	

(f) *p*-values from pairwise *t*-Tests



(g) 7 examples of the stimuli used for testing **H5**, corresponding to (a) and (b) from left to right

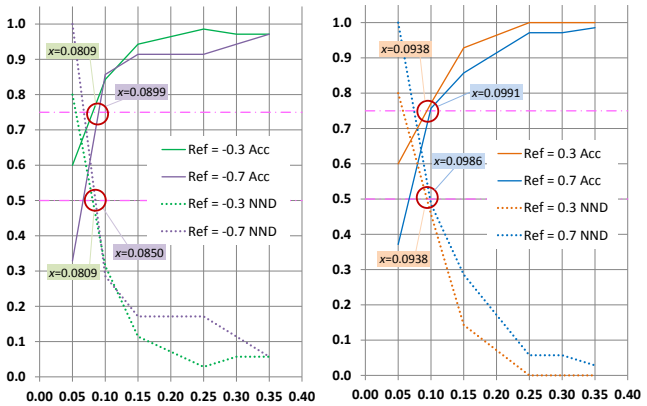
Figure 9: *Summary statistics for **H5: Spatial Distribution**. In (d) and (f), comparisons with MSOs in **H1** indicate significant variations.*

rection). The individual pairwise comparisons among the difference values obtained for the 7 samples show that $ndp = 0:80:0$ and $ndp = 5:70:5$ contributed main variations. In other words, the other five samples show a similar as well as a significant amount of deviations from the mean offsets for **H1**. Essentially, almost all participants were not able to estimate the impact of the two additional clusters on the PPMCC value. In the cases of $ndp = 10:60:10, 15:50:15, 20:40:20$ where the actual PPMCC values became negative ($C_R = -0.1, -0.3, -0.5$, most still returned positive estimations. In the cases of $ndp = 25:30:25, 30:20:30$ when the actually PPMCC values would suggest strong negative correlation ($C_R = -0.7, -0.9$), most participants returned estimations indicating no correlation. Clearly the null hypothesis for **H5** can be rejected.

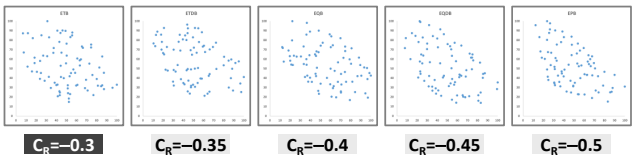
H6: Varying JND References. Let Δ be the distance between two PPMCC values, $C_{R,a}$ and $C_{R,b}$. The JND trials were divided into two sets. The first set was designed to make a coarse examination of JND at the very unlikely range $\Delta \geq 0.2$. There were 4 variations $\Delta = 0.2, 0.4, 0.6, 0.8$. Each group consisted of 3 trials with varying $C_{R,a}$ and $C_{R,b}$. There were only a few errors. 100% accuracy was attained for nine trials, 97% for two trials ($\Delta = 0.2$, $\Delta = 0.4$), and 94% for one trial ($\Delta = 0.2$).

The second set of trials focused on the detection of JND in relation to four reference points where $C_R = -0.7, -0.3, 0.3, 0.7$. The goal was to observe the possible JND points by varying the distances at a fine interval of 0.05. As defined by Eq. 1, we classify each pair of responses to a JND trial as "correct", "same", or "wrong". At $\Delta = 0.05$, *t*-Test (unpaired, two-tailed) indicates that 4 (out of 6) two-sample comparisons show significant differences between some reference points, i.e., $-0.7 \bowtie \{-0.3, 0.3\}$ ($p < 0.01$), and $0.7 \bowtie \{-0.3, 0.3\}$ ($p = 0.01, 0.02$). At $\Delta \geq 0.1$, all differences are insignificant. The result offers a weak rejection to the null hypothesis of **H6**. However, further examination with finer intervals between $\Delta = 0.05$ and $\Delta = 0.15$ will be necessary to provide an absolute confirmation.

Because we used *direct rating* (DR) [Sjö85] for our JND trials, we classify each pair of responses to a JND trial as "correct", "same", or "wrong". We can thus estimate JND using two estimation methods, for both *forced binary choice* (FBC) and *direct rating* (DR) [Kru89]. Given a pair of stimuli in a JND trial, let N_c, N_s, N_w denote the numbers of "correct", "same", and "wrong" answers from all participants. For calculating *accuracy* typically used in conjunction with FBC, we assign half of N_s to N_c and half to N_w by assuming a random choice when a "same" answer is given. For calculating *not-notifiable-difference* in DR, we assume that all "wrong" answers result from random choices, and the number of



(a) $C_R = -0.7, -0.3$ (b) $C_R = 0.3, 0.7$



(c) 5 examples of the stimuli used for testing **H6**, where $C_R = -0.3$ is a reference point

Figure 10: *Summary statistics for **H6**. (a) The results of two sets of JND trials for two negative reference points. (b) Results for two positive reference points.*

"correct" answers resulting from random choices is $\leq N_w$. Thus, the percentages of both can be computed as:

$$P_{Acc} = \frac{N_c + 0.5N_s}{N_c + N_s + N_w}$$

$$P_{NND} = \frac{N_s + N_r}{N_c + N_s + N_w}, \text{ where } N_r = \begin{cases} 2N_w & N_w \leq N_c \\ N_w + N_c & N_w > N_c \end{cases}$$

Figs. 10(a,b) show the P_{Acc} and P_{NND} for the four reference points. Each polyline represents trials with $\Delta = 0.05, 0.1, 0.15, 0.25, 0.3, 0.35$. As 75% is the commonly-used JND threshold for FBC, and 50% for DR, we can easily observe that P_{Acc} and P_{NND} are close to each other for all four reference points.

7. Discussions and Conclusions

This study has provided strong evidence for rejecting the null hypotheses **H1** (varying PPMCC), **H4** (varying progressive symmetry), and **H5** (varying spatial distribution). It has also shown that the null hypothesis of **H2** is partially incorrect (i.e., for large variations of density), and that of **H3** (varying reflective symmetry) is unlikely but requires confirmation from further studies. Judging from the numbers and magnitudes of under or overestimation, we must conclude that humans' estimation of PPMCC values when observing arbitrarily-given scatterplots is rather unreliable.

This leads to an anxious reflection and a new hypothesis as to the benefit of scatterplots. In fact, it is inappropriate to use PPMCC in most of the conditions of **H5**, and highly risky in those of **H4**. Without visualization, a user would not know the multiple clusters as exemplified by **H5** or **H4**, and using PPMCC would lead to incorrect inferences.

To echo what was found in [KARC15], the main benefits of visualization may not be value retrieval tasks. Information-theoretically, the benefit of scatterplots can be explained using the cost-benefit metric in [CG16]. The input alphabet to PPMCC is the data space of numerous bivariate datasets. In most application contexts, the number of letters (i.e., possible datasets) is huge. On the other hand, the output from the PPMCC is a much smaller alphabet. For example, there are 2001 letters for $C_R \in [-1, 1]$ with an accuracy at three decimal places, which is normally sufficient for most correlation analysis. Hence the scale of *alphabet compression* using PPMCC is huge. Meanwhile the *potential distortion* is also very high.

Given a C_R value, many people would normally imagine an elliptical point cloud. However, experienced analysts know that $C_R \simeq 0$ does not always imply a circular point cloud, and there are many corresponding possible scatterplots that feature strong correlation patterns (e.g., points distributed along two opposite diagonal lines). Similarly, given $C_R = 0.7$, there are many corresponding scatterplots where the suggested correlation is rather doubtful (e.g., some stimuli for **H4**). Viewing a scatterplot can quickly reduce such potential distortion, though using it to estimate C_R would lead to numerical distortion and high cognitive load. Hence, the best practice would be to have both for correlation analysis, and viewing both a scatterplot and the corresponding PPMCC value does not cost much more than viewing either. This also suggests that it is highly risky to include PPMCC calculations in fully-automated processes without any normality test. When no reliable normality test is available, it is necessary to involve humans to validate the necessary conditions for PPMCC calculations using scatterplots. In fact, this is exactly what Anscombe tried to enlighten us through his quarte [Ans73].

The JND part (**H6**) of the study confirmed the effect of reference points as shown in [RB10, HYFC14]. Our estimated JND values at $C_R = \pm 0.3$ are close to [HYFC14] but much lower than [RB10]. Those at $C_R = \pm 0.7$ are close to [RB10], but higher than [HYFC14]. Although this study was not intended to evaluate the question on linearity [HYFC14, KH16], it shows that such functional approximation should be restricted to the conditions that the bivariate dataset satisfies bivariate normal distribution in both directions. At the moment, we do not know if JND values are affected by conditions in **H2**∼**H4**. The results of this study indicate such a possibility. In addition, JND values could also be affected by other factors such as training [LAK85, MS92]. When we divided the 35 participants into four bands of mean NND, we observe noticeable differences in NND values (from 0.11 to 0.75). An additional visualization is given in the supplementary materials. The non-trivial numerical differences in terms of JND values among these three studies also suggest that further empirical studies will be desirable.

References

[Ans73] ANSCOMBE F. J.: Graphs in statistical analysis. *The American Statistician 27* (1973), 17–21. 11

[BK79] BOBKO P., KARREN R.: The perception of pearson product mo-

ment correlations from bivariate scatterplots. *Personnel Psychology 32*, 2 (1979), 313–325. 1, 2, 6

[BS66] BEACH L. R., SCOPP T. S.: Inferences about correlations. *Psychpnomic Science 6* (1966), 253–254. 2

[CDM82] CLEVELAND W. S., DIACONIS P., MCGILL R.: Variables on scatterplots look more highly correlated when the scales are increased. *Science 216* (1982), 1138–1141. 1, 2

[CG16] CHEN M., GOLAN A.: What may visualization processes optimize? *IEEE Transactions on Visualization and Computer Graphics 22*, 12 (2016), 2619–2632. 11

[CH17] CORRELL M., HEER J.: Regression by eye: Estimating trends in bivariate visualizations. In *Proc. CHI* (2017). 2

[Cha67] CHAPMAN L. J.: Illusory correlation in observational report. *Journal of Verbal Learning and Verbal Behavior 6* (1967), 151–155. 2

[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. 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

[Cum13] CUMMING G.: The new statistics: why and how. *Psychological Science 25*, 1 (2013), 7–29. 1

[DAAK07] DOHERTY M. E., ANDERSON R. B., ANGOTT A. M., KLOPFER D. S.: The perception of scatterplots. *Perception & Psychophysics 69*, 7 (2007), 1261–1272. 1, 2, 8

[EDF08] ELMQVIST N., DRAGICEVIC P., FEKETE J.-D.: Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. *IEEE Transactions on Visualization and Computer Graphics 14*, 6 (2008), 1539–1148. 2

[Erl66] ERLICK D. E.: Human estimates of statistical relatedness. *Psychonomic Science 5* (1966), 365–366. 2, 6

[FD05] FRIENDLY M., D D.: The early origins and development of the scatterplot. *Journal of the History of the Behavioral Sciences 41 2* (2005), 103–130. 2

[HG76] HAMILTON D. L., GIFFORD R. K.: Illusory correlation in interpersonal perception: A cognitive basis of stereotypic judgments. *Journal of Experimental Social Psychology 12* (1976), 392–407. 2

[HYFC14] HARRISON L., YANG F., FRANCONERI S., CHANG R.: Ranking visualizations of correlation using Weber's law. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (2014), 1943–1952. 1, 2, 11

[KARC15] KANJANABOSE R., ABDUL-RAHMAN A., CHEN M.: A multi-task comparative study on scatter plots and parallel coordinates plots. *Computer Graphics Forum 34*, 3 (2015), 261–270. 2, 11

[KH16] KAY M., HEER J.: Beyond Weber's law: A second look at ranking visualizations of correlation. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (2016), 469–478. 1, 2, 11

[Kon05] KONARSKI R.: Judgments of correlation from scatterplots with contaminated distributions. *Polish Psychological Bulletin 36*, 1 (2005), 51–61. 2

[Kru89] KRUEGER L. E.: Reconciling Fechner and Stevens: Towards a unified psychophysical law. *Behavioral and Brain Sciences 12* (1989), 251–320. 4, 10

[LAK85] LANE D. M., ANDERSON C. A., KELLAM K. L.: Judging the relatedness of variables: The psychophysics of covariation detection. *Journal of Experimental Psychology: Human Perception and Performance 11*, 5 (1985), 640. 2, 11

[LMVW10] LI J., MARTENS J.-B., VAN WIJK J. J.: Judging correlation from scatterplots and parallel coordinate plots. *Information Visualization 9*, 1 (2010), 13–30. 1

[LP89] LAUER T. W., POST G. V.: Density in scatterplots and the estimation of correlation. *Behaviour & Information Technology 8*, 3 (1989), 235–244. 2, 8

[LS89] LEWANDOWSKY S., SPENCE I.: The perception of statistical graphs. *Sociological Methods & Research 18*, 2-3 (1989), 200–242. 2

[Mor78] MORRISON D. G.: A probability model for forced binary choice. *The American Statistician 32*, 1 (1978), 23–25. 4

[MS92] MEYER J., SHINAR D.: Estimating correlations from scatterplots. *Human Factors: The Journal of the Human Factors and Ergonomics Society 34*, 3 (1992), 335–349. 1, 2, 11

[Pea95] PEARSON K.: Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London 58* (1895), 240–242. 2

[PKF*16] PANDEY A. V., KRAUSE J., FELIX C., BOY J., BERTINI E.: Towards understanding human similarity perception in the analysis of large sets of scatter plots. In *Proc. CHI* (2016). 2

[Pol60] POLLACK I.: Identification of visual correlational scatterplots. *Journal of Experimental Psychology 59* (1960), 351–360. 2

[RB10] RENSINK R. A., BALDRIDGE G.: The perception of correlation in scatterplots. *Computer Graphics Forum 29*, 3 (2010), 1203–1210. 1, 2, 11

[Ren13] RENSINK R. A.: On the prospects for a science of visualization. In *Handbook of Human Centric Visualization*. Springer, 2013, pp. 147–175. 2

[SH78] STRAHAN R. F., HANSEN C. J.: Underestimating correlation from scatterplots. *Applied Psychological Measurement 2*, 4 (1978), 543–550. 1, 2

[Sjö85] SJÖBERG L.: A study of four methods for scaling paired comparisons data. *Scandinavian Journal of Psychology 6*, 1 (1985), 173–185. 3, 4, 10

[WCHB10] WICKHAM H., COOK D., HOFMANN H., BUJA A.: Graphical inference for infovis. *IEEE Transactions on Visualization and Computer Graphics 16*, 6 (2010), 973âĂŞ979. 2