Ranking Visualizations of Correlation Using Weber's Law

Lane Harrison, Fumeng Yang, Steven Franconeri, Remco Chang

Abstract— Despite years of research yielding systems and guidelines to aid visualization design, practitioners still face the challenge of identifying the best visualization for a given dataset and task. One promising approach to circumvent this problem is to leverage perceptual laws to quantitatively evaluate the effectiveness of a visualization design. Following previously established methodologies, we conduct a large scale (n=1687) crowdsourced experiment to investigate whether the perception of correlation in nine commonly used visualizations can be modeled using Weber's law. The results of this experiment contribute to our understanding of information visualization by establishing that: 1) for all tested visualizations, the precision of correlation judgment could be modeled by Weber's law, 2) correlation judgment precision showed striking variation between negatively and positively correlated data, and 3) Weber models provide a concise means to quantify, compare, and rank the perceptual precision afforded by a visualization.

Index Terms—Perception, Visualization, Evaluation

1 Introduction

The theory and design of information visualization has come a long way since Bertin's seminal work on the Semiology of Graphics [1]. Years of visualization research has led to systems [17, 18] and guidelines [5, 24] that aid the designer in choosing visual representations based on general data characteristics such as dimensionality and datatype. Unfortunately, many aspects of visualization design still remain more art than science. For example, given a set of a data characteristics, there are almost always multiple visualizations (*forms*) that are theoretically valid and therefore difficult to choose from [18]. In addition, beyond selecting a visualization form, the designer must also take into account many other aspects of the visualization. Examples include design elements such as color, shape, glyph size, as well as usage considerations such as context, and user profile. With so many options, it is tremendously difficult for even experienced designers to identify the most accurate and appropriate visualization given a dataset.

One method for objectively identifying the "best" visualization is to conduct multi-factor human-subject experiments. In these experiments, each design or usage consideration is incorporated as an experimental factor, often resulting in a large number of conditions. While these experiments produce actionable results, they are difficult to generalize beyond the scope of the experiment and provide limited explanation as to why one visualization is better than another. As visualization becomes more widely adopted and diversified, it is clear that exhaustive comparative experimentation cannot fully address the growing needs of the infovis community.

What is needed then are quantitative and robust models for visualizations that are generalizable beyond one-off comparative studies while still providing designers with actionable information about tradeoffs between "valid" visualization forms. Although such models challenge conventional wisdom in information visualization design [13], recent research has suggested that the use of perceptual laws from psychology and cognitive science [12, 3] can be applied to model how humans perceive certain data properties given a visualization. In particular, Rensink et al. successfully demonstrated that the perception of positive correlation in scatterplots can be modeled using Weber's law [22], which indicates that the human perception of differences in correlation and the objective differences in data correlation has a linear relationship:

$$dp = k\frac{dS}{S} \tag{1}$$

where dp is the differential change in perception, dS is the differential increase in the data correlation (change in stimulus), and S is the overall correlation in the data (stimulus). k is known as the Weber fraction, and is derived experimentally. Taken together, this equation and experimentally-inferred parameter k form a *Weber model* for the perception of positive correlation in scatterplots.

What is significant about this finding by Rensink et al. is that it describes the perception of correlation in a concise, quantitative manner via the derived Weber model. The authors hypothesize that if other visualizations of correlation could be shown to follow Weber's law, then it might be possible to compare them without exhaustive empirical testing [22]. Another significant benefit of establishing perceptual models for visualization is that it provides a predictive and falsifiable baseline to investigate the effect of design elements *within* a visual form. For example, in followup studies Rensink et al. used the original Weber model to study whether the effectiveness of scatterplots was impacted by changes in design elements such as point color, brightness, size, canvas aspect ratio, and others [21]. With the Weber model for scatterplots as a baseline, the authors demonstrated that it was possible to determine when the perception of positive correlation in scatterplots was invariant to design elements, and did so without resorting exhaustive multi-factor testing. Therefore, if the perception of correlation in commonly used visualizations can be shown to follow Weber's law, we gain the ability to quantitatively compare and rank the effectiveness of visualizations, as well as a baseline to explore the effect of design elements on the basis of perceptual laws.

In this paper, we confirm the hypothesis of Rensink et al. by demonstrating that nine commonly used visualizations also follow Weber's law, and that the perception of correlation across multiple valid visual forms can be quantified, compared, and ranked using the derived Weber models. After adapting the experimental methodology used by Rensink et al. for a crowdsourcing environment, we first validate our approach by replicating their original findings with scatterplots on Amazon's Mechanical Turk. We then apply this methodology to eight other visualizations commonly used in the infovis community and commercial software, including parallel coordinates plots, stacked area charts, stacked bar charts, stacked line charts, line charts, ordered line charts, radar charts, and donut charts (see Figure 3).

The results of this experiment contribute to our understanding of information visualization in several ways:

- We demonstrate that the perception of correlation in several commonly used visualizations can be modeled using Weber's law.
- We provide evidence that the effectiveness of most visualizations tested depends significantly on whether it depicts positively or

- *Lane Harrison, Fumeng Yang, and Remco Chang are with Tufts University. E-mail: [lane,fyang,remco]@cs.tufts.edu.*
- *Steven Franconeri is with Northwestern University. E-mail: franconeri@northwestern.edu.*

Manuscript received 31 Mar. 2014; accepted 1 Aug. 2014. Date of publication 11 Aug. 2014; date of current version 9 Nov. 2014.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.
Digital Object Identifier 10.1109/TVCG.2014.2346979

1077-2626 © 2014 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

negatively correlated data (asymmetric performance), implying that many visualizations may require two Weber models to be described completely.

- Using the derived Weber models, we rank the effectiveness of visualizations for representing correlation.

In the following section, we discuss related work including perceptual studies for visualization and recent advances in evaluation methodologies. We then describe our first experiment, where we adapt the methodology of Rensink et al. for a crowdsourced environment and replicate their original findings. Following the replication, we present our full experiment evaluating eight other visualizations, including the resulting analyses and models. Our discussion highlights the implications of our findings, such as how our results make it possible to quantify, compare, and rank the effectiveness of visualizations for depicting correlation on the basis of a perceptual law. We also point out several surprising differences and similarities in the visualizations tested, offering possible explanations based on recent work in vision science.

2 Related Work

A key component of modeling a perceptual process using Weber's law is the need to experimentally determine how much a given stimulus must increase/decrease before humans can reliably detect changes, a quantity called the just-noticeable difference (JND) [6]. Although they are a key part of the methodology we use in this paper, the notion of using JNDs to advance visualization design is not entirely new. Color models that approach perceptual-uniformity, such as the CIELAB space, are the result of years of experiments that examine perceptual distances between colors [23]. In these color spaces, two colors that have different RGB values but are perceived as being the same are said to be within one JND of each other. Perceptually-driven color models have led to important advances in infovis, for example the popular ColorBrewer tools [9], and more recently the development of algorithms that automatically select effective color schemes for visualization [15]. Motivated by the success of these approaches, which quantify the perceptual space of color, our work similarly seeks to explore and quantify the perceptual space of visualization forms.

