Correlation Judgment and Visualization Features: A Comparative Study

Fumeng Yang, Lane T. Harrison, Ronald A. Rensink, Steven L. Franconeri, and Remco Chang

Abstract—Recent visualization research efforts have incorporated experimental techniques and perceptual models from the vision science community. Perceptual laws such as Weber's law, for example, have been used to model the perception of correlation in scatterplots. While this thread of research has progressively refined the modeling of the perception of correlation in scatterplots, it remains unclear as to *why* such perception can be modeled using relatively simple functions, e.g., linear and log-linear. In this paper, we investigate a longstanding hypothesis that people use *visual features* in a chart as a proxy for statistical measures like correlation. For a given scatterplot, we extract 49 candidate visual features and evaluate which best align with existing models and participant judgments. The results support the hypothesis that people attend to a small number of visual features when discriminating correlation in scatterplots. We discuss how this result may account for prior conflicting findings, and how visual features provide a baseline for future model-based approaches in visualization evaluation and design.

Index Terms—Information visualization, Perception and psychophysics, Evaluation/methodology, Weber's law, Power law.

1 Introduction

In a recent study, Rensink and Baldridge demonstrated that the perception of correlation in scatterplots can be mathematically modeled using Weber's law [1]. In followup experiments, Rensink showed that this law is robust to changes in data characteristics and scatterplot design choices [2]. Based on these findings, Harrison et al. replicated the original study by Rensink and Baldridge, moving beyond scatterplots to measure and compare the effectiveness of a range of visualizations [3]. Their results indicate that the perception of correlation in all of these bivariate visualizations can be modeled using Weber's law. Together, these studies sparked a renewed interest in the information visualization community towards better understanding the underlying mechanics of visualization and modeling approaches, such as Kay and Heer's followup analysis of Harrison et al.'s released experimental data [4].

Beyond the information visualization community, researchers in perceptual psychology have also studied scatterplots at length, in particular attempting to develop models that capture how people estimate correlation from them. For example, Boynton studied the perceptual dimensions of covariation estimate, producing a model that used elongation ratio and standard error as factors [5]. Meyer et al. fit the perception of correlation in scatterplots to a power function [6]. Others studies include Pollack [7], Jennings et al. [8], and Cleveland et al. [9], all of whom attempted to formally model the relationship between perceived and objective correlation in scatterplots.

A recurring hypothesis in these studies is that peoples' perception of correlation in scatterplots is related to *visual features* in the visualization. The intuition is that participants are not directly perceiving correlation per se, but rather, visual features produced *by* the visualization technique (i.e., scatterplot) that are related to correlation. Meyer and Shinar, for example, suggested that estimates of correlations from scatterplots are partly based on perceptual processes influenced by "visual properties" of the charts and are unrelated to participants' formal statistical training [10]. Lauer and Post included some of these factors in their models, such as regression slope and point dispersion, along with factors such as screen size [11]. More recent work from Rensink showed that correlation judgments can be made within just a few hundred milliseconds, suggesting a "heuristic" approach to discriminating correlation [2]. These studies all suggest that visual features of some sort may underlie human's perception of correlation.

The goal of this paper is to bridge these two sides of research, bringing findings from perceptual psychology to large-scale approaches for modeling perception in the information visualization community. Such an approach could help explain the extent to which visualizations such as scatterplots are effective for judging correlations and provide explanations for when they might become ineffective.

The core concept in this paper is the use of *visual features* in modeling of the perceptual process. As used here, the term *feature* refers to a visual feature refers to the perceivable and distinguishable properties (e.g., shape, dispersion, and orientation) in a 2D image or a part of an image. Outfitted with this concept, our paper takes a computational approach towards evaluating how visual features manifest in models of the perception of correlation in scatterplots, including the approaches proposed by Rensink and Baldridge, Harrison et al., and Kay and Heer.

- *Fumeng Yang is with the Department of Computer Science, Brown University, Providence, RI, 02906. E-mail: fy@brown.edu.*
- *Lane T. Harrison is with the Department of Computer Science, Worcester Polytechnic Institute, Worcester, MA, 01609. E-mail: lane@cs.wpi.edu.*
- *Ronald A. Rensink is with the Departments of Computer Science and Psychology, University of British Columbia, Vancouver, BC V6T 1Z4, Canada. E-mail: rensink@cs.ubc.ca.*
- *Steven L. Franconeri is with the Psychology Department, Northwestern University, Evanston, IL, 60208. E-mail: franconeri@northwestern.edu.*
- *Remco Chang is with the Department of Computer Science, Tufts University, Medford, MA, 02155. E-mail: remco@cs.tufts.edu.*

Manuscript received MM DD, YYYY; revised MM DD, YYYY.

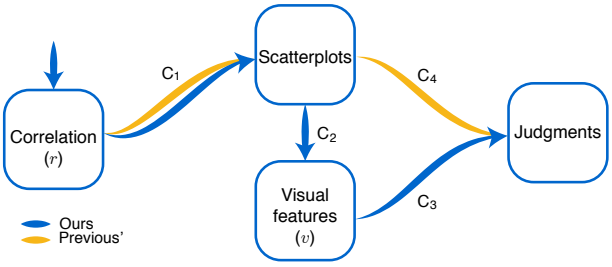


Fig. 1. An overview of this paper: we propose to investigate the visual features in the scatterplots to study the perception of correlation (the blue path, $C_1 \Rightarrow C_2 \Rightarrow C_3$), as opposed to a direct study approach using correlation to model the perception by Rensink and Baldrige, Harrison et al., and Kay and Heer (the yellow path, $C_1 \Rightarrow C_4$).

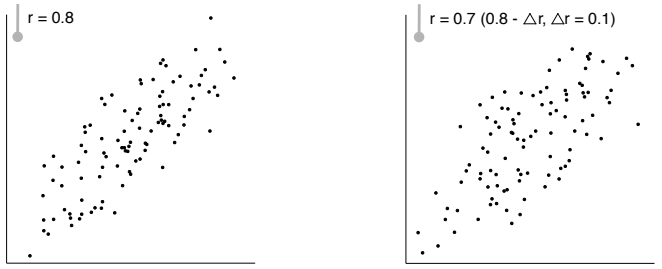


Fig. 2. An example of approaching a target correlation level from *below* in the experiments. Two side-by-side scatterplots (without any indication of actual correlation value or the regression line) are shown to the participant in the experiment. The participant chooses which of the two appears to be more correlated.

Figure 1 illustrates an overview of our proposed research. Here, the yellow path shows the general research methodology proposed by Rensink and Baldrige and adopted by Harrison et al. and Kay and Heer. In this approach, a dataset with a known correlation value (r) is mapped to a scatterplot (C_1); participants are asked to compare the correlation values between two scatterplots in a judgment (C_4). From participants' judgments, the perceptual model of correlation is built. In contrast, our approach moves further to examining whether visual features in scatterplots tackle the participants' judgments ($C_1 \Rightarrow C_2 \Rightarrow C_3$, the blue path in Figure 1).

Toward this goal, we begin by replicating the methodology and experiments from Rensink and Baldridge [1], [2] and Harrison et al. [3] (Section 2) to collect a set of judgment data. To create a set of visual features, we broadly examine the perceptual psychology, visualization, and computational geometry literature to collect a set of candidate features that can be computed from scatterplots. In total, we identify and extract 49 candidates from scatterplots (Section 3).

Out of the 49 initial candidates, our analysis shows that the participants' judgments highly correlate with four, such as the dispersion of the point cloud around the regression line (Section 4). We evaluate their performance against several model metrics. We find that models using these top-performing features are at least as precise as existing models (Section 5). Building on top of these analyses, we examine power transformation, a fundamental part of modeling in perceptual psychology, to create a new model of the perception of correlation in scatterplots. The resulting model outperforms the original models in precision, and is also more easily understandable, as it directly relates to visual features commonly inferred from scatterplots (Section 6).

As such, this paper contributes a new perspective on modeling the perception of correlation in scatterplots. Our findings indicate that the use of visual features can lead to more precise mathematical models of behavior, while suggesting plausible theories about how people perceive scatterplots and extract information from visualization. More specifically, our work contributes to the field of visualization in three ways:

- We evaluate the longstanding hypothesis that participants use visual features instead of correlation itself when judging correlation in scatterplots;
- We establish that visual features can be integrated into the approaches proposed by Rensink and Baldridge, Harrison et al., and Kay and Heer without loss of precision;
- We develop a new, more precise model based on these existing models by using power transformation, which has an additional benefit of linking models to existing work in perceptual psychology.

2 Replication: Discrimination Thresholds and Judgments for Scatterplots

In this section, we introduce our experiment with three goals in mind: 1) collect data for our modeling approach, 2) faithfully replicate the prior results[1], and 3) familiarize the readers with the terminology used in this paper.

The experiments by Rensink and Baldridge, and Harrison et al. are based on the discrimination of correlation, and have three components: 1) side-by-side comparison of two scatterplots with data of different correlation values, 2) the use of above and below *approaches* to estimate discrimination thresholds from both sides of a target correlation value, and 3) a staircase method [12] to modulate the difference in correlation values between the two scatterplots.

More specifically, this kind of experiment presents scatterplots with underlying data having regression lines along the 45° axis $y = x$ (see Figure 2). In a *judgment*, a participant must indicate which of the two side-by-side scatterplots appears to have the *higher* correlated dataset: one with a fixed correlation value (r), the other generated from a dataset with a different correlation value ($r \pm \Delta r$). The fixed correlation can be one of [0.3, 0.4, 0.5, 0.6, 0.7, 0.8], and remains so until a stable discrimination threshold is reached. The sign of Δr is positive or negative depending on the *approach*: in the *above* approach, the sign is always positive (i.e., plus); in the *below* approach, the sign is always negative (i.e., minus). The value of Δr typically changes as a *trial* progresses, as determined by the staircase method. The value of Δr decreases by 0.01 if the participant makes a correct judgment, and increases by 0.03 if the participant makes an incorrect judgment (see Figure 4a), so that steady-state behavior corresponds to 75% correct.

We make one minor modification to the experimental design used by both Rensink and Baldridge and Harrison et al. In the prior experiments, a trial terminates when Δr converges over the course of last 24 judgments via a successful F test (α=0.1), or 50 judgments have been

1. Although Harrison et al. had published their experimental data, and the same data was used by Kay and Heer, a new experiment is necessary due to a change to allow our inclusion of visual features.

