Beyond Weber's Law: A Second Look at Ranking Visualizations of Correlation

Matthew Kay and Jeffrey Heer

Abstract—Models of human perception – including perceptual "laws" – can be valuable tools for deriving visualization design recommendations. However, it is important to assess the explanatory power of such models when using them to inform design. We present a secondary analysis of data previously used to rank the effectiveness of bivariate visualizations for assessing correlation (measured with Pearson's *r*) according to the well-known Weber-Fechner Law. Beginning with the model of Harrison *et al.* [1], we present a sequence of refinements including incorporation of individual differences, log transformation, censored regression, and adoption of Bayesian statistics. Our model incorporates all observations dropped from the original analysis, including data near ceilings caused by the data collection process and entire visualizations dropped due to large numbers of observations worse than chance. This model deviates from Weber's Law, but provides improved predictive accuracy and generalization. Using Bayesian credibility intervals, we derive a partial ranking that groups visualizations with similar performance, and we give precise estimates of the difference in performance between these groups. We find that compared to other visualizations, scatterplots are unique in combining low variance between individuals and high precision on both positively- and negatively- correlated data. We conclude with a discussion of the value of data sharing and replication, and share implications for modeling similar experimental data.

Index Terms—Weber's law, perception of correlation, log transformation, censored regression, Bayesian methods

1 Introduction

Perceptual laws, such as Weber's Law, offer the tantalizing potential to explain differences in the performance of visualizations, simplify design recommendations, and drive automated visualization systems. However, many such laws have long been studied – not without controversy [2] – using simple models that aggregate individual differences before modelling. Harrison *et al.* [1] follows from this tradition, investigating the relationship between the correlation of two variables (measured using Pearson's *r*) and the precision of people's estimates of that correlation using different visualizations (measured using just-noticeable differences). In accordance with Weber's Law, they fit linear regressions to the means of the just-noticeable differences for each value of *r* in each condition, not to the individual observations directly.

By removing a large portion of the variance in the data (individual differences), they complicate the use of their results in deriving design recommendations. Even if we establish that one visualization is better than another on average, we need to understand how much individual variation plays a part. Given visualization *B* that is slightly worse than visualization *A* on average, but which is more consistently good for a range of people, we might be better to recommend the use of visualization *B* for a broad audience. In other words, a visualization that is slightly better *on average* may still be much worse for some subset of the population. From a design perspective, this is not unlike an architect who designs every home for the average family of 2.6 people. Individuals, not group means, digest visualizations.

In this paper, we conduct a secondary analysis of the data in Harrison *et al.* [1]. Much of this paper focuses on understanding and accounting for individuals' differences in precision of estimation in order to derive parametric models that can predict the expected precision of each visualization technique. Given an appropriate parametric model, we can estimate interpretable differences between visualization types (e.g., as a ratio of just-noticeable differences) in order to judge whether these differences have practical significance.

- *Matthew Kay is with University of Washington. E-mail: mjskay@uw.edu.*
- *Jeffrey Heer is with University of Washington. E-mail: jheer@uw.edu.*

Manuscript received 31 Mar. 2015; accepted 1 Aug. 2015; date of publication xx Aug. 2015; date of current version 25 Oct. 2015.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org.

We begin by revisiting the experimental setup and subsequent data analysis of Harrison *et al*. [1]. We then progress through a series of model refinements, starting with a basic linear model. We first address problems of non-constant variance, presenting evidence that a log-linear model – which does not follow Weber's Law – better describes the relationship between just-noticeable differences and objective correlation. We then augment our model with censored regression to include all observations in the analysis, including outliers, data near ceilings resulting from features of the data collection process, and entire visualizations originally dropped due to large numbers of data points worse than chance. Finally, we adopt a Bayesian model with linear mixed effects, which allows us to account for correlated observations within participants and incorporate knowledge from previous work into our analysis.

This model allows us to directly and quantitatively answer questions left largely unaddressed by the original paper: given a dataset with unknown correlation, how well would we expect each visualization technique to perform (and what is the uncertainty associated with this estimate)? What are the expected differences in performance? Which visualizations are effectively equivalent? We identify clusters of visualizations with similar precision and quantify the expected difference in precision between clusters, yielding a comprehensive set of practical recommendations in the form of a partial ranking of visualizations of correlation. This partial ranking provides concrete guidance to practitioners by grouping visualizations with similar performance and by giving precise estimates of the difference in performance between groups of visualizations. Most concretely, we find that scatterplots are unique in yielding high precision of estimation of correlation for both positively- and negatively- correlated data while also having low variation in performance between individuals. This yields a straightforward design recommendation grounded in data.

Finally, we discuss the applicability of similar models to other problems of estimating the perceptual performance of visualizations from experimental data. Censored regression offers a flexible way to account for a class of experimental artifacts likely to be found in other perceptual experiments in visualization, and examination of individual differences in general yields models with greater explanatory power. Our Bayesian approach also facilitates future work by providing a principled way to build upon our results.

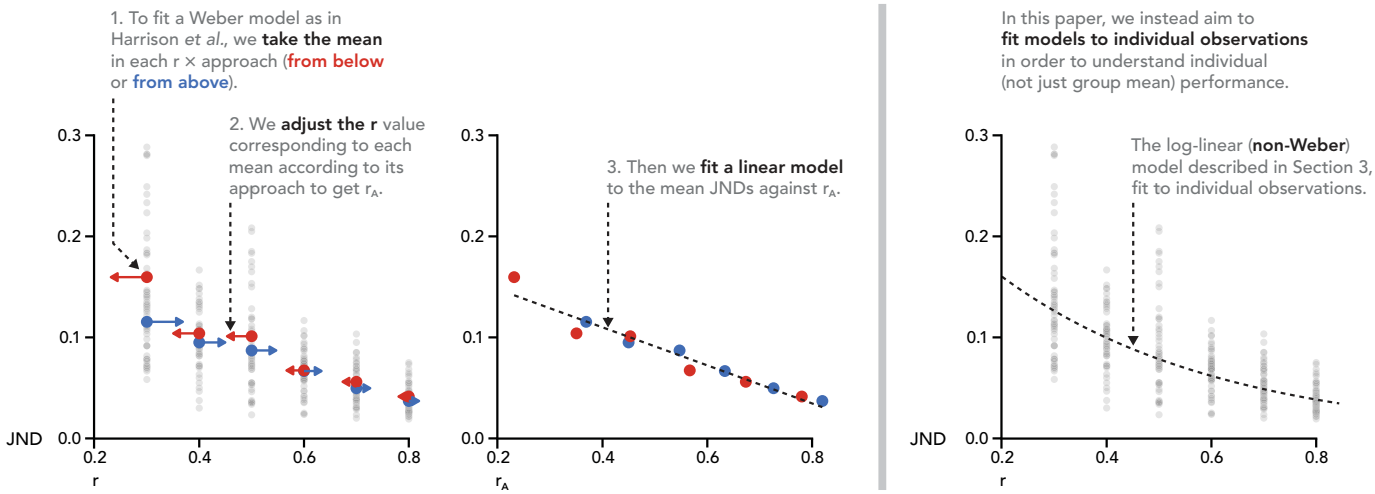

Fig. 1. High-level comparison of the Weber's Law-based modelling approach of Harrison *et al.* [1] and the approach taken in this paper as applied to the scatterplot–positive condition. The approach adjustment (part 2) for the Weber model is necessary because when the approach is from above, JND is underestimated (as higher values of *r* tend to have lower JND), and vice versa when the approach is from below. The correction moves *r* up by half the mean JND at that value of *r* when from above, and down by half the mean JND when from below. See Section 3 for more detail on this problem and the alternative approach we take to addressing it in our model.

2 Background

2.1 Harrison *et al.* experimental setup

Harrison *et al.* [1] describe an experiment in which they model the relationship between the correlation of a pair of variables, measured using Pearson's *r*, and viewers' precision in estimating this correlation in a variety of different visualization types. This experiment employs a staircase procedure: participants are shown pairs of visualizations of correlation (e.g., two scatterplots) and asked to choose which has the higher correlation. Through successive choices, the procedure hones in on each participant's *just-noticeable difference* (JND) for a given *r*: the minimum difference in *r* at which they can notice a difference between the correlations of two visualizations 75% of the time. The experiment includes the following variables:

- 9 *visualizations*: scatterplot, donut, line, ordered line, parallel coordinates, radar, stacked area, stacked bar, and stacked line (see Figure 3 in Harrison *et al.* [1])
- 2 *directions*: positive or negative correlation.
- 2 *approaches*: *from above*, indicating that the reference value of *r* was compared to values of *r* above it in the staircase procedure; or *from below*, indicated it was compared to values of *r* below it.
- 6 values of *r*: [0.3, 0.4, 0.5, 0.6, 0.7, 0.8].

Visualizations and directions were analyzed as *visualization* × *direction pairs*, of which there are 18 (e.g., scatterplot–negative, scatterplot–positive, parallel coordinates–positive, etc.). 1687 participants were recruited using Mechanical Turk. Each participant completed 4 staircase tasks corresponding to one visualization × one direction × two values of *r* × both values of approach. For example, a single participant might be assigned to complete 4 staircase tasks for scatterplot–negative: [$r = 0.3$ from above], [$r = 0.3$ from below], [$r = 0.7$ from above], and [$r = 0.7$ from below].

2.2 Harrison *et al.* analysis

Classic work on perceptual laws commonly takes the approach of first averaging individual responses over groups before fitting models. This includes work employing Weber's Law, as in Rensink & Baldridge [3], but also work employing Stevens' Power Law [4]. This approach has been criticized for failing to account for individual differences [2,5], finding, for example, that such laws may not fit as originally described [5] or that variation between individuals complicates straightforward application of the laws in practice [2].

Harrison *et al.* [1] used such an approach, following after Rensink & Baldridge [3]. First, they took the mean of all individual observations of JND within each condition (where each condition is defined as a unique combination of visualization × direction × approach × *r*). They then modelled the relationship between the value of *r* and that within-condition mean JND (Fig. 1). Thus, their model describes the relationship between *r* and the mean performance of a group of people from the population, but not the performance of any individual.

What this omits is any sense of the variance in individual performance, which diminishes the explanatory power of such models. For example, it may be that visualization *A* exhibits high precision of estimation (low JND) in the average case – but that its variance is higher than visualization *B*, which performs slightly worse on average but is more consistent across individuals. Without considering variance, we have no way of knowing whether such differences exist, and we may be led, for example, to choose to deploy a visualization that has slightly better average-case performance but which elicits much worse performance for some substantial portion of the population. This is exactly the problem of *bias-variance tradeoff,* well-known in machine learning [6]. Indeed, by modelling individual differences, Cleveland *et al.* [2] found that variance undermines recommendations to transform areas according to parameters derived by a similar mean-fitting procedure (Stevens' power law) in visualizations of circles on maps.

Analyzing group means only also obscures problems with model fit by discarding large portions of the variance (essentially all individual variation) and reducing a large sample of data to comparatively few data points. This explains why Harrison *et al.* [1] (like Rensink & Baldridge [3]) found very high R^2 values describing the fit of their models (as high as 0.98 for one visualization). But when we attempt to interpret these values of R^2 – for example, as the percent of variation explained by the model – something is missing. 98% of individual variation is not explained by this model, as individual variation was discarded before the model was fit. We might instead interpret this as indicating 98% of the variation in the location of the mean was explained, but this is a much less useful thing to know if we wish to understand how *individuals* perceive visualizations. As we will see below, if we try to fit linear models to individual observations directly, the linear model does not exhibit the best fit.

Finally, by discarding individual observations, we cannot use the error from these models to estimate significant differences between conditions. Harrison *et al.* do not use their parametric models to estimate differences between conditions; they use the nonparametric Wil-

coxon rank-sum test instead. In this paper we propose a model of sufficient specificity that we can conduct parametric estimation of differences; this allows us to not only examine the differences between conditions but to clearly describe the expected magnitude of those differences (i.e., effect sizes) using parameters from the model. By employing parametric models, we can derive interpretable effect sizes – for example, ratios of just-noticeable differences, from which we can say, "visualization *A* is *x* times more precise than visualization *B*".

3 Model 1: Linear Model

We begin our secondary analysis by incorporating individual differences to model just-noticeable differences directly on raw values of *r*. A first pass at this would be to simply use a linear regression. Such a model might look like:

$$y_{i,v} = \beta_{v,1} + \beta_{v,2} r_i + \epsilon_i$$
$$\epsilon_i \sim \mathcal{N}(0, \sigma_v^2)$$

This is a fairly standard linear regression. For each visualization × direction pair *v*, each JND ($y_{i,v}$) is equal to a linear function of r_i with intercept $\beta_{v,1}$ and slope $\beta_{v,2}$ plus some normally-distributed error ϵ_i. Note that the intercept, slope, and variance of the error (σ_v) are all dependent on *v,* the particular visualization × direction pair.

Unfortunately, this straightforward model leaves out consideration of *approach* – half of the JNDs were determined by a procedure having people compare the reference *r* to higher values of *r* (an approach *from above*), and half compared to lower values of *r* (*from below*). When the approach is from above, the values of JND are underestimated (because higher values of *r* tend to have lower JND), and when the approach is from below, JND is overestimated. This effect is visible in Fig. 2: note the two systematically different estimates of JND depending on approach. Harrison *et al.* used the correction described by Rensink & Baldridge [3] to address this: they adjusted the value of *r* by moving it up by half the mean JND at that value of *r* when from above, and down by half the mean JND when from below (see Fig. 1).

However, this adjustment is only well-defined if we are using the within-condition means of *r* as our unit of analysis. When fitting a model to individual observations, we must find another way to account for approach. Again consider Fig. 2: because each condition causes a bias in the opposite direction, we could take the average of the two fit lines to approximate the outcome *y* for each *r* (the black line in Fig. 2). Such a model can be fit by including *approach* and its interaction with *r* in the regression. We code *approach* as a sum-to-zero contrast, defined by the variable a_i:

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

We then add the effects of *approach* and *approach* × *r* to the model (new terms in red):

$$y_{i,v} = \beta_{v,1} + \beta_{v,2} r_i + \beta_{v,3} a_i + \beta_{v,4} a_i r_i + \epsilon_i$$
$$\epsilon_i \sim \mathcal{N}(0, \sigma_v^2)$$

Because these are sum-to-zero contrasts, the overall slope ($\beta_{v,1}$) and intercept ($\beta_{v,2}$) for *r* are defined with respect to the mean of the two levels of a_i – in other words, the slope and intercept describe the mean of the slope and intercept of the *above* and *below* levels, exactly the black line in Fig. 2.