Many perceptual studies have examined the perception of correlation in scatterplots. Early work from Cleveland et al. suggested that subjective estimations of correlation could be biased by increasing the scale of the axes in a scatterplot [4]. Combining perceptual studies of scatterplots and visualization design, Fink et al. integrated participant preferences to develop an automatic method for selecting effective aspect-ratios for scatterplots [8]. Li and van Wijk examined differences in the subjective judgments of correlation in scatterplots and parallel coordinates, finding scatterplots to perform better [14]. The key difference between these approaches and ours is that they do not explicitly link their results to underlying perceptual laws, limiting the generalizability of their results.

One of the primary goals of this work is to provide a means to evaluate and compare visualization effectiveness on the basis of perceptual laws. It is useful, therefore, to situate our contributions in the context of existing approaches to visualization evaluation. To better bridge the gap between design goals and evaluation methodologies, Munzner's Nested Model arranges the design and evaluation space into four interrelated levels [20]. More recently, Lam et al. conducted an extensive survey of infovis publications, distilling them into seven evaluation scenarios [13]. Subsuming both qualitative and quantitative approaches, these evaluation models allows researchers and practitioners to better identify the appropriate level(s) at which a given visualization should be evaluated. Carpendale has pointed out, however, that an inherent limitation of many evaluation approaches is that they are difficult to generalize to different usage contexts [2]. A recent proposal from Demiralp et al. seeks to address these limitations by developing visualization generation and evaluation methods that map perceptual distances between visual elements to similar structures in data [7]. Our work contributes to these evaluation research directions in two ways.

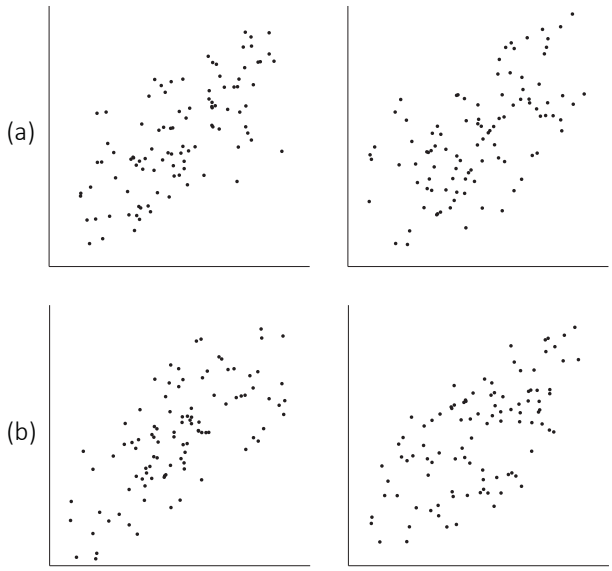

Fig. 1: a) A sample starting comparison from the experiment: $r = 0.7$ on the left and $r = 0.6$ on the right. Participants were asked to choose which of the two appeared to be more highly correlated. b) The staircase procedure hones in on the just-noticeable difference by gradually making comparisons more difficult: $r = 0.7$ on the left and $r = 0.65$ on the right.

First, we demonstrate that visualization effectiveness can be quantified and evaluated using perceptual laws. Additionally, we provide Weber models describing the perception of correlation in several commonly used visualizations, which can be used as a baseline to investigate the impact of individual design elements.

3 Experiment 1: Replication and Crowdsourcing Validation

Given our goal of testing a wide range of visualizations, we turn to a crowdsourcing platform to recruit the necessary number of participants. However, since our goal is to leverage experiment methodologies from vision science, which have not been previously validated for crowdsourcing [16], it is necessary to first replicate the original experiment by Rensink and Baldridge on modeling the perception of positive correlation in scatterplots using Weber's law [22].

In particular, our experiment seeks to confirm the "precision" portion of Rensink and Baldridge's original experiment. As the authors note in their paper, "precision" and "accuracy" are examined through different experiment methodologies. Precision, in their experiment, refers to the ability of participants to detect differences between two correlations, even if they are blind to the actual numerical correlation values. Accuracy, on the other hand, corresponds to participants' bias towards systematically over- or under-estimating correlation values (see [22] for more on these differences). The results of their experiments demonstrated, however, that precision and accuracy for the perception of correlation in scatterplots are "systematically linked" via Weber's law. Given this result, we restrict the scope of our experiment to investigate whether in-lab results for inferring *precision* can be replicated using crowdsourcing.

3.1 Materials

Following the experimental design of Rensink and Baldridge [22], scatterplots in this experiment were all 300×300 pixels, contained 100 normally distributed points along the 45 degree line, used the same point size of 2 pixels, and displayed both the left and bottom axes.

To generate correlated data for a target correlation value r, $n = 100$ data points were first taken a standard normal distribution within 2

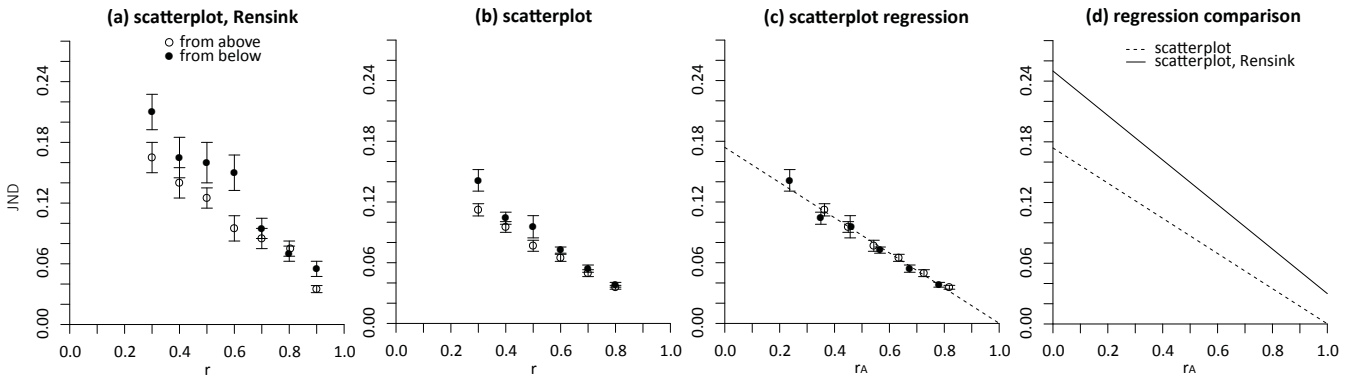

Fig. 2: Experiment 1 replicated the results from [22] on Amazon's Mechanical Turk, validating the platform for our larger experiment. (a) shows the raw data from previous studies [22]. (b) shows the raw data in Experiment 1. (c) shows the data after adjustment, where JNDs are drawn as functions of adjusted correlation r_A. (d) compares previous regression results [22] and the regression from Experiment 1.