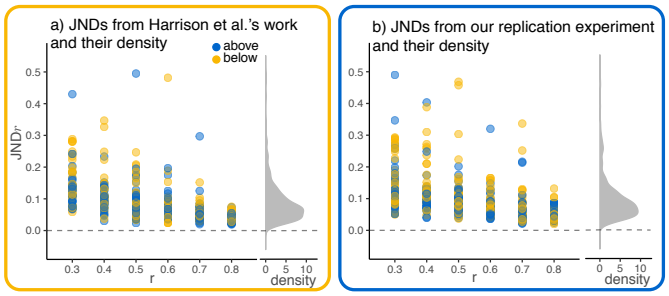


Fig. 3. The two sets of JND from Harrison et al.'s [3] and our replication experiment. The distributions of two datasets are similar, indicating that they are similar and comparable.

made (see Figure 4a). To avoid premature convergence and allow the computation for a set of visual features, in our experiment, a trial is always comprised of 50 judgments. The inclusion of 50 judgments makes it possible to compare visual features with correlation and also makes it necessary to first compare the result of our replication experiment to the original experiment.

For this modified experiment, we recruited 95 participants (33 female) via Amazon Mechanical Turk, with participants receiving $2.20 for their time (commensurate with the U.S. minimum wage). The experiment collected participants' judgments with the two datasets, participants' answers, the correct answers, and the experiment conditions (e.g., approach). In total, 19,000 judgments were collected.

To validate the results from this experiment, we compute *Just-noticeable Difference (JND)* of correlation from the data and compared it with the dataset published by Harrison et al. [3]. JND (see Section 5) is the measurement of discrimination used in Rensink's, Harrison et al.'s, and Kay and Heer's work. These two sets of JNDs are plotted in Figure 3, following the style of comparisons made by Kay and Heer's [4]. When compared using the Kolmogorov-Smirnov test [13], the difference is a marginally significant (D=0.094, p=.059). Along with the visual similarity between our data and the data by Harrison et al., it validates our modified experimental design and the resulting data.

3 Visual Features in Scatterplots

To examine whether visual features are used by participants in judging correlation, we first conduct a survey of visual features commonly used in visualization, perceptual psychology, statistics, and computational geometry.

This survey is a summary based on literature of candidates of visual features that might represent correlation in scatterplots. Work in perceptual psychology suggests twelve visual features, including the dispersion of points [6], [10] and the prediction ellipse of Cleveland et al. [9], [19]. The visualization literature suggests ten more, including several features related to correlation from Wilkinson's Scagnostics [16]. Statistics and data science literature suggests twenty-five visual others, such as density [17]. Computational geometry suggests the convex hull, which used to describe the general shape and size of the point clouds.

These visual features can be clustered into eight concepts across four *groups*. The first group are features that pertain to *length*, such as the length or width of the bounding box that surrounds the points. The second group is based on *area*, such as the area of a convex hull. The third group is based on *shape*, which is made of dimensionless quantities, such as the ratio of two length features. The final group includes those features that similar to *density*, such as the average distance of all points to the regression line. Table 1 shows the set of all 49 features. More precise definitions of the visual features can be found in Appendix A.

This list is intended to be broad, but not necessarily exhaustive, as further work may yield new candidate features. In the section below we describe how we evaluate each visual feature and use it in various perceptual models.

4 Identifying Visual Features

In this section, we investigate the relationship between visual features and participants' judgments of correlation using regression analysis. The earlier studies by Rensink and Baldridge, Harrison et al., and Kay and Heer found that the discrimination threshold (i.e., JND) varies with as the base correlation level. In particular, higher correlations were found to have smaller JNDs than low correlations (i.e., were more easily discriminated). The critical difference in our approach is to modulate not only correlation values, but also candidate features and use these results to determine which models and features best align with participants' judgments.

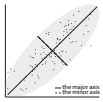
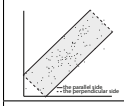
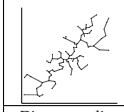
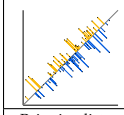
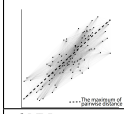
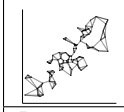
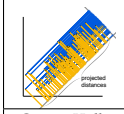
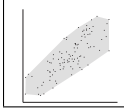
Figure 4 illustrates our study approach. The intuition is that visual features can better explain participants' judgments than correlation values. Figure 4a shows two example trials from the replication experiment. The experiment modulates the difference in correlation between the scatterplots (Δr, the y-axis) based on the correctness of participants' judgments (the x-axis). Intuitively, participants should be better at justifying a difference when Δr is larger. However, Δr may not fully predict whether a participant will judge correctly. When Δr is the same in two different judgments, the participant may make a correct judgment for one, and an incorrect judgment for another. Part of this may be due to random chances, for example, a participant may sometimes choose one or the other by mistake. Another possible cause is the existence of misleading visual features produced in the scatterplots, which make a particular scatterplot pair to be difficult to discriminate.

At a conceptual level, visual features may align more closely with participants' judgments than the actual correlation values presented in the scatterplots. Consider a hypothetical feature that perfectly predicts participants' judgments (Figure 4b). For such a feature, when its difference is above a certain threshold, the participant should always make a correct judgment and vice versa (excluding the small chance for random mistakes). Thus, visual features that highly correlate with the participants' judgments are likely candidates employed by the participants to compare correlation in scatterplots.

4.1 Pair Judgment Data

The judgment data used contains the following attributes: difference in correlation between the two scatterplots (Δr, magnitude), differences in visual features between the two scatterplots (Δv, magnitude), along with the base correla-

TABLE 1
The concepts of candidate visual features

Concept		Visual Feature	Category
Prediction Ellipse 	Cleveland et al. suggested "an elliptical contour" [9] and Jennings et al. proposed an isoprobability ellipse [8], similar to the proposal of using the width of the 2D probability density function by Rensink [14]. The core idea here is a *prediction ellipse*, which is a statistical measure that predicts the location of a new observation, under the assumption of a bivariate normal distribution [15].	The major axis of the prediction ellipse The minor axis of the prediction ellipse The area of the prediction ellipse [8], [9] The ratio of the major axis to the minor axis The ratio of the minor axis to the major axis [9]	length length area shape shape
Bounding Box 	The *bounding box* of a scatterplot is a rectangle parallel to the regression line ($y = x$), and includes all points in the plot. A *confidence box* excludes extreme outliers in the data. The bounding box of all the points within the confidence box is referred to as the *confidence bounding box*.	The side parallel to the regression line The side perpendicular to the regression line The area of the box The ratio of the perpendicular side to the parallel side The ratio of the parallel side to the perpendicular side	length length area shape shape
MST 	Wilkinson et al. used the *Minimum Spanning Tree (MST)* to identify interesting patterns in scatterplots [16]. We use features based on MST, as well as other graph-based measures of linear correlation from Wilkinson et al.'s work. The figure on the left side shows an MST of a scatterplot.	The average length of the edges on MST The standard deviation of the edges on MST The skewness of the edges on MST [16].	length density density
Distance to line 	The *perpendicular distance to the regression line* was used as a measurement of dispersion in Meyer et al. [6]. This feature is similar to Cleveland et al.'s, Jennings et al.'s, and Rensink's proposal of the width of the point cloud. We use "absolute perpendicular distance from the regression line" [6] as a visual feature. The figure on the left side shows all perpendicular distances to the regression line.	The average of the distances [6] The average of the inverted distances The standard deviation of the distances The standard deviation of the inverted distances The skewness of the distances The skewness of the inverted distances	length density density density shape shape
Pairwise distance 	*Pairwise distance* is usually used to estimate density and clustering [17]. We use all pairwise distances as the measure of global density. The figure on the left side shows all pairwise distances in a scatterplot and a maximum distance.	The maximum and percentiles of pairwise distance The average of the inverse of pairwise distance The standard deviation of pairwise distance The skewness of pairwise distance	length density density shape
kNN 	Inspired by Wilkinson et al.'s work [16], we take local density into consideration. A general measurement of this is based on *k-Nearest Neighbors* (kNN, k = 3, 5, 7, 9). The average distance to kNN is used as the measure of local density. The figure on the left side shows all points that are connected to their 3-Nearest Neighbors.	The average of all local density The standard deviation of all local density The skewness of all local density	length density shape
Projections 	The *projections* on $y = x$ and $y = -x$ directions are taken as candidate features. The lines in the figure on the left side show how to project in both directions, the dots show the projected points, and the projection is the distance to the first projected point.	The standard deviation of projections on $y = x$ The standard deviation of projections on $y = -x$	shape shape
Convex Hull 	The *convex hull* is an important geometric property in the Euclidean plane [18]. Wilkinson et al. identified convex hull as a measure of interestingness in scatterplots.	The area of convex hull	area

tion r ([0.3, 0.4, 0.5, 0.6, 0.7, 0.8]), approach ([above, below]), and judgment correctness ([correct, incorrect]).

These data are visually inspected using scatterplot matrix and correlation matrix, with two goals in mind: first, remove extreme values that could indicate possibly spurious judgments; second, resolve collinearity between different visual features that can significantly affect the outcome of a regression analysis [20].

In all, we remove 4 out of 19,000 judgments that stem from participants' erroneous input during the experiments (see Figure 5a). Collinearity is investigated by computing pairwise linear dependence between all visual features and correlation. Most features used in our data exhibit some amount of collinearity with the correlation value (r) and other visual features (see Figure 5b and c). We remove 5 linearly dependent features that can be trivially derived from one another, resulting in a final set of 44 features.

4.2 Modeling Judgments using Standardized Weighted Logistic Regression

To determine the relation between participants' judgments and the visual features, we apply a technique known as standardized weighted logistic regression for four reasons: 1) Logistic regression can model dependent variables that are dichotomous (binary), and 2) it does not assume particular distributions about the independent variables [21], [22]. In our data, judgments are the binary dependent variable, being either *correct* or *incorrect*. 3) Weighted models compensate for the imbalance between judgment counts (i.e., 75% of the judgments are *correct* and 25% *incorrect*) to avoid skewed results. 4) Standardized models transform models with different value ranges to the same so that all model coefficients are comparable.

