The Risks of Ranking: Revisiting Graphical Perception to Model Individual Differences in Visualization Performance

Russell Davis, Xiaoying Pu, Yiren Ding, Brian D. Hall, Karen Bonilla, Mi Feng, Matthew Kay, and Lane Harrison

Abstract—Graphical perception studies typically measure visualization encoding effectiveness using the error of an “average observer”, leading to canonical rankings of encodings for numerical attributes: e.g., position > area > angle > volume. Yet different people may vary in their ability to read different visualization types, leading to variance in this ranking across individuals not captured by population-level metrics using “average observer” models. One way we can bridge this gap is by recasting classic visual perception tasks as tools for assessing individual performance, in addition to overall visualization performance. In this paper we replicate and extend Cleveland and McGill’s graphical comparison experiment using Bayesian multilevel regression, using these models to explore individual differences in visualization skill from multiple perspectives. The results from experiments and modeling indicate that some people show patterns of accuracy that credibly deviate from the canonical rankings of visualization effectiveness. We discuss implications of these findings, such as a need for new ways to communicate visualization effectiveness to designers, how patterns in individuals’ responses may show systematic biases and strategies in visualization judgment, and how recasting classic visual perception tasks as tools for assessing individual performance may offer new ways to quantify aspects of visualization literacy. Experiment data, source code, and analysis scripts are available at the following repository: https://osf.io/8ub7i/

Index Terms—visualization, graphical perception, individual differences

1 INTRODUCTION

Visualization continues to be created for, and read by, a broad and diverse audience. Advances in visualization authoring tools alongside the rise of social media have contributed to a greater saturation of visualizations in peoples’ daily lives. One challenge posed by this increased exposure is that people may vary in their ability to perform fundamental visualization tasks, such as estimating and comparing values, judging correlations, or identifying trends and outliers. While there are some situations in which people might change a visualization to suit their needs (e.g. shared spreadsheets), there are numerous everyday scenarios such as digital journalism, television, newspapers/magazines, and public settings where the decisions designers make about visual encodings cannot be changed. Complicating the problem is the observation that we know little about the extent to which and in what ways people can vary in visualization performance, as many studies focus on the performance of the “average observer” rather than on the variance in participants themselves.

Graphical perception studies are one of the primary means through which visualization research has developed a better understanding of how people perform fundamental tasks with visualizations, stretching back to the classic work of Cleveland and McGill [1]. Such studies have yielded several longstanding results, such as canonical rankings of visualization effectiveness [1], which form the basis of visualization guidelines [2], [3] and recommendation systems [4], [5]. Other graphical perception studies have validated the use of crowdsourcing to evaluate visualizations [6], and explored how social information can influence visualization judgments [7]. Yet because the majority of graphical perception studies assess quantitative performance based on the “average observer”, it is difficult or impossible to gain an understanding of how people vary in performance at the individual level.

As Cleveland and McGill acknowledge in their original paper on graphical perception [1], modeling the variance in individual performance across a population is a substantial undertaking:

Because each subject judged all of the experimental units in an experiment, the judgments of one unit are correlated with those of another, and modeling this correlation would have been a substantial chore.

Equipped with recent advances in Bayesian methods (such as modern Markov chain Monte Carlo samplers [8]) that have made it easier to fit complex hierarchical models, we aim to undertake that chore. These techniques make it possible to fit models to individual-level observations—estimating individual-level parameters for the mean and variance of a person’s observations and the correlation between those parameters—while simultaneously estimating population-level means and variance. Such models are sometimes called mixed effects location-scale models (MELSM) [9], or multilevel distributional regression [10].
Recently, researchers have applied MELSM techniques to understand variance in individuals’ performance on some specific visualization types, such as uncertainty visualizations \[11\] or visualizations of correlation \[12\]. A recent study from McColeman et al. targets ratio perception tasks that underlie the comparison task in Cleveland and McGill’s study, finding deviations in performance from traditional ranks of visual channels \[13\]. However, the fundamental comparison tasks from Cleveland and McGill \[1\], which underlie many replication efforts in visualization as well as visualization design recommendations, have not been examined from the perspective of individual variance. This raises a question: might individual people substantially vary in performance on the same tasks that form the basis of widespread guidance in visualization design?

In this paper, we investigate individual differences in graphical perception through a crowdsourced experiment and Bayesian modeling. We replicate and extend the comparison task from Cleveland and McGill’s study, making necessary adjustments to facilitate person-to-person comparison, such as consistent stimuli sets and repeated trials. Using a total of 130,800 judgments from 109 crowdsourced participants, we progress through a series of models, beginning with mean log error approaches from prior studies and ending with a model of absolute error in individual observations with participant-level effects for each visualization type and correlations between them. Our contributions include:

- **Evidence of substantial differences in peoples’ graphical perception performance.** Results from the proposed hierarchical models suggest that people can vary considerably compared to the “average participant”. On certain chart types, some perform consistently better (up to 5pp), while others perform worse (up to 20pp), see Figure 9. This pattern holds across the tested population. Results show that expected differences in performance across pie, bubble, and stacked-bar charts (1-1.5pp) is smaller than the expected differences between people (1.5-3pp), see Figure 7.

- **Positive correlation in performance across visualization types.** Fulfilling the “chore” of correlation modeling as described by Cleveland and McGill \[1\], we estimate correlations in individuals’ performance between chart types. We find that performance is generally positively correlated across all chart types \(r \approx 0.5 - 0.7\), though this correlation is weak for some pairs, e.g. **Stacked Bar and Bubble** \(r \approx 0.3\), see Figure 8.

- **What is ranked best for the average participant may be not ranked best for a substantial portion of people.** While **Bar** charts elicit the best performance on average, results show that around 20-25% of people consistently perform better with another chart type. Less than 40% of people are expected to share the “canonical” ranking of **Bar** best and **Pie** as second-best, with some instead performing better with **Stacked Bar** or **Bubble**, see Figure 9. This calls into question whether we should be using rankings at all to derive design guidance. Instead, we might encourage the use of effect sizes and their uncertainty to make judgments about practical differences in encodings for a given visualization design context.