standard deviations of the mean and normalized. The correlation coefficient of this starting dataset is then computed and noted as r_z. Then, each point (x_i, y_i) is transformed using the same transformation in [22]:

$$y_i' = \frac{\lambda x_i + (1-\lambda) y_i}{\sqrt{\lambda^2 + (1-\lambda)^2}} \tag{2}$$

where λ is defined as follows:

$$\lambda = \frac{(r_z - 1)(r^2 + r_z) + \sqrt{r^2 (r_z^2 - 1)(r^2 - 1)}}{(r_z - 1)(2r^2 + r_z - 1)} \tag{3}$$

Note that our equation for λ differs from that of Rensink and Baldridge [22]. Specifically, rather than using an estimation approach for computing λ, we instead incorporate the correlation value of the starting dataset r_z. Our extension of this method uses the same transformation and parameters but a) converges more quickly, and b) eliminates error in the original approach (target $r \pm 0.005$). After the dataset is generated, it is re-normalized and transformed to have a mean of 0.5 and standard deviation of 0.2 (following [22]).

3.2 Methodology

Following Rensink and Baldridge [22], we use the same adaptive psychophysical method, a staircase procedure, to infer just-noticeable differences (JNDs) for the perception of correlation. This experimental procedure has a 6×2 design in that there are six correlation r values (0.3, 0.4, 0.5, 0.6, 0.7 and 0.8) and two *approach* conditions (above and below).

In the staircase procedure, given a target value for correlation, r, participants are given two visualization stimuli side-by-side (two scatterplots in this case) and asked to choose which they perceive to have a higher correlation (see Figure 1). With an "above" approach, the participant is given one visualization with the target r, and another with an r value higher than the target. For example, if the target r is 0.7, then the second r value would be 0.8 (assuming a starting distance of 0.1). Conversely, with an "below" approach, the participant would be given a visualization with the target r, and another that has an r value lower than the target.

In both cases, if a participant chooses correctly, the distance in correlation between the two visualizations is decreased by 0.01 while keeping the target r constant (e.g. 0.7 versus 0.79 in the "above" condition, or 0.7 versus 0.61 in the "below" condition). If a participant chooses incorrectly, the distance in correlation between two visualizations is increased by 0.03, making the next judgment easier. The staircase procedure "hones in" on the JND by penalizing incorrect choices more than correct choices. These distance changes (0.01, 0.03) correspond to inferring "75%" JNDs, or the minimum difference in correlation required to be reliably discriminated 75% of the time. After each selection is made, the position of the target and variable visualization is randomized (i.e., whether the target appears as the left vs. right visualization), and new datasets are generated for both stimuli.

The staircase procedure ends when either 50 individual judgments are reached or when a convergence criteria is met. Following [22], the convergence condition is determined based on the last 24 user judgments, and is designed to determine if the user's ability to discriminate between the two correlation values depicted has stabilized. Specifically, the 24 user judgments are divided into 3 subgroups, and convergence is reached when there is no significant difference between these three subgroups via an F-test ($F(2,21)$; $\alpha = 0.1$). Finally, after the staircase procedure ends, the average distance in correlation value in these subgroups is used as the JND for the tested r values and approach (above/below).

There are two important limitations to the staircase procedure that we include when reporting results: ceiling effects and the "chance" boundary.

A ceiling effect occurs when the participant fails to perceive correlation reliably and the adaptive algorithm reaches an upper limit for correlation ($r = 1.0$). For example, if the base r value is 0.7 and the r value of the second stimuli reaches 1.0, yet the participant still answers randomly for the remainder of the judgments (roughly 50% accuracy), the resulting JND for $r = 0.7$ will be 0.3. This upper limit (0.1 for 0.9, 0.2 for 0.8, etc.) will be illustrated in our results.

The chance boundary is defined by the parameters of the staircase procedure (convergence criteria, starting distance, number trials, etc.) To obtain this boundary, we ran a simulation of the staircase procedure 10,000 times, simulating a participant guessing at chance (50% accuracy). The resulting boundary in our procedure was $JND = 0.45$, meaning that any resulting JNDs at or above this boundary would indicate that the participant did not reliably perceive correlation. There are two possible cases where a participant performs at chance throughout an experiment. The first is when a participant is simply making random choices, which is possible given that we are using a crowdsourcing platform [11, 19]. The second possibility occurs when the actual JNDs for a given stimuli are at or near 0.45, forcing the participant to guess throughout most of the judgments. This chance boundary ($JND = 0.45$) is illustrated all of our results figures, and will be used to establish an exclusion criteria for poorly performing visualizations.

3.3 Procedure

The conditions in this experiment include the six correlation value tested, and whether the target correlations were approached from above or below. Informed by early pilot testing, participants were randomly assigned to two correlation values: one from [0.3, 0.4, 0.5], and one from [0.6, 0.7, 0.8]. These groups roughly correspond to "hard" and "easy", since high correlations are more easily discriminated than low correlations. For the two correlation values chosen, participants

