The Perceptual Proxies of Visual Comparison

Nicole Jardine, Brian D. Ondov, Niklas Elmqvist, *Senior Member, IEEE*, Steven Franconeri
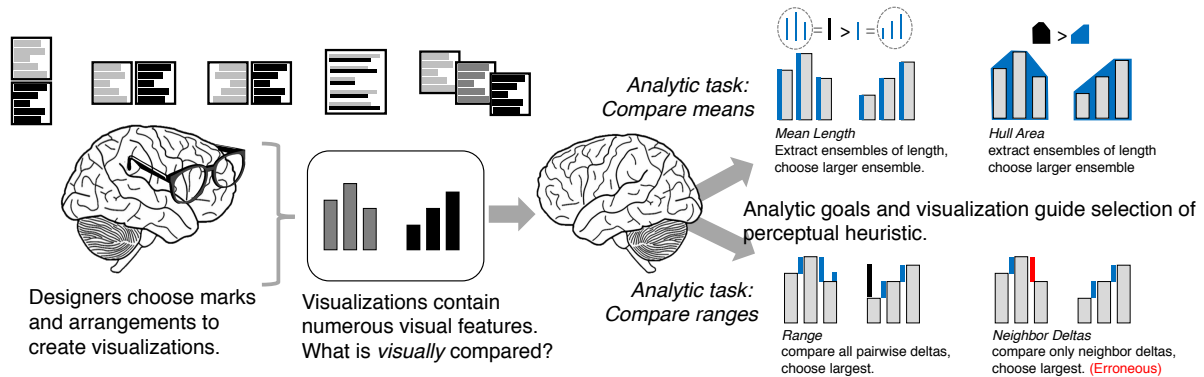

Fig. 1: What's visual in visual comparisons, such as finding the larger mean value? We identify mark arrangements that allow for better performance across comparison tasks. Combining previous results with the results of two new tasks fails to produce a clean ranking of arrangement effectiveness across tasks. We argue that to explain these complex patterns of performance, we first need a perceptual explanation of how visual comparison actually unfolds. