Evaluating Perceptual Bias During Geometric Scaling of Scatterplots

Yating Wei, Honghui Mei, Ying Zhao, Shuyue Zhou, Bingru Lin, Haojing Jiang, and Wei Chen

Abstract—Scatterplots are frequently scaled to fit display areas in multi-view and multi-device data analysis environments. A common method used for scaling is to enlarge or shrink the entire scatterplot together with the inside points synchronously and proportionally. This process is called geometric scaling. However, geometric scaling of scatterplots may cause a perceptual bias, that is, the perceived and physical values of visual features may be dissociated with respect to geometric scaling. For example, if a scatterplot is projected from a laptop to a large projector screen, then observers may feel that the scatterplot shown on the projector has fewer points than that viewed on the laptop. This paper presents an evaluation study on the perceptual bias of visual features in scatterplots caused by geometric scaling. The study focuses on three fundamental visual features (i.e., numerosity, correlation, and cluster separation) and three hypotheses that are formulated on the basis of our experience. We carefully design three controlled experiments by using well-prepared synthetic data and recruit participants to complete the experiments on the basis of their subjective experience. With a detailed analysis of the experimental results, we obtain a set of instructive findings. First, geometric scaling causes a bias that has a linear relationship with the scale ratio. Second, no significant difference exists between the biases measured from normally and uniformly distributed scatterplots. Third, changing the point radius can correct the bias to a certain extent. These findings can be used to inspire the design decisions of scatterplots in various scenarios.

Index Terms—Evaluation, scatterplot, geometric scaling, bias, perceptual consistency

1 Introduction

Scatterplots are a common type of visualization that supports the presentation and exploration of multi-dimensional data plotted on a two-dimensional (2D) plane [56]. At present, various display devices [10, 16, 46] are increasingly used in modern data analysis scenarios, and thus scatterplots are frequently scaled to fit different displays. In using multi-viewed business intelligence (BI) or visual analytics (VA) systems, analysts often expand a scatterplot of interest in thumbnail views to the main view to conduct in-depth exploration. Likewise, sharing a scatterplot across various displays of different sizes has nearly become a routine operation for communicating findings in collaborative data analysis [7, 17] (Figure 1).

The most straightforward method for scaling a scatterplot is to enlarge or shrink the entire plot together with the objects inside it synchronously and proportionally. This process is called geometric scaling. However, is geometric scaling always effective? Considering our practical experience and relevant research, we assume that the answer is no. Geometric scaling may cause a bias in visual perception, which is supposed to affect perceptual consistencies of data characteristics (e.g., numerosity, correlation, and clusters); therefore, this is detrimental to interactive data exploration. In Figure 1, two clusters are distinguished in Bob's scatterplot. However, they are indistinguishable when shown on mobile devices. In Figure 2(a), two scatterplots have the same amount of points, but people may feel that the right one has more points than the other. Moreover, existing research in psychology has demonstrated that a patch of points is recognized as sparse when the patch size increases [69], which supports our work.

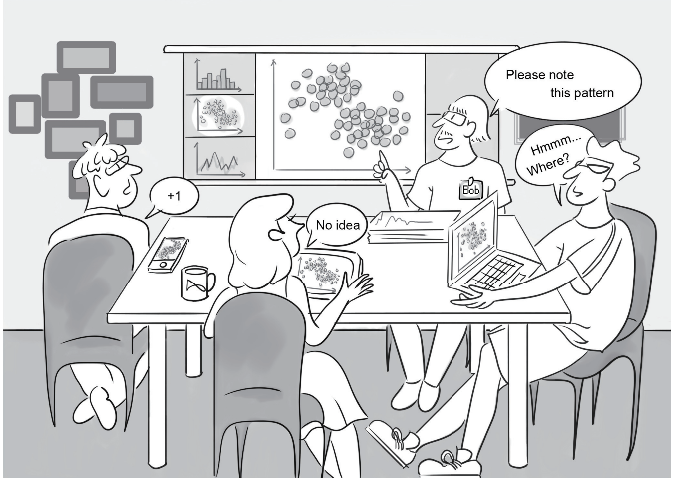


Figure 1. Scenario of scatterplot scaling. Bob finds a pattern of interest in a small scatterplot view and expands the scatterplot to a large view for further analysis. To exchange findings, he shares the scatterplot, and other collaborators examine the scatterplot using their own devices. The potential perceptual bias of scatterplots in various displays caused by scaling may lead to understanding inconsistency.

Extensive studies have devoted to the design decisions of scatterplots (e.g., point color [66, 75], aspect ratio [20, 33], and animation [12, 57]) in terms of empirical evaluations [52, 58, 72] or quantitative metrics [9, 40, 62] to achieve improved information presentation in various scenarios. Moreover, the usage issues of scatterplots in mobile devices [16], high-resolution displays [46], and immersive environments [13] have been extensively studied. However, scatterplot scaling, which is considered a special scenario for the usage and design decisions of scatterplots, has received minimal attention. To the best of our knowledge, no investigation has been systematically conducted on the potential perceptual bias during scatterplot scaling.

To address this research gap, we conducted controlled experiments to study such a bias. Specifically, the bias refers to the deviation between the perceived and physical values of visual features. We emphasized on three visual features (i.e., numerosity, correlation, and cluster separation, as shown in Figure 2(c)) and proposed three hypotheses: geometric scaling may cause a bias; the bias can be affected by data distribution; and changing the radius of points may reduce the bias.

- *Y. Wei, H. Mei, S. Zhou, B. Lin, and W. Chen are with The State Key Lab of CAD & CG, Zhejiang University, Hangzhou, Zhejiang 310058, China. E-mail: {weiyating, meihonghui, zhoushuyue, linbingru} @zju.edu.cn, chenwei@cad.zju.edu.cn.*
- *Y. Zhao, and H. Jiang are with School of Computer Science and Engineering, Central South University, Changsha, Hunan 410083, China. E-mail: {zhaoying, gallows}@csu.edu.cn*
- *Corresponding authors: Wei Chen and Ying Zhao*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

We adopted a two-interval forced choice (2IFC) method with a two-way staircase (2WS) design to simulate scatterplot scaling scenarios and collect participants' choice patterns of comparing a series of original-and-scaled scatterplot-pairs based on their subjective experience. Seven levels of scale ratio and point radius were selected, and scatterplots were generated using well-prepared synthetic data with 13 levels of visual feature, 2 distributions (normal and uniform). After several pilot studies to determine the experiments, we recruited 20 participants and conducted 3 rounds of experiments on each visual feature. Each round consisted of preparation, introduction and tutorial, pre-experiment, formal experiment, and subjective questionnaire.

We recorded the participants' choice patterns and subjective questionnaire answers as experimental results. We conducted a point of subjective equality (PSE) analysis to derive the quantitative biases from these choice patterns and carried out a series of statistical analyses on these derived biases. The analysis results showed that the first hypothesis was fully confirmed, the second hypothesis was fully negated, and the third hypothesis was partially confirmed. We also obtained other interesting findings from subjective questionnaires. We summarized these outputs systematically and discussed the limitations of this work and future directions.

In summary, we present the first attempt to understand the perceptual bias of scatterplots caused by geometric scaling. We contribute a carefully designed evaluation and a series of instructive findings, which may bring new considerations to the design decisions of scatterplots in various creation, exploration, and sharing scenarios [41, 74, 84, 85].

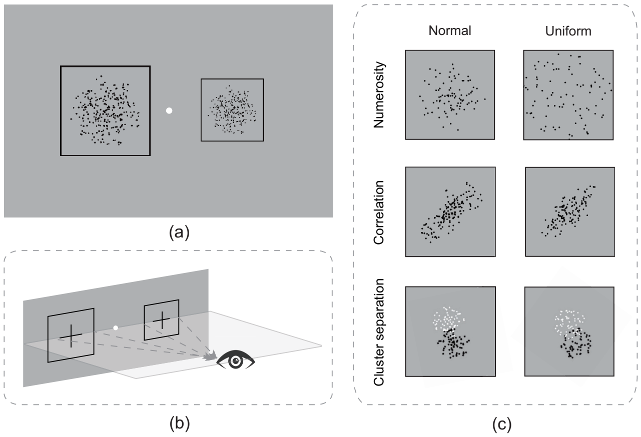


Figure 2. Stimuli and visual features. (a) Stimulus-pair consists of two scatterplots. In this example, the left figure is the original scatterplot and the right one is scaled by 63%. (b) Positioning of the participant and stimuli. (c) Examples of synthetic scatterplots with three visual features and two distributions.

2 Related Work

2.1 Perceptions of Scatterplot

Scatterplots are popular charts with a long history [15, 25]. Research has shown that scatterplots are more effective than many other diagrams, especially when data are 2D [39, 42, 59, 80]. The popularity and usability of scatterplots are due to their simplicity and capability to allow users to easily perceive different visual features [56], which can be divided into three levels based on cognitive complexity. Low-level perceptions are obtained from direct visual stimuli (e.g., position, color and density/numerosity), which are called preattentive processing [32, 77]. Mid-level perceptions are understandings gained from stimuli patterns, such as correlation [51, 52] and cluster separation [62, 67, 78]. High-level perceptions, such as memorability [11], aesthetics [30], and engagement [54], are user cognition built upon the understandings. In this work, we investigate the effect of scatterplot scaling on perceived visual features. We examine one low-level (numerosity, which is perceived equally as density in most scenarios of visual analytics [5]) and two mid-level (correlation and cluster separation) visual features.

2.2 Visual Encodings of Scatterplot

Many works have studied design evaluation and automatic decision on the visual encodings of scatterplots. Design evaluation investigates the visual encodings of scatterplots in particular scenarios and provides empirical guidance on many facets, such as size and color of points [18, 66, 77], aspect ratio [20, 33, 76], selection of dimensions [9, 68], amenities [14], and interaction [49, 55] and animation [12, 57]. Furthermore, many quality metrics have been developed for automatic design decision [8, 43, 45]. Quality metrics, such as those for clutter reduction [48] and cluster detection [35, 68, 79], measure the ability of diagrams to perform certain analytical tasks. However, existing studies have mainly focused on the design decisions when scatterplots are fixed and unchanged, whereas we are interested in the scenarios of changing display sizes.

The requirements for analyzing data on different displays, rather than only on desktop monitors, have increased with the development of display devices. Researchers have noted that display sizes strongly affect the usability of scatterplots [4]. They have studied the effect of display sizes on design decisions of scatterplots, such as interactions on mobile phones [23, 29], cognitive processes on large high-resolution displays [3], and graph layouts in immersive environments [38]. Moreover, the advances in display technology invoke new scenarios for exploratory analysis. On the one hand, large displays allow many diagrams to be presented simultaneously, thereby forming multi-view interfaces [4, 50, 82]. On the other hand, given the application of spatially-aware and wireless connection techniques, multi-device combinations are increasingly used to provide collaborative data analysis involving multiple people. These new scenarios have led to frequent scatterplot scaling. However, few studies have investigated the effect of scatterplot scaling. In this work, we aim to find possible solutions to alleviate the effect of scatterplot scaling on visual consistency.

2.3 Visual Biases of Scatterplot

Perceptual bias is an important issue that has been extensively studied [1, 24, 73, 86]; it considers the perceptual differences between situations, which include three cases. The first case refers to the differences between perceived visual features and quantitative statistics. Rensink et al. [51, 52] found that the perceived correlation tends to underestimate the physical correlation; the correlation perception in scatterplots remarkably fits the Weber's law and is stable under the varying densities, aspect ratios, and distributions. Sedlmair et al. [60] proposed that many factors may differentiate the perceived cluster separation from defined quality metrics. Alexander et al. [1] found that using font size as a data encoding can influence the perception of underlying values. The second case is the inter-individual variations of perception. Many studies [27, 36, 83] have proven that scatterplots have minimal inter-individual differences compared with other diagrams, such as radar graph [21, 61, 64] and parallel coordinates [36]. The third case refers to the differences between the perceived visual features of plots with various designs. Cleveland et al. [19] found that the correlation perception can be enhanced by expanding the blank margins of scatterplots. Sophian and Chu proposed that the perceived numerosity decreases when dots cluster together [63]. Valdez et al. [73] proved that priming and anchoring effects exist in perception tasks of scatterplots under certain circumstances. Yang et al. [81] indicated that people tend to use a small number of visual features when judging the correlation in scatterplots. To the best of our knowledge, no study has systematically evaluated the perceptual bias caused by scatterplot scaling.

3 Problems and Hypotheses

A scatterplot often needs to be scaled to fit displays of different sizes in multi-view and multi-device data analysis environments. To maintain perceptual consistency among displays, the most straightforward method is to use geometric scaling, which either enlarges or shrinks the entire original scatterplot together with visual forms inside it proportionally. However, it is unclear that whether geometric scaling is effective in preserving perceptual consistency. Our basic observation is that geometric scaling will cause a perceptual bias. Such a bias can affect perceptual inconsistency in interactive data exploration.

According to Jerit [34], *"Perceptual bias occurs when factual beliefs deviate from reality."* Empirically, people frequently mismatch the perceived and actual values of visual features when a scatterplot is scaled. For example, if a scatterplot on a laptop is projected onto a curtain, then people may feel that the enlarged scatterplot has fewer points than the original one. Similarly, if a scatterplot is transferred from a desktop computer to a phone, then the clusters on the original scatterplot occasionally become blurred and unclear on the phone. Therefore, we suspect that geometric scaling can lead to a perceptual bias. Moreover, we are interested in investigating the factors that are affected by the bias. Is a certain law involved? Can we find a way to reduce the bias? In this work, we conduct controlled experiments to explore these issues. To guide our experimental design, we formulate the following hypotheses:

H1: We assume that geometric scaling causes a bias in the perceived visual features of scatterplots, including numerosity, correlation, and cluster separation. We also believe that the bias has a linear relationship with scale ratio ($Size_{scaled-one}/Size_{original-one}$). Thus, we seek to verify the existence of the bias and its inherent law related to the scale ratio. The three visual features are the most fundamental and important data characteristics that scatterplots allow users to perceive, and they have been thoroughly investigated by numerous studies (Section 2.1). The assumed linear relationship is based on our practical experience that the bias will be superior as the scale ratio increases or decreases.

H2: We assume that the bias caused by geometric scaling can be affected by data distribution. This hypothesis is used to explore whether the bias is related to data characteristics. We assume that scatterplots with various data patterns may cause different degrees of the bias. We select data distribution as our research object because it is a fundamental data characteristic. Specifically, we select normal and uniform distributions to perform a comparative analysis. Normal distribution that grows dense toward the center is treated as a different data pattern with uniform distribution.

H3: We assume that changing the radius of points can reduce the bias. We use this hypothesis to explore whether the bias can be reduced by changing visual encodings. The point radius, which is an important visual channel, is scaled proportionally in geometric scaling. We suppose that changing the radius in another way can reduce the bias.

4 Experimental Design

To test these hypotheses, we carefully designed three experiments. The first experiment (E1) measured the perceptual bias on the three visual features (H1). The second experiment (E2) examined whether the bias is affected by data distribution (H2). The third experiment (E3) aimed to find the effect of changes in point radius on the bias (H3). In this section, we introduce the experimental design in detail.

4.1 Stimuli

The stimuli were 2D scatterplots that shared some visual encodings and rendering options. The aspect ratio of all scatterplots was set to 1 to eliminate the interference caused by aspect ratio. All scatterplots were rendered with gray backgrounds and black thin borders [6], as illustrated in Figure 2(a). The domains of x and y dimensions in all scatterplots were normalized to [0,1]. Points in the scatterplots were represented by black circles. One exception was when testing cluster separation, in which the circles of two colors presented two clusters. The radii of circles in a scatterplot were the same. When a scatterplot was scaled, the visual encodings, including the width and height of the scatterplot, the length of axis, and the radii of points, were scaled proportionally. Our experiments used seven scale ratios, namely, 25%, 40%, 63%, 100%, 159%, 252%, and 400%, which formed a geometric sequence for simulating the geometrical size changes of scatterplots. The scale ratio of the original scatterplot was maintained at 100%. Other scatterplots were scaled by different scale ratios, of which three were enlarging and three were shrinking. We selected these scale ratios because the scale ratios of scatterplots can cover most scenarios considering common displays, including 4−6 *in* mobile phones, 7−12 *in* tablets, and 25−32 *in* desktop monitors. To simplify the study, we used a unified desktop monitor to render the scaled scatterplots.

4.2 2IFC & 2WS Method

We used a two-interval forced choice method with a two-way staircase design (2IFC & 2WS) to simulate scatterplot scaling scenarios and measure people's subjective experience.

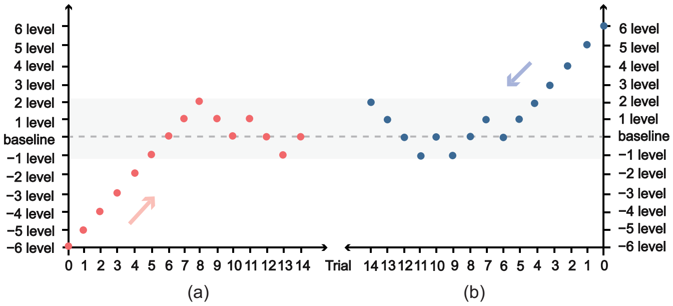


Figure 3. Example of a 2WS sequence. The y-axis shows visual feature levels. The x-axis shows the number of trials. (a) Fifteen trials of the forward-staircase. (b) Fifteen trials of the backward-staircase. The gray background shows the fluctuating interval of both staircases.

The 2IFC is a method used to measure the subjective experience of a person through his/her choice pattern. In a 2IFC trial, the person is asked to judge between a reference and test stimulus-pair. The judgment is a subjective comparison to obtain the stimulus with a larger value of the visual feature. Figure 2(a) depicts an example. The left scatterplot with 256 points is the reference/original stimulus, and the right scatterplot with the same amount of points is the test stimulus that is scaled by 63%. The person needs to select the one with a larger numerosity. Suppose the person selects the right one because he/she thinks that it has a larger numerosity, which deviates from the physical amount. Such a deviation may be caused by a perceptual bias. We can collect multiple judgments within a certain value range of the testing visual feature, and then measure the bias from the choice pattern. Therefore, a sequence of trials is required in a 2IFC task for each combination of visual feature and scale ratio.

The construction of a reasonable 2IFC sequence involves two important considerations. (1) The levels of the testing visual feature must cover a suitable range. Too few levels make bias measurement less precise, and too many levels lead to an excessive number of trials. (2) The order of appearance of feature levels in a sequence of trials must be carefully designed. A random sequence requires many trials to obtain an accurate bias, whereas participants may find the regularity when levels are sorted.

Given that we had seven scale ratios, we used more levels than conventional settings (e.g., 7 in [69] and 10 in [52]) to cover a large value range of visual features. We defined 13 feature levels for each visual feature, that is, {−6 level, −5 level, ..., 0 level, ..., +6 level}, 6 below and 6 above the baseline (0 level). In addition, referring to previous studies [69, 70], the levels of numerosity were defined through octave method [69], and the levels of the two other features were defined as an arithmetic series [51] (Section 4.3).

We used a 2WS design to determine the order of appearance of feature levels. A staircase design [26, 47] is that, the feature level of the test scatterplot varies to make the next trial difficult or simple, and the reference scatterplot is maintained at the baseline. The sequence can begin from either the −6 level (forward-staircase) or +6 level (backward-staircase); both terminate near the baseline, thereby forming a two-way staircase. Figure 3 demonstrates the experimental result of a participant in a 2WS sequence. In the forward- or backward-staircase, the trials fluctuate within a certain interval around the baseline (the highlighted area) after several monotonic changes (the arrows) because the feature levels of the reference and test scatterplots are close to each other, thus making judgments difficult. Such a fluctuating choice pattern is the aim of our 2IFC & 2WS method, through which the participant's subjective experience can be measured using PSE [37, 71] to derive the bias (Section 5.1). In a word, the 2WS design provides a two-pass verification and eliminates the effect caused by the direction of level change. Moreover, the number of trials in each direction was set to 15 for a total of 30. This number of trials was greater than twice the number of feature levels, thereby guaranteeing the occurrence of fluctuation patterns.

In Figure 2(b), we positioned the center of the reference and test scatterplots symmetrically on the left and right parts of the center of the participant's viewpoint. Considering the displays of different sizes, we specified the width and height of a scatterplot in degrees of visual angle. The reference scatterplot extended 5° of the visual angle vertically and horizontally. The scatterplot sizes of 7 scale ratios ranged from 1.25° to 20° of the visual angle. These sizes received comfortable feedback in our pilot studies and fit inside the near-peripheral vision [28].

4.3 Experimental Variable and Data

We used synthetic data to generate scatterplots, as shown in Figure 2(c). The variables and data used in the three experiments were different.

4.3.1 Experiment 1

E1 was designed to measure the perceptual bias caused by geometric scaling on the three visual features of scatterplots: numerosity, correlation, and cluster separation. The variables of E1 included scale ratio, feature, and feature level. For each combination of scale ratio and feature, each participant was asked to finish a 2IFC & 2WS sequence with 30 trials, thereby obtaining the choice pattern for bias measurement. Corresponding to 7 scale ratios, each participant completed 7 sequences for a total of 210 trials to detect the variation trend when the scale ratio changes. Three features resulted in 630 trials for each participant.

We generated a series of point sets for each feature. In each series, the corresponding visual feature was a variable with 13 levels of values to control the generation of 13 point sets, whereas other features remained unchanged. For each level, we generated a group of point set candidates and randomly selected one candidate for each trial to avoid participants' learning or fatigue effects. Specifically, the point sets for the three features were generated as follows.

Numerosity: Point sets for numerosity were normally distributed point clouds with 13 numerosity levels. The baseline was 89 points. By extending the baseline numerosity in two ways, we obtained all 13 levels, which formed a 2-octave range [69] as follows: 11, 15, 22, 31, 44, 63, 89, 127, 180, 256, 363, 515, and 730 points. The points were sampled from a 2D normal distribution with

$$\mu = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}, \Sigma = \begin{bmatrix} 0.25 & 0 \\ 0 & 0.25 \end{bmatrix}$$

The positions of all points were limited in a constraint circle with (0.5, 0.5) as the center and 0.5 as the radius. If any point exceeded this constraint circle, then this point was removed and regenerated. Thus, the point cloud was normally distributed in the center of the scatterplot within a moderate range and with minimal overlap.

Correlation: Point sets for correlation were also normally distributed point clouds but with 13 levels of correlation coefficients, which formed an arithmetic series from 0.15 to 0.9 with a step of 0.0625 (baseline is 0.53). We generated point clouds for each correlation level by using a method similar to the data for numerosity, except that the number of points remained unchanged at 128, and the parameters of normal distribution were changed with

$$\mu = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}, \Sigma = \begin{bmatrix} 0.2 & r \\ r & 0.2 \end{bmatrix}$$

where r is the target correlation coefficient. However, the correlation coefficient of generated point sets may float around target r. We only preserved those with correlation within $r \pm 0.005$.

Cluster separation: For cluster separation, we considered the simple case of two clusters. Thus, we generated point sets that contained two normally distributed point clusters. We used a silhouette index s to measure the cluster separation because it only depends on the actual partition of points rather than the clustering algorithm. We selected 13 levels of silhouette indices that formed an arithmetic series ranging from 0.11 to 0.65 with a step of 0.045 (the baseline is 0.38). Each point set contained 128 points, which were divided into two clusters of 64 points. However, directly generating a point set for a given silhouette index was difficult. Initially, we randomly generated two constraint circles, which centered at (x_1, y_1) and (x_2, y_2) with a radius of 0.4, for the two clusters. x_1, y_1, x_2 and y_2 were random numbers to ensure that the two constraint circles were completely inside the scatterplot. Then, we sampled points from two 2D normal distributions with

$$\mu_1 = \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}, \mu_2 = \begin{bmatrix} x_2 \\ y_2 \end{bmatrix}, \Sigma_1 = \Sigma_2 = \begin{bmatrix} 0.2 & 0 \\ 0 & 0.2 \end{bmatrix}$$

Lastly, we computed all silhouette indices of generated point sets and preserved these sets with silhouette indices within $s \pm 0.005$. The points of the two clusters were rendered black and white on a gray background. Such color settings reduced the participants' visual burden and enabled them to focus on cluster separation.

4.3.2 Experiment 2

Given that E2 focused on the effect of data distribution, it shared the same variable settings with E1, and its data generation was similar to E1 in most settings. The difference was that the point sets were sampled from uniform distributions. Note that, for correlation tests in E2, we constructed a constraint ellipse and then sampled from a uniform distribution within the ellipse. The constraint ellipse had a center point at (0.5, 0.5), and the major axis went along the diagonal direction from the lower left to the upper right of the scatterplot. The length of the major axis was set to 1, whereas the length of the minor axis was a random number. Thus, we generated several candidate point sets and computed their correlation coefficients. Subsequently, we selected the candidates that satisfied the value requirements of each level with an error less than 0.005.

4.3.3 Experiment 3

E3 was designed to investigate the effect of changing the point radius on the perceived visual features. It investigated three variables: point radius, feature, and feature level. Moreover, we selected two scale ratios as constants: 63% and 252%, that is, the -1 level and $+2$ level of 7 scale ratios, to test enlarging and shrinking cases. For the enlarging case, we selected the moderate $+2$ level. The shrinking cases used the -1 level because identifying small scatterplots (≤ -2 level) was difficult for participants. The radius ranged from 1.16 to 3.46, increasing geometrically and resulting in 7 levels (1.16, 1.39, 1.67, 2, 2.4, 2.88, and 3.46). For each combination of radius and visual feature, each participant was asked to finish a 2IFC & 2WS sequence with 30 trials including 13 feature levels, thereby obtaining the bias of a certain feature for a given radius. For each feature, each participant was asked to complete 7 sequences of different radii to detect the variation trend of the bias when radius changed. Three features and two scale ratios resulted in 1,260 trials for each participant. The data used in E3 were the same as those in E1. The only difference was that the points were rendered with various radii.

4.4 Participants and Apparatus

Given that our experimental tasks were time consuming and visually strenuous, we recruited 20 participants (8 females and 12 males; undergraduate and graduate students with experience in data analysis using scatterplots; aged 19−25 years, median age: 21; with normal or corrected-to-normal vision) rather than using crowdsourcing approaches such as Amazon Mechanical Turk. We can observe the mental state of the participants and control the experimental process. Moreover, we provided an independent, quiet laboratory with a minimal external interference. We asked the participants to sleep early for at least three days before the experiments so that they would be fully rested and have full physical strength. We also prepared coffee and snacks to help the participants relax during breaks. The participants who completed the study were compensated with $6 per hour.

All experiments were conducted on a Dell 25 *in* U2515H monitor whose screen size is 21.8 *in*×12.3 *in* and resolution is 2560×1440 pixels, with a 60 Hz refresh rate and 50 cd/m^2 mean luminance. The participants viewed the stimuli at 59 *cm* for the visual angle of approximately 1.25 arcmin per pixel. A standard wireless mouse and wired keyboard were used.

4.5 Procedure

4.5.1 Pilot Study

We invited 5 participants to conduct pilot studies for 3 times. These studies helped determine the details of the experiment design. In the pilot studies, we found that the stability of 2IFC judgments was much higher during the day than at night. This observation indicated that the human spirit and experimental environment had a great impact on the formulation of 2IFC judgments. Therefore, we arranged all experiments during the day.

To determine the experimental order, the first pilot followed the order of E1, E2, and E3. Each experiment contained all three features. We noticed that the participants had to refocus between features in a single experiment. Therefore, we used a feature-major order in the second pilot. Only one feature was tested in one round of E1−E3 to maintain the thinking-pattern continuity of the participants.

We originally planned to use a two-alternative forced choices (2AFC) method [69] to display a scatterplot-pair (i.e., the left and right scatterplots appear concurrently), but we finally selected the 2IFC method (i.e., the left and right scatterplots appear sequentially) because the method enabled the participants to focus on comparing the scatterplot-pair for making a judgment, without worrying whether they could view two scatterplots simultaneously.

Moreover, we repeatedly adjusted the data generation parameters (e.g., the value ranges and intervals of features), the interface and the procedure (e.g., adding practice trials and pre-experiment phase and adjusting the time span of breaks) through the pilot studies.

4.5.2 Formal Study

We designed a three-round experimental process for the formal study. The core variable in the three rounds was visual feature. That is, each participant performed the three experiments (E1, E2, and E3) of one feature in one round and completed all features (numerosity, correlation, and cluster separation) by performing three rounds. Four experimental phases were included in each round.

Preparation: An experiment instructor first led a participant into the laboratory room, and then assisted the participant in adjusting the position and height of the monitor to ensure that the point in the center of the screen was at eye level, as displayed in Figure 2(b).

Introduction and tutorial: This phase familiarized the participant with the experimental procedures, tasks, visual features, and data. This phase generally took approximately 10 min and comprised four steps.

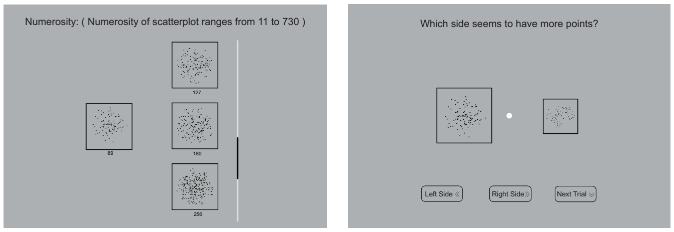
STEP 1. The instructor provided a brief description of the research purposes, related concepts, experimental procedure, and tasks.

STEP 2. The instructor selected a visual feature from three features following Latin squares. Then, the instructor showed the prepared data samples of the selected feature and asked the participant to observe them for 2−3 min, as presented in Figure 4(a). Thus, the participant gained an intuitive presentation of the feature.

STEP 3. The instructor explained the interactions of the system interface to the participant. The system interface is illustrated in Figure 4(b). Based on the 2IFC method, the left scatterplot was first presented for 250 ms and then disappeared. After displaying a center fixation point for 500 ms blank time, the right scatterplot was presented for 250 ms and then disappeared. Subsequently, the participant judged on the basis of his/her subjective experience. For numerosity, correlation, and cluster separation, the questions were "Which side seems to have more points?", "Which side looks more relevant?", and "Which side seems to be more obvious in cluster separation?", respectively. The interface supported two types of interactions to report the judgment: pressing the "left/right arrow" keys on the keyboard or clicking the "Left/Right Side" buttons on the screen using a mouse. The participant could modify the choice until he/she presses the "down arrow" key or click the "Next Trial" button to proceed to the next trial.

STEP 4. The instructor provided 10 practice trials of the current visual feature and asked the participant to complete the trials to be familiarized with the interactions and rhythm of making judgments.

Pre-experiment: After finishing the tutorial, the participant was required to perform a short pre-experiment for the current feature to test whether the participant had been fully prepared. The pre-experiment contained 2 (levels) × 30 (trials of a sequence) × 4 (E1, E2, E3 [scale ratio = 63%], and E3 [scale ratio = 252%]) trials. After completing the pre-experiment, the instructor analyzed the result immediately. If the result did not reach the standard (Section 5.1), then the experiment was suspended. In general, a poor pre-experiment result was due to poor focus or imperfect task understanding. If the participant expressed fatigue, then he/she would be asked to rest for at least 2−3 hours. If the participant was confused with tasks, then back to the previous phase. This phase took approximately 6 min.


Numerosity: (Numerosity of scatterplot ranges from 11 to 730)

Which side seems to have more points?

(a)

(b)

Figure 4. Interface of the experiment system. (a) Interface of the tutorial. In this example, the visual feature is numerosity. The scatterplot on the left is the reference. The 13 scatterplots on the right show 13 numerosity levels in ascending order, which can be scrolled through to browse. (b) Interface of the formal experiment. The judgment question is shown on the top. The scatterplot-pair and a center fixation point are located in the middle. Buttons for reporting judgments and proceeding to the next trial are found at the bottom.

Formal experiment: The participant clicked the "Start" button to begin making judgments individually. After completing all trials of one experiment, the participant rested for 2−3 min and then proceeded to the next experiment. The order of three experiments (E1−E3) was fully counterbalanced across the participants. Each participant was required to complete 840 trials in one round (7 (levels) × 30 (trials of a sequence) × 4 (E1, E2, E3 [scale ratio = 63%], and E3 [scale ratio = 252%])). A full round of formal experiments, including breaks, took approximately 90 min to complete.

After one round, the participant was required to fill out the subjective questionnaire of the current round. A short interview was conducted subsequently to obtain additional subjective information. Then, the participant was given 1 h of rest. He/she was allowed to leave the laboratory room to take coffee and snacks to relax and relieve fatigue. Afterward, the participant proceeded to the next round. After three rounds, a formal interview was conducted. All three rounds had 2,520 trials (840 (trials of a round) × 3 (features)), and generally took approximately 6.5 h. Moreover, 7 scale ratios/point radii were fully counterbalanced using Latin squares among the participants, and the positions of reference and test scatterplots were completely random.

4.5.3 Subjective Questionnaire

The participants were asked to fill out a questionnaire to collect their subjective feelings on the current examined visual feature after completing a round of experiments. The questionnaire was set with two types of questions. (1) The first type inquired about the subjective feeling of the overall difficulty level (DL) of making 2IFC judgments. For example, "Do the changes of scatterplot size make your judgments on the perception of normal-distributed numerosity difficult? If so, then how difficult is it?" The participants rated it using a five-point Likert scale ranging from 1 (the lowest difficulty) to 5 (the highest difficulty). (2) The second type inquired about the subjective experience of the difficulty tendency (DT) of making 2IFC judgments with different scale ratios or point radii. For example, "How does the difficulty of making judgments change on the perception of normal-distributed numerosity with the changes in scatterplot size from small to large?" The participants were given five options: A. increasing, B. decreasing, C. increasing after decreasing first, D. decreasing after increasing first, and E. not sure. In the questionnaire, four specific questions were used for each of the two question types with respect to E1, E2, E3 (scale ratio = 63%), and E3 (scale ratio = 252%). The four specific questions were slightly different in experimental variables. Specifically, we asked about the normal distribution for E1, uniform distribution for E2, and point radius for E3. In summary, each participant was asked to solve 8 subjective questions for a visual feature and 24 questions for the three visual features.

5 Experimental Results

5.1 Analysis Approach

We collect all objective and subjective results from the experiments. Objective results record the information of the participants who complete each 2IFC trial, including choice, location of reference scatterplot, experimental variables, and completion time. Subjective results are extracted from the questionnaires, including 24 questions and interview records for each participant. We analyze the results in the following three main aspects.

Abnormal choice patterns identification. Some participants could be distracted occasionally within hours of 2IFC judgments, resulting in abnormal choice patterns in a sequence of trials. Two representative examples are depicted in Figure 5(a−b). Such abnormalities would have a serious effect on the accuracy of the subsequent bias measurement. We perform a variance analysis to identify fiercely fluctuating choice patterns. We compute the variance of the feature levels of 10 essential trials (the last five trials of forward- and backward-staircases) of a sequence. A large variance (> 2) typically indicates the occurrence of fiercely fluctuating choice patterns on the basis of the feedback from the participants in the pilot studies. In addition, we manually examine a small amount of sequences with early fluctuating choice patterns by observing whether the sequence fluctuates from the second or third trial. Finally, we identify 16% abnormal sequences. We mark these sequences as outliers and replace the biases measured from them with the mean of biases of all participants with the same experimental variables.

Bias measurement. Based on the human psychometric function [22] and Weber's law [36], the bias of a certain participant on a visual feature at a scale ratio/point radius can be measured quantitatively from his/her choices of a sequence of trials using point of subjective equality (PSE) and point of objective equality (POE). In this case, when the test and reference scatterplots of a trial appear subjectively to have the same feature value, the participant makes a random choice. Thus, the PSE is the 0.5 probability point of selecting "test $<$ reference" in a sequence of trials. The POE is the baseline feature value in which the actual feature values of the test and reference scatterplots are objectively equal. Therefore, if scatterplot scaling causes a perceptual bias, then the PSE in a 2WS sequence is set apart from the POE with a certain interval. The interval is exactly the bias, which equals the PSE minus the POE. As shown in Figure 5(c), the POE is 89 and the PSE is 65.86. At the PSE, the participant perceived that the numerosity values of two scatterplots were equal, yielding a negative bias ($-23.14 = 65.86 - 89$). However, the actual numerosity value (65.86) of the test scatterplot was less than that (89) of the reference one. This finding indicates that the shrunk test scatterplot (scale ratio = 40%) was perceived with a relatively large value of numerosity.

Significance analysis. We conduct two significance analyses for each experiment (E1, E2, and E3 [ratio = 63% and 252%]) to examine the significance of the objective results. The first is for the biases among the participants. We first use the Shapiro−Wilk test to examine the normality and find that most results did not follow the normal distributions ($p < 0.05$). Then, we use non-parametric Friedman and ANOVA tests for examination. The results show that, for each experiment and visual feature, no significant differences are found in the biases of various participants. Thus, the biases of different participants can be treated equally. The second significance analysis is for the biases among different scale ratios/point radii. The Shapiro−Wilk test shows that all results do not follow the normal distributions ($p < 0.05$). Thus, we use non-parametric Friedman tests to examine whether significant differences occur among the biases of 7 scale ratios/point radii for each visual feature.

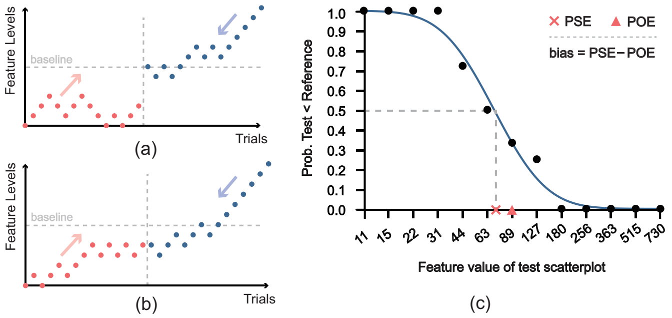


Figure 5. Examples of abnormal choice patterns: (a) fiercely fluctuating and (b) early fluctuating. (c) Illustration of the bias measurement. The plot is drawn on the basis of the choices of a sequence of trials made by a certain participant on numerosity at scale ratio = 40%, where the x-axis is the actual feature value of the test scatterplot, and the y-axis is the probability of selecting "the feature value of test scatterplot is perceived to be smaller than the feature value of the reference scatterplot." Furthermore, a data point represents the proportion of "test $<$ reference" choices made by the participant at a feature value of the test scatterplot, and the blue curve is the best-fitting cumulative Gaussian function of all data points. Thus, the POE is the baseline feature value point on the x-axis, whereas the PSE is the point on the best-fitting curve where y = 0.5. The bias is the value of the PSE minus the POE.

5.2 Objective Result Analysis

In this section, we test against the three hypotheses with the results of the bias measurement (Figure 6) and significance analysis. We take Plot 1 in Figure 6 as an example to explain how to read the bias measurement results. The bias is close to 0 when the scale ratio is 100%, indicating that the participants were nearly unbiased on the perception of numerosity when the test scatterplot was not scaled. The biases are negative and gradually move away from 0 bias when the scale ratio decreases (from 100% to 63%, 40%, and 25%), denoting that the participants perceived a higher numerosity than the physical numerosity in the shrunk test scatterplot and the difference between the physical and perceived numerosity increased gradually. The biases are positive and gradually move away from 0 bias when the scale ratio increases (from 100% to 159%, 252%, and 400%). This result signifies that the participants perceived a lower numerosity in the enlarged test scatterplot and the difference increased gradually.

Hypothesis 1. We assume that geometric scaling causes a bias in perceiving the three visual features, and the bias may have a linear relationship with the scale ratio. This hypothesis is fully confirmed.

With the existence of the bias, as shown in Plots 1, 2, and 3 in Figure 6, all features show the biases to a certain extent when the scale ratio is not at 100%, whereas the biases are very close to 0 when the scale ratio is at 100%. This observation indicates that enlarging and shrinking the test scatterplot affect the perceptual consistency of the three features.

In terms of the relationship between the bias and scale ratio, the results of significance analysis show significant differences in the biases of 7 scale ratios for all features (numerosity: χ^2 = 85.987, $p = 0.000 < 0.05$; correlation: χ^2 = 47.244, $p = 0.000 < 0.05$; cluster separation: χ^2 = 58.140, $p = 0.000 < 0.05$), indicating that a relationship must occur between the bias and scale ratio. We conduct a linear regression analysis on the biases of 7 scale ratios by feature. The results of fitting index R^2 (numerosity: $R^2 = 0.940 > 0.9$; correlation: $R^2 = 0.969 > 0.9$; cluster separation: $R^2 = 0.920 > 0.9$) denote that the bias has a significant linear relationship with the scale ratio for each feature.

We analyze further the variation trend of the bias by feature when the size of the test scatterplot gradually increases from the smallest scale ratio (25%) to the largest (400%) and obtain the following findings.

Numerosity: The bias presents a positive growth trend with the increase in the scale ratio (Plot 1 in Figure 6). This result means that the perceived numerosity decreases while the test scatterplot is gradually enlarged. On the basis of the occupancy model [2] and Steven's power law [65], the perceived numerosity in a scatterplot can be identified with

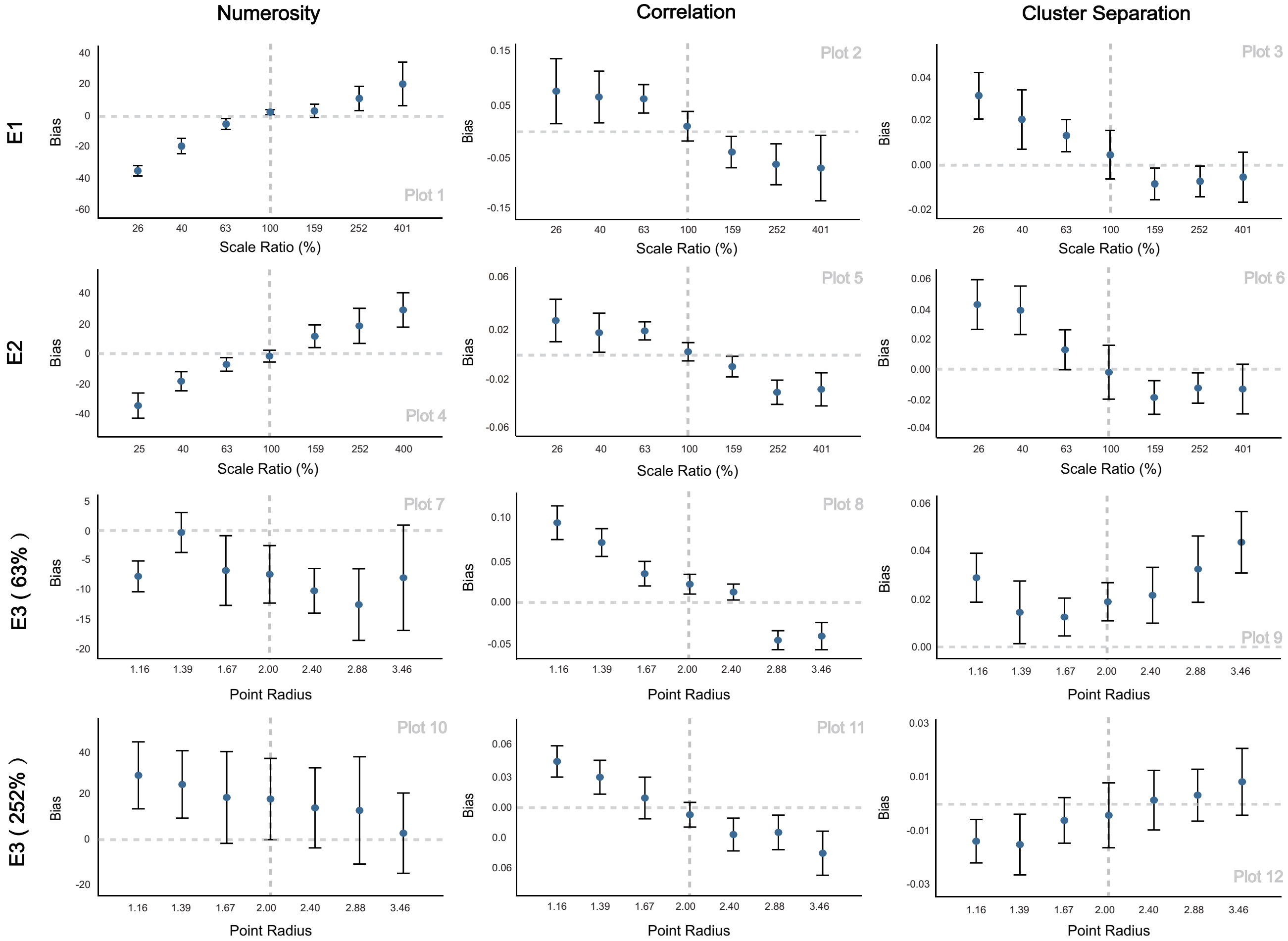


Figure 6. Objective results of the bias measurement with regard to experiments (E1, E2, E3 [scale ratio = 63%], and E3 [scale ratio = 252%]) and three visual features (numerosity, correlation, and cluster separation). A plot presents the biases of 7 scale ratios/point radii for a certain experiment and feature, in which the x-axis is a logarithmic (base 10) axis of scale ratio for E1 and E2 or point radius for E3 because 7 levels of scale ratios/point radii form a geometric sequence. The y-axis is the value of a bias. A point represents the mean of biases of all participants at a certain scale ratio/point radius. The error bar of a point encodes a 95% confidence interval, and two dashed lines in a plot represent the baselines of 0 bias and 100% ratio/2.0 radius.

the area occupied by points, and the perceived change in the point area is generally less significant than the perceived change in the scatterplot size when the scatterplot is scaled. That is, the perceived increase in the point area cannot keep pace with the perceived increase in the scatterplot size when the test scatterplot is enlarged, thereby leading to an underestimation of numerosity in the enlarged test scatterplot. The test scatterplot requires a high physical numerosity to be perceived as the same numerosity of the reference one.

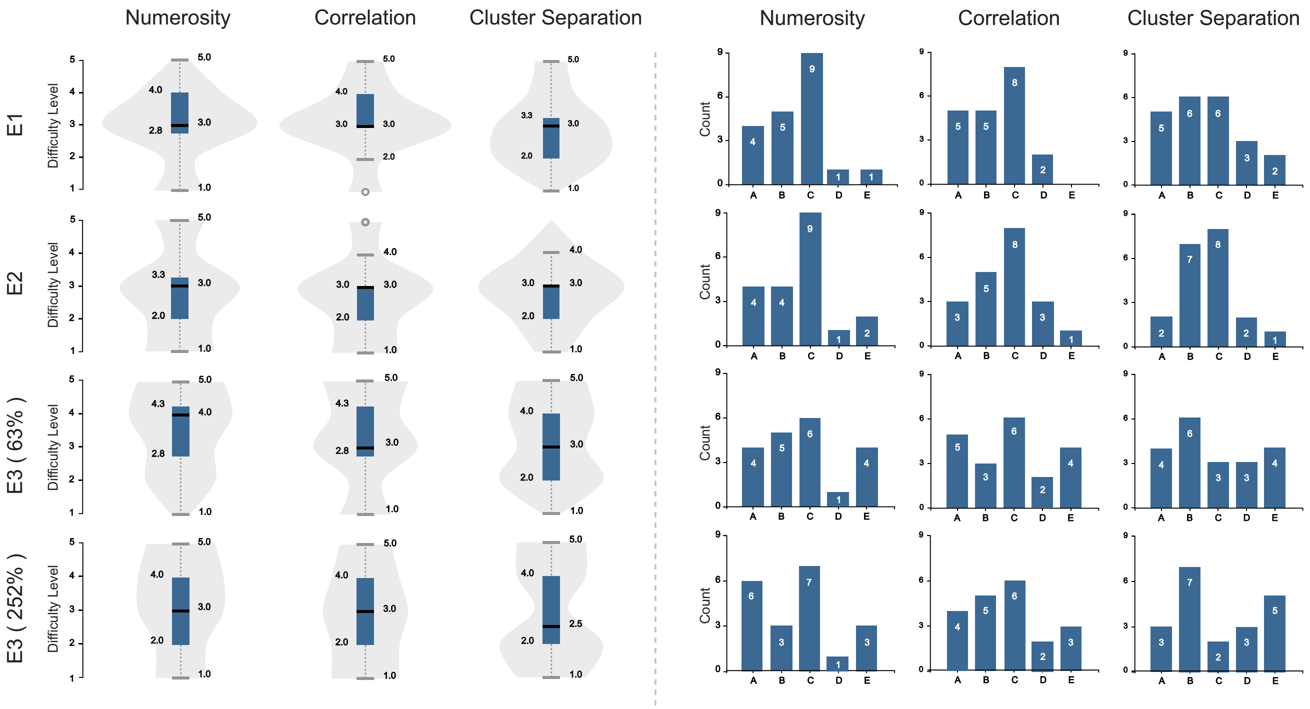
Correlation: The bias has a negative growth trend with the increase in the scale ratio (Plot 2 in Figure 6). This result means that the perceived correlation increases while the test scatterplot is gradually enlarged. The participants said that they perceived a correlation mainly by observing the area and major axis length of the ellipse formed by points [81]. According to Steven's power law [65], the perceived change in the ellipse area is less significant than the perceived change of the ellipse length when a scatterplot is scaled. That is, the perceived ellipse is relatively thin in the enlarged test scatterplot, and the test scatterplot requires a small physical correlation to be perceived as correlated as the reference one.

Cluster separation: The bias has a negative growth trend with the increase in the scale ratio (Plot 3 in Figure 6). The participants expressed that they perceived cluster separation mainly by observing the number of points in the overlap area of two clusters. Similar to numerosity, the perceived change in the overlap area is less significant than the perceived change in the scatterplot size when a scatterplot is scaled. That is, two clusters are perceived with a relatively large level of cluster separation in the enlarged test scatterplot. Moreover, the negative growth trend of the bias becomes flat on the right side of the 100% scale ratio. The participants likely encounter less difficulty in identifying the points in the overlap area when the size of the test scatterplot is sufficiently large.

Hypothesis 2. We assume that the bias can be affected by data distribution. We conduct pairwise t-tests to examine the significant differences between the biases of E1 (normal distribution) and E2 (uniform distribution) by feature and scale ratio. Then, we apply a Bonferroni correction to reduce the significance level from $p = 0.05$ to 0.00714 (0.05/7) because we divide the results of the bias measurement of each feature by 7 scale ratios in the t-test. No significant differences ($p < 0.00714$) are found between the biases of E1 and E2 at 7 scale ratios for the three features. This hypothesis is fully negated. We think that the visual patterns formed by a normal distribution may have inefficient difference from those formed by a uniform distribution.

Hypothesis 3. We assume that changing the point radius can reduce the bias. In E3, the test scatterplot was scaled at a fixed scale ratio, and its point radius changed in 7 levels, whereas the reference scatterplot had an unchanged baseline radius ($r = 2$), which was the point radius for all scatterplots in E1. Two scale ratios (63% and 252%) were selected to test the enlarging and shrinking cases. This hypothesis is partially confirmed.

For the bias at the baseline radius ($r = 2$), we find a certain degree of bias in each of Plots 7−12 in Figure 6. This result is reasonable because the test scatterplot is consistently scaled in E3. The results of



(a)

(b)

Figure 7. Subjective questionnaire results. (a) Results of DL questions. A violin plot, which consists of a box plot and an area plot, presents the results of all participant ratings on the DL questions for an experiment and feature. The y-axis indicates the judgment difficulty levels ranging from 1 (lowest difficulty) to 5 (highest difficulty). A box plot presents the median, upper quartile, lower quartile, 1.5 IQR of the lower quartile, and 1.5 IQR of the upper quartile of ratings. An area plot presents the kernel density distribution of ratings on the five difficulty levels. (b) Results of DT questions. A bar chart presents the results of all participants who select the tendency options of the DT question for an experiment and feature. The x-axis encodes the five options (A. increasing, B. decreasing, C. increasing after decreasing first, D. decreasing after increasing first, and E. not sure) and the y-axis encodes the counts of the participants selecting a certain option.

significance analysis show that significant differences exist in the biases of 7 point radii for each feature (for scale ratio = 63%, numerosity: χ^2 = 19.330, $p = 0.004 < 0.05$, correlation: χ^2 = 89.688, $p = 0.000 < 0.05$, and cluster separation: χ^2 = 30.021, $p = 0.000 < 0.05$; for scale ratio = 252%, numerosity: χ^2 = 37.104, $p = 0.000 < 0.05$, correlation: χ^2 = 53.314, $p = 0.000 < 0.05$, and cluster separation: χ^2 = 17.593, $p = 0.007 < 0.05$).

We then analyze whether changing the point radius can reduce the bias for the three features.

Numerosity: For scale ratio = 63%, when the point radius decreases from 2 to 1.39 step by step, the bias gradually approaches 0 from -6.23 to -0.0712, as exhibited in Plot 7, thereby indicating that the negative bias can be corrected by reducing the point radius when the test scatterplot is shrunk. This correction trend is consistent with the results of E1 that the perceived numerosity is relatively large when the test scatterplot is shrunk, and the radius must be reduced to correct the bias. Notably, the bias largely deviates from 0 suddenly when the point radius becomes 1.16, indicating that such a correction trend only exists in a certain range of point radii. For scale ratio = 252%, when the point radius increases from 2 to 3.46 step by step, the bias gradually reduces from 16.02 to 3.204, as demonstrated in Plot 10, denoting that the positive bias can be corrected by increasing the point radius when the test scatterplot is enlarged. This correction trend is consistent with the results of E1 that the perceived numerosity is relatively small in the enlarged test scatterplot, and increasing the point radius can reduce the bias. In addition, the bias is not very close to 0 even at $r = 3.46$ likely because we do not sufficiently test the levels of point radius. In summary, changing the point radius within a certain range can slightly correct the perceptual bias of numerosity.

Correlation: For scale ratio = 63%, when the point radius increases from 2 to 2.4, the bias reduces from 0.021 to 0.0105, as illustrated in Plot 8. For scale ratio = 252%, when the point radius decreases from 2 to 1.67, the bias changes from -0.00735 to 0.0047, which approaches 0, as demonstrated in Plot 11. These results indicate that changing the point radius can correct the perceptual bias of correlation. However, the correction trend of correlation only occurs within two adjacent levels of point radius, whereas such a trend of numerosity occurs within three adjacent levels. The reasons are that the biases of correlation at the baseline radius are highly close to 0, and the correlation perception may be insensitive to the change in point radius.

Cluster separation: For scale ratio = 63%, the bias reduces from 0.0186 to 0.0129 when the point radius decreases from 2 to 1.67, as shown in Plot 9, but the bias suddenly increases to 0.028 when the point radius continues to decrease to 1.16. For scale ratio = 252%, the bias changes from -0.00418 to 0.0019 when the point radius increases from 2 to 2.4, as presented in Plot 12. However, as the point radius continues to increase, the bias also increases. These results indicate that the bias of cluster separation can be corrected by changing the point radius within a certain range. Similar to correlation, the correction effect of cluster separation only occurs within two adjacent feature levels.

5.3 Subjective Result Analysis

In this section, we analyze the results of the subjective questionnaires.

Experiment 1. For the DL questions, Level 3 (moderate difficulty) obtains the most ratings for all features. A comparison of the ratings among the three features shows that cluster separation obtains a relatively low overall rating. This condition reflects that geometric scaling has a stronger influence on the perception of numerosity and correlation than that of cluster separation. Many participants commented that, *"In the trials of cluster separation, two clusters are marked with different colors, making the judgments easy."* For the DT questions, Option C gains the most selections for numerosity and correlation; this finding is consistent with the bias trend analysis of H1 (Section 5.2). In terms of cluster separation, Options B and C obtain the same amount of selections. Both options are reasonable. In Plot 3 of Figure 6, the bias tends to decrease first and then increases with the scale ratio. However, the rate of increase decreases after the scale ratio reaches 100%.

Experiment 2. The subjective results of E2 are similar to those of E1. With distributions, the overall judgment difficulty level is slightly lower in E2 than in E1. This result can be observed by comparing the box plots of the two experiments by feature in Figure 7(a). A participant reported, *"The boundary of points of uniform distribution is clearer than that of normal distribution, which is beneficial to the judgments."*

Experiment 3. For the DL questions, more participants rated Levels 4 and 5 for all features compared with E1 and E2. We speculate that the participants had difficulty focusing on various point radii and scaled scatterplots simultaneously in E3. This speculation can be verified by the results of the DT questions. The number of participants who selected Option E for all features was clearly more than that of E1 and E2. According to a participant, *"I had difficulty knowing exactly whether the point radius becomes larger or smaller as the scatterplot itself has been scaled."* Moreover, the overall difficulty level is higher in the 63% scale ratio than in the 252% one, as depicted in Figure 7(a). Many participants reported the ease of making judgments in a large-scale ratio. As a participant said, *"The details are clear enough in an enlarged scatterplot, and I am more confident about my judgments."*

5.4 Summary

The objective results reflect three main findings. First, geometric scaling causes a bias on the perception of the three visual features (numerosity, correlation, and cluster separation). The bias has a linear relationship with the scale ratio; that is, the absolute value of bias is enlarged when the scale ratio increases or decreases. Second, no significant difference occurs between the biases measured from normally and uniformly distributed scatterplots. Third, changing the point radius can correct the bias. Such a correction only appears in a certain radius range, and the correction effect is large in the cases of numerosity and at a large scale ratio.

Our findings from the results of subjective questionnaires are also threefold. First, the participants were aware of the effect of geometric scaling on the perceived three features, of which numerosity was most affected. The participants could roughly realize the relationship between the bias and scale ratio. Second, several participants reported that the biases were slightly less affected by geometric scaling in the uniformly distributed scatterplots than in the normally distributed ones. Third, the participants were ambiguous about the bias correction effect of changing the point radius, but some participants were able to easily judge the enlarged scatterplots.

6 Discussion

In this section, we discuss the lessons learned from this study. We also summarize the limitations and suggest directions for further work.

Hypothesis formulation. We studied the low- and mid-level perceptions of scatterplot (H1), except high-level. We believe that the bias is ubiquitous during scatterplot scaling but can be presented in various forms. We preliminarily investigated the effect of normal and uniform distributions (H2). Diverse data distributions are worth exploring in the future. We only investigated the effect of point radii on the bias (H3). Other visual channels require further exploration, such as color, contrast [69], opacity [44], and luminance [53].

Experimental design. The most important lesson learned from the experiment design was the importance of pilot studies. The 2AFC method, the random sequence of trials, and the experiment order from E1 to E3, were verified to be unreasonable in our pilot studies. We found that the performance of the participants would decrease over time if the experiments were conducted continuously. To minimize the effect of fatigue on the experimental results, we set up a variety of rigorous rest mechanisms. After completing all trials in one experiment, the participants were asked to rest for 2−3 min before proceeding to the next experiment. After a round of experiments, the participants were given an hour of rest, and a pre-experiment was conducted before the next round to test the participants' mental state. The participants were allowed to complete all experiments within a few days if they reported a strong feeling of fatigue or encountered unexpected situations during the experiments. Moreover, all scatterplots in the experiments were simplified to be shown on a unified desktop monitor. This practice benefited controllable experiments but reduced the ecological validity. We can further explore how different display devices and screen resolutions influence the bias caused by geometric scaling.

Experimental result analysis. The biases measured from a small number of abnormal sequences were replaced with the mean of biases of all participants with the same experimental variables. We carefully examined these sequences to ensure all of them have obviously abnormal choice patterns, but such replacements may still have a potential effect on the experimental results. In terms of E1, we conducted a linear regression and trend analysis on the mean of biases of all participants in each scale ratio because no significant differences existed among the biases of participants. Rather than investigating the biases of all individuals, we focused on the mean performance of the population, which was similar to the study of Harrison et al. [31]. For E3, we had not yet obtained an accurate model to guide the bias correction. The main reason may be that the tested levels of point radii and scale ratios remained insufficient. Moreover, error bars in Figure 6 are large because we scaled the range of y-axis to make the trends easy to read.

We recorded the time spent on each trial, but we did not perform any analysis on the time because the participants were allowed to repeatedly modify their answers until they were satisfied. No time limit was set on a trial. Moreover, we found in the significance analysis that most experimental results did not conform to a normal distribution. This may be caused by the small number of participants. In addition, Figure 6 shows that the effect sizes of biases for numerosity seem substantial, and those for correlation and cluster separation seem small. However, effect sizes of biases for the three visual features could not be compared directly because of different perceptions and metrics.

Potential future directions. Further study on the breadth and depth can be conducted. For the former, we plan to investigate whether additional visual features can be affected by geometric scaling. We can also examine the influence of other distributions and visual channels on the bias. For the latter, we plan to explore the inner relations between the perceptions of high-level and low-level tasks. An example [24] shows why the cluster separation is perceptually biased from the visual density. We can also explore the relationships between features and biases. Experimental conditions, such as additional scale ratios and visual feature levels, should be enriched. An extensive range of participants and realistic multi-device environments should be involved. Our ultimate goal is to find a quantitative model to describe the bias, which can be used to formulate recommendations for automatic scaling.

7 Conclusion

In this study, we investigated the bias on the perceived visual features of scatterplots caused by geometric scaling. We proposed three hypotheses based on practical experience, and conducted a series of controlled experiments to test against the hypotheses. By analyzing the experimental results, we found that such a bias occurred and was linearly related to the scale ratio, thereby affecting the perception inconsistency of the scatterplots. We also found that no significant difference existed between the biases measured from normally and uniformly distributed scatterplots, and the bias could be partially corrected by changing the point radius. Our work is the first exploration of the bias caused by scatterplot scaling. We hope that this work will inspire other researchers to further study the perceptual process of scatterplot scaling, which should be a meaningful direction for enhancing the scatterplot design and promoting collaborative data exploration. We also expect that similar phenomena can be studied on other visualization technologies, such as bar and line charts, which should result in many interesting findings.

Acknowledgments

We wish to thank all the anonymous reviewers for their valuable comments, and all the participants for their active participation. The work is supported by the National Key Research and Development Program of China (No. 2018YFB1700403) and the National Natural Science Foundation of China (No. 61772456, 61761136020, U1609217, 61672538 and 61872388).

REFERENCES

[1] E. Alexander, C.-C. Chang, M. Shimabukuro, S. Franconeri, C. Collins, and M. Gleicher. Perceptual biases in font size as a data encoding. *IEEE Transactions on Visualization and Computer Graphics*, 24(8):2397–2410, 2017.

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

[3] C. Andrews, A. Endert, and C. North. Space to Think: Large High-Resolution Displays for Sensemaking. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 55–64. ACM, 2010.

[4] C. Andrews, A. Endert, B. Yost, and C. North. Information visualization on large, high-resolution displays: Issues, challenges, and opportunities. *Information Visualization*, 10(4):341–355, 2011.

[5] G. Anobile, G. M. Cicchini, and D. C. Burr. Separate Mechanisms for Perception of Numerosity and Density. *Psychological Science*, 25(1):265–270, 2014.

[6] G. Anobile, M. Turi, G. M. Cicchini, and D. C. Burr. Mechanisms for perception of numerosity or texture-density are governed by crowding-like effects. *Journal of Vision*, 15(5):4–4, 2015.

[7] C. Ardito, P. Buono, M. F. Costabile, and G. Desolda. Interaction with Large Displays: A Survey. *ACM Computing Surveys (CSUR)*, 47(3):46, 2015.

[8] M. Behrisch, M. Blumenschein, N. W. Kim, L. Shao, M. El-Assady, J. Fuchs, D. Seebacher, A. Diehl, U. Brandes, H. Pfister, et al. Quality Metrics for Information Visualization. *Computer Graphics Forum*, 37(3):625–662, 2018.

[9] E. Bertini, A. Tatu, and D. Keim. Quality Metrics in High-Dimensional Data Visualization: An Overview and Systematization. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2203–2212, 2011.

[10] T. Blascheck, L. Besançon, A. Bezerianos, B. Lee, and P. Isenberg. Glanceable Visualization: Studies of Data Comparison Performance on Smartwatches. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):630–640, 2019.

[11] M. A. Borkin, Z. Bylinskii, N. W. Kim, C. M. Bainbridge, C. S. Yeh, D. Borkin, H. Pfister, and A. Oliva. Beyond Memorability: Visualization Recognition and Recall. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):519–528, 2016.

[12] M. Brehmer and T. Munzner. A Multi-Level Typology of Abstract Visualization Tasks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2376–2385, 2013.

[13] S. Butscher, S. Hubenschmid, J. Müller, J. Fuchs, and H. Reiterer. Clusters, Trends, and Outliers: How Immersive Technologies Can Facilitate the Collaborative Analysis of Multidimensional Data. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 90–100. ACM, 2018.

[14] Y.-H. Chan, C. D. Correa, and K.-L. Ma. Flow-based Scatterplots for Sensitivity Analysis. In *Proceedings of the 2010 IEEE Symposium on Visual analytics science and technology (VAST)*, pp. 43–50. IEEE, 2010.

[15] H. Chen, W. Chen, H. Mei, Z. Liu, K. Zhou, W. Chen, W. Gu, and K.-L. Ma. Visual abstraction and exploration of multi-class scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1683–1692, 2014.

[16] L. Chittaro. Visualizing Information on Mobile Devices. *Computer*, 39(3):40–45, 2006.

[17] H. Chung, S. L. Chu, and C. North. A comparison of Two Display Models for Collaborative Sensemaking. In *Proceedings of the 2nd ACM International Symposium on Pervasive Displays*, pp. 37–42. ACM, 2013.

[18] W. S. Cleveland and W. S. Cleveland. *The Elements of Graphing Data*, vol. 2. Wadsworth Advanced Books and Software Monterey, CA, 1985.

[19] W. S. Cleveland, P. Diaconis, and R. McGill. Variables on Scatterplots Look More Highly Correlated When the Scales are Increased. *Science*, 216(4550):1138–1141, 1982.

[20] W. S. Cleveland, M. E. McGill, and R. McGill. The Shape Parameter of a Two-variable Graph. *Journal of the American Statistical Association*, 83(402):289–300, 1988.

[21] C. Conati and H. Maclaren. Exploring the Role of Individual Differences in Information Visualization. In *Proceedings of the working conference on Advanced visual interfaces*, pp. 199–206. ACM, 2008.

[22] S. C. Dakin, M. S. Tibber, J. A. Greenwood, M. J. Morgan, et al. A common visual metric for approximate number and density. *Proceedings of the National Academy of Sciences*, 108(49):19552–19557, 2011.

[23] S. M. Drucker, D. Fisher, R. Sadana, J. Herron, et al. TouchViz: A Case Study Comparing Two Interfaces for Data Analytics on Tablets. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2301–2310. ACM, 2013.

[24] G. Ellis. *Cognitive Biases in Visualizations*. Springer, 2018.

[25] M. Friendly and D. Denis. The Early Origins and Development of The Scatterplot. *Journal of the History of the Behavioral Sciences*, 41(2):103–130, 2005.

[26] M. A. Garcıa-Pérez. Forced-choice staircases with fixed step sizes: asymptotic and small-sample properties. *Vision Research*, 38(12):1861–1881, 1998.

[27] M. Gleicher, M. Correll, C. Nothelfer, and S. Franconeri. Perception of Average Value in Multiclass Scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2316–2325, 2013.

[28] C. Gutwin, A. Cockburn, and A. Coveney. Peripheral Popout: The Influence of Visual Angle and Stimulus Intensity on Popout Effects. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 208–219. ACM, 2017.

[29] C. Gutwin and C. Fedak. Interacting with Big Interfaces on Small Screens: a Comparison of Fisheye, Zoom, and Panning Techniques. In *Proceedings of Graphics Interface 2004*, pp. 145–152. Canadian Human-Computer Communications Society, 2004.

[30] L. Harrison, K. Reinecke, and R. Chang. Infographic aesthetics: Designing for the first impression. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 1187–1190. ACM, 2015.

[31] L. Harrison, F. Yang, S. Franconeri, and R. Chang. Ranking Visualizations of Correlation Using Weber's Law. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1943–1952, 2014.

[32] C. Healey and J. Enns. Attention and Visual Memory in Visualization and Computer Graphics. *IEEE Transactions on Visualization and Computer Graphics*, 18(7):1170–1188, 2012.

[33] J. Heer and M. Agrawala. Multi-Scale Banking to 45 Degrees. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):701–708, 2006.

[34] J. Jerit and J. Barabas. Partisan Perceptual Bias and the Information Environment. *The Journal of Politics*, 74(3):672–684, 2012.

[35] E. Kandogan. Just-in-time Annotation of Clusters, Outliers, and Trends in Point-based Data Visualizations. In *Proceedings of the 2012 IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pp. 73–82. IEEE, 2012.

[36] M. Kay and J. Heer. Beyond Weber's Law: A Second Look at Ranking Visualizations of Correlation. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):469–478, 2016.

[37] S. Kim, M. Al-Haj, S. Chen, S. Fuller, U. Jain, M. Carrasco, and R. Tannock. Colour vision in ADHD: Part 1-Testing the retinal dopaminergic hypothesis. *Behavioral and Brain Functions*, 10(1):38, 2014.

[38] O.-H. Kwon, C. Muelder, K. Lee, and K.-L. Ma. A Study of Layout, Rendering, and Interaction Methods for Immersive Graph Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 22(7):1802–1815, 2016.

[39] J. Li, J.-B. Martens, and J. J. Van Wijk. Judging correlation from scatterplots and parallel coordinate plots. *Information Visualization*, 9(1):13–30, 2010.

[40] J. Li, J. J. van Wijk, and J.-B. Martens. A Model of Symbol Lightness Discrimination in Sparse Scatterplots. In *Proceedings of the 2010 IEEE Symposium on Pacific Visualization (PacificVis)*, pp. 105–112. IEEE, 2010.

[41] H. Liao, Y. Wu, L. Chen, and W. Chen. Cluster-based visual abstraction for multivariate scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, 24(9):2531–2545, 2017.

[42] P. Lv, Z. Zhang, C. Li, Y. Guo, B. Zhou, and M. Xu. Crowd behavior evolution with emotional contagion in political rallies. *IEEE Transactions on Computational Social Systems*, 6(2):377–386, 2018.

[43] Y. Ma, A. K. Tung, W. Wang, X. Gao, Z. Pan, and W. Chen. Scatternet: A deep subjective similarity model for visual analysis of scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2018.

[44] J. Matejka, F. Anderson, and G. Fitzmaurice. Dynamic Opacity Optimization for Scatter Plots. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 2707–2710. ACM, 2015.

[45] H. Mei, Y. Ma, Y. Wei, and W. Chen. The design space of construction tools for information visualization: A survey. *Journal of Visual Languages & Computing*, 44:120–132, 2018.

[46] T. Ni, G. S. Schmidt, O. G. Staadt, M. A. Livingston, R. Ball, and R. May.

A Survey of Large High-Resolution Display Technologies, Techniques, and Applications. In *Proceedings of the 2006 IEEE Conference on Virtual Reality*, pp. 223–236. IEEE, 2006.

[47] B. Ondov, N. Jardine, N. Elmqvist, and S. Franconeri. Face to Face: Evaluating Visual Comparison. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):861–871, 2019.

[48] W. Peng, M. O. Ward, and E. A. Rundensteiner. Clutter Reduction in Multi-Dimensional Data Visualization Using Dimension Reordering. In *Proceedings of the 2004 IEEE Symposium on Information Visualization*, pp. 89–96. IEEE, 2004.

[49] A. M. Reach and C. North. Smooth, Efficient, and Interruptible Zooming and Panning. *IEEE Transactions on Visualization and Computer Graphics*, 25(2):1421–1434, 2019.

[50] K. Reda, A. E. Johnson, M. E. Papka, and J. Leigh. Effects of Display Size and Resolution on User Behavior and Insight Acquisition in Visual Exploration. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 2759–2768. ACM, 2015.

[51] R. A. Rensink. The nature of correlation perception in scatterplots. *Psychonomic Bulletin & Review*, 24(3):776–797, 2017.

[52] R. A. Rensink and G. Baldridge. The Perception of Correlation in Scatterplots. *Computer Graphics Forum*, 29(3):1203–1210, 2010.

[53] J. Ross and D. C. Burr. Vision senses number directly. *Journal of vision*, 10(2):10–10, 2010.

[54] B. Saket, A. Endert, and J. Stasko. Beyond Usability and Performance: A Review of User Experience-focused Evaluations in Visualization. In *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization*, pp. 133–142. ACM, 2016.

[55] B. Saket, A. Srinivasan, E. D. Ragan, and A. Endert. Evaluating Interactive Graphical Encodings for Data Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 24(3):1316–1330, 2018.

[56] A. Sarikaya and M. Gleicher. Scatterplots: Tasks, Data, and Designs. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):402–412, 2018.

[57] H.-J. Schulz, T. Nocke, M. Heitzler, and H. Schumann. A Design Space of Visualization Tasks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2366–2375, 2013.

[58] M. Sedlmair and M. Aupetit. Data-driven Evaluation of Visual Quality Measures. *Computer Graphics Forum*, 34(3):201–210, 2015.

[59] M. Sedlmair, T. Munzner, and M. Tory. Empirical Guidance on Scatterplot and Dimension Reduction Technique Choices. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2634–2643, 2013.

[60] M. Sedlmair, A. Tatu, T. Munzner, and M. Tory. A Taxonomy of Visual Cluster Separation Factors. *Computer Graphics Forum*, 31(3):1335–1344, 2012.

[61] Y. Shi, Y. Zhao, F. Zhou, R. Shi, and Y. Zhang. A novel radial visualization of intrusion detection alerts. *IEEE Computer Graphics and Applications*, 38(6):83–95, 2018.

[62] M. Sips, B. Neubert, J. P. Lewis, and P. Hanrahan. Selecting good views of high-dimensional data using class consistency. *Computer Graphics Forum*, 28(3):831–838, 2009.

[63] C. Sophian and Y. Chu. How do people apprehend large numerosities? *Cognition*, 107(2):460–478, 2008.

[64] B. Steichen, G. Carenini, and C. Conati. User-adaptive information visualization: using eye gaze data to infer visualization tasks and user cognitive abilities. In *Proceedings of the 2013 international conference on Intelligent user interfaces*, pp. 317–328. ACM, 2013.

[65] S. S. Stevens and G. Stevens. Psychophysics: Introduction to its perceptual, neural, and social prospects. New Brunswick, 1986.

[66] D. A. Szafir. Modeling Color Difference for Visualization Design. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):392–401, 2018.

[67] A. Tatu, G. Albuquerque, M. Eisemann, J. Schneidewind, H. Theisel, M. Magnork, and D. Keim. Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. In *Proceedings of the 2009 IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pp. 59–66. IEEE, 2009.

[68] A. Tatu, P. Bak, E. Bertini, D. Keim, and J. Schneidewind. Visual quality metrics and human perception: an initial study on 2D projections of large multidimensional data. In *Proceedings of the International Conference on Advanced Visual Interfaces*, pp. 49–56. ACM, 2010.

[69] M. S. Tibber, J. A. Greenwood, and S. C. Dakin. Number and density discrimination rely on a common metric: Similar psychophysical effects of size, contrast, and divided attention. *Journal of Vision*, 12(6):1–19, 2012.

[70] M. S. Tibber, G. S. Manasseh, R. C. Clarke, G. Gagin, S. N. Swanbeck, B. Butterworth, R. B. Lotto, and S. C. Dakin. Sensitivity to numerosity is not a unique visuospatial psychophysical predictor of mathematical ability. *Vision Research*, 89:1–9, 2013.

[71] M. Tokita and A. Ishiguchi. How might the discrepancy in the effects of perceptual variables on numerosity judgment be reconciled? *Attention, Perception, & Psychophysics*, 72(7):1839–1853, 2010.

[72] M. Tory, D. Sprague, F. Wu, W. Y. So, and T. Munzner. Spatialization Design: Comparing Points and Landscapes. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1262–1269, 2007.

[73] A. C. Valdez, M. Ziefle, and M. Sedlmair. Priming and Anchoring Effects in Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):584–594, 2018.

[74] H. Wang, M. Xu, F. Zhu, Z. Deng, Y. Li, and B. Zhou. Shadow traffic: A unified model for abnormal traffic behavior simulation. *Computers & Graphics*, 70:235–241, 2018.

[75] Y. Wang, X. Chen, T. Ge, C. Bao, M. Sedlmair, C.-W. Fu, O. Deussen, and B. Chen. Optimizing color assignment for perception of class separability in multiclass scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):820–829, 2019.

[76] Y. Wang, Z. Wang, C. W. Fu, H. Schmauder, O. Deussen, and D. Weiskopf. Image-based aspect ratio selection. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):840–849, 2019.

[77] C. Ware. *Information Visualization: Perception for Design*. Elsevier, 2012.

[78] J. Xia, L. Gao, K. Kong, Y. Zhao, Y. Chen, X. Kui, and Y. Liang. Exploring linear projections for revealing clusters, outliers, and trends in subsets of multi-dimensional datasets. *Journal of Visual Languages & Computing*, 48:52–60, 2018.

[79] J. Xia, F. Ye, W. Chen, Y. Wang, W. Chen, Y. Ma, and A. K. Tung. Ldsscanner: exploratory analysis of low-dimensional structures in high-dimensional datasets. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):236–245, 2017.

[80] M. Xu, C. Li, P. Lv, N. Lin, R. Hou, and B. Zhou. An efficient method of crowd aggregation computation in public areas. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):2814–2825, 2017.

[81] F. Yang, L. T. Harrison, R. A. Rensink, S. L. Franconeri, and R. Chang. Correlation Judgment and Visualization Features: A Comparative Study. *IEEE Transactions on Visualization and Computer Graphics*, 25(3):1474–1488, 2019.

[82] B. Yost and C. North. The Perceptual Scalability of Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):837–844, 2006.

[83] Y. Zhao, F. Luo, M. Chen, Y. Wang, J. Xia, F. Zhou, Y. Wang, Y. Chen, and W. Chen. Evaluating Multi-Dimensional Visualizations for Understanding Fuzzy Clusters. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):12–21, 2019.

[84] F. Zhou, X. Lin, C. Liu, Y. Zhao, P. Xu, L. Ren, T. Xue, and L. Ren. A survey of visualization for smart manufacturing. *Journal of Visualization*, 22(2):419–435, 2019.

[85] Z. Zhou, L. Meng, C. Tang, Y. Zhao, Z. Guo, M. Hu, and W. Chen. Visual abstraction of large scale geospatial origin-destination movement data. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):43–53, 2018.

[86] E. Zimmermann and G. R. Fink. Numerosity perception after size adaptation. *Scientific Reports*, 6(1):32810–32817, 2016.