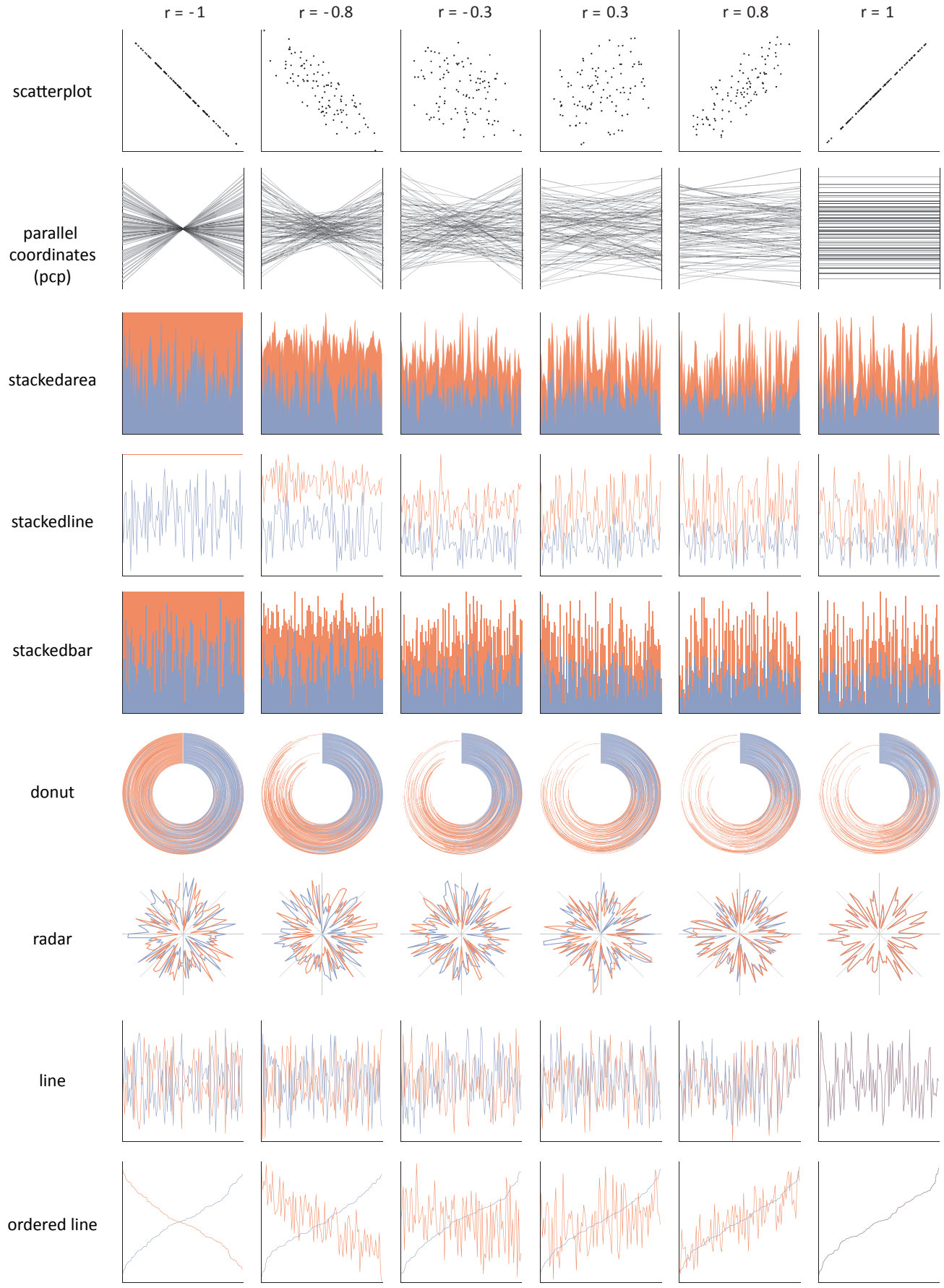

Fig. 3: The nine visualizations tested in our experiment, at several correlation values. Because many of these visualizations appear differently when depicting negatively versus positively correlated data, we test both in our experiment. The visualizations were larger (300×300 pixels) when presented to participants. The color scheme used is colorblind-safe, chosen from ColorBrewer.

complete both the above and below approach, resulting in a total of four trials (easy×above, easy×below, hard×above and hard×below). This corresponds to collecting up to 200 individual judgments for each participant.

Since participants on Amazon's Mechanical Turk (AMT) come from diverse educational/statistical backgrounds [19], both a training and practice session were added before the main trials began. The training session consisted of a short definition of correlation, including a grid of ten scatterplots showing correlations ranging from 0.1 to 1.0. After the training session, participants were given a practice session consisting of 30 individual judgments. In the first 15 judgments, participants were shown the "easy" (high) correlation they would be working with; in the second 15 they were shown the "hard" (low) correlation condition. After each judgment, participants were given feedback on whether they chose correctly.

After completing the training and practice sessions, participants began the four main trials. The order of the correlation-approach pairs was randomized in this session. Upon completing a trial set (either by reaching the convergence criteria or 50 individual judgments), participants were given the option to take a short break, and notified as to how many experiment trials remained. Following the completion of all four trials, a demographics questionnaire was given, which included a question that asked participants to describe the strategy they used to assess correlation. Finally, participants were given a short debrief explaining the purpose of the experiment.

3.4 Results

We recruited $n = 88$ participants (36 female) for this experiment via Amazon's Mechanical Turk (AMT). It took approximately two days to gather all responses. Participants were paid $2.10 for their time, commensurate with the US minimum wage. To avoid possible confounds from participants using mobile devices or tablets, such devices were programmatically blocked from accessing the experiment. This experiment adhered to a between-subjects design, since participants were randomly assigned to complete two correlation values (out of six) for both above and below approaches. While there were $n = 20$ participants for each correlation-approach in [22], in our experiment we recruited approximately $n = 30$ for each pair in order to account for the inherent variability in AMT worker responses [11, 19].

Our results indicate that crowdsourcing effectively replicates measurements obtained in-lab. Individual JND and error data were not published in [22], eliminating the possibility for direct statistical comparison. However, we estimated JNDs and errors from figures, and compare them with our results in Figure 2. The general trends between results obtained in-lab and via crowdsourcing are similar, including both the higher error-bars for lower correlation values, and small JNDs for higher correlations.

To determine if our results also can be modeled using Weber's law, we follow the model-fitting procedure in [22]. Specifically, each correlation value r was moved by half of the average JND from the above and below approach. For the above approach, the correlation r was moved towards $r = 1$, while the r from the below approach was moved towards $r = 0$. Specifically, correlation r is transformed into adjusted-correlation r_A by:

$$r_A = r \pm 0.5\, jnd(r) \tag{4}$$

Figure 2 illustrates this intermediate step, showing all points in our data after adjustment. Linear models were then fit to the data, and errors were computed based on the square root of the mean squares of the residuals (RMS error).

Following this procedure, we find the model to be a good fit ($r^2 = 0.98$), indicating that our crowdsourcing results also follow Weber's law. The resulting regression lines are shown in Figure 2. Although our results have a slightly lower intercept (0.21 cf. 0.25), indicating better performance overall, the slopes are also similar (-0.17 cf. -0.22). Given the similarity in both the JND and regression results, we find this to be evidence that the AMT crowdsourcing platform is appropriate for our experiments.

4 Experiment 2: Extension to Other Visualizations

Experiment 1 established that results obtained via crowdsourcing replicate Rensink and Baldridge's original experiment that modeled the perception of positive correlation in scatterplots using Weber's law. In Experiment 2, we extend this experiment to investigate whether the perception of correlation in other commonly used visualizations can also be modeled using Weber's law.

4.1 Materials

We chose nine visualizations for this experiment based on two main criteria: a) they must be commonly used in either infovis or commercial software (external validity), and b) they must be viable within the constraints of the experiment methodology. The nine visualizations chosen include: scatterplots, parallel coordinates plots, stacked area charts, stacked bar charts, stacked line charts, line charts, ordered line charts, radar charts, and donut charts.

Scatterplots were included both because of their widespread use in the scientific community, and to serve as a baseline for replicating the results of Rensink and Baldridge [22]. Parallel coordinates plots were included because of their continued widespread use in the infovis community.

Line charts, stacked line charts, stacked area charts, and stacked bar charts were included based on the top recommendations when viewing one of our datasets (100 points, 2 dimensions) in Microsoft Excel. Despite their similarities, all of the stacked charts (line, area, bar) were included since any significant differences in these charts might shed light on the underlying perceptual processes people use when judging correlation.

Donut charts and radar charts were included since they are essentially radial transforms of stacked bar charts and line charts, respectively. Comparing these may allow us to understand the effect of coordinate transforms on the perception of correlation.

Note that, of all the visualizations tested, scatterplots and parallel coordinates plots are the only two that are truly *bivariate*, in that the two quantitative variables in the data displayed (X, Y) determine the exact position of the graphical elements. In contrast, all of the other visualizations contain an additional explicit variable *order*. In other words, scatterplots and parallel coordinates plots are ordered by default, while any of the other visualizations can be ordered in different ways. To test the effect that manipulating order may have on the perception of correlation, ordered line charts (sorted on the X-axis) were also added to the experiment.

Finally, we hypothesize that the performance of these visualizations may be impacted by the fact that many of them appear very differently when depicting positively versus negatively correlated data (see Figure 3). To test this hypothesis, we test each of these nine visualizations were twice: once with positively correlated data, and once with negatively correlated data.

As in experiment 1, all visualizations were 300×300 pixels, contained 100 data points and displayed datasets generated from same algorithm. For visualizations that required more than one color, we chose a single color scheme from ColorBrewer [9] [1]. All visualizations used in this experiment are illustrated for several correlation values in Figure 3.

4.2 Procedure

The procedure for this experiment follows that of Experiment 1, except for the following two changes. To mitigate possible confounds from exposing some participants to negative (versus positive) correlations in the training session, we label the correlation grid (0.1 to 1.0) with the same labels regardless of the correlation direction, and provide a short disclaimer explaining the change for participants familiar with correlation. Secondly, since early pilot testing showed that many of the visualizations had higher JNDs than scatterplots for low correlations, we allow the staircase procedure to move above or below $r = 0.0$ if necessary.

[1] http://colorbrewer2.org/?type=qualitative&scheme=Set2&n=3

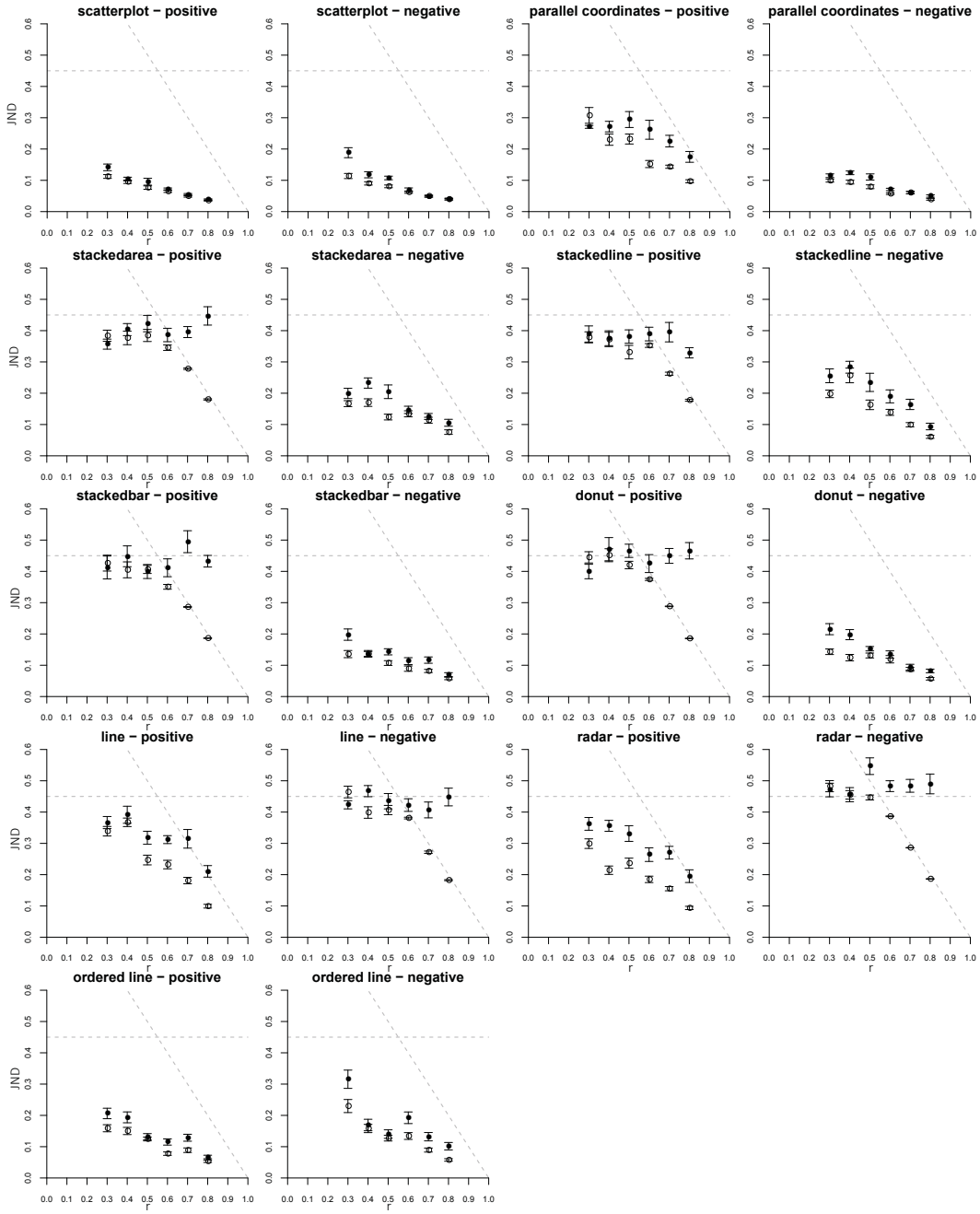

Fig. 4: JND plotted as a function of correlation r for both above (light points) and below (dark points) approaches. Error bars show the standard-error of the mean (SEM). Broken lines show the chance and ceiling boundaries defined in Section 3.2. The x-axis is correlation value r, the y-axis is JND.

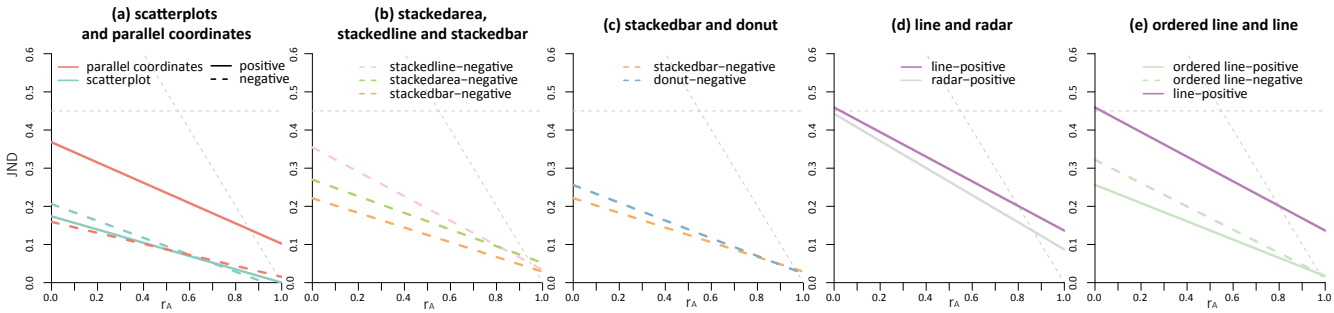

Fig. 5: Regression results for several paired comparisons. JNDs are modeled as functions of adjusted correlation r_A. The x-axis is adjusted correlation value r_A, the y-axis is JND.