The variation in performance between individuals across different chart types may imply that the role of psychophysical explanations for differences in error rates has been overstated. In addition to the prior points, many participants perform very similarly with chart types that are supposed to possess very different psychophysical properties, and the effects of variance across individuals can dwarf the mean effects of chart type. We explore how these findings suggest that individual-level factors might be more important to attend to than suggested by prior research, yet the precise nature and origin of these factors are unknown.

Considering visualization design, we discuss how graphical perception research can be made more accessible by shifting reporting more towards constructs such as the variance in people’s performance and the use of raw error measures (instead of log error). There is also a need to improve how we measure and communicate visualization effectiveness to designers beyond the “average” participant. We take a step towards this goal by exploring model-driven probabilistic rank representations (see Figure 9). Finally, we discuss how these findings and the modeling approaches that drive them may provide needed support for the expanding visualization literacy efforts in our community.

2 Background

Graphical perception studies evaluating how people perform basic visualization tasks are a common refrain in visualization research. For example, these studies have been used to investigate particular visualization techniques, like treemaps \[14\], variants of bar charts
or aspects of peoples’ behavior with visualization, like social bias \cite{7}. We draw on prior graphical perception studies, as well as work in visualization and human-computer interaction that has moved beyond population-level analyses.

2.1 Graphical Perception Studies in Visualization

In a study design common in graphical perception research, experimenters vary visualization types (e.g., Bar, Pie, and Bubble charts) to measure the effect of different ways of encoding data on participants’ error in reading that data \cite{15, 6, 16}. Cleveland and McGill’s seminal graphical perception study examined several “elementary perceptual tasks” across different visualization types \cite{1}. Using 95% bootstrapped confidence intervals on trimmed means of log error, they derived design recommendations based on participants’ average performance across different visualization types.

Heer et al. \cite{6} and others (e.g. \cite{7}) have replicated parts of Cleveland and McGill’s original study for various purposes, including validating that online platforms such as Mechanical Turk (MTurk) are a viable testbed for conducting graphical perception studies \cite{6}. Cleveland and McGill’s results have also been cited in the development of tools for automatically creating effective visualizations \cite{4}, thereby having influence on visualization design practice.

Other studies using similar visualization types but different tasks contest the visual encoding rankings from Cleveland and McGill’s original study. Hollands and Spence, for example, demonstrate that the original ranking is not necessarily suitable in explaining the average participants’ performance when it comes to “discrimination” tasks, i.e. discerning the larger of two quantities \cite{17}. Similarly, Yuan et al. found that the ranking also does not apply in tasks requiring a comparison of multiple values, e.g. averages across various data points \cite{18}. Chung et al. investigate and rank the perceptual orderability of visual channels such as hue, size, texture, etcetera. \cite{19}.

As these prior studies rank visualizations based on means (representing the “average observer”), other informative measures, such as variance—how people may differ across visualization types, and how consistently individuals make the same judgment—are less apparent. One aim of our work is to advocate for more analysis of individual-level phenomena in graphical perception, as opposed to only drawing conclusions about the average of a population. As argued by Ziemkiewicz and Kosara, a better understanding of the possible disparities and differences between how people perform with visualizations could lead to more thorough incorporation of individual differences into the design of systems used to create visualizations \cite{20}.

2.2 Individual Differences in Visualization

The term “individual differences” in visualization typically refers to differentiating visualization performance on the basis of factors like personality traits, scores on cognitive ability tests, cognitive states, etcetera. However, individual differences might also refer to examining variance between participants themselves. Binning participants by personality factors or spatial ability scores, for example, will not necessarily reveal performance differences between individuals when groups are the focus of the analysis. In the present work, we focus on measuring between-participant variance and making performance comparisons at the individual level, while contributing statistical models that realize this goal.

2.2.1 Individual Differences as Subgroup Analyses

Several studies have focused on “individual differences” as “traits or stable tendencies to respond to certain classes of stimuli or situations in predictable ways” \cite{21, 22}. For example, Ziemkiewicz and Kosara explore personality factors such as extraversion and openness, finding impacts on task performance measures such as speed and accuracy \cite{20}. Further studies from Green and Fisher, Ziemkiewicz et al., and others explore the impact of personality traits in a variety of visualization contexts \cite{23, 24}. Similar effects are found for spatial ability by Micallef et al. and Ottley et al. in Bayesian reasoning tasks with visualizations \cite{25, 26}. Liu et al. provides a comprehensive survey of these and more individual differences-focused studies in visualization \cite{22}.

2.2.2 Average vs. Individual Level Analyses in Visualization

Mean performance, such as the perceptual error of an “average observer”, is a concise measure that can be extrapolated over a population and can easily be transformed into a guideline, i.e. “just pick the visualization type with lowest average error!” Yet if we only rely on mean performance to rank the effectiveness of visualizations for a given task, there is a possibility that trends for the average observer may deviate at the individual level. Fortunately, recent studies in visualization have begun to use hierarchical modeling and Bayesian techniques to better account for individual variance in population level measurements, laying a foundation for thinking beyond the average.

Harrison et al. modeled peoples’ perception of correlation differences in several bivariate visualization types using Weber’s law, providing a means to quantitatively evaluate (and thus rank) the effectiveness of each tested type \cite{27}. Kay et al. built on Harrison et al.’s work by applying Bayesian modeling to incorporate variation between individuals using random intercepts \cite{12}. Kay et al. proposed a new ranking of visualizations for correlation discrimination that incorporated model-derived uncertainty, and in evaluating this issue at the individual level, they concluded that canonical rankings in visualization may unintentionally “overstate the strength of the evidence” they are based on.

Beecham et al. quantified peoples’ confidence in drawing conclusions from geospatial data visualizations—using a similar methodology as Harrison et al.’s \cite{27}’s JND experiments \cite{28}. Following Kay et al. \cite{12}, Beecham et al. also included a random intercept to account for differences in performance between participants, improving model fit in one of their conditions. Such hierarchical models have also been applied to understanding variance in peoples’ performance on different types of uncertainty visualizations \cite{11, 29}. Additionally, Lu et al. have extended these modeling efforts to explore differences in discriminability for bar, bubble, and pie charts \cite{30}.