3.1 Problems with the linear model

The linear model[1] exhibits several issues of fit that indicate violations of model assumptions, illustrated in Fig. 3A. Two such issues in particular are *non-constant variance* and *skewed residuals*, both of which

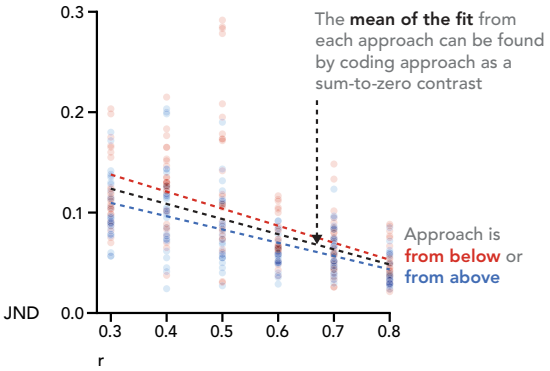

Fig. 2. Example of data and linear regression fits for the different values of *approach* for parallel coordinates–negative.

are violations of the assumptions inherent in the distribution of the error term ϵ_i.

Non-constant variance (heteroscedasticity). As defined in the model, the variance of the error, σ_v^2, is constant with respect to *r*. That is, for a given visualization × direction pair *v*, the variance of *y* (the JND) is assumed to be the same no matter what the value of *r* is. Fig. 3A.1 shows a sample plot of the model fit for scatterplot–negative: we can see that as *r* increases, the variance of the residuals gets smaller, violating this assumption.

We can assess this assumption for all visualization × direction pairs simultaneously by examining the differences between the observed JNDs and their predicted values (the *residuals*). If the constant variance assumption holds, the scale of the residuals should be the same for all predicted values of JND. Fig. 3A.2 shows that at low values of JND, the variance of the residuals is lower than at higher values of JND. This is consistent with the example fit of scatterplot–negative, as low values of JND correspond to high values of *r*. Non-constant variance is common in data with a well-defined lower bound: here, JND cannot be less than 0, and as we approach 0, performance tends to cluster together more tightly.

Skewed residuals. Data with a lower bound also often exhibits the second model violation seen here: skewed residuals (more generally, non-normal residuals). We can think of JND as "bunching up" the closer it gets to 0; besides resulting in less variance, this also explains the skew in the residuals seen in Fig. 3A.3. The residuals do not follow a normal distribution, which is not unexpected given the bounded nature of the data. While it is sometimes the case that we can get away with assuming bounded data is normally-distributed, such simplifications tend to break down the closer we get to the boundaries; here, the assumptions are clearly violated and suggest we should consider other models. This makes sense: looking at Fig. 3A.1, JND gets quite close to the 0 boundary.

4 Model 2: Log-Linear Model

Fortunately, a log transformation of the response is often sufficient in cases of non-constant variance and skewed residuals to solve both problems simultaneously, and often shows up in models of human performance [7]. The applicability of such a transformation is hinted at here, as the residual distribution has the approximate appearance of a log-normal distribution. We can more systematically justify this transformation by fitting a Box-Cox transformation [8] to the data, whose parameter λ describes a power transformation of JND that stabilizes variance. The Box-Cox procedure for this data estimates $\lambda = 0.0292$ with a 95% confidence interval of $[-0.005, 0.0635]$, which includes 0 (the log transform) and excludes 1 (identity; i.e. the linear model) at $p < 0.00001$ (LR $\chi^2(1) = 2756.77$).

[1] This and all other non-Bayesian models in this paper were fit using the `gamlss` procedure in R [18].

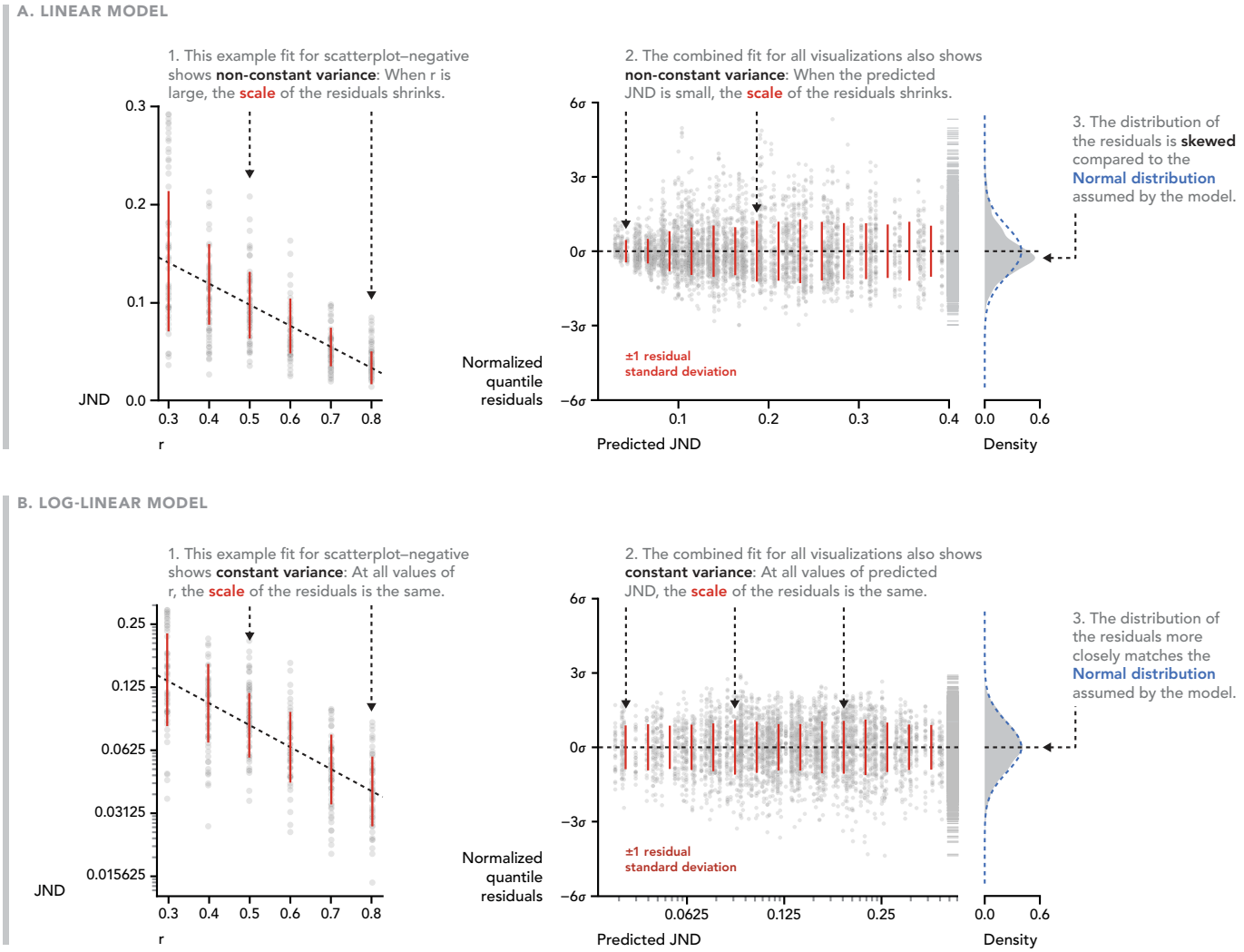

Fig. 3 Comparison of fits of the linear model (Section 3) and the log-linear model (Section 4). Example fits of each model to scatterplot–negative are shown in A.1 and B.1. Plots of normalized residuals for all visualization × direction pairs are shown in A.2 and B.2. Density plots of normalized residuals with comparison to the standard normal distribution are shown in A.3 and B.3.

Log transformation also has the useful property that the resulting model retains some interpretability: coefficients of this model that describe additive differences on the log-scale correspond to multiplicative differences on the original data scale (in other words, we will be able to use this model to make claims like, "visualization A yields x times the precision of visualization B for estimating correlation"). The log-linear model, which deviates from Weber's Law, is as follows:

$$\log(y_{i,v}) = \beta_{v,1} + \beta_{v,2} r_i + \beta_{v,3} a_i + \beta_{v,4} a_i r_i + \epsilon_i$$
$$\epsilon_i \sim \mathcal{N}(0, \sigma_v^2)$$

Comparing the residual fit of the log-linear model to the linear model (Fig. 3B), we can see that the fit no longer suffers from problems of non-constant variance or highly-skewed residuals. This fit also exhibits lower AIC than the linear model (-11683 versus -10037), indicating greater predictive validity.[2] The residual distribution more closely matches the normal distribution assumed by the model: Its residuals exhibit less skewness than the linear model (-0.29 versus 0.96) and less excess kurtosis (0.18 versus 2.05), where the normal distribution is 0 for both measures. In addition, because all values in $(-\infty, +\infty)$ are mapped onto $(0, +\infty)$ by the log transformation, we have solved another problem for free: the linear model can make nonsensical predictions, such as JNDs that are less than 0, that the log-linear model does not.[3] Thus, the log-linear model more accurately describes the observed distribution of JND for a given r, visualization, and direction than the linear model, and should be preferred.

4.1 Data dropped from the analysis so far

So far, we have restricted our analyses to those data points analyzed in the original work. The original work used two criteria to exclude data from analysis:

Outliers. Within each condition (visualization × direction × approach × r), observations outside of 3 median absolute deviations from the median were dropped from analysis. The original paper justified this as a way to address non-normality in the data (although as we have seen above, it did not). Since we have addressed the issue of normality

[2] Model comparison by the Akaike Information Criterion (AIC) is asymptotically equivalent to leave-one-out cross validation [19]. The log-linear model was fit using a log-normal error distribution (rather than the equivalent log transformation of responses with a normal error distribution shown here) so that its AIC can be compared to the linear model.

[3] An alternative to the log-linear model might be a linear model with variance proportional to r. This addresses non-constant variance but does not address skewed residuals. Such a model has AIC of -10668, skewness of 0.95, and excess kurtosis of 1.36 (i.e. it exhibits worse fit and less-normal residuals compared to the log-linear model). It also does not gain the advantage of the log transform in restricting predicted JNDs to be positive.

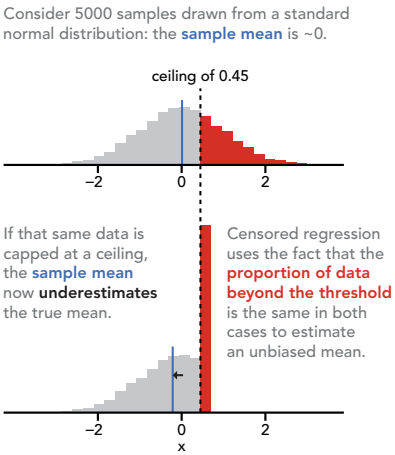

Fig. 4. An example of the use of censored regression to estimate a model when some of the data has been capped at a ceiling.

through log transformation, this criteria is no longer particularly relevant. Since our goal is to explain as much of the data as possible, we believe there is no additional need to drop outliers from the analysis.

Data worse than chance. In Harrison *et al.*, visualization × direction pairs with more than 20% of JNDs greater than 0.45 were dropped (6 out of 18 pairs). The 0.45 threshold represents the *chance* threshold for this experiment: values of JND near or beyond this threshold indicate a failure on a participant's part to judge degree of correlation better than could be done by answering at random. However, removing visualizations with large numbers of observations worse than chance addresses only part of the problem. Many of the remaining tested visualization × direction pairs still have observations at or beyond the chance boundary. The problem is that we have excluded certain visualization × direction pairs for having too many observations worse than chance, but have done nothing to address those observations worse than chance that remain in the visualizations we *do* analyze.

Importantly, when points are near or beyond this boundary, we can say that they probably represent JNDs of 0.45 or worse, but that we do not know the exact JND due to the constraints of the experiment. This type of data can be analyzed using censored regression.

5 Model 3: Censored Log-Linear Model

Censored regression can be used when some of the observed data points do not have a known value, but instead are known to lie above (or below) a certain threshold [8,9]. While we do not know the exact value of points beyond the threshold, we still know how many points were observed beyond the threshold, and it is this information that we can use to fit the model (see example in Fig. 4). While we cannot reliably observe certain values of JND – either because the setup of the experiment makes them indistinguishable from chance, or because of ceilings in observable JND due to the bounds on r – we can use observations close to or beyond those thresholds to estimate the proportion of values we might expect to see above them.

There are three potential ceilings in this experiment. The first two were discussed in Harrison *et al.* but not addressed by their modelling procedure, and the third is one we identified in our own analysis:

- **Chance = 0.45:** Above the chance threshold, random guessing is equally as effective, so we should not expect values of JND much higher than this (see Fig. 5 and Fig. 6). The chance boundary was determined by simulating a participant guessing randomly in the staircase procedure; see Section 3.2 of Harrison *et al* [1].
- **Ceiling from above = 1 – *r*:** When the approach is *from above*, JND cannot be higher than $1 - r$, as we cannot generate a plot with an r value greater than 1 (see Fig. 5).
- **Ceiling from below = *r*:** When the approach is *from below*, JND *can* be less than r (indicating comparison to some $r < 0$, which was allowed in the experimental setup [1]). However, the data suggests that the threshold at 0 in some visualizations may nevertheless have caused a ceiling in JND (perhaps due to some perceptual difference in positive or negative correlations) – see Fig. 6. We therefore take the conservative approach to censor this data when the approach is from below.

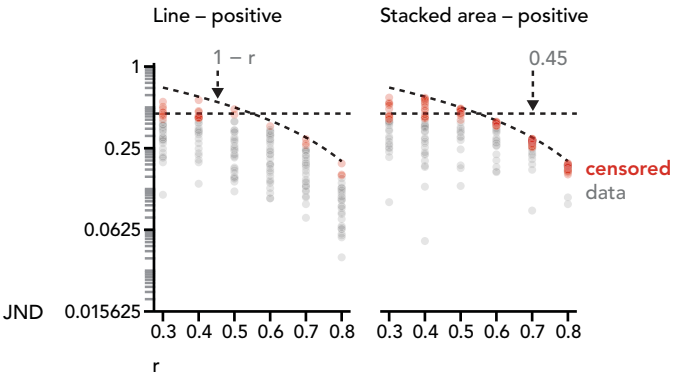

Fig. 5. Data in two visualization × direction pairs for approach *from above*. The ceilings used to derive censoring thresholds for this approach are shown. Note how data bunches up near those thresholds.

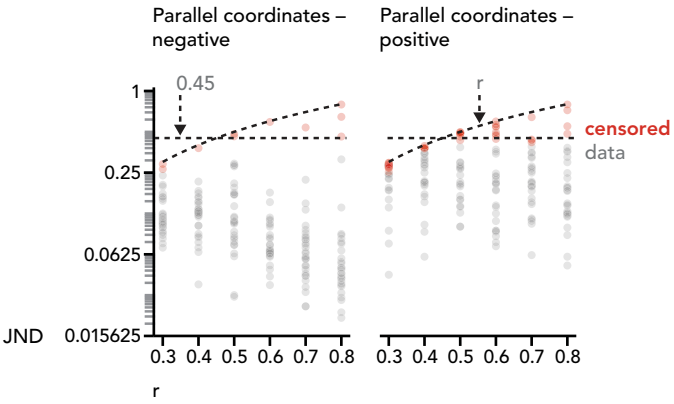

Fig. 6. Data in two visualization × direction pairs for approach *from below*. The ceilings used to derive censoring thresholds for this approach are shown. Note how data bunches up near those thresholds.

Finally, it is worth noting that we cannot simply censor at the thresholds described, as the data tends to bunch up *just below* these thresholds. We therefore censor at 0.05 less than these thresholds, which was chosen based on examining plots like Fig. 5 and Fig. 6 to ensure that the dense set of observations just below the thresholds are censored. We also tried censoring further below (at 0.1 less than these thresholds) and saw similar results from the model, suggesting 0.05 is a reasonable offset here.

To incorporate the censoring described, we first define a censoring threshold $c_{i,v}$ that varies depending on r and the approach:

$$c_{i,v} = \begin{cases} \min(0.95 - r_i, 0.4), & a_i = -1 \\ \min(r_i - 0.05, 0.4), & a_i = 1 \end{cases}$$

Then we change the log-linear model to predict a latent variable y^* instead of y:

$$\log(y^*_{i,v}) = \beta_{v,1} + \beta_{v,2} r_i + \beta_{v,3} a_i + \beta_{v,4} a_i r_i + \epsilon_i$$
$$\epsilon_i \sim \mathcal{N}(0, \sigma_v^2)$$

Finally, we redefine y as being equal to the censoring threshold at the corresponding value of r if its observed value is greater than that

threshold.[4] The model then predicts y based on the latent variable y^* and the censoring threshold c:

$$y_{i,v} = \begin{cases} y^*_{i,v}, & y^*_{i,v} \leq c_{i,v} \\ c_{i,v}, & y^*_{i,v} > c_{i,v} \end{cases}$$

5.1 Bias in uncensored model

The censored model allows us to address problems of bias caused by JND being underestimated near the ceilings described above. See Fig. 7, which compares censored models for line–positive and donut–positive to uncensored models fit to the same data. Note that where large amounts of observations are worse than chance, the uncensored model estimates people as having *higher* precision (lower JND) than we should expect. This bias conspires to make low-performing visualizations seem better than they are, motivating our use of censored regression here. This underscores the problem with excluding some visualization × direction pairs based on the chance criteria without accounting for chance in the pairs we do analyze, and demonstrates how censoring lets us include conditions excluded from the original analysis. Censoring addresses these issues without sacrificing the quality of the fit for conditions that are not affected by these issues, and thus is preferable to the uncensored model.

6 Model 4: Bayesian Censored Log-Linear Model

In this section we describe a Bayesian variant of the censored log-linear model. In Bayesian modelling, we specify our *prior beliefs* about a model as probability distributions, and then *update* our beliefs based on observed evidence (the data collected in an experiment) [11]. These updated beliefs are called *posterior distributions*.

This approach yields a richer estimation of the parameters of interest – complete posterior probability distributions of all parameters – instead of point estimates and confidence intervals. Such posteriors offer an easy way for others to build on our work by using our posterior estimates to inform prior distributions in future work. As we will see, Bayesian estimation also provides a straightforward way to derive the expected performance of a visualization (with uncertainty) on any hypothetical dataset of correlations that can be expressed as a probability distribution over r. We largely adopt Kruschke's [12] approach to Bayesian experimental statistics by using 95% credibility intervals[5] of posterior distributions to estimate differences between parameters.

6.1 Participant effects

As a final refinement to the model, we also incorporate linear mixed effects modelling [12,13]. Specifically, we add a varying-intercept *random effect* dependent on participant.[6] This effect helps account for the fact that we have taken multiple measurements from each participant in the experiment (4 each) by modelling each participant's average performance as an offset from the fit line. Without accounting for this, we effectively are treating our data as having 4 times the number of independent observations as we actually have, causing us to overestimate the precision of our parameters (a problem known as *pseudoreplication* [15], which motivates related modelling approaches for repeated measures, such as within-subjects ANOVAs). By incorporating random effects, we improve the generalizability of our estimates of other parameters by accounting for the correlation between observations from the same participant.

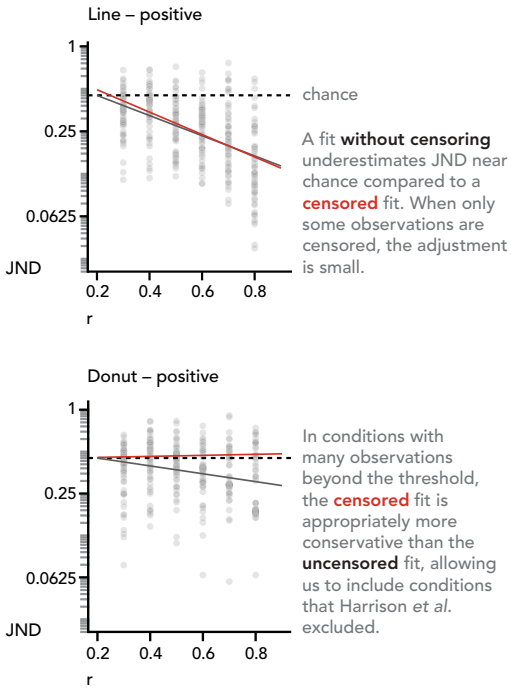

Fig. 7. Comparison of models fit to all data for line–positive and donut–positive with and without censoring, showing how censoring responds when some observations or many observations are censored. In conditions with few censored observations (not shown), the censored model fit is virtually identical to the uncensored fit.

We include a random intercept by estimating an offset U_k from the intercept for each participant k:

$$\begin{aligned} \log(y^*_{i,v}) &= \beta_{v,1} + \beta_{v,2} r_i + \beta_{v,3} a_i + \beta_{v,4} a_i r_i + \epsilon_i + U_k \\ \epsilon_i &\sim \mathcal{N}(0, \sigma^2_v) \\ U_k &\sim \mathcal{N}(0, \tau^2_v) \end{aligned}$$

U_k is called a random effect because each value of it is assumed to be randomly drawn from the distribution $\mathcal{N}(0, \tau^2_v)$. That is, each participant comes from some broader population, and their differences in average performance are distributed normally (on the log scale). The parameter τ^2_v is the variance of participants' average performance for visualization × direction pair v. This allows us to estimate how variable participants' performance is within each visualization × direction pair, which tells us how similar different individuals' estimations are to each other for each visualization. As noted earlier, understanding how individuals vary compared to the rest of the group is an important consideration when deriving design recommendations meant to be applied broadly.

6.2 Priors

In order to fit a Bayesian model, we must also provide *prior distributions* for all unknown parameters. These represent our belief in the location of each parameter prior to running the experiment. In this paper we use weakly-informed priors derived from the results of Rensink & Baldridge [3]. We use weakly-informed priors because we do not have data on all plot types, but can use knowledge of performance on

[4] As a result of this transformation of the responses and the inclusion of data not included in previous models, the censored model cannot be compared to the previous models using AIC. However, we believe the theoretical justification based on ceilings caused by the structure of the experiment and the ability of these models to accommodate data dropped previously motivate the use of censored regression here, and for visualization × direction pairs far from those ceilings the fit is similar to the uncensored log-linear model.

[5] Though we use 95% quantile intervals instead of 95% highest-density intervals since these are invariant under the log transform.

[6] Several models with varying slopes and intercepts failed to show convergence within 1,000,000 iterations, likely due to the small number of observations per participant and the use of censoring. Visual inspection of per-participant slopes suggested low variance in slope, and since our question of interest here is that of variance in average performance more than sensitivity to r, we believe a varying-intercepts model suffices.

scatterplots to infer what range of performance we should expect on other plot types. The high-level goal of our priors is to express some skeptical, but informed, initial belief. For example, our priors on the slope and intercept:

$$\beta_{v,1} \sim \mathcal{N}(\log(0.45),1)$$
$$\beta_{v,2} \sim \mathcal{N}(0,20)$$

Our prior on the location of the intercept ($\beta_{v,1}$) is chance (log(0.45)), and our prior on the location of the slope ($\beta_{v,2}$) is flat (0). In other words, our prior mode is that each condition has no relationship between *r* and JND and is indistinguishable from chance.

However, this is only the mode: we can use Rensink & Baldridge's data to specify the prior variance of these parameters as encompassing a set of reasonable models by ensuring that believable models are within 1 or 2 standard deviations of the mean of the prior. While Rensink & Baldridge did not fit log-linear models to their data, we can approximate a log fit to the data in their Figure 4 [3], giving an intercept of ~ -1 and slope of ~ -2. Since $|-1-\log(.45)| \approx 0.20$, a standard deviation of 1 (variance of 1) will easily cover models having intercepts 2 or 3 times as extreme as the scatterplot condition. If we wish our prior to include all models with an intercept even twice as steep as the scatterplot within 1 standard deviation, a standard deviation of $|-2 \times 2| = 4$ (variance of 16; conservatively we round up to 20) should suffice.

We use a similar examination of Rensink & Baldridge's Figure 4 to estimate priors on the effect of *approach*, which was ~0.2, meaning a variance of 0.25 easily covers values of approach twice as extreme:

$$\beta_{v,3} \sim \mathcal{N}(0,0.25)$$
$$\beta_{v,4} \sim \mathcal{N}(0,0.25)$$

Finally, we use relatively uninformed priors for variance parameters:[7]

$$\sigma_v^2 \sim \text{InverseGamma}(1,1)$$
$$\tau_v^2 \sim \text{InverseGamma}(1,1)$$

We fit the model using MCMC sampling in JAGS [16].[8]

6.3 Performance on a hypothetical set of datasets

We can use our model to derive the expected precision of estimation of a typical individual on an unknown dataset. Rensink & Baldridge [3] proposed doing this (equation 8 in that paper) by integrating the fitted line over a probability distribution of values of *r* one might expect to encounter in a given domain. Applied to the models in that paper, this method has the disadvantage that it cannot derive the uncertainty associated with the calculated average performance, making it impossible to determine differences between visualizations.

However, if we adopt the same approach in a Bayesian framework on models derived from individual observations, uncertainty is straightforward to derive in the form of the posterior distribution of expected JND. We can do this by drawing *r* values from a hypothetical distribution, for example a uniform distribution over the same space sampled in the experiment (we could use a more specific distribution if we had knowledge of expected correlation values in some domain):

$$r \sim \mathcal{U}(0.3,0.8)$$

We can then use MCMC sampling[9] to obtain a posterior distribution of μ_v, the expected JND for the average person given for each visualization × direction pair:

$$\log(\mu_v) = \beta_{v,1} + \beta_{v,2} r$$

We can then rank visualizations by their expected performance on an unknown dataset (see results, below). Given a problem space with datasets having some known/estimated distribution of *r*, we can easily re-compute rankings from the model.

7 Results of Final Model

Fig. 8.1 shows the results of our model in log space for each visualization. Harrison *et al.* used their model to derive a total ranking of all visualizations analyzed (i.e., a ranking that explicitly places each visualization either above or below every other visualization). However, they did not take the error in their model into account when deriving this ranking – given how close the estimates of participants' precision is for many of the visualization types, it is likely that their relative positions in a total ranking are often simply due to chance.

Because of this, we instead focus our results on a partial ranking, admitting that based on the available evidence there is little practical difference between certain visualization × direction pairs (though a total ranking is easily derived from our model, available in our supplementary materials). Based on the method outlined in Section 6.3, we roughly group visualization × direction pairs into a partial ranking based on the expected average person's performance integrated over the fit lines (Fig. 8.2). We can see four groups emerge. We then take the difference in expected precision between each successive group. This difference in means (on the log scale) corresponds to a ratio of geometric means on the original data scale; here we see that the visualizations in each successive group yield at least 1.5x better precision (lower JND) than the previous group on average (Fig. 8.3).

Note how the model accommodates the fact that several of the visualizations have many observations worse than chance (none of the models in the "indistinguishable from chance" group were included in Harrison *et al.*'s final analysis). Using censoring, we were able to formulate a model such that we did not have to drop these conditions; instead, we can derive estimates of their performance, just with comparatively higher uncertainty – the posterior distributions for those conditions near chance are more diffuse than those with higher precision, and the difference between that group and the low precision group has much more uncertainty associated with it (it should be noted that "indistinguishable from chance" in Fig. 8.2 should be read "indistinguishable from chance so far as this model and data can tell"). Rather than dropping these visualizations as in Harrison *et al.*, we simply *learn less about them* from the model. Given a future experiment designed to be more sensitive to JNDs in this range, we might still use these posteriors as priors in such an analysis.

Finally, it is worth considering the expected variance in performance between individuals. In the best-performing group, we find that variance is fairly similar between conditions (Fig. 9). While parallel coordinates–negative may be slightly more variable, the difference between it and scatterplot–negative is not credible. However, this is worth investigating further: with more data, we could estimate this difference more precisely, and also judge whether it has practical significance for design implications. For now, all evidence seems to support a general recommendation for the use of scatterplots in nearly all cases – it is in the highest-performing group of visualizations for both negative and positively-correlated data, and its individual variance is comparable to (and likely slightly better than) its nearest contender in

[7] The inverse-gamma distribution is also the conjugate prior here, which facilitates convergence.

[8] Two pilot chains were run, and the Raftery-Lewis diagnostic [20] used to estimate a minimum chain length to convergence of ~70,000. We then ran two chains with burn-in of 100,000 and sample length of 100,000 each, thinned by 10, for final sample size of 10,000 per chain. Convergence was assessed by visual inspection of trace plots, density plots, and autocorrelation plots, and all parameters passed the Gelman-Rubin diagnostic [21] (multivariate potential scale reduction factor < 1.05).

[9] Conducted in the same sampling run used to fit the model.

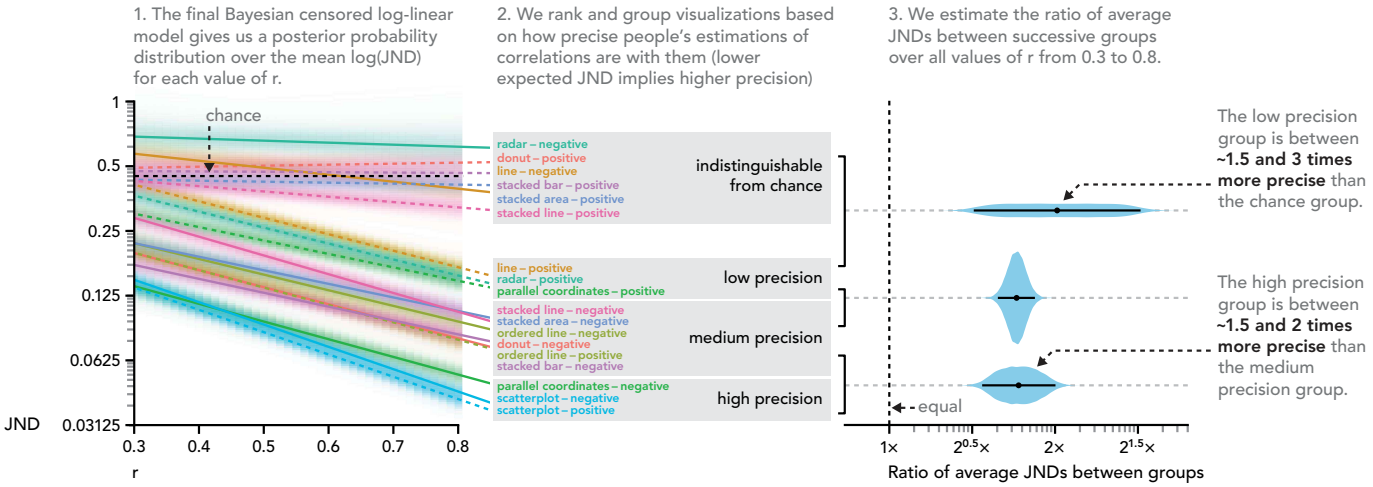

Fig. 8. Final model and partial ranking of visualizations. Part 1 may be compared to Figure 6 from Harrison *et al.* [1], with several notable differences: our results are on a log scale (suggesting a different fit shape), our results provide uncertainty (even if standard errors had been given in Harrison *et al.* they would not have been valid due to the mean-fitting procedure), and we include all tested visualizations in the analysis. Part 2 may be compared to Figure 7 from Harrison *et al.*, except a total ranking from our results would not be the same (e.g., parallel coordinates–negative and scatterplot–negative swap positions), and we provide and emphasize a partial ranking (instead of a total ranking) consistent with the available evidence. Part 3 has no direct analog in Harrison *et al.* Posterior densities in Part 3 are augmented with median and 95% quantile credibility intervals.

average performance (parallel coordinates). Parallel coordinates, on the other hand, has the disadvantage that it does not work equally well for negatively- and positively- correlated data.

8 Discussion

In this section we discuss several aspects of our work, beginning with modelling perception according to "laws" as compared to more exploratory analysis, touching on the broader applicability and limitations of our modelling approach, discussing replication and the value of Bayesian analysis to it, and concluding with design implications.

8.1 Perceptual laws

One implication of our work concerns the use of perceptual laws, such as Weber's and Stevens' Laws. While valuable, these models should of course be subject to skepticism. Avoiding premature theoretical commitment is a core tenet of exploratory data analysis [17]. In our case, exposure to the data led us to a more accurate and actionable model, but which does not conform to Weber's Law. Moreover, these classic laws stem from research conducted over a half-century ago, and were likely shaped by the modelling methods available at the time (often by necessity amenable to calculation by hand). We can now bring more powerful statistical and computational methods to bear. In particular, we need not limit ourselves to analysis of averages only, and can instead account for individual differences.

It is worth noting that Weber's law is most likely to break down at extremes – such as r of .1 or .9. It is at these extremes where the log-linear model disagrees most with the linear model; unfortunately the experiment analyzed in this paper collected data only in the range of .3 to .8. However, given the better description of the distributions of residuals, we expect a log-linear model to better describe these extreme regions if tested in future work.

8.2 Wide applicability of log transformation and censoring

We believe that log transformations and censoring both have wide applicability in modelling human perception in visualization. Indeed, log transformations have been identified as both widely applicable and sorely underused in many areas, including human and biological models [7], despite the fairly approachable interpretation of them in our and others' opinions [7]. In this work, we found that a log-normal distribution – rather than a normal distribution – better describes just-

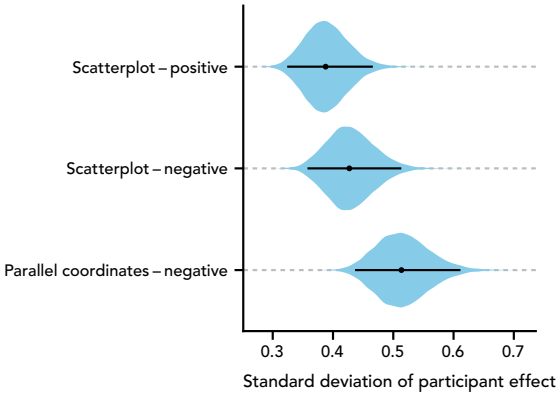

Fig. 9. Posterior distributions of the standard deviation of random effects for participants in the high precision group (τ_v). Higher standard deviation here indicates greater variance in performance between participants. As we do not see evidence that the scatterplot is more variable than its next closest contender for lowest JND (indeed, there is some evidence of the opposite), we can recommend scatterplots as a widely-applicable visualization for correlation.

noticeable differences in visualizations of correlation at a given value of r (yielding a model with lower AIC and less-skewed residuals).

We believe that censored regression is broadly applicable in this space as well. Censoring allowed us to account for artifacts of the experimental design, such as the chance threshold, that doubtless occur in other studies of perception. These models also provide a principled way to learn *something* – but not *too much* – about conditions that have large numbers of observations that cross such thresholds (conditions that had to be excluded from Harrison *et al.*). We see this in the higher uncertainty in the model's estimates for conditions near chance – censoring accounts for this in a principled way. Had we conducted the analysis without censoring (but still included those conditions), the estimates of those parameters would have had less uncertainty associated with them, giving false precision. As we saw, censoring also reduces bias in estimation for conditions with only some observations crossing the threshold. Finally, since the fit is identical when no observations cross the threshold, we do not sacrifice quality of fit for conditions fully below the threshold, making the censored model strictly better than the uncensored one.

8.3 Limitations

Of course, no model is without limitations. The introduction of the log-linear model in Section 4 notes that we can use the Box-Cox transformation to estimate a parameter λ describing a power transform of the data. While this parameter is not significantly different from 0 (the log model), its maximum likelihood estimate is not exactly zero (0.0292, 95% CI: [-0.005, 0.0635]). Using an estimated value of λ (instead of "rounding to log") might have yielded slightly better fit, but also sacrifices both parsimony and the interpretability of coefficients used to derive ratios of precision between groups of visualizations in Section 7. By contrast, we believe that the log-linear model yields equally interpretable results to a linear model with substantially better fit, motivating its preference.

The censoring thresholds derived in Section 5 required us to "fudge" the boundary in order to capture observations just below the ceilings. It is worth noting that it is consistent with the requirements for censoring if we censor conservatively below the ceiling: for example, if the threshold in Fig. 4 is moved to the left (censoring more data), it retains the property that the expected proportion of observations to its right is the same as that in the underlying distribution. That said, since our thresholds were ultimately derived from an exploratory analysis of the data, complete validation of those thresholds can only be made by testing them against data collected in future studies.

In addition, we have focused squarely on issues of statistical analysis to robustly evaluate visualization designs. What remains is to articulate credible perceptual processes driving the observed data.

8.4 Replication and building knowledge with Bayes

This paper would not have been possible without the public release of data from Harrison *et al.* [1]. That release of data contributes to a broader conversation not only about the results of any particular study, but the analysis of data, and the accumulation of datasets and shared knowledge. The analysis in this paper was made not only possible, but straightforward by Harrison *et al.*'s release of data and previous analyses in an easily digested form (a git repository with CSV data, R code, and a clear README [http://github.com/TuftsVALT/ranking-correlation]). It is worth noting that releasing data is much easier than releasing *easy-to-use* data, a practice we hope the community continues to encourage. In that spirit, we also release our analysis code as supplemental material to this paper and as a fork of Harrison *et al.*'s repository [http://github.com/mjskay/ranking-correlation].

We also believe that the Bayesian approach we have taken has some attractive properties with respect to building a body of knowledge. Included in our dataset is a complete posterior sample from our Bayesian model, in the hope that others might use it to derive priors in future work. In this way, the Bayesian framework offers an easy way to build knowledge across studies, one that is perhaps more amenable to the publishing incentives of a field centered around conference publications and which has fewer incentives to conduct traditional meta-analyses. This also allows others to use our model results to calculate their own rankings in domains with some known distribution of values of r (as in Section 6.3), or to use the model to drive automated visualization selection depending on a known r.

8.5 Implications for design

Harrison *et al.* provided a total ranking of the precision of visualizations of correlation for all values of r in (0.1, 0.3, 0.5, 0.7, 0.9) – Fig. 7 in that paper [1]. This ranking implies, for example, that parallel coordinates might be a better visualization of correlation at r = .1, .3, and .5 than a scatterplot for negatively-correlated data. We believe that these design recommendations overstate the strength of evidence. By contrast, our model finds that the performance of scatterplots and parallel coordinates are virtually indistinguishable at those values of r for negatively-correlated data, and that when considering variance between individuals, scatterplots may even be better.

We believe our model yields design recommendations that more faithfully reflect the strength of evidence in the data collected than a total ranking does. The partial ranking of visualizations of correlation in Fig. 8.2-3 communicates the practical differences between visualizations of correlation to designers without overstating small differences. Finally, given the unique advantages of scatterplots – low variance between individuals, high precision on both positively- and negatively- correlated data – we can give a clear recommendation for designers in the vast majority of circumstances: use scatterplots to visualize bivariate correlation, regardless of the value of r.

9 Conclusion

In this work we build upon Harrison *et al.* [1] in modelling the precision of estimation of correlation. We present a series of refinements to their model: incorporation of individual differences, log transformation, censoring, and Bayesian modelling. Besides better modelling the relationship between r and just-noticeable differences, we incorporate a notion of the uncertainty of the effects not estimable in the models of Harrison *et al.* Thus, we are able to derive a partial ranking of visualizations of correlation concordant with the available evidence and which does not follow Weber's Law. Ultimately, we find that scatterplots offer both high precision and low individual variation, making them an attractive technique for visualizing bivariate correlation.

Acknowledgments

We would like to acknowledge the original authors of Harrison *et al.* [1] for the quality of their experiment, their release of data (without which this paper would not be possible), and for encouraging us to pursue a secondary analysis of that data. We also thank the reviewers for their constructive comments in improving this paper. This work was funded by the Intel Science and Technology Center for Pervasive Computing and the Moore Foundation.

References

[1] L. Harrison, F. Yang, S. Franconeri, and R. Chang, "Ranking Visualizations of Correlation Using Weber's Law," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 12, pp. 1943–1952, 2014.

[2] W. S. Cleveland, C. S. Harris, and R. McGill, "Judgments of circle sizes on statistical maps," *J. Am. Stat. Assoc.*, vol. 77, no. 379, pp. 541–547, 1982.

[3] R. A. Rensink and G. Baldridge, "The perception of correlation in scatterplots," *Comput. Graph. Forum*, vol. 29, no. 3, pp. 1203–1210, 2010.

[4] S. S. Stevens, "On the psychophysical law," *Psychol. Rev.*, vol. 64, no. 3, pp. 153–181, 1957.

[5] D. M. Greem and R. Duncan Luce, "Variability of magnitude estimates: A timing theory analysis," *Percept. Psychophys.*, vol. 15, no. 2, pp. 291–300, 1974.

[6] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference and prediction*, Second Edi. Springer, 2009.

[7] E. Limpert, W. a. Stahel, and M. Abbt, "Log-normal Distributions across the Sciences: Keys and Clues," *Bioscience*, vol. 51, no. 5, p. 341, 2001.

[8] G. E. P. Box and D. R. Cox, "An analysis of transformations," *J. R. Stat. Soc. Ser. B*, vol. 26, no. 2, pp. 211–252, 1964.

[9] T. James, "Estimation of Relationships for Limited Dependent Variables," *Econometrica*, vol. 26, no. 1, pp. 24–36, 1958.

[10] T. Amemiya, "Tobit models: A survey," *J. Econom.*, vol. 24, pp. 3–61, 1984.

[11] J. K. Kruschke, "Bayesian data analysis," *Wiley Interdiscip. Rev. Cogn. Sci.*, vol. 1, no. 5, pp. 658–676, Apr. 2010.

[12] J. K. Kruschke, *Doing Bayesian Data Analysis*. Elsevier Inc., 2011.

[13] D. Bates, M. Maechler, B. Bolker, and S. Walker, "lme4: Linear mixed-effects models using Eigen and S4, R package version 1.1-7." 2014.

[14] D. J. Barr, R. Levy, C. Scheepers, and H. J. Tily, "Random effects structure for confirmatory hypothesis testing: Keep it maximal," *J. Mem. Lang.*, vol. 68, no. 3, pp. 255–278, 2013.

[15] S. H. . Hurlbert, "Pseudoreplication and the Design of Ecological Field Experiments," *Ecol. Monogr.*, vol. 54, no. 2, pp. 187–211, 1984.

[16] M. Plummer, “JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling,” *Proc. 3rd Int. Work. Distrib. Stat. Comput. (DSC 2003)*, 2003.
[17] J. W. Tukey, *Exploratory Data Analysis*. Addison-Wesley, 1977.
[18] R. A. Rigby and D. M. Stasinopoulos, “Generalized additive models for location, scale, and shape,” *Appl. Stat.*, vol. 54, no. 3, pp. 507–554, 2005.
[19] M. Stone, “An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike’s Criterion,” *J. R. Stat. Soc. Ser. B*, vol. 39, no. 1, pp. 44–47, 1977.
[20] A. E. Raftery and S. M. Lewis, “Comment: One Long Run with Diagnostics: Implementation Strategies for Markov Chain Monte Carlo,” *Stat. Sci.*, vol. 7, no. 4, pp. 493–497, 1992.
[21] A. Gelman and D. B. Rubin, “lnference from Iterative Simulation Using Multiple Sequences,” *Stat. Sci.*, vol. 7, no. 4, pp. 457–472, 1992.