Table 1: *Mann-Whitney-Wilcoxon Tests Results for condition pairs. Significant values denoted by * with* $\alpha = 0.0036$.

visualization - direction 1	visualization - direction 2	W	p-value
scatterplot - negative	scatterplot - positive	51165.5	0.54
scatterplot - negative	parallel coordinates - positive	10885.5	$< 0.001^{\star}$
scatterplot - positive	parallel coordinates - positive	8623	$< 0.001^{\star}$
parallel coordinates - negative	scatterplot - negative	51291	0.42
parallel coordinates - negative	scatterplot - positive	51491	0.16
parallel coordinates - negative	parallel coordinates - positive	8641.5	$< 0.001^{\star}$
stacked bar - negative	stacked line - negative	34421	$< 0.001^{\star}$
stacked bar - negative	stacked area - negative	33348.5	$< 0.001^{\star}$
stacked bar - negative	donut - negative	43361	0.037
stacked line - negative	stacked area - negative	66646	0.014
line - positive	radar - positive	73775.5	$0.0017^{\star}$
line - positive	ordered line - positive	104163.5	$< 0.001^{\star}$
line - positive	ordered line - negative	101883	$< 0.001^{\star}$
ordered line - negative	ordered line - positive	66292	0.0075

Table 2: *Intercepts, Slopes, Correlation Coefficients r,* r^2*, and RMS for JNDs modeled as functions of adjusted correlation* r_A.

visualization - direction	intercept-b	slope-k	correlation-r	r^2	RMS
scatterplot - positive	0.17	-0.17	-0.99	0.98	0.0041
scatterplot - negative	0.21	-0.22	-0.95	0.90	0.013
parallel coordinates - positive	0.37	-0.27	-0.86	0.74	0.032
parallel coordinates - negative	0.16	-0.14	-0.95	0.90	0.0085
stacked line - negative	0.35	-0.32	-0.92	0.84	0.027
stacked area - negative	0.27	-0.22	-0.93	0.86	0.016
stacked bar - negative	0.22	-0.19	-0.95	0.90	0.011
donut - negative	0.26	-0.23	-0.96	0.93	0.012
line - positive	0.46	-0.32	-0.86	0.74	0.043
radar - positive	0.44	-0.36	-0.95	0.91	0.024
ordered line - positive	0.26	-0.24	-0.95	0.91	0.014
ordered line - negative	0.32	-0.31	-0.88	0.78	0.031

4.3 Results

We recruited 1,687 participants through AMT (834 female) for this experiment. It took approximately two weeks to gather all responses. Our study used a total of nine visualizations, two correlation directions (positive/negative), and six correlation values (0.3 to 0.8) yielding 54 main groups. Since each participant was assigned to one visualization, one correlation direction, and two correlation values (above and below), roughly 30 participants were assigned to each visualization×direction×r-value group.

The resulting data were non-normally distributed, so to mitigate the effect of outliers, JNDs that fell outside 3 median-absolute deviations from the median (within one of the 54 groups) were excluded from the following analyses. Because the staircase methodology penalizes incorrect responses and controls for guessing by defining a convergence criteria (see Section 3.2), this exclusion criteria also mitigates the effect of "click-through" responses that often impact crowdsourced experiments [11, 19]. Figure 4 shows average JNDs for all visualizations, correlation values, and approaches tested after filtering.

An exlusion criteria was also enforced for visualization×direction pairs that exceeded 20% of values falling on or outside the "chance" boundary of $JND = 0.45$ established previously. Six of the eighteen pairs met this exclusion criteria: stacked area-positive, stacked bar-positive, stacked line-positive, donut-positive, radar-negative, and line-negative (see Section 5.2 for more details on the exclusion criteria). The following analyses include the remaining twelve visualization×direction pairs.

4.4 Weber Model Fit

Following the same model fitting procedure as Experiment 1 (and [22]), slopes, intercepts, correlation coefficients and root-mean-square errors (RMS) were computed. These statistics are included in Table 2 and corresponding regression lines illustrated in Figure 6. All visualizations follow a linear relationship between JNDs and adjusted correlation r_A, based on the high correlation coefficient and small RMS for each. Regression lines for the following statistical comparisons are shown in Figure 5.

4.5 Statistical Analyses

Examining the JND data alone, there appear to be large differences in performance between many of the visualizations, as well as asymmetries between many of the positive/negative pairs. In order to confirm these observations, an overall Kruskal-Wallis test was conducted on the raw JNDs to evaluate whether there is an interaction between visualization and correlation direction conditions. This test confirmed there was an overall effect for visualization×correlation direction ($\chi^2(17) = 3147.70, p < 0.001, \alpha = 0.05$).

To explore further, several visualization×direction pairs were compared via Mann-Whitney-Wilcoxon tests. Rather than compare all possible pairs, we instead investigate 14 pairings reflecting the original motivations for choosing the visualizations tested (see the Materials section 4.1). We use Bonferonni correction to address the problem of multiple comparisons, resulting in an $\alpha = 0.0036$ required for rejecting the null hypothesis. All tests results and parameters are reported in Table 1. To aid visual comparisons, we provide regression lines corresponding to these comparisons in Figure 5.

Examining the comparison results for scatterplots in Table 1, we find no significant difference between scatterplots depicting positively and negatively correlated data ($p = 0.54$, see Figure 5.a). In addition to their symmetric performance, scatterplots appear to be among the best performing visualizations (see Figure 6).

We find clear evidence of asymmetry in parallel coordinates plots, with those depicting negatively correlated data significantly outperforming those depicting positively correlated data ($p < 0.001$, see Figure 5.a). Futhermore, we find that parallel coordinates plots depicting negatively correlated data were not significantly different from scatter-

better

r = 0.1 *	r = 0.3	r = 0.5	r = 0.7	r = 0.9 *	overall
pcp–negative	pcp–negative	scatterplot–positive	scatterplot–negative	scatterplot–negative	scatterplot–positive
scatterplot–positive	scatterplot–positive	pcp–negative	scatterplot–positive	scatterplot–positive	pcp–negative
scatterplot–negative	scatterplot–negative	scatterplot–negative	pcp–negative	pcp–negative	scatterplot–negative
stackedbar–negative	stackedbar–negative	stackedbar–negative	stackedbar–negative	ordered line–positive	stackedbar–negative
ordered line–positive	ordered line–positive	ordered line–positive	ordered line–positive	donut–negative	ordered line–positive
donut–negative	donut–negative	donut–negative	donut–negative	ordered line–negative	donut–negative
stackarea–negative	stackarea–negative	stackarea–negative	ordered line–negative	stackedbar–negative	stackarea–negative
ordered line–negative	ordered line–negative	ordered line–negative	stackarea–negative	stackedline–negative	ordered line–negative
stackedline–negative	stackedline–negative	stackedline–negative	stackedline–negative	stackarea–negative	stackedline–negative
pcp–positive	pcp–positive	pcp–positive	pcp–positive	radar–positive	pcp–positive
radar–positive	radar–positive	radar–positive	radar–positive	pcp–positive	radar–positive
line–positive	line–positive	line–positive	line–positive	line–positive	line–positive

Fig. 7: Using the inferred Weber models, we can produce a perceptually-driven ranking for individual correlation (r) values, as well as an overall ranking (right column). Performance is ordered from the best (top) to the worst (bottom). The columns denoted by * are predicted responses using the fit models shown in Figure 6.

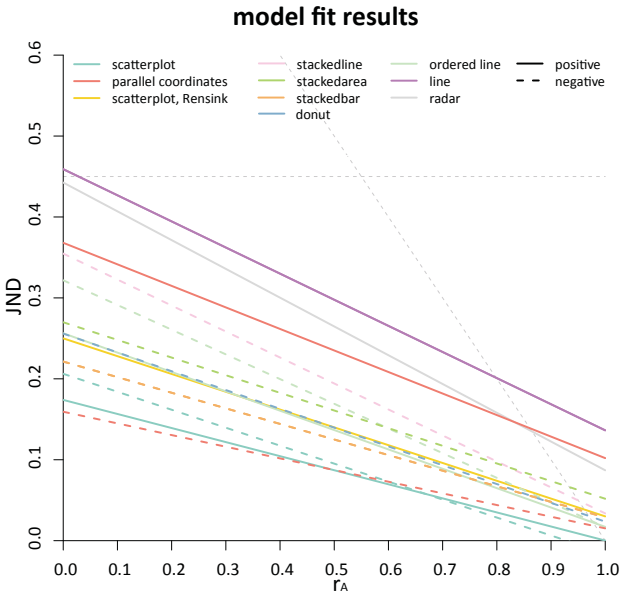

Fig. 6: Regression results from Experiment 2 (including results from [22]). JNDs are modeled as functions of adjusted correlation r_A.

plots depicting either positively ($p = 0.16$) and negatively ($p = 0.42$) correlated data.

Between the stacked chart variants (stacked bar, stacked area, stacked line) depicting negatively correlated data, we find that stacked bar charts significantly outperform both stacked area and stacked line charts (both $p < 0.001$, see Figure 5.b), and that there was no significant difference between stacked area and stacked line charts ($p = 0.014$). We also find no significant difference between the stacked bar chart and the donut chart ($p = 0.037$, see Figure 5.c).

Between the line chart variants (line chart, radar, ordered line) depicting positively correlated data, the ordered line chart significantly outperformed both the line chart ($p < 0.001$) and the radar chart ($p < 0.001$), (see Figure 5.d and 5.e). In fact, the ordered line chart is the only other chart besides scatterplots to show symmetric performance (positive and negative difference $p = 0.0075$). The radar chart also performed significantly better than the line chart ($p = 0.0017$).

5 Discussion

Our results demonstrate that there are significant differences in the perception of correlation across visualizations, and that these results often vary significantly when depicting positively versus negatively correlated data.

Between scatterplots and parallel coordinates, we find that using scatterplots to depict correlation results in better performance overall. However, this performance difference only occurs when depicting positively correlated data. In fact, parallel coordinates depicting negatively correlated data appear to perform as well as scatterplots (see Figure 5.a).

Among the stacked chart variants (stacked bar charts, stacked area charts, and stacked line charts), the stacked bar significantly outperformed both the stacked area and stacked line (see Figure 5.b). This finding suggests that although these visualizations appear to be similar, the underlying perceptual processes that participants use when judging correlation in them may differ substantially (for more discussion, see Section 5.3).

While there was no difference between the stacked bar chart and the donut chart, the radar chart significantly outperformed the line chart, indicating that coordinate transforms may yield inconsistent performance implications.

Some of these findings can be directly applied to inform visualization design. For example, because parallel coordinates plots depicting negatively correlated data significantly outperform those depicting positively correlated data, new layout algorithms could be developed to maximize the number of negative correlations depicted by flipping and re-arranging axes. However, since our experiment results establish that the perception of correlation in these visualizations can be modeled using Weber's law, it also becomes possible to rank the effectiveness of the tested visualizations.

5.1 Ranking

One of the primary questions this experiment sought to explore is whether the effectiveness of visualizations for depicting correlation can be quantified, compared, and ranked on the basis of a perceptual

law. Since we have inferred Weber models for each of the nine visualizations tested, this ranking becomes possible.

Recall that we tested each of the nine visualizations with both positively and negatively correlated data, for a total of eighteen visualization×correlation-direction pairs. However, since six of the eighteen pairs met our exclusion criteria, we include the remaining twelve models in our ranking.

Using the inferred Weber models, we produce a ranking for six correlation values (see Figure 7). Each of the visualization×correlation-direction pairs is ordered by performance, with the best in the top row, and the worst in the bottom row. Note that using Weber's law allows us to make predictions for the perception of correlation r values that were not explicitly tested in the experiment (e.g. 0.1 and 0.9). The ranking order for each of correlation value varies due to crossings in the Weber models (see Figure 6). While ranking visualizations within individual correlation values can be useful for design, an overall "best" ranking is also desirable. One straightforward way to obtain an overall ranking is to identify the visualization which has the lowest JND on average. This can be computed by calculating the area of the regions between the regression lines and the r_A-axis [10] (see the rightmost column in Figure 7).

This ranking has many potential applications. One possible direction is to define a "tolerable" range of effectiveness, in order to restrict the possible design space to fewer visualizations. For example, if a designer needs to reliably communicate correlation for a given dataset, they can refer to this model and obtain a precise ranking based on the correlation values in their dataset. Another possible application is in visualization system design, where it may be helpful to use either the overall ranking, or to obtain a custom ranking based on a range of correlation values (e.g. a scientific application may require identifying the most effective visualizations for correlations above 0.6).

5.2 Limitations of the Methodology

Although we were able to demonstrate that each of the nine visualizations followed Weber's law, we found that six visualization×correlation-direction pairs produced unreliable results and therefore were excluded from the rest of the analyses. Specifically, the six pairs excluded were: stacked area-positive, stacked bar-positive, stacked line-positive, donut-positive, radar-negative, and line-negative charts.

The exclusion criteria was based on the upper limit ("chance") boundary defined in Section 3.2. Recall that this boundary ($JND = 0.45$) is a function of the staircase procedure parameters such as the starting distance (0.1) and correct/incorrect penalties (0.01 and 0.03 respectively). Since the JNDs for the excluded visualization×correlation-direction pairs frequently met the upper limit of our staircase procedure, it is possible that either correlation is not reliably perceived in these visualization×correlation-direction pairs, or that a larger starting distance (and corresponding penalties) is required to infer reliable JND results for these visualizations.