2.3 Perception Studies

Studies in perceptual psychology and vision science have also introduced modeling results involving graphical elements common in visualizations. Early work in psychophysics, such as Fechner’s introduction of Weber’s law \cite{31}, was cited as partial inspiration for Cleveland and McGill’s graphical perception experiments, and used directly in recent work in the visualization field for modeling the perception of correlation in scatterplots (e.g. \cite{12, 27, 32}). Other psychophysics research has dealt with ratio estimation problems directly. Stevens reviews several methodologies and experiment paradigms in psychophysics, including production and estimation
Experiment Procedure

For each participant we collect:

60 Data Collection which shows the higher ratio [38]. However, we note that the primary motivation of this study is to explore how people might vary in their ability to perform basic visualization tasks. To do so, we adapt and extend the graphical comparison experiment asking how much individuals vary in their ability to perform basic graphical perception tasks with visualizations.

We aim to pursue a more principled, flexible, and robust approach to person variance, adopting more sophisticated statistical methods such as Bayesian hierarchical regression, and returning to the techniques for ratios and magnitudes [33]. Ekman investigates ratio estimation procedures and discusses model fitting procedures to account for potential biases in participant responses [34]. Ekman et al. also investigate interindividual differences in perceptual ratio estimation in stimuli such as weight, brightness, and area, with results suggesting that participants can vary systematically at the perceptual level [35]. Baird reports a bias towards multiples (e.g. 1, 5, 10) in free-response ratio estimation tasks [36], an effect which has been reported in Talbot et al.’s partial replication of Cleveland and McGill’s graphical perception study [37].

A common goal throughout these studies is investigating peoples’ responses to various visual stimuli. We draw on these experiment methodologies and modeling considerations for the current experiment. In contrast to prior perception-focused experiments, we intentionally adopt Cleveland and McGill’s task because of its perceptual and cognitive dimensions. A perception focused study, for example, might use the method of quadruples where two pairs of charts are shown and the participant is asked to identify “is the smaller element % of the larger for a = [5, 100]?”

For each vis type, we make the smaller element % of the larger for a = [5, 100]. Then we add distractors.

For each unique (vis type, dotsize) pair, there were 15 repeats.

1200 Intermixed trials with no feedback, and breaks every 60 trials

1200 trials shuffled & split into 60 trial blocks

20 datasets

15 repeats

Experiment Overview: To facilitate person to person comparison in a graphical perception experiment, we make necessary changes to the stimuli and data generation processes, illustrated here. The experiment procedure and data collection details are also shown.

Figure 2. Experiment Overview: To facilitate person to person comparison in a graphical perception experiment, we make necessary changes to the stimuli and data generation processes, illustrated here. The experiment procedure and data collection details are also shown.

Stimuli Generation

Experiment Procedure

Data Collection

3 METHODOLOGY

The primary motivation of this study is to explore how people might vary in their ability to perform basic visualization tasks. To do so, we adapt and extend the graphical comparison experiment from Cleveland and McGill [1]. We then use a hierarchical Bayesian model to examine variance in performance both between individuals and between visualization types, as well as correlations between participants’ performance across visualization types. This graphical comparison experiment has been previously replicated in a range of studies targeting different visualization types and participant scenarios (e.g. [6], [7], [14], [37], [39]). As such, it also serves as a touch point for broader discussion on modeling approaches in visualization.

In addition, close examination of the methodologies in these prior experiments reveals that they are not directly suitable for person-to-person performance comparison. For example, some participants might see tasks from only one chart type, requiring between-subjects comparison. To address these issues and to facilitate our goal of comparing person-to-person performance, we highlight several minor but necessary changes to the experiment protocol (see Figure 2 for an overview).

3.1 Experiment Stimuli and Data Generation

The majority of the experiment protocol is adapted from Cleveland and McGill’s original design [1] and Heer and Bostock’s crowd-sourced replication of it [6]. Specifically, participants still see plain, black and white visualizations of five data points—two marked for comparison—and give a numerical answer to “what percentage is the smaller of the larger?” One difference in our experiment are the visualization types tested, which needed to be chosen to facilitate performance comparisons between participants while not excessively expanding the experiment length. In Cleveland and McGill’s comparison experiment, five variations of chart types were tested, derived from Bar and Stacked Bar charts. Heer and Bostock’s replication extended this count to 9 types, adding Pie charts, Bubble charts, stand-alone rectangular areas, and treemaps.

To select charts that would allow us to compare performance between people, we analyzed the results reported in the studies, with two selection criteria in mind. The first goal was to choose a visualization that could serve as a baseline across all participants.
The Bar chart, given its consistently lower error relative to other chart types in these studies and other replications, met this criterion. The second goal was to choose a small set of visualizations that were similar in average performance, to raise the possibility of identifying people who perform consistently better or worse on them in contrast to the canonical ranking (shown in Figure 1).

Other factors were weighed in the decision-making process, such as evidence from Kosara et al. [10] that people may use different strategies when judging Pie charts, which could lead to systematic performance differences, and Kong et al.’s study showing that orientation issues lead to difficulties in interpreting treemaps [14] (we exclude treemaps for this reason).

The data generation process follows prior studies [1], [6], with modifications to accommodate repeated trials and consistent stimuli across participants for participant-to-participant comparison. Five values are generated in each trial, including 1 smaller comparison value (S), 1 larger comparison value (L), and 3 distractor values. To evenly cover the domain of possible answers, the proportion of the smaller value to the larger value is separated by 5%, ranging from 5% to 95% (19 values), plus a 99% for a total of 20 different proportions. The 5% differences in stimuli are informed by prior studies, in particular Talbot et al., who find that participants typically give answers ending in 5 or 0 [37]. For the Bar, Bubble, and Stacked Bar charts, the 3 distractor values are randomly generated within a normalized 0 to 1 range. The Pie chart, given its part-to-whole arrangement, is handled differently: all 5 values are constrained to sum to 1 to represent the whole of the chart. Figure 3 shows an experiment task with the pie chart stimuli. In total, for each visualization-dataset pair (4x20), there were 15 repetitions. Thus each participant answered 1200 trials in total (4 visualizations × 20 datasets × 15 repetitions). After the set of trials was created, all were shuffled to appear in random order.

3.2 Experiment Procedure

The experiment procedure included three phases: Training, Trials, and Demographics.

Training: Participants were given eight practice trials, covering all four visualization types. After answering, participants were shown feedback text with the correct answer. For example, in a Bar chart practice trial, participants might see: The correct answer is 42%. The smaller bar is just over 2/5 the size of the larger one. Training trials included a mixture of rounded and non-rounded answers, to emphasize to the participant that the answers could feasibly take on any value between 1 and 100.

Trials: The trials phase began with a paragraph that reminded the participant that their answers were being recorded. As in prior experiments [6], participants were encouraged to make a “quick visual judgment” and to avoid physically measuring the stimuli to get exact answers. Breaks were given after every 60 trials (20 breaks in total). The participants were encouraged to take a break as long as needed before resuming.

Demographics: After participants finished all training and trials, they were asked to provide basic demographic information, including reported gender, age, country of origin, and highest degree obtained. Participants were also asked to self-rate (on a 1-7 scale) their experience with visualization and statistics.

3.3 Resulting Participants & Exclusion Criteria

Participants were recruited in an IRB-approved study on Amazon’s Mechanical Turk, which has been shown to be a reliable testbed for graphical perception experiments [6]. 200 participants started the experiment, and 118 of them completed all 1200 trials. A survival chart is shown in the data collection section of Figure 3 indicating that the experiment had a 59% completion rate in total. Among the participants who did not finish all trials, 57.3% dropped before 10 trials and 80.5% dropped before 100 trials. Each participant received the same set of trials, making this a within-subjects design. Each participant that completed all trials was compensated $22. The average completion time was 2 hours 40 minutes, for an hourly rate $8.3/h, exceeding U.S. minimum wage.

Conditions where the true proportion was 100% (both elements were the same size) and 5% are additionally used as attention checks. While a range of errors can be expected, repeated large errors on these trials likely represent a participant not paying attention, not trying to be accurate, or misunderstanding the experimental instructions. We only excluded participants where extreme errors of greater than 50pp occurred on more than 25% of trials in either of the two conditions, which equates to 16 or more such occurrences out of 60 trials per condition. A total of 9 participants met these criteria, and were thus excluded from all further analyses, leaving 109 participants in all. Data with and without such exclusions are available in supplemental material.

4 Analysis Approaches

We will use two approaches to analyzing the data: a partial replication of Cleveland & McGill’s analysis [1] which focuses on means, then a Bayesian regression model that allows us to also examine within- and between-participant variance. We will use a model expansion to develop a model that describes, as well as we are able, the phenomena in question, rather than proposing and testing specific hypotheses about the data [41]. Before describing the results (Section 5), we will first describe both approaches in detail.

4.1 Replicating Cleveland & McGill

We replicate Cleveland & McGill’s analysis [1] to put our study in context with their seminal findings. In addition, since Heer & Bostock closely replicates Cleveland & McGill, we will compare...
findings from both papers in Section [5]. For the replication analysis, we use the same formula for log error as Cleveland & McGill: \[ \log_{2}(|\text{judged percent} - \text{true percent}| + 1/8). \] At a high level, the analysis uses bootstrapping to calculate means and confidence intervals of midmean of log errors (see Section 4.3 and 4.4 in the original paper [1]). First, we create 1000 bootstrapped samples. Each sample has 109 participants resampled from our data with replacement (109 \times 1200 = 130800 observation). For each sample:

1. Calculate log error = \log (|\text{error}| + \frac{1}{2}) for each observation.
2. In Cleveland & McGill [1] and Heer & Bostock [6], each participant completes one trial per condition combination (true proportion and visualization type), while we ask each participant to complete 15 repeated trials per combination. Thus, to be most comparable to previous approaches, we compute the mean response from the 15 repeated trials for each combination of true proportion, visualization, and participant.
3. Take mid-means within each combination of visualization type and true proportion, yielding 80 mid-means (4 visualizations \times 20 true proportions).
4. Group by visualization type and take the mean in each group, yielding four means of midmeans, one for each visualization type.

The result of the above procedure is a bootstrapped sampling distribution of 1000 means of midmeans of log errors for each visualization type, from which we can calculate 95% confidence intervals.

4.2 Building a more complete model of errors

To build up to a more complete model—one which describes absolute errors at the visualization and participant level, allowing individuals’ abilities (and consistencies) to vary between each other and across visualization types—we will first start with a simplified model. For interpretability, we will also aim to have a model that describes errors on the original percentage response scale, not on a log scale as in Cleveland and McGill [1]. To develop our final model, we followed a model expansion approach [41], gradually adding more complexity to the model to describe the assumed data generation process in more detail. Throughout this process, we assessed model fit and quality using posterior predictive checks [42] to understand in what ways a given model failed to describe the data generation process. We then used these checks to determine in what direction to expand the model until it was able to adequately describe the data. (In addition to the walkthrough here, we provide annotated source files for replicating this analysis and experiment in the supplemental material.)

We will begin with a model of mean absolute errors, assuming that errors have been averaged within participant \times vis. In other words, for each participant, we calculate their mean absolute error on each visualization. Such a model allows us to look at mean absolute error at the visualization level (much like the Cleveland and McGill analysis), but does not permit us to analyze individual-level performance, as this is averaged out in advance. This is a common approach in the visualization literature (e.g. [6], [7], [14], [37], [39]).

To build a Bayesian model, we also need a likelihood. The likelihood is a distribution that we assume observations to be drawn from, conditional on the predictors in the model. Each observation in this model is the mean absolute error for one participant on a particular visualization condition. Traditional linear regression, for example, typically assumes a Normal likelihood:

\[
V = 4 \text{ : number of visualization types}
\]
\[
P = 109 \text{ : number of participants}
\]
\[
i \in \{1 \ldots VP\} \text{ : index of observations}
\]
\[
\text{mean errors within participant} \times \text{vis}
\]
\[
\text{vis}[i] \in \{1 \ldots V\} \text{ : visualization associated with the } i\text{th observation}
\]
\[
\text{mean_abs_error}[i] \sim \text{Normal}(\mu[i], \sigma) \quad \text{likelihood}
\]
\[
\mu[i] = \beta[\text{vis}[i]] \quad \text{mean submodel}
\]

This model says that each observation, mean_abs_error[i], is Normally distributed with a mean of \( \mu[i] \) and standard deviation of \( \sigma \). Given a specific visualization type \( v \in \{1 \ldots V\} \), \( \beta[v] \) is the average mean absolute error for that visualization type. Thus, \( \beta[\text{vis}[i]] \) is the average mean absolute error for the visualization type associated with observation \( i \) in the dataset.

This model fails to capture individual-level differences; e.g., that some participants might be better or worse on some visualization types, or even systematically better or worse on these tasks in general. The Normal likelihood also fails to capture key constraints of the data generating process: e.g., we know that absolute error on a percentage response scale must be between 0% and 100% (i.e., 0 and 1).

Let’s tackle the latter problem first: adjusting the likelihood. Instead of the Normal likelihood, we’ll adopt a zero-inflated Beta distribution as the assumed distribution of errors, described by a mean parameter (\( \mu \)), precision parameter (\( \phi \)); also called the sample size parameter, this gets larger as variance gets smaller), and a zero probability parameter (\( \pi \)). The Beta distribution is defined on (0, 1), and so is commonly used to model bounded data [43], [44]. Unfortunately, the Beta distribution does not allow zeros, yet our data contains zeros at the individual observation level (when a participant gets a response exactly correct). The zero-inflated Beta distribution is a modified Beta distribution that allows zeros by modeling the probability of a zero being present as a separate process, as follows:

\[
y \sim \text{ZeroInflatedBeta}(\mu, \phi, \pi)
\]
\[
\implies y = \begin{cases} 
0 & \text{if } z = 1 \\
y^* & \text{if } z = 0 
\end{cases}
\]
\[
y^* \sim \text{Beta}(\mu \phi, (1 - \mu) \phi)
\]
\[
z \sim \text{Bernoulli}(\pi)
\]

Like the observations in a Beta distribution, its mean parameter (\( \mu \)) must also be between 0 and 1. Thus, to adjust our model to use the zero-inflated Beta distribution, we must ensure that \( \mu \) is bounded between 0 and 1. We can do this by using a link function that takes mean absolute errors in (0, 1) and translates them onto \( (-\infty, +\infty) \); thus the inverse of this link function ensures \( \mu \) is in (0, 1). The logit function does this (changes from the previous specification are highlighted in red):

\[
\text{mean_abs_error}[i] \sim \text{ZeroInflatedBeta}(\mu[i], \phi, \pi) \quad \text{likelihood}
\]
\[
\text{logit}(\mu[i]) = \beta[\text{vis}[i]] \quad \text{mean submodel}
\]
To model individual errors directly, one might naively think to simply fit the above model to participant-level errors instead of mean errors. However, doing so would result in artificially inflating the number of samples in the model, leading to overconfident estimates (pseudo-replication [45]). To address this problem, we must ensure that the model accounts for differences in individuals’ performance. We can do this using random effects by adding an offset $U[v, p]$ that describes how the participant associated with observation $i$ deviates from the average for the visualization associated with that same observation $i$. Now we can model individual errors directly, without averaging first:

$$V = 4: \text{number of visualization types}$$
$$P = 109: \text{number of participants}$$
$$K = 300: \text{number of repetitions}$$
$$i \in \{1 ... VPK\}: \text{index of observations}$$

We build this model in brms [43], a modeling library in the R statistical programming language [46], which fits models using Stan [8], a probabilistic programming language and Markov chain Monte Carlo sampler. The model can be specified using the brms function in modified Wilkinson-Pinheiro-Bates syntax [47, 48]:

```r
brm(
  abs_error ~ vis + 0 + (vis + 0 | pt | participant), # mean
  phi ~ vis + 0 + (vis + 0 | pt | participant), # precision
  zi ~ vis + 0 + (vis + 0 | pt | participant) # zeros
)
```

This syntax closely parallels the equations for the likelihood and the mean ($\mu[i]$), precision ($\phi[i]$), and zeros ($\pi[i]$) submodels. The use of + 0 tells brms to use a one-hot-coding of the vis variable: each level of vis has its own coefficient. Otherwise, the default dummy coding would be used for categorical variables, which would use an intercept for one level of vis and coefficients for the offset from that intercept for the other three levels of vis. Dummy coding makes it difficult to set the same prior on each visualization type (because one is an intercept and the other three are offsets), whereas one-hot-coding makes it straightforward (because each is its own intercept). Finally, | pt | (where the identifier pt is arbitrary) is used in brms in place of a | from standard Wilkinson-Pinheiro-Bates syntax to indicate that the covariance matrix for random effects ($\Sigma$) is shared across submodels.

### 4.2.1 Priors

To fit the Bayesian model we must supply priors for unknown parameters. We used weakly-informed priors [49]: that is, priors which are set to cover reasonable ranges of values a priori, rather than priors set tightly around a best estimate from previous literature (an informed prior). Ideally, this conservative approach allows our priors to regularize [50] estimates and improve model fit without unduly biasing estimates towards previous results. Our priors are as follows:

- $\beta_v \sim \text{Normal}(-2.1)$: Prior for the mean error in percentage points on a log-odds scale for each visualization. This prior covers roughly $[-4.0]$ in log-odds space in its 95% central interval, which is roughly [1.7, 50] in percentage points; in other words, we do not expect people to make errors larger than 50 percentage points on average, which would be quite a large error in a proportion judgment task. Prior work also has not found mean absolute error to be credibly less than 1 on the adjusted log-2-absolute-error scale of Cleveland & McGill
We begin by replicating analysis methods from prior work on our data as a point of comparison, before using our models to investigate individual differences in performance.

5 Results

We begin by replicating analysis methods from prior work on our data as a point of comparison, before using our models to investigate individual differences in performance.

5.1 Replicating Analyses from Cleveland & McGill and Heer & Bostock

To compare our data to previous work, we extracted the reported results from two previous studies: the original Cleveland & McGill study [1] (their Figure 16; upper panel of our Figure 4), and Heer & Bostock’s crowdsourced experiment [6] (their Figure 4; middle panel of our Figure 4). The lower panel of Figure 4 shows an analysis of our data. For the purposes of this comparison, we used methods that match as closely as possible to those of the previous authors: means of midmeans of adjusted log error with (stratified) bootstrapped confidence intervals.

By replicating these analysis methods, we arrive at error estimates and relative visualization rankings that broadly agree with prior work. Visualization type T3 of prior work is most similar to our Bar type, and all 3 studies arrive at estimates that are essentially identical. Our Pie results match with T6 of [6], and our Bubble results closely match their T7 type. Finally, our Stacked Bar type is most similar to T5 in the two prior studies, and our error estimate is approximately centered between the results estimates from those studies.

Across the three studies, the primary ambiguity is in the exact relative rankings of the visualization types other than Bar. While Bar has the lowest error overall, the remaining rankings are roughly T1 (most like Bar) < T2 (like Stacked Bar) < T3 (like Bar) < T4 (like Stacked Bar) < T5 (most like Stacked Bar). Our results closely match their results from two previous studies: the original Cleveland & McGill study [1] (their Figure 16; upper panel of our Figure 4), and Heer & Bostock.

5.2 Distributions: A Step Beyond Averages

Before considering further analysis, it is important to distinguish between two sources of information available to us: the observed data, which are the responses collected directly from individual participants; and the posterior distribution of the fitted model, which encodes our understanding that people will vary in their responses across repeated trials and conditions, that people vary from each other, that the participants in this experiment are only a sample, and that there may be correlations both between and within people. The model attempts to quantify all of these sources of
uncertainty. We show observed data together with model estimates whenever feasible.

With the model and data, we inspect performance based on error distributions, rather than average error alone. The top row of Figure 5 shows the empirical cumulative distribution functions (CDFs) of participant errors as solid, opaque lines. CDFs can provide answers to questions like, what proportion of the observed errors were less than 5 percentage points (pp)? What proportion were less than 10pp? etc. The shaded bands show the 66% (dark) and 95% (light) credible intervals for the CDF of the average participant as estimated by the model.

The CDFs of observed data have several sharp jumps, like stair-steps. This artifact has two causes: 1) of the twenty levels of true proportion examined in this study, eighteen were evenly divisible by 5; 2) roughly 70% of participant responses were also evenly divisible by 5, i.e. participants tended to round their responses to end in 0 or 5, similar to prior studies (e.g. Talbot et al. [37]). For example, when the true proportion for a task was 60%, 373 responses were exactly 55% (5pp of error), but only 214 responses were between 56% and 59%. The model does not directly model participant rounding behavior, but the top-center of Figure 5 shows how the model effectively smooths the response curve in a way that still closely fits the data. Since rounding is an artifact of the elicitation procedure, and our model did a good job of otherwise recovering the shape of the distributions, we did not consider it important to model in higher fidelity.

Figure 5 shows that the median error (50% on the y axis) is around 5pp for all conditions in the observed data. The distributions are also skewed: the top quartile of errors (75% on the y-axis) are as much as 2-4 times larger than those in the bottom quartile (25% on the y-axis). Finally, the vast majority of errors are less than 35pp (about 98.5% below 35pp), regardless of visualization type.
5.3 Pairwise Comparisons Between Visualization Types

One issue with the CDFs shown in the top row of Figure 5 is that it is difficult to make comparisons between visualization types to determine their performance relative to each other. The bottom half of Figure 5 shows pairwise differences between CDFs of each visualization type. For example, where the top row of Figure 5 shows that the average participant is estimated to make zero-error responses about 25% of the time with Bar, the pairwise graph labelled Pie - Bar shows that the proportion of zero-error responses for Pie will be about 5–7pp less than Bar. Bar generally dominates Pie, having a higher proportion of smaller errors at every magnitude of error. At error levels higher than 15pp, the difference between Pie and Bar shrinks to less than 5% of cumulative errors. By an error level of 35pp, the predicted errors for Pie and Bar charts are virtually indistinguishable.

Overall, Figure 5 shows Bar has the lowest-error responses for the average participant. Pie is ranked second, receiving proportionally less error than Bubble and Stacked Bar, though primarily at error levels below 10pp. There is effectively a tie between Bubble and Stacked Bar, with a small probability favoring Stacked Bar over Bubble at error levels below 20pp; however, the large overlap of zero in the Bubble - Stacked Bar graph shows that there can be no certainty that either is better than the other. In any case, the cumulative difference in the two is never estimated to be greater than 5% at any given error level, so there is likely little practical difference between them for the average participant. These results largely agree with canonical rankings of these visualization types in prior graphical comparison, while adding perspective about the relative rates of occurrence and magnitudes of errors people make.

5.4 Between-Person Variance in Mean Error

While our results at the population level agree with previous population-level analyses, our model also allows us to examine individual-level performance, asking questions like: do some people perform at their best on a visualization type that is the worst type of chart for other people? Are the canonical visualization rankings really universal?

The top of Figure 6 shows model-estimated means for all participants as gray lines, with the average participant in black. The large between-person variance can be seen as the wide spread of the positions of the gray lines relative to the difference in conditions: in many ways individual differences dwarf the effects of specific chart types. We also highlight three participants with highly “non-canonical” rankings: one person who is good at every chart type, one who is particularly bad at Stacked Bar, and one who is bad at Bubble and performs best on Stacked Bar (better even than Bar). These participants are also not unusual: note the characteristic crossing pattern in the bottom part of the parallel coordinates chart, indicating a large proportion of reversals of the average-participant ranking of Stacked Bar < Bubble.

Our hierarchical model allows us to simulate new participants while accounting for the correlation between people’s performance across conditions (e.g., that people who are better at one chart type are likely better at others; see 5.5). We can quantify the between-person variance in mean errors by simulating a sample of 6,000 new people within each one of the 6,000 draws from the posterior distribution of our model, then taking the standard deviation of the mean error from each simulated sample of people. This yields 6,000 draws from a distribution describing our uncertainty in the standard deviation of people’s mean error.

This provides an estimate of how much variance in error there is expected to be between people. Figure 7 compares the differences in mean absolute error between charts types (top) to these standard deviations of between-person mean absolute errors. For example, the difference in mean error between Bar and Pie is about 2pp (top panel); yet, the standard deviation of between-person mean absolute errors is also about 2pp (bottom panel). Thus, the change in error to be expected by randomly selecting a different participant is about the same as the change we would expect by switching from Bar to Pie with an average participant. For visualizations besides Bar, the mean differences are even smaller (less than 1pp; top panel); for these visualization types, it appears that differences between people are larger than differences between conditions.

5.5 Correlation in Individuals’ Mean Error

Examining correlations in the parallel coordinates chart in Figure 6 suggests another way of looking at the data: correlations between mean error across individuals; e.g., do people who perform well on Stacked Bar perform tend to perform well on Bubble? As the model estimates correlations between all individual-level parameters in the mean submodel as part of the covariance matrix of hierarchical parameters (Σ), we can extract these correlations and their uncertainty directly from the model (Figure 8).
Mean error with all visualization types are positively correlated: e.g., if a person does better (worse) than average with Bar, they will likely do better (worse) than average with Pie. The Bar to Pie (~0.65) and Pie to Stacked Bar (~0.62) correlations are highest, while Bubble to Stacked Bar have much lower correlation. This matches with a qualitative assessment of individual means in Figure 5, where there is substantial variance from the average ranking of best-to-worst mean error across individuals, particularly in Stacked Bar and Bubble.

5.6 Ranking Individuals’ Mean Error
Finally, it is traditional to include an attempt to rank chart types in order to derive actionable design guidance from empirical visualization papers. This task is made difficult by our desire to account for between-person variance; thus, rather than providing a single ranking based on population means, we start by calculating a distribution over rankings. We use the same approach to simulating new samples of people within each draw from our posterior distribution as in 5.4, but within each draw we calculate the proportion of people having each possible ranking of the four visualization types in terms of mean error. This gives us both the proportion of people expected to have each ranking and the uncertainty in that proportion (Figure 5).

The “canonical” ranking of Bar < Pie < Stacked Bar < Bubble is indeed one shared by the most people—but still only about 22% of the population. The top-6 most common rankings all have Bar as best, but these account for only about 75–80% of people. There is considerable variance in rankings (particularly for 2nd to 4th place). As our model gives very precise estimates of population-level means, this variance is likely largely driven by large between-person variance as compared to the size of differences between chart types, as we saw in 5.4 These results suggest that in the face of large between-person variances, rankings are an inadequate way to summarize visualization performance or derive design guidelines. We discuss alternative approaches to deriving design recommendations in 6.2.

6 DISCUSSION
At face value, graphical perception studies aim to identify the best among a set of visualizations for a given task. Yet visualization recommendations that build on prior studies are often sparse and conflicting. One recommendation might suggest that Stacked Bar charts are a superior choice to Pie charts because our perception of angles is worse than our perception of position. Another might discourage the use of Bubble charts for similar reasons, citing difficulties in comparing areas. The results of our replication/extension and modeling efforts suggest that the reality of graphical perception performance is more complex.

6.1 Between-Person Visualization and Visualization Tasks, Broadly
One of the primary aims of Cleveland and McGill’s study was to investigate whether psychophysical differences predicted differences in chart effectiveness— which is supported by their aggregate level analysis [1]. However, more recent work on more complex chart types has suggested that differences in people’s performance cannot simply be attributed to differences in the psychophysical properties in charts: for example, Kale et al. [29] found that the strategies different people use to interpret different uncertainty visualizations had a profound effect on their performance, beyond their ability to estimate the relevant psychophysical quantities (such as ratios of areas in a probability density chart). One could argue that this was an artifact of the complexity of uncertainty visualizations, and that this finding might not replicate for simpler, fundamental tasks like reading ratios in Bar or Pie charts.
Individual-level rankings of visualization effectiveness

While the “canonical” bar < pie < stacked bar < bubble ranking is the most likely for any individual to have...

...our model estimates that only about 20–25% of people will have that “canonical” ranking...

...and only about 75–80% of people will have bar as the best visualization.

The variability in rankings below bar reflects the closeness of mean absolute errors of all visualizations except bar; individual variability dwarfs the relatively small differences between the 2nd, 3rd, and 4th place visualization types.

This calls into question the usefulness of rankings themselves: judging usefulness in terms of error on an interpretable scale (like percentage points) may yield more appropriate design recommendations—

—for example, that pie, stacked bar, and bubble are practically identical, relative to the amount of variability across people.

Figure 9. Proportions of the population predicted to have each ranking in terms of mean absolute error, listed in order from most to least common ranking. All proportions added together equal 1 (100% of the population).

Our work suggests this is not the case: even in these fundamental tasks, there is substantial variation in individual performance, and the magnitude of this between-person variance is large compared to the average differences between chart types (see Figure 6). Some people are even able to perform equally well (and much better than the population average) on all chart types. Perhaps, as Kale et al. [29] argue for uncertainty visualizations, the particular strategies (or proxies [52, 53]) people use to accomplish visualization tasks play a larger role in performance, on par with (or in some cases more important than) psychophysical properties of different chart types. Studies on pie charts from Kosara et al. underscore this possibility, by establishing that people may be using one of many strategies to perform comparison tasks [40]. With the provided modeling approach, such signals for strategy could be investigated as part of future work on an individual, rather than aggregate, basis. Furthermore, the provided models could be extended to investigate correlations of individual performance against measures thought to impact visualization performance such as visualization literacy [54, 55] or spatial ability [26].

6.2 Design Recommendations at the Individual Level on the Raw Data Scale

Fitting models to data on the raw error scale which account for individual differences in visualization performance enables us to explore design recommendations for graphical perception across multiple perspectives and scales.

Moving beyond averages to CDF-based approaches (Figure 5) allows us to answer questions about possible magnitudes of errors across visualization types and their respective rates of occurrence, along with uncertainty about such estimates. This may enable more nuanced visualization design as it relates to task accuracy. For example, if the true proportion between two elements of a chart was 50% (one is half the size of the other), our results suggest that about 98% of people’s responses would be between 15 and 85%. Whether this magnitude of accuracy could be called “good enough” or “greatly concerning” would need to be determined by a designer or subject-matter expert on a case-by-case basis.

For example, if a designer is interested in “all of the errors people are likely to make”, then we might want to look at something like the 75th percentile (or higher) of the distribution. Or a designer might say, “I want to maximize the likelihood of highly accurate judgments”, in which case we could compare the 25th percentile of each distribution, as this would tell us which visualization is most likely to elicit errors that are very small. Or a designer might say, “I just want to pick the visualization that is most likely to result in the lowest error judgments most of the time”, which puts us back to comparing medians (50th percentile). Such considerations are all but impossible if only summarized averages are made available.

As we saw in 3.6, considering between-person variance also calls into question the value of “ranking” visualizations by effectiveness. Looking at between-person variance of errors on the raw scale offers an effective alternative: results suggest that for most, Bar will be about 2 percentage points better than the other three visualization types, and that the remaining differences between visualization types are likely washed out by between-person variance. Thus, a simple recommendation to designers might be: use Bar if an extra 2pp of precision is needed; otherwise, the chart type is unlikely to make much difference when factoring in the larger differences between people.

Analysing error on the raw scale thus makes it easier for designers to incorporate results from the literature into a design process where “best encoding” has to compete with other concerns (aesthetics, use of metaphor or rhetoric [56], memorability [57].
It is very hard to make an informed design decision if one just has a ranking of effectiveness: as a designer, one needs to know the magnitude of differences in effectiveness on an understandable scale (ideally the data scale) to make these tradeoffs carefully, and log error abstracts away this understandability. Log error was adopted by Cleveland and McGill as a data analysis convenience, not because it particularly aids interpretation or generalization; we suggest the field abandon log error in favor of measures more easily translated into practice.

6.3 Between-Person Variance as a Form of Visualization Literacy

Looking at distributions of individual-level mean error, as in Figure 6, could allow designers to make judgments about variance between people. How many people are likely to be “left behind” by a particular encoding choice? Further research is vital to fully understanding the variance of individuals’ performance across the range of visualization types that exist in the literature, in order to understand just how broadly accessible a visualization is. Work on visualization literacy has begun to address this problem [53], [55], [59]–[62]; our work suggests some effects of individual differences may even dwarf effects of different chart types.

The ability to quantify between-person variance in visualization performance raises new possibilities for ongoing efforts in visualization literacy. For example, chart interpretation strategy may be one explanation for the observed credible differences in participant performance, such as people who perform poorly with the Stacked Bar but well with all other chart types (e.g., Figure 9). Another potential literacy-focused application of the models described here would be to use them to drive feedback and educational interventions, to make people aware of opportunities to improve their skill and reliability in interpreting visualizations. Efforts in improving “low-level” visualization literacy might leverage prior work exploring visualization interpretation strategies, such as arcs, angles, areas for Pie charts [40], or perceptual proxies for Bar charts [52], [53].

Beyond improving design recommendations, combined models of variance in individual performance and visualization literacy might also be used to drive adaptive user interfaces (user interfaces which adapt to individual characteristics). Adaptive user interfaces typically employ models targeting specific personal attributes, taking into account a range of ways in which people may differ—whether it be in abilities (both physical and perceptual) [63], [64], preferences as influenced by culture [65], the users’ surroundings and personal characteristics [64], [66], [67], or their demographic profile [68]. In the visualization community, the Draco project from Moritz et al. could plausibly be extended to take as input probabilistic instead of deterministic rules, opening up new avenues for visualization recommendation that accommodate model-based individual differences [5]. Probabilistic representations of rank data, potentially drawing on techniques such as hypothetical outcome plots (HOPs) [69], might also be explored as a means for presenting ranks of visualization performance to designers while faithfully representing individual variance like those modeled here. Future work might aim to better understand individual traits and their relationship with individuals’ performance, developing visualization systems that better optimize accessibility for wider audiences.

7 Limitations

While the resulting model reflects error patterns we observed on a set of common visualizations, it is not without limitations, raising questions of validity and reliability as we move beyond the particular visualizations and task considered. For example, a new visualization type with the same task could lead to markedly different patterns of errors, which could require changes in the model to accommodate. We ultimately used a Bayesian zero-inflated beta model, which in prior simulation studies has been shown to be more reliable than other models on similar types of responses [70]. We also compared this response distribution to other reasonable alternatives (e.g., hurdle lognormal models) and found our substantive conclusions were not sensitive to the response distribution. That said, it is possible some unmodelled (or mis-specified) aspects of people’s behavior could have an impact on error in ways not accounted for by our model. At the same time, further refining resulting models might be considered a strength of model-based analysis, as it would suggest differences in chart reading require different statistical constructs to be adequately captured.

Another limitation is that model-based methods may require additional effort, in part due to its unfamiliarity in research communities that have established norms, practices, and expectations surrounding analyses and reporting. One of the goals of the present paper is to explore these differences by replicating a familiar visualization study, and to demonstrate some of the possible benefits of model-based approaches for answering research questions that would be difficult or impossible to address through traditional methods.

8 Conclusion

In this paper, we undertook the “substantial chore” [1] of modeling variance and correlation in individuals’ performance on classical graphical perception tasks. Our work identifies problems with two common practices in visualization research: (1) modeling or reporting visualization rankings only for the “average observer” and (2) reporting only log error. The problem with the first practice is revealed by our finding that substantial between-individual variance exists for even these elementary visualization tasks; e.g., as much as 30% of people are likely not “best with the Bar”, and different people may depart substantially from the canonical ranking of visualization type effectiveness. The problem with the second practice comes from the increasing consideration of factors other than raw effectiveness—aesthetics, metaphor, etc.—in visualization design. We now have the tools to analyse errors on the raw scale, laying groundwork for future studies evaluating how visualization designers interpret effect sizes on different scales (e.g. log versus original units), and how designers reason about the cost/benefits of optimizing for one measure over another. Ultimately, we believe the field should move beyond the use of rankings prevalent in prior work, building a more complete picture of the spectrum of human performance on visualization tasks so that we can create more practically-applicable recommendations for visualization designers, and support the important work of measuring and promoting visualization literacy.

Acknowledgments

This work was supported in part by a grant from the US National Science Foundation (#1815587, #1815790).


Matthew Kay is an Assistant Professor jointly appointed in Computer Science and Communications Studies at Northwestern University. He works in human-computer interaction and information visualization, and particularly in uncertainty visualization, personal health informatics, and the design of human-centered tools for data analysis. He co-directs the Midwest Uncertainty Collective (http://mucollective.co) and is the author of the tidybayes (https://mjskay.github.io/tidybayes/) and ggdist (https://mjskay.github.io/ggdist/) R packages for visualizing Bayesian statistical model output and uncertainty.

Lane Harrison is an Associate Professor in the Department of Computer Science at Worcester Polytechnic Institute. Prior to joining WPI, he was a postdoctoral fellow in the Department of Computer Science at Tufts University. He obtained his Bachelor’s and PhD degrees in computer science from the University of North Carolina at Charlotte. Lane directs the VIEW group at WPI, where he and his students leverage computational methods to understand and shape how people use visualizations and visual analytics tools.