Specifically, our logistic regression has the form:

$$g = \beta_0 + \beta_1 a_i + \beta_2 r_i + \beta_3 \Delta x_i \quad (1)$$

where g represents the logit function, r is the fixed correlation level in the experimental procedure, a is the *approach* (*above* or *below*, see Section 2), Δx represents the difference in the stimuli (i.e., Δr or Δv, the y-axis in Figure 4), β is the model coefficients, and i represents each of the 18,996 judgments. Note that the inclusion of r and a follows the work of Rensink and Baldridge [1] where the authors show that the perception of correlation is affected by the amount of correlation as well as the approach used in the study.

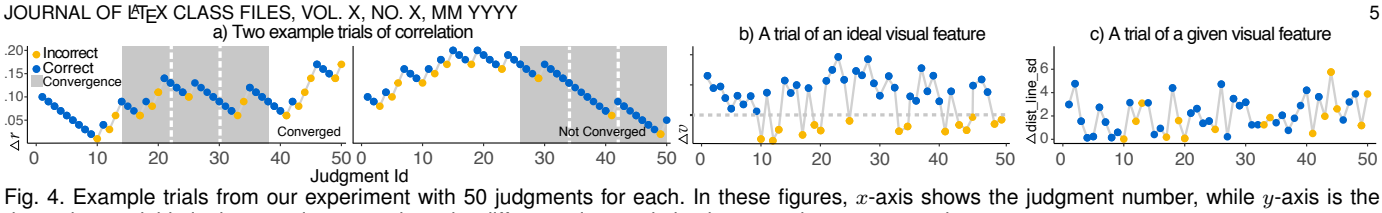


Fig. 4. Example trials from our experiment with 50 judgments for each. In these figures, x-axis shows the judgment number, while y-axis is the dependent variable in the experiment, such as the difference in correlation between the two scatterplots.

a) Two trials from our replication experiment. Note that Δr increases for an incorrect judgment and vice versa. The *left* and *right* show examples of converged and not converged trials, respectively. The difference is whether there is a window of 24 judgments that have similar differences in correlation. In the original experiment by Rensink and Baldridge and Harrison et al., the 24 judgments inside the gray rectangles were used to compute JND, and the trial terminates if it is converged. In our new experiment, the trials continue anyway until 50 judgments have been made.

b) An example scenario assuming an ideal visual feature that perfectly predicts the participant's judgments. The grey dashed line represents the JND of that visual feature. Note that all correct judgments are above the grey line and all incorrect judgments are below.

c) The changing of the difference in the visual feature *the standard deviation of all perpendicular distances to the regression line* over the course of a trial. Note that, while not perfectly following the ideal visual feature in b), this visual feature highly correlates with the participant's judgments.

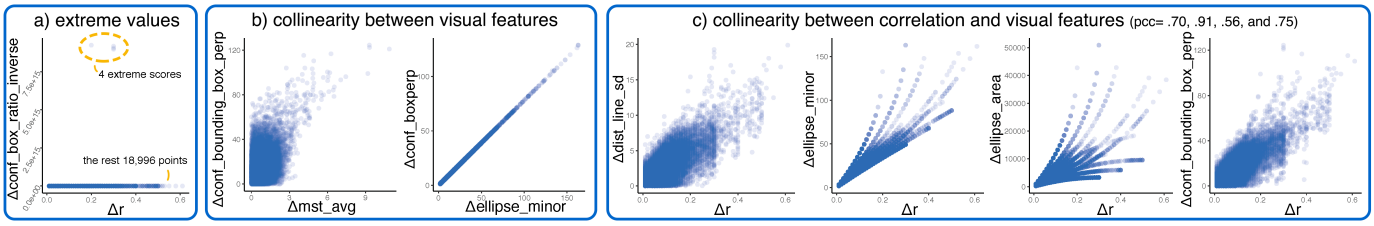


Fig. 5. Judgments data: extreme values and the collinearity between visual features. We present a few examples. The full scatterplot matrix and correlation matrix can be found in Appendix B, and notions can be found in Appendix A. We use the term Pearson correlation coefficient (pcc) instead of "correlation" to avoid confusion.

Using standardized weighted logistic regression, we first build the null model using a constant as the independent variable. A second step is to construct the model based on correlation (r). This sets a baseline for eliminating visual features that are less predictive of participants' judgments. The third step is to build a model for each visual feature and compare it against the model of correlation. The separate modeling avoids issues raised by collinearity and allows a comparison of models using multiple statistics metrics.

4.3 Model Metrics

We apply three types of metrics commonly used in evaluating logistic regression models [23]. First, using odds ratios [24], we analyze the effectiveness of each independent variable (e.g., the difference in a visual feature Δv) when explaining participants' judgments. Second, we examine the quality of the regression model using the Akaike Information Criterion (AIC). Lastly, we compare the regression model of a visual feature to the regression model of correlation using a Cox test [25]. The Cox test evaluates two non-nested models by fitting the regressors of one model into the fitted values of the other, and it is measured by explanatory values: 1) we use the Bonferroni correction [26] and set p=.0011 as the critical value; 2) we use a relaxed view where larger z-scores are expected, since the sample size of 18,996 may result in many p-values becoming significant.

Taken together, these metrics provide a means to evaluate the candidate visual features, and identify the ones that best account for the participants' judgments.

4.4 Results

Table 2 shows the results of modeling each of the 44 visual features, using the three metrics described above. The numeric results and the results of additional statistical metrics can be found in Appendix C.

4.4.1 The Null Model

The first line of Table 2 presents the null model. The model has an odds ratio of 1, indicating that the independent variable is not associated with any change in the dependent variable (i.e., judgment correctness).

4.4.2 The Baseline Model

The next three lines of Table 2 show the results of the baseline model, using only correlation (r) and its difference (Δr). It does not contain any visual feature.

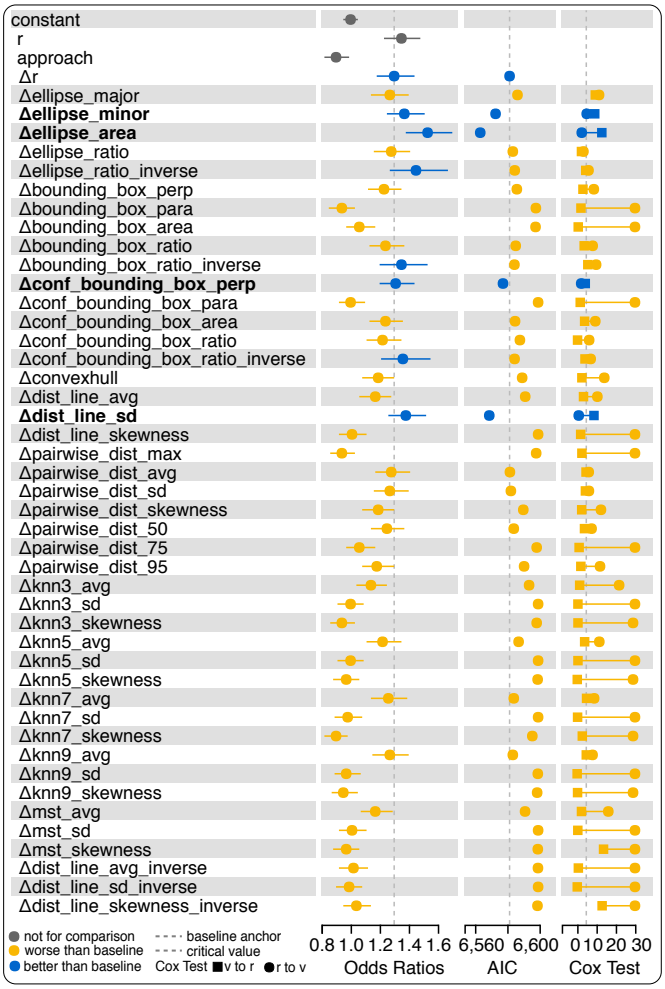
These results confirm that the difference in correlation is closely associated with the participants' judgments. The variable r has an odds ratio of 1.35 (95% CI: [1.23, 1.48]), indicating that one unit increase in r is 1.35 times (i.e., more) likely to obtain a correct judgment. Approach, denoted as a, has an odds ratio of 0.90 (95% CI: [0.82, 0.99]), indicating that the approach variable can be 0.90 times (i.e., less) likely to obtain a correct judgment. This confirms that the magnitude of correlation (r) itself impacts the judgments correctness, and the approach factor has a smaller impact as originally reported by Rensink and Baldridge [1]. More importantly, in the fourth line, the variable Δr has an odds ratio of 1.30 (95% CI: [1.18, 1.44]), providing the baseline for evaluating all our candidate visual features.

4.4.3 Visual Feature Models

We then compare the effectiveness of models built from the visual features against this baseline model. If a visual feature model is "better" than this baseline model, the implication is that the visual feature is very likely to be used by participants when judging correlation in scatterplots.

The remainder of Table 2 reports the performance of each of the visual feature models. A visual feature model is colored in blue if the model outperforms the baseline model in that metric (e.g., odds ratios, AIC, and the Cox Test). As a result, the four top-performing visual features are:

TABLE 2
The results of modeling judgments with correlation and visual features



- *the standard deviation of all perpendicular distances to the regression line* (*dist_line_sd*),
- *the area of the prediction ellipse* (*ellipse_area*),
- *the length of the minor axis of the prediction ellipse* (*ellipse_minor*), and
- *the length of the perpendicular side of the confidence bounding box* (*conf_bounding_box_perp*).

These outperform the baseline model (i.e., the model of correlation) across all the metrics. In particular, the feature *the standard deviation of all perpendicular distances to the regression line* and *the area of the prediction ellipse* have similar effectiveness while outperforming the other two regarding odds ratios, AIC, and the Cox Test.

4.5 Summary and Discussion

These four visual features, based on modeling metrics, are more predictive of participants' judgments than correlation itself. Coincidentally, these four visual features come from different categories defined in Section 3: *length*, *area*, and *density*. It may suggest that participants use several visual features when judging correlation. This finding suggests a step towards exploring multi-factor models in future work.

The top-performing visual features support existing hypotheses in prior research in perceptual psychology and information visualization. For example, Meyer et al. identified the mean of the geometric distance between points and the regression line as impacting participants' ability to perceive correlation in scatterplots [6], which is synonymous with the feature *the standard deviation of all perpendicular distances to the regression line*.

These four features together all suggest that participants seek dispersion measures along the regression line. For example, the feature *the standard deviation of all perpendicular distances to the regression line* uses standard deviation to measure the density around the regression line; the feature *the length of the minor axis of the prediction ellipse* only relies on the minor axis, a confidence measure of the length of the point cloud along the regression line. These observations support the findings from perceptual psychology, as Eades [27] (cited in Lane et al. [28]), Cleveland et al. [9], and Meyer and Shinar [10] that the density and dispersion of data points in scatterplots affect participants' judgments.

5 Modeling Perception of Correlation using Visual Features

In this section, we examine the use of the visual features in modeling the perception of correlation in scatterplots. We investigate whether substituting visual features into the existing models of perception of correlation results in performance similar to those of the original models. In Section 5.1, we describe the three models used in our study: a linear model using mean observations (used by Rensink and Baldridge [1] and Harrison et al. [3]), a linear model using individual observations (by Kay and Heer [4]), and a log-linear model using individual observations (by Kay and Heer [4]). In Section 5.2, we propose a substitution method to verify the effectiveness of the visual features.

5.1 Background and Overview

Rensink and Baldridge [1], Harrison et al. [3], and Kay and Heer [4] introduced perceptual models that capture peoples' ability to judge correlation in scatterplots. These models are based on the concept of *Just-Noticeable Difference (JND)*, a measure of discrimination threshold. The JND describes the minimum amount of change in a stimulus needed for a person to reliably perceive a difference between two stimuli. The relation between JND and the stimulus can be described using Weber's law [29], [30]:

$$dP = k\frac{dI}{I} \tag{2}$$

where dP is the differential change in perception, dI is the differential increase in the stimulus, and I is the intensity of the baseline stimulus. The parameter k is known as the *Weber fraction* and is estimated via perceptual experiments. Given a specific I and Weber fraction k, the JND corresponds to the smallest increase of dI that will produce a noticeable difference in perception.

Based on the concept of JND, Rensink and Baldridge [1] as well as Harrison et al. [3] proposed a model for the perception of correlation. They measured JNDs from in-lab and crowdsourced experiments, and aggregated participants' JNDs into a Weber (linear) model. Kay and Heer, re-analyzing the experimental data from Harrison et al. based on individual observations, propose a non-linear model using multi-level Bayesian statistics and logarithmic transformation, which improved the fit and generalizability.

Weber's law generally applies to low-level perceptual properties [29], such as discriminating line lengths. Line lengths are closely related to two *length* visual features we identified, as they explain judgments better than correlation values. This observation implies that the modeling of the perception of correlation in scatterplots, as Rensink and Baldridge and Harrison et al. proposed, may be partially explained by people using visual features as proxies of correlation in the judgment process.

We use two techniques to determine the potential interchangeability between visual feature and correlation:

1) Extend the existing models to the use of visual features: we fit the data from visual features into the three models proposed by Rensink and Baldridge, Harrison et al., and Kay and Heer. A successful fit would indicate a similarity between existing models of correlation and models that use visual features.
2) Algebraic substitution: we use algebraic techniques to determine whether visual feature can reproduce the original models of correlation. This analysis has two purposes: (i) it may provide evidence that the visual features are used as proxies of correlation judgment, and as such, (ii) it may explain that why low-level perceptual laws apply to the perception of correlation.

Without loss of generality, the analysis below uses the visual feature: *the standard deviation of all perpendicular distances to the regression line* (denoted as ***dist_line_sd***), one of the best-performing features from our previous experiment. Analysis of the other three visual features yields similar results, which are included in Appendix G.

5.2 The Analytics Pipeline

This section presents our modeling procedure and the substitution technique (see Figure 6). The modeling procedure allows us to replicate extant models and extend them to include the use of visual features. The substitution technique is used to validate that these features can be used in lieu of correlation in the perceptual models.

We first generalize the relationship between perception and level of correlation into the following form:

$$JND_r = f(r) \tag{3}$$

The r subscript represents correlation r (i.e., I in Weber's law). The equation states that the JND of correlation (JND_r) is a function (f) of correlation (r). The function f can have various forms. For the linear model by Rensink and Baldridge and Harrison et al., f is a linear function (Weber's law). In the case of Kay and Heer's log-linear model, f is a log-linear function. We replicate both of these forms as the first step of our analysis.

Using a similar notation, we can likewise describe the perception of a visual feature as:

$$JND_v = f(v) \tag{4}$$

The v subscript represents visual feature (v). It states that the JND of a visual feature (JND_v) is a function (f) of the magnitude of that visual feature (v).

Although these two functions appear disparate, our reasoning is that if in fact visual features are proxies used by participants to judge correlation (r) in scatterplots, then these two functions would be interchangeable. Our substitution technique builds on these equations and is simply a series of operations that transform Equation 4 (Box 2 in Figure 6) into Equation 3 (Box 1 in Figure 6), and evaluates the equality between them.

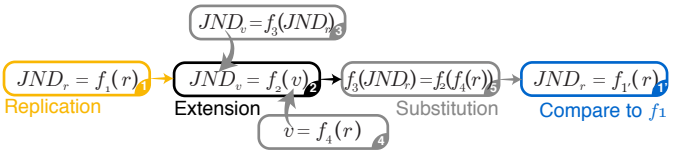


Fig. 6. Our analytic pipeline: replication (Box 1), extension (Box 2), and substitution (Box 5). Box 1 represents the JND model of correlation based on the experimental data; this can be extended to the visual feature (Box 2). Boxes 3 and 4 present the relation of correlation and the visual feature and the relation of their JNDs, respectively. Substituting Boxes 3 and 4 into Box 2 yields Box 5, simplified into Box 1', and compared to Box 1. This comparison validates whether the visual feature can reproduce the model of correlation.

These procedures can be integrated into a single analytics pipeline with the following steps (see Figure 6):

1) Box 1 (f_1): replicate the original model of correlation, including the models by Rensink and Baldridge, Harrison et al., and Kay and Heer[2].
2) Box 2 (f_2): extend the original model to include the use of the visual feature. Again, the example visual feature used is *the standard deviation of all perpendicular distances to the regression line* (***dist_line_sd***).
3) Box 3 (f_3): model the relation between the JND of correlation (JND_r) and the JND of the visual feature (JND_v).
4) Box 4 (f_4): model the relation between correlation (r) and the visual feature (v).
5) Box 5→1' ($f_{1'}$): derive a new model of the perception of correlation based on the visual feature.
6) Box 1' vs. 1: compare the derived model of correlation with the original model estimated from the experimental data. This step will validate whether the visual feature can replace correlation in the model.

The forms of f_1, f_2, and f_3 are consistent with each other and vary, depending on different modeling techniques (e.g., linear, log-linear). We use a linear form for f_4 uniformly to simplify computation and avoid discrepancies between different forms. The derivation of the substitution and a discussion of this assumption can be found in Appendix D.

5.3 Model Metrics

We employ two sets of evaluation metrics for our regression analysis. First, we use metrics similar to those used to evaluate the judgments models in Section 4, including p-value, R^2, and Akaike information criterion (AIC)[3]. Second, we perform regression diagnostics, including testing normality of residuals using the Shapiro test [33], skewness [34], kurtosis [34], and homoscedasticity of residuals using the Levene's test [35]. Skewness and kurtosis measure different aspects of the distribution, with a sign for direction (i.e.,

2. In both Rensink and Baldridge's and Harrison et al.'s work, JND is calculated using the average difference in correlation in last 24 judgments over a trial (see Section 2 and Figure 4a). Our modified experiment uses 50 judgments with no convergence criteria. We instead use weighted logistic regression to estimate JND based on all 50 judgments in a trial. This approach is implicitly validated by the similarity between the previous results and ours.

3. We use `R` package `gamlss` [31], [32] to fit all models. The `gamlss` procedure fit models by fitting residuals using different residual distributions. This makes it valid to compare AIC for models based on transformed data.

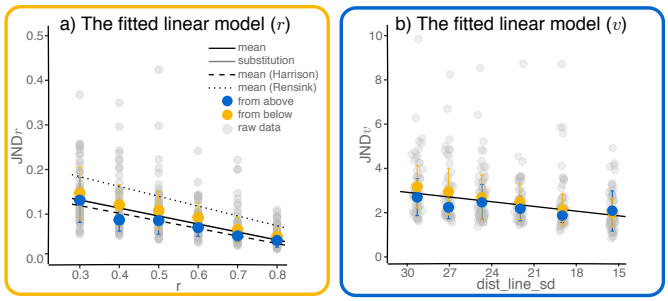


Fig. 7. The linear models using mean observations for a) correlation and b) the visual feature ($dist\_line\_sd$). Error bars are the standard deviation in aggregation.

TABLE 3
The results of the linear models using mean observations

Model	Correlation Coefficient	Coefficients				R^2	AIC
		β0	p	β1	p		
f_1: $JNDr \sim r$	-.9778	0.1860	<.001	-0.1791	<.001	.9561	-79.7720
f_2: $JNDv \sim v$	.9011	0.7975	.0101	0.0708	<.001	.8119	-3.0762
f_3: $JNDv \sim JNDr$	.9950	1.4717	<.001	10.8154	<.001	.9900	-40.7950
f_4: $v \sim r$	-.9962	38.0295	<.001	-27.4850	<.001	.9924	18.6965
f_1': *Substitution*	-	0.1864	-	-0.1798	-	-	-

*The inferred model (the last line) is similar to the original model (the first line). Since these numbers are small, we round to 4 decimal places in the tables to enable readers to reproduce our substitution results.

left- or right-skewed). We also illustrate residuals using detrended Q-Q plots [36], a means to present residual distribution, and in the case of normality, the difference between normalized residual and unit normal quantile fall into with the confidence band. Last, we visually inspect our results and compare them to the results from the extant works.

5.4 Linear Model using Mean Observations (Rensink and Baldridge; Harrison et al.)

Proposed by Rensink and Baldridge and replicated by Harrison et al., the first model in our analysis is based on the mean discrimination thresholds to approximate a linear function for correlation perception (i.e., Weber's law).

5.4.1 Modeling

To replicate Harrison et al.'s results, we follow their approach to mitigate large variations in individual performance. We exclude participant averages outside 3 Median Absolute Deviations [37] for a fixed correlation level (e.g., r=0.5). Within the given correlation level and approach (e.g., approaching from below), participants' data are averaged to obtain an estimation of mean JND and further combined using an adjustment [1]. The model has the form:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (5)$$

where JND is represented as y, written as a linear function of the adjusted baseline intensity of the stimuli x (i.e., r or v), with an overall slope β_1, an intercept β_0, and an error term ε. In this equation, i represents the mean observations with 12 data points ($i = [1..12]$, $r \times approach = [0.3, 0.4, 0.5, 0.6, 0.7, 0.8] \times [above, below]$).

5.4.2 Results

Figure 7 and Table 3 report the results of the replication, extension, and substitution pipeline. Due to the small sample size (n=12), we omit the results of regression diagnostics as they are less meaningful [38]. As a result, we have

1) Box 1 (f_1): the linear regression fit of the correlation data is satisfactory (R^2=.96) when replicating the existing models (see Figure 7a). In addition, the coefficients of the regression slope (β_1=-0.18) and intercept (β_0=0.19) are close to those of Harrison et al. (β_1=-0.17, β_0=0.17) [3], and not far from the results from Rensink and Baldridge (β_1=-0.20, β_0=0.22) [1]. These comparisons establish that our replication experiment data are consistent with previous findings, inviting a further comparison to the models using visual features.
2) Box 2 (f_2): we observe a decent fit (R^2=.81) when extending the model to the visual feature ($dist\_line\_sd$, see Figure 7b).
3) Box 3 (f_3): JND_r and JND_v can be fit by a linear function (R^2=.99).
4) Box 4 (f_4): correlation (r) and the magnitude of the visual feature (v) are linearly correlated (R^2=.99).
5) Box 5→1' ($f_{1'}$): we derive a new model of correlation perception using the visual feature. This model has the coefficients β_1=-0.18 and β_0=0.19.
6) Box 1' vs. 1: we compare the new derived model with the original one. We see that the two models are nearly identical (see Figure 7a and Table 3).

In sum, we replicate the linear model using mean observations from our replicate experiment, and extend it to the use of the example visual feature. The result confirms that the visual feature can replace correlation in this model without loss of precision. The slightly higher intercept and slope compared to Harrison et al.'s may be due to fatigue effect [39], since our experimental duration was longer.

5.5 Linear Model using Individual Observations (Kay and Heer)

Next, we validate the use of visual features in the techniques proposed by Kay and Heer [4]. Note that Kay and Heer used a series of models and techniques: 1) a linear model using individual observations, 2) a log-linear model using individual observations, 3) censoring method for observations without a known value, 4) Bayesian statistics, and 5) a log-linear model with a random intercept.

A key observation from Kay and Heer is that the aggregated model does not take into account the non-constant variance between the individuals [4]. Instead, they started with a linear model based on individuals observations, which allows to include all individual variance. It offers a principled way for including outliers, as each observation is assigned a likelihood, and outliers are assigned a relatively low weight. Another technique employed by Kay and Heer is the inclusion of random effect to improve the generalizability of the model coefficients. This resolves the correlation between observations from the same participant.

In this section, we replicate the linear model of correlation based on individual observations (the first model from Kay and Heer) and extend it to visual features. This linear model sets a baseline for comparison based on individual observations and bridges between the model by Rensink and Baldridge and Harrison et al. and the further models by Kay and Heer. Following Kay and Heer, we include all individual observations in all the models (95 participants × 4 observations per participant = 380 observations).

5.5.1 Modeling

Following Kay and Heer, we first construct a regression model that incorporates individual observations:

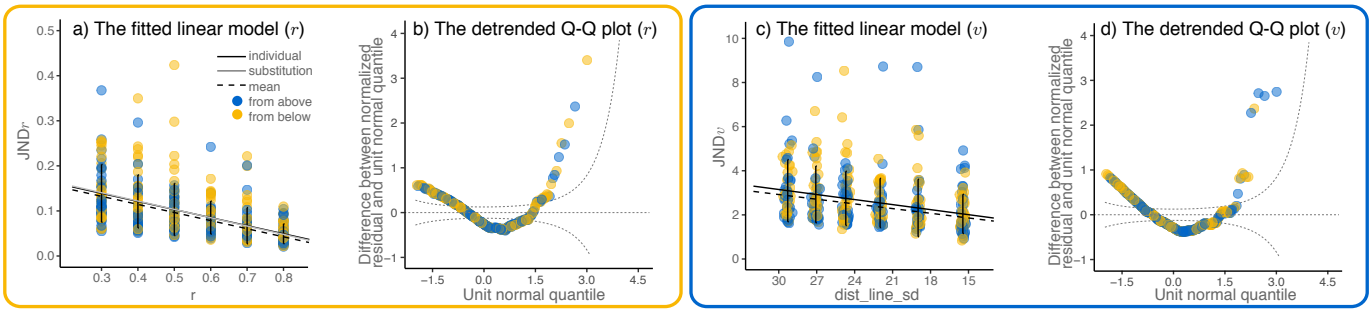


Fig. 8. The linear models using individual observations and their regression diagnostics: a) and b) correlation, c) and d) the visual feature ($dist\_line\_sd$). Error bars are root mean square errors. The detrended Q-Q plots illustrate the deviation from normality and the skewness in residuals, aligning with the analysis by Kay and Heer.

TABLE 4
The coefficients, substitution, R^2, and regression diagnostics of the linear models using individual observations

Method	Coefficients								R^2	AIC	Normality of residuals	Skewness	Kurtosis	Homoscedasticity
	β0	p	β1	p	β2	p	β3	p						
f_1: $JNDr \sim r$	0.1927	<.001	-0.1792	<.001	0.0202	.0167	-0.0195	.1809	.3076	-1213.7160	p < .001	2.1075	7.4827	p < .001
f_2: $JNDv \sim v$	0.8672	.0046	0.0761	<.001	-0.2949	.3314	0.0155	.2345	.0876	1228.8330	p < .001	2.3604	8.1541	p = .4596
f_3: $JNDv \sim JNDr$	1.5412	<.001	11.3588	<.001	-	-	-	-	.9648	-937.7772	p < .001	-0.9502	0.5097	p < .001
f_4: $v \sim r$	38.0921	<.001	-27.6028	<.001	-	-	-	-	.9915	455.0915	p < .001	-0.3038	-1.2431	p = .0011
f_1': *Substitution*	0.1959	-	-0.1849	-	-	-	-	-	-	-	-	-	-	-

*The inferred linear model using visual features (the last line) is similar to the model using correlation, which was estimated directly from the experimental data (the first line). Regression diagnostics show that the residuals are skewed and not normally distributed.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 a_i + \beta_3 a_i x_i + \varepsilon_i \quad (6)$$

where y represents JND similar to that of the linear model using mean observations (Equation 5), a represents approach, x represents the stimulus with an error term of ε, and a_x is the interaction between the two. According to Kay and Heer, a_i is defined as

$$a_i = \begin{cases} -1, & \text{if } approach \text{ is } from\ above \\ 1, & \text{if } approach \text{ is } from\ below \end{cases}$$

In these equations, i is from $1..n$, where n is the number of individual observations (n=380), differentiated from the *mean observations* used by Rensink and Baldridge and Harrison et al. where n=12.

5.5.2 Results

The results of modeling with individual observations using a linear model are presented in Figure 8 and Table 4. Specifically, we have:

1) Box 1 (f_1): we first present the linear model of correlation based on individual observations, which has similar coefficients with the previous aggregated model using mean observations (see Figure 8a, β_1=-0.18, β_0=0.19 vs. β_1=-0.18, β_0=0.19).
2) Box 2 (f_2): we extend the model to the visual feature and find a drawback in goodness-of-fitness (R^2=.09 vs. .31).
3) Box 3 (f_3): we find a strong linearity between the two sets of JNDs from correlation and the visual feature ($dist\_line\_sd$, R^2=.96).
4) Box 4 (f_4): we find a strong linearity between the visual feature and correlation based on individual observations (R^2=.99).
5) Box 5→1' ($f_{1'}$): combining the three equations above, we derive a new model for correlation.
6) Box 1' vs. 1: the resulting model from substitution (β_1=-0.18, β_0=0.20) is very similar to the original model of correlation (β_1=-0.18, β_0=0.19) in both shape and form (see Figure 8a).

These results are consistent with Kay and Heer's findings in the following ways. First, in the regression diagnostics for the models of both correlation and the visual feature, the residuals are not normally distributed (p<.001), with non-zero skewness and kurtosis (see Figure 8b and d). Second, the models also do not hold the assumption of homoscedasticity for residuals (p<.001). These are two important findings from Kay and Heer, which leads to a step of further refining the model.

5.6 Log-Linear Model with a Random Intercept, using Individual Observations (Kay and Heer)

Kay and Heer pointed out that, because the linear model violates the key assumptions of normality and homoscedasticity in regression analysis, it may result in a biased model and an overestimated goodness-of-fit. They, therefore, proposed a log-linear model to transform the data into one that meets the assumptions for regression analysis. They also incorporated random effect to account for observations from the same participant. The focus of this section is to replicate the log-linear model with a random intercept, to extend it to the use of the visual feature ($dist\_line\_sd$), and validate whether the visual feature can reproduce the log-linear model of correlation proposed by Kay and Heer.

5.6.1 Modeling

Following Kay and Heer, the log-linear model has the form:

$$log(y_i) = \beta_0 + \beta_1 x_i + \beta_2 a_i + \beta_3 a_i x_i + \varepsilon_i + U_k \quad (7)$$

The log-transformed JND ($log(y)$) is modeled as a linear function of the baseline intensity x (i.e., r or v), approach a, an interaction between them (ax), and an offset (U_k) for each participant k. The difference is the use of a log transformation on the individual observations to correct for skewed residuals and an U_k comprising a treatment factor effect [40] (see Kay and Heer [4]).

5.6.2 Results

Similar to the previous section, we first replicate the model, extend it to the visual feature, and derive the substitution from the visual feature.

1) Box 1 (f_1): Table 5 shows the fit of the log-linear model

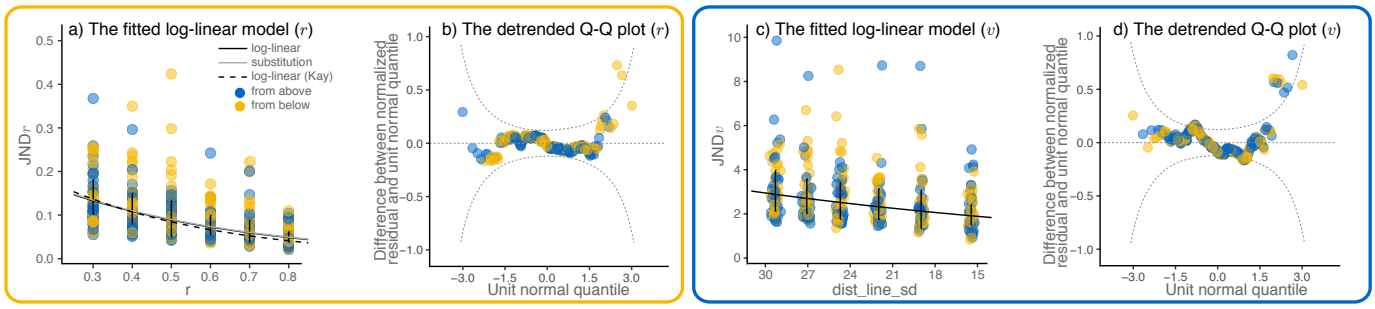


Fig. 9. Log-linear models using individual observations and their regression diagnostics: a) and b) correlation, c) and d) the visual feature ($dist\_line\_sd$). Error bars are root mean square errors.

TABLE 5
The coefficients, substitution, R^2, and regression diagnostics of the log-linear models using individual observations

Method	Coefficients								R^2	AIC	Normality of residuals	Skewness	Kurtosis	Homoscedasticity
	β0	p	β1	p	β2	p	β3	p						
f_1: $log(JNDr) \sim r$	-1.4137	<.001	-2.0152	<.001	0.1365	.0021	-0.0815	.2837	.7941	-1724.3940	p = .0586	0.1170	0.5619	p = .9961
f_2: $log(JNDv) \sim v$	0.1903	<.001	0.0297	<.001	-0.1074	.0516	0.0059	.0130	.7104	748.3163	p < .001	0.4009	0.8595	p = .2798
f_3: $log(JNDv) \sim log(JNDr)$	1.8886	<.001	0.4038	<.001	-	-	-	-	.9917	-1460.4100	p < .001	-0.8475	-0.2552	p < .001
f_4: $v \sim r$	38.0921	<.001	-27.6028	<.001	-	-	-	-	.9915	455.0915	p < .001	-0.3038	-1.2431	p = .0011
$f_{1'}$: *Substitution*	-1.4064	-	-2.0285	-	-	-	-	-	-	-	-	-	-	-

*The residual analysis shows that the residuals of the log-linear model deviate from a normal distribution. The substitution results in a model (the last line) that is similar to the original model estimated directly from the experimental data (the first line).

to our data (R^2=.79) and the coefficients (β_1=-2.02, β_0=-1.41), which are similar to those of Kay and Heer's (β_1=-2.39, β_0=-1.27, from the results released online).

2) Box 2 (f_2): we confirm that extending to the visual feature yields a decent fit ($dist\_line\_sd$, R^2=.71).
3) Box 3 (f_3): we find a favorable fit between the two sets of log-transformed JNDs (R^2=.99).
4) Box 4 (f_4): the relation between the visual feature and correlation remains the same (R^2=.99).
5) Box 5→1′ ($f_{1'}$): combining the results above turns into a new model of correlation (see the last line in Table 5).
6) Box 1′ vs. 1: the derived correlation model is very similar to the original model (see Figure 9, β_1=-2.02, β_0=-1.41 vs. β_1=-2.03, β_0=-1.41).

Similar to Kay and Heer's results, we find that AIC of the log-linear model is an improvement over the linear model (e.g., -1724.39 vs. -1213.72). However, we still observe that the log transformation leaves some skewness in the residuals, and the residuals are non-normally distributed (e.g., p<.001), especially for the visual feature.

5.7 Summary and Discussion

Thus far, we have examined three existing modeling techniques of correlation perception in scatterplots: a linear model using mean observations, a linear model using individual observations, and a log-linear model with random intercepts to account for individual observations. For each of these, we have replicated the original model using our experimental data, extended the model to include the visual feature, then substituted the visual feature for correlation. Our analysis indicates that the use of the visual feature generally performs similarly to the use of correlation. As a result, we find that models using visual features can successfully reproduce the original model of correlation.

The visual feature that best explains participants' judgments yields at least the same effectiveness in modeling the perception of correlation in scatterplots. This supports our two speculations: 1) the perceptual laws for low-level perception may apply to the perception of correlation because 2) visual features are possible proxies of correlation.

Our demonstration of an interchangeable relationship between the visual features and correlation is not a "proof" of that participants in fact use visual features as a proxy to judge correlations in scatterplots. Instead, this demonstration provides evidence that the visual features are possible proxies for correlation. Future perceptual and cognitive experiments will be necessary to verify this claim. For example, we observe that adding a random intercept (in the case of the log-linear model by Kay and Heer) to allow the inclusion of individual difference enhances the goodness-of-fit, especially when using the visual feature in the model. This finding may imply that different participants utilize different visual features, although the same people may use the same features across judgments.

6 Improving Perceptual Models Using Power Transformation

The models explored thus far (Rensink and Baldridge, Harrison et al., and Kay and Heer), still have room for improvement. Several issues remain, such as non-normality in the residuals in the log-linear model and a general need to improve the overall fit. Inspired by the power function widely used in perceptual psychology, we propose a straightforward power transformation (instead of a linear or a log-linear model) to better model the perception of correlation and the visual feature in scatterplots.

Power functions are commonly used in perceptual psychology for modeling our perception of physical stimuli. The most well-known use of power functions in this area is Stevens' power law, which was introduced as a means to extend Weber's law to describe a wider range of stimuli [41]. Specifically, Stevens' power law states that the subjective magnitude of sensation is proportional to the intensity of the stimulus raised to a certain power a. For example, as reported by Stevens, a is 0.7 for the perception of area of projected square, where as a is 1.2 for the sensation of lightness using the reflectance of gray papers [41].

Power functions have also been used in modeling the perception of correlation in scatterplots in several existing studies. For example, Pollack used a square function [7] to

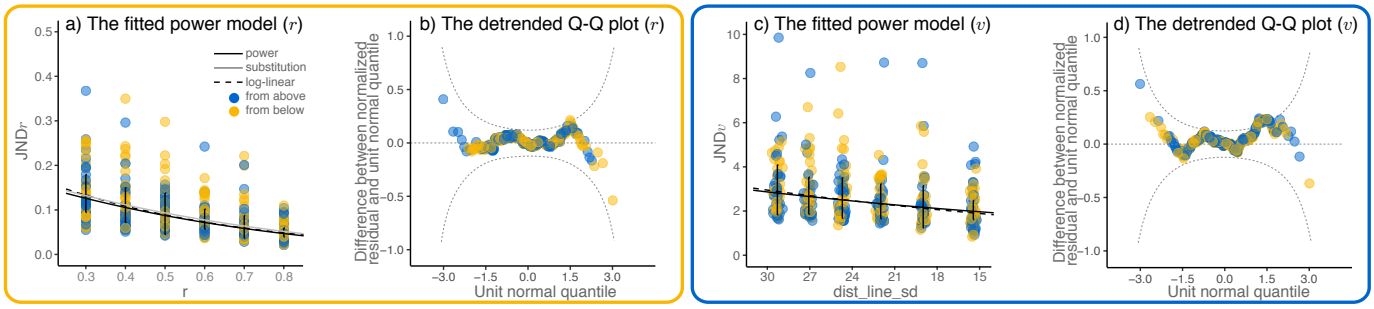


Fig. 10. The power transformation models and residual analyses: a) and b) correlation, c) and d) the visual feature ($dist\_line\_sd$). In contrast to the previous models, the power transformation models have the desirable properties of normal-like and constant-like residuals.

TABLE 6
The coefficients, substitution, R^2, AIC, and regression diagnostics of the power transformation models

Method	BCT	Coefficients								R^2	AIC	Normality of residuals	Skewness	Kurtosis	Homoscedasticity
	ω	β0	p	β1	p	β2	p	β3	p						
f_1: $JND_r^{\omega_1} \sim r$	0.26	0.6629	<.001	-0.2646	<.001	0.0186	<.001	-0.0155	.0615	.8099	-1742.0840	p = .2866	-0.0298	-0.2223	p = .3326
f_2: $JND_v^{\omega_2} \sim v$	-0.17	0.9465	<.001	-0.0037	<.001	0.0187	<.001	-0.0010	<.001	.7387	703.7296	p = .0702	0.1061	-0.2850	p = .0607
f_3: $JND_v^{\omega_2} \sim JND_r^{\omega_1}$	-	1.0722	<.001	-0.3985	<.001	-	-	-	-	.9907	-4124.9540	p < .001	0.7579	-0.9203	p < .001
f_4: $v \sim r$	-	38.0921	<.001	-27.6028	<.001	-	-	-	-	.9915	455.0915	p < .001	-0.3038	-1.2431	p = .0011
f_1': *Substitution*	-	0.6672	-	-0.2549	-	-	-	-	-	-	-	-	-	-	-

*Suggested by R^2 and AIC, the power transformation model is an improvement over the log-linear model. In the substitution, the inferred model (the last line) closely resembles the model estimated empirically from the experimental data (the first line).

model a relationship between perceived correlation (sensation) and objective correlation (stimulus); Jennings et al. proposed a square root function [8]; Cleveland et al. used a square root function and double-power functions with two free parameters [9]; Boynton proposed a power function with one or two free parameters [5].

What is common in all these works is the use of the power function. However, the data collected in these experiments overwhelmingly come from experiments where participants were instructed to directly estimate the correlation of a given chart. A key difference in the approach that we adopt from Rensink and Baldridge, as pointed out in their work, was the use of psychophysical techniques which mitigate estimation bias and variance by only requiring participants to indicate which plot appears more correlated. Such techniques align more with the original experiments described by Stevens [41].

6.1 Power Transformation and Evaluation Metrics

The power transformation model is generally considered to be more flexible than a linear or a log-linear model because of the use of the exponent. We use the power transformation to transform both JNDs of correlation and the visual feature. Specifically, we utilize the Box-Cox t distribution, which is a generalization of the Box-Cox normal distribution [42] that can model both skewness and kurtosis, and has been confirmed to surpasses Box-Cox normal distribution [43]. We also choose a power function for the link function of location (median) [44]. The exponent (denoted as ω) in the link function is chosen based on considerations from both statistical metrics and perceptual psychology:

1) Given that the exponent in Stevens' power law related to human vision is commonly around 0.3-1.5, we sample all possible values from [-5, 5] at a step of 0.01.
2) The classic Box-Cox transformation results in exponents of -0.06 (95% CI:[-0.17, 0.05]) and -0.41 (95% CI:[-0.57, -0.24]), validating that sampling from [-5, 5] is reasonable.
3) Steps 1) and 2) yield multiple exponential terms that lead to a model outperforming the log-linear model, and we present the one that has the best tradeoff between the statistical metrics, including R^2, AIC [45], skewness etc.

6.2 Power Transformation with Individual Observations

The power transformation model has the general form:

$$y_i^{\omega} = \beta_0 + \beta_1 x_i + \beta_2 a_i + \beta_3 a_i x_i + \varepsilon_i + U_k \quad (8)$$

This equation is similar to the linear model and the log-linear model by Kay and Heer, except the use of y_i^{ω} instead of $log(y_i)$, where y represents JND. The exponent ω indicates the power term from link function for location, while a and x represent the approach and the stimulus (i.e., r or v) respectively, β represents model coefficients, U_k accounts for random intercept, and ε is the error term.

6.3 Results

The results of power transformation for both correlation and the visual feature ($dist\_line\_sd$) are presented in Table 6.

Compared to the log-linear model, the two models appear similar (see Table 6 and Figure 10). We find that the transformed JNDs of correlation and the visual feature can be represented as linear models (R^2=.99). When substituting the visual feature into the power transformation model, the visual feature exactly reproduces the original model of correlation (β_1=-0.26, β_0=0.66 vs. β_1=-0.25, β_0=0.67).

The power model, however, outperforms the log-linear model in a variety of ways:

1) The power model shows a better goodness-of-fit (e.g., R^2=.81 vs. R^2=.79) and an improved AIC (e.g., 703.73 vs. 748.32) over the log-linear model. These indicate that the overall quality of the model is improved.
2) The power model also shows an improvement in regression diagnostics. It has a residual distribution that is normally distributed compared to the log-linear model (p=.07 vs. p<.001, p=.29 vs. p=.06). The model appears to contain a slight drawback in homoscedrasticity over the log-linear model, but the value is still acceptable and surpasses the linear model (p=.06 vs. p<.001). This model generally exhibits less skewness and kurtosis in

the residuals than the log-linear model (skewness: -0.03 vs. 0.12, kurtosis: -0.29 vs. 0.86).

The power model improves upon the log-linear model in that it has a higher R^2 and improved AIC, skewness, and kurtosis. In conclusion, in the areas where the log-linear model had room for improvement, the power model fills these gaps, resulting in a model that by many measures can be considered a more faithful representation of the participants' perception of correlation in scatterplots.

7 Discussion and Conclusion

In this paper, we examine the longstanding hypothesis that people perceive *visual features* related to correlation when judging correlation in scatterplots. Toward this end, we collect 49 visual features, spanning literature in perceptual psychology to visualization, statistics, and computational geometry. We analyze them at several levels including both aggregate and individual judgments and across several recently proposed mathematical models of behavior (i.e., [1], [3], [4]). The results at the individual level indicate that visual features are more predictive of participants' judgments than correlation. At the model level, the results of analyses indicate that the extant models can be successfully extended to visual features. Finally, drawing inspiration from decades of perceptual psychology, we move beyond current models (linear, log-linear) to show that a power transformation of observed JNDs produces a more precise model than existing approaches, resulting in better performance when using either correlation or the non-transformed visual features.

7.1 Power and Log Transformations

The "power transformation vs. log transformation" debate [46] has spanned a variety of research fields, including image enhancement [47] (e.g., Gamma correction) and biology [48], many of which have suggested that power transformations are more robust than log transformations for fitting data [46], and have desirable properties for explaining the underlying behavior of phenomena.

Given these considerations, the available transformations should also be evaluated as candidate explanations for the underlying judgment behavior, viewed through the lens of perceptual psychology alongside their mathematical properties. The results in this work indicate that a power transformation better describes the observed data than a log transformation. Similarly, the log transformation by Kay and Heer was reportedly chosen on the basis that it addresses modeling concerns such as skewness [4].

Beyond the intuition that the power transformation is connected to studies that model perceived stimuli, its mathematical inference is via integration. Using our modeling of correlation perception in scatterplots as an example, the linear, log transformation, and power transformation can be written as $\Delta I = k(I+b)$, $log(\Delta I) = k(I+b)$, and $(\Delta I)^{\omega} = k(I+b)$, respectively. In these equations, I is the magnitude of the stimulus, and ΔI is the smallest change resulting in a unit step in the perception (i.e., JND). On Fechner's assumption [1], [2], [49], we can infer the perceived stimulus (P) from these three equations by integrating on both sides (see Appendix F): $P = \frac{1}{ck} log(I+b) + C$, $P = -\frac{1}{ck} e^{-k(I+b)} + C$, and $P = \frac{k^{-a}}{c(1-a)} (I+b)^{1-a} + C$.

The linear model results in a logarithmic function (Fechner's law [50]), and this log function has been found to best describe mean observations of perceived correlation [14]. Our results also show that the basic linear models using mean observations can capture observed behavior, and thus can be descriptive and predictive when understanding phenomena at the population level.

At the individual level, linear models might not be able to explain and describe the variance and differences across people. In this case, transformations might help. A log transformation, as we showed above, results in an exponential function for the perceived stimulus. In other perceptual studies for visualizations, however, results appear to use non-exponential functions (e.g., [51], [52]). The log transformation is widely used in biomedicine, psychosocial research, and physics [53], [54] as phenomena like life growth, information spread, and radioactive decay are all modeled at using exponentials.

The power transformation results in a function that models perceived correlation as a multiplicative factor of objective correlation, and therefore can be an *instance* of Guilford's [55] and Stevens' [41] power laws, which is consistent with literature using power functions to model the perceived correlation in scatterplots [5], [6], [9], [56] and beyond [41]. Some literature suggested that Weber's law may have been superseded by Stevens' power law as the standard modeling for understanding the mapping between stimulus and perception [57], [58], though debates [59] and exceptions [60] remain.

When viewed from a statistical standpoint, the use of a power transformation allows more modeling generality and precision by adjusting the exponential term. It addresses the consideration as to whether the model can be generalized to a wide range of stimuli when choosing a model for a perceptual process. The log transformation, in contrast, can be viewed as a special case of the power transformation. Though the quantitative improvements of using the power transformation model are relatively small compared to the log transformation models, the power transformation better describes the underlying perceptual data by providing both the modeling advantages of the log transformation, evidenced by our results, with the added benefit of a link to existing research in perceptual psychology. However, it should be noted that there may be cases that the two transformations perform indistinguishably. Our discussion focuses on transforming the dependent variables, while other research might find transforming the independent variables, such as the visual features themselves, yields new avenues towards understanding the relationships between people and visualizations.

Finally, while some of our analyses use individual observations and partially measure individual perception, it may be worth considering future studies that employ a rigorous comparison of models (e.g., linear, log-linear, power, exponential) that use observations and judgments from the same participant across different correlation levels. Such investigations will help us better understand variation in individual perception and visualization performance.

7.2 Perceptual Science and Visualization

One goal of this paper is to ground recent research efforts in InfoVis on perceptual modeling of statistical visualizations in the concepts and practices of the perceptual science community. While the InfoVis community can apply sophisticated statistical methodologies towards fitting psychophysical experimental data into models, we argue that we must also incorporate methodologies and techniques from perceptual scientists to help understand *why* the resulting models and the observed phenomenon occur.

One notable example is that the perception of "Pearson correlation" is an abstract mathematical concept, has shown to be modeled using Weber's law, a simple linear model used for hundreds of years in psychological research [1], [3]. In particular, because Weber law's is generally only applied to low-level physical stimuli (e.g., sound, weight, length, etc.), it is surprising that the abstract notion of "Pearson correlation" falls in the same family of models. If Rensink and Baldridge and Harrison et al. had used general model-fitting methods without making the connection to Weber's law, it is possible that an important link to prior psychological research would have been missed.

Our investigation of visual features in this paper seeks to fill this knowledge gap. Although the result is positive, it is far from conclusive. For one, we are working under the assumption that the participants use a single visual feature (and the same feature) to estimate correlation. The results suggest room for improvement here, as our models using visual features performed well, but not as well as those using the difference in correlation, which can be considered an amalgam of visual features. Intuitively, given a difficult correlation judgment participants may switch "strategies" (i.e., the use of different visual features or combinations of features) to make a final judgment.

This paper focuses on the perception of correlation in scatterplots, we posit that our experimental and analytical methods based on perceptual features may extend to the study of other multivariate visualizations for other perceptual tasks (e.g., detecting outlier, mean, trend). Quantifying and modeling such features for commonly used visualizations can be an important area for future work, as it will enable more predictive models and shed light on how people perceive information from visualizations. Recent work by Rensink suggests that a model of correlation perception based on entropy is possible (i.e., that people can perceive how "random" a visualization appears) [14]. These and other interdisciplinary efforts that cut across both visualization and perceptual psychology can begin to develop theories of visualization [2] that can serve as the foundation for the next generation of information visualization research, design, and practices.

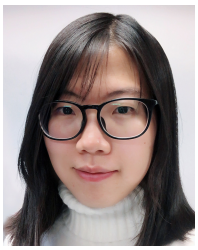
ACKNOWLEDGMENTS

The research is supported in part by DARPA FA8750-17-2-0107, NSF awards IIS-1452977 and IIS-1162067. We thank all the anonymous reviewers for their thoughtful feedback. We thank Megan Van Welie and R. Jordan Crouser for their help with the manuscript.

REFERENCES

[1] R. A. Rensink and G. Baldridge, "The perception of correlation in scatterplots," *Computer Graphics Forum*, vol. 29, no. 3, pp. 1203–1210, 2010.
[2] R. A. Rensink, "On the prospects for a science of visualization," in *Handbook of Human Centric Visualization*, W. Huang, Ed. Springer, 2014, pp. 147–175.
[3] L. Harrison, F. Yang, S. Franconeri, and R. Chang, "Ranking visualizations of correlation using Weber's law," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1943–1952, 2014.
[4] M. Kay and J. Heer, "Beyond Weber's law: A second look at ranking visualizations of correlation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 469–478, 2016.
[5] D. M. Boynton, "The psychophysics of informal covariation assessment: Perceiving relatedness against a background of dispersion," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 26, no. 3, pp. 867–876, 2000.
[6] J. Meyer, M. Taieb, and I. Flascher, "Correlation estimates as perceptual judgments," *Journal of Experimental Psychology: Applied*, vol. 3, no. 1, pp. 3–20, 1997.
[7] I. Pollack, "Identification of visual correlational scatterplots," *Journal of Experimental Psychology*, vol. 59, no. 6, pp. 351–360, 1960.
[8] D. Jennings, T. M. Amabile, and L. Ross, "Informal covariation assessment: Data-based vs. theory-based judgments," in *Judgment Under Uncertainty: Heuristics and Biases*, 1982, pp. 211–230.
[9] W. S. Cleveland, P. Diaconis, and R. McGill, "Variables on scatterplots look more highly correlated when the scales are increased," *Science*, vol. 216, no. 4550, pp. 1138–1141, 1982.
[10] J. Meyer and D. Shinar, "Estimating correlations from scatterplots," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 34, no. 3, pp. 335–349, 1992.
[11] T. W. Lauer and G. V. Post, "Density in scatterplots and the estimation of correlation," *Behaviour & Information Technology*, vol. 8, no. 3, pp. 235–244, 1989.
[12] T. N. Cornsweet, "The staircase-method in psychophysics," *The American Journal of Psychology*, vol. 75, no. 3, pp. 485–491, 1962.
[13] W. W. Daniel, "Applied nonparametric statistics," 1990.
[14] R. A. Rensink, "The nature of correlation perception in scatterplots," *Psychonomic Bulletin & Review*, vol. 24, pp. 776–797, 2017.
[15] P. Schubert and M. Kirchner, "Ellipse area calculations and their applicability in posturography," *Gait & Posture*, vol. 39, no. 1, pp. 518–522, 2014.
[16] L. Wilkinson, A. Anand, and R. Grossman, "Graph-theoretic scagnostics," in *Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization*. Washington, DC, USA: IEEE Computer Society, 2005, pp. 157–164.
[17] H.-P. Kriegel, P. Kröger, J. Sander, and A. Zimek, "Density-based clustering," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 3, pp. 231–240, 2011.
[18] M. De Berg, M. van Kreveld, M. Overmars, and O. C. Schwarzkopf, *Computational Geometry*. Springer, 2000.
[19] W. S. Cleveland and R. McGill, "Graphical perception: Theory, experimentation, and application to the development of graphical methods," *Journal of the American Statistical Association*, vol. 79, no. 387, pp. 531–554, 1984.
[20] G. A. Seber and A. J. Lee, *Linear Regression Analysis*. John Wiley & Sons, 2012, vol. 936.
[21] S. J. Press and S. Wilson, "Choosing between logistic regression and discriminant analysis," *Journal of the American Statistical Association*, vol. 73, no. 364, pp. 699–705, 1978.
[22] J. C. Stoltzfus, "Logistic regression: a brief primer," *Academic Emergency Medicine*, vol. 18, no. 10, pp. 1099–1104, 2011.
[23] "Interpreting logistic regression coefficients," in *Logistic Regression*, 0th ed., F. C. Pampel, Ed. SAGE Publications, Inc., 2000, pp. 19–40.
[24] N. Scotia, "Explaining odds ratios," *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, vol. 19, no. 3, pp. 227–229, 2010.
[25] W. H. Greene, *Econometric Analysis*. Pearson Education India, 2003.
[26] R. A. Armstrong, "When to use the Bonferroni correction," *Ophthalmic and Physiological Optics*, vol. 34, no. 5, pp. 502–508, 2014.
[27] S. Eade, "The effect of magnitude of criterion validity and positive cue validity upon human inference behavior," Master's thesis, Ohio State University, 1967.

[28] D. M. Lane, C. A. Anderson, and K. L. Kellam, "Judging the relatedness of variables: The psychophysics of covariation detection," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 11, no. 5, pp. 640–649, 1985.
[29] E. H. Weber, *De Pulsu, Resorptione, Auditu Et Tactu: Annotationes Anatomicae Et Physiologicae, Auctore...* Prostat apud C.F. Koehler, 1834.
[30] E. H. Weber, H. E. Ross, and D. J. Murray, *E.H. Weber On The Tactile Senses*. Psychology Press, 1996.
[31] R. Rigby and D. Stasinopoulos, *A flexible regression approach using GAMLSS in R*. University of Lancaster: Lancaster, UK, 2010.
[32] D. M. Stasinopoulos, R. A. Rigby *et al.*, "Generalized additive models for location scale and shape (GAMLSS) in R," *Journal of Statistical Software*, vol. 23, no. 7, pp. 1–46, 2007.
[33] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3/4, pp. 591–611, 1965.
[34] I. B. Weiner, J. A. Schinka, and W. F. Velicer, *Handbook of Psychology, Research Methods in Psychology*. John Wiley & Sons, 2003.
[35] M. B. Brown and A. B. Forsythe, "Robust tests for the equality of variances," *Journal of the American Statistical Association*, vol. 69, no. 346, pp. 364–367, 1974.
[36] S. van Buuren and M. Fredriks, "Worm plot: a simple diagnostic device for modelling growth reference curves," *Statistics in Medicine*, vol. 20, no. 8, pp. 1259–1277, 2001.
[37] D. C. Howell, "Median absolute deviation," in *Wiley StatsRef: Statistics Reference Online*. John Wiley & Sons, Ltd, 2014.
[38] W. D. Dupont and W. D. Plummer, "Power and sample size calculations for studies involving linear regression," *Controlled Clinical Trials*, vol. 19, no. 6, pp. 589–601, 1998.
[39] J. Liu, J. R. Allspach, M. Feigenbaum, H.-J. Oh, and N. Burton, "A study of fatigue effects from the new SAT®," *ETS Research Report Series*, vol. 2004, no. 2, pp. i–13, 2004.
[40] M. D. Stasinopoulos, R. A. Rigby, G. Z. Heller, V. Voudouris, and F. De Bastiani, *Flexible Regression and Smoothing: Using GAMLSS in R*. CRC Press, 2017.
[41] S. S. Stevens, "On the psychophysical law," *Psychological Review*, vol. 64, no. 3, pp. 153–181, 1957.
[42] G. E. Box and D. R. Cox, "An analysis of transformations," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 26, no. 2, pp. 211–252, 1964.
[43] R. A. Rigby and D. M. Stasinopoulos, "Using the Box-Cox t distribution in GAMLSS to model skewness and kurtosis," *Statistical Modelling*, vol. 6, no. 3, pp. 209–229, 2006.
[44] J. A. Nelder and R. W. M. Wedderburn, "Generalized linear models," *Journal of the Royal Statistical Society. Series A (General)*, vol. 135, no. 3, pp. 370–384, 1972.
[45] R. A. Rigby and D. M. Stasinopoulos, "Generalized additive models for location, scale and shape," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 54, no. 3, pp. 507–554, 2005.
[46] M. Mitzenmacher, "A brief history of generative models for power law and lognormal distributions," *Internet Mathematics*, vol. 1, no. 2, pp. 226–251, 2004.
[47] R. Maini and H. Aggarwal, "A comprehensive review of image enhancement techniques," *Journal of Computing*, vol. 2, no. 3, 2010.
[48] X. Xiao, E. P. White, M. B. Hooten, and S. L. Durham, "On the use of log-transformation vs. nonlinear regression for analyzing biological power laws," *Ecology*, vol. 92, no. 10, pp. 1887–1894, 2011.
[49] E. B. Goldstein, *Sensation and Perception*. Thomson Brooks/Cole Publishing Co, 1996.
[50] G. Fechner, *Elements of Psychophysics*. Holt, Rinehart and Winston, 1966.
[51] Ç. Demiralp, M. S. Bernstein, and J. Heer, "Learning perceptual kernels for visualization design," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1933–1942, 2014.
[52] D. A. Szafir, "Modeling color difference for visualization design," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 392–401, 2018.
[53] E. Limpert, W. A. Stahel, and M. Abbt, "Log-normal distributions across the sciences: Keys and clues," *BioScience*, vol. 51, no. 5, pp. 341–352, 2001.
[54] C. Feng, H. Wang, N. Lu, T. Chen, H. He, Y. Lu *et al.*, "Log-transformation and its implications for data analysis," *Shanghai Archives of Psychiatry*, vol. 26, no. 2, pp. 105–109, 2014.
[55] J. Guilford, "A generalized psychophysical law." *Psychological Review*, vol. 39, no. 1, pp. 73–85, 1932.
[56] J. Meyer and D. Shinar, "Perceiving correlations from scatterplots," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 35, no. 20. SAGE Publications, 1991, pp. 1537–1540.
[57] L. E. Krueger, "Reconciling Fechner and Stevens: Toward a unified psychophysical law," *Behavioral and Brain Sciences*, vol. 12, no. 02, pp. 251–267, 1989.
[58] A. K. Myers, "Psychophysical scaling and scales of physical stimulus measurement," *Psychological Bulletin*, vol. 92, no. 1, pp. 203–214, 1982.
[59] G. S. Wasserman, G. Felsten, and G. S. Easland, "The psychophysical function: harmonizing Fechner and Stevens," *Science*, vol. 204, no. 4388, pp. 85–87, 1979.
[60] H. M. Ditz and A. Nieder, "Numerosity representations in crows obey the Weber–Fechner law," in *Proceedings of the Royal Society B*, vol. 283, no. 1827. The Royal Society, 2016.

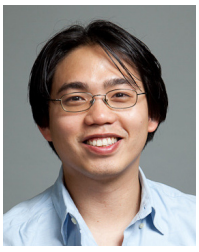


Fumeng Yang is a PhD student in the Department of Computer Science at Brown University. Her research interests include information visualization, virtual reality, and human computer interaction.

Lane T. Harrison is an assistant professor in the Computer Science Department at Worcester Polytechnic Institute. He received his PhD in Computer Science from University of North Carolina at Charlotte and was formerly a postdocoral fellow at Tufts University. His research interests focus on information visualization and visual analytics research.

Ronald A. Rensink is an associate professor in the Departments of Computer Science and Psychology at the University of British Columbia. He received his PhD in Computer Science from University of British Columbia. His research interests include human vision, computational vision, and information visualization.

Steven L. Franconeri is a professor of psychology at Northwestern University, and director in the Northwestern Cognitive Science Program. He received his PhD in Psychology from Harvard University. He studies visuospatial thinking and visual communication, across psychology, education, and information visualization.



Remco Chang is an associate professor in Computer Science at Tufts University. He received his PhD in Computer Science from the University of North Carolina Charlotte. His research interests include visual analytics, information visualization, human computer interaction, and databases.