While each visualization tested was shown to follow Weber's law for least one correlation-direction (positive and/or negative), we cannot say for sure whether the six excluded visualization×correlation-direction pairs also follow Weber's law. Examining the underlying perceptual processes involved in judging correlation for each of these visualizations, however, may allow us to identify the reasons for the observed performance variations.

5.3 Visual Features

Reviewing Figure 3, we observe that many of the visualizations tested vary significantly when depicting positively versus negatively correlated data. In most cases, the visualizations appear to have different *visual forms* for the two correlation directions but the same absolute correlation value ($|r|$). For example, in parallel coordinates, when $r = 1$, the visualization depicts a set of parallel lines and has the shape of a square; whereas when $r = -1$, the set of lines intersect at a single point and appear as two triangles.

One exception to this asymmetric relationship in visual forms is the scatterplot. For both positive and negatively correlated data, the visual *form* of a scatterplot converges to a single line when $|r|$ approaches 1. In fact, past studies have hypothesized that the visual feature viewers attend to when making correlation judgments in scatterplots is the width of the bounding box (or ellipse) that surrounds the points [4, 14]. For example, when $|r| = 1$, the width of this bounding box is essentially 0. While further testing is needed to confirm exactly what visual features are perceived in scatterplots, it has long been established that perceptual judgments of line length follow Weber's law [12]. Based on these observations, our results suggest that **the reason the perception of correlation in scatterplots follows Weber's law is because the underlying visual features that vary with correlation follow Weber's law**.

Reasoning about the underlying visual features of visualizations may also explain the significant performance differences between the stacked chart variations. Our results demonstrate that, for negative correlations, stacked bar charts significantly outperform both stacked area charts and stacked line charts. This finding is surprising since the visual *forms* of these three charts are similar. However, based on participant feedback, we see that the visual features employed when judging correlation in these charts might in fact be different:

> "I looked for which average value of the orange line was higher, also taking into account which orange line had fewer peaks/valleys..."

> "It seemed like the less-spiky charts were more correlated."

In contrast, a participant in the stacked bar condition noted:

> "I mostly compared how much white and orange were compared to the blue on each chart. Usually the one with less white was more correlated."

While the *visual forms* for the stacked charts variants are similar (see Figure 3), the *visual features* that convey correlation differ. Since visual features (rather than visual forms) may be the underlying cause that significantly impacts the effectiveness of visualizations, we believe there are several important areas for future work. These include comparing visual features produced by visualizations, identifying how perceptual laws apply to other common visualization tasks, and investigating these findings with real-world datasets and datasets with different characteristics and distributions[2].

6 Conclusion

In this paper, we described a large scale (n=1687) crowdsourced experiment to investigate whether the perception of correlation in nine commonly used visualizations can be modeled using Weber's law. The results of this experiment indicate that all visualizations tested can be modeled using Weber's law, but that the effectiveness of many visualizations varies when depicting negatively or positively correlated data. Furthermore, using the learned Weber models, we rank the effectiveness of the tested visualizations on the basis of a perceptual law. We also introduce the notion of perceptual symmetries (or asymmetries) that emerged from observing significant performance differences in visualizations depicting positively versus negatively correlated data, and suggest that these symmetries might be related to the visual features participants attend to when judging correlation.

Acknowledgments

This work was supported in part by a grant from NSF (IIS-1218170). Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

[2]To facilitate future work, all experiment materials, participant data, and analyses scripts are available on the author's website: http://valt.cs.tufts.edu/papers/ranking-correlation

REFERENCES

[1] J. Bertin. Semiology of graphics: diagrams, networks, maps. 1983.

[2] S. Carpendale. Evaluating information visualizations. In *Information Visualization*, pages 19–45. Springer, 2008.

[3] H. Choo and S. Franconeri. Enumeration of small collections violates weber's law. *Psychonomic bulletin & review*, 21(1):93–99, 2014.

[4] W. S. Cleveland, P. Diaconis, and R. McGill. *Variables on scatterplots look more highly correlated when the scales are increased*. Defense Technical Information Center, 1982.

[5] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):531–554, 1984.

[6] S. Coren, L. Ward, and J. Enns. Sensation and perception. 1999. *Hartcourt Brace, Forth Worth*.

[7] C. Demiralp, C. E. Scheidegger, G. L. Kindlmann, D. H. Laidlaw, and J. Heer. Visual embedding: A model for visualization. *Computer Graphics and Applications, IEEE*, 34(1):10–15, 2014.

[8] M. Fink, J.-H. Haunert, J. Spoerhase, and A. Wolff. Selecting the aspect ratio of a scatter plot based on its delaunay triangulation. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2326–2335, 2013.

[9] M. Harrower and C. A. Brewer. Colorbrewer. org: An online tool for selecting colour schemes for maps. *The Cartographic Journal*, 40(1):27–37, 2003.

[10] S. Hecht. The visual discrimination of intensity and the weber-fechner law. *The Journal of general physiology*, 7(2):235–267, 1924.

[11] J. Heer and M. Bostock. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 203–212. ACM, 2010.

[12] V. A. C. Henmon. *The time of perception as a measure of differences in sensations*. Number 8. Science Press, 1906.

[13] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. Empirical studies in information visualization: Seven scenarios. *Visualization and Computer Graphics, IEEE Transactions on*, 18(9):1520–1536, 2012.

[14] 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.

[15] S. Lin, J. Fortuna, C. Kulkarni, M. Stone, and J. Heer. Selecting semantically-resonant colors for data visualization. In *Computer Graphics Forum*, volume 32, pages 401–410. Wiley Online Library, 2013.

[16] G. Little, L. B. Chilton, M. Goldman, and R. C. Miller. Turkit: human computation algorithms on mechanical turk. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 57–66. ACM, 2010.

[17] J. Mackinlay. Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics (TOG)*, 5(2):110–141, 1986.

[18] J. Mackinlay, P. Hanrahan, and C. Stolte. Show me: Automatic presentation for visual analysis. *Visualization and Computer Graphics, IEEE Transactions on*, 13(6):1137–1144, 2007.

[19] W. Mason and S. Suri. Conducting behavioral research on amazon's mechanical turk. *Behavior research methods*, 44(1):1–23, 2012.

[20] T. Munzner. A nested model for visualization design and validation. *Visualization and Computer Graphics, IEEE Transactions on*, 15(6):921–928, 2009.

[21] R. A. Rensink. On the prospects for a science of visualization. In *Handbook of Human Centric Visualization*, pages 147–175. Springer, 2014.

[22] R. A. Rensink and G. Baldridge. The perception of correlation in scatterplots. In *Computer Graphics Forum*, volume 29, pages 1203–1210. Wiley Online Library, 2010.

[23] G. Sharma. *Digital color imaging handbook*, volume 11. CRC, 2002.

[24] J. Zacks and B. Tversky. Bars and lines: A study of graphic communication. *Memory & Cognition*, 27(6):1073–1079, 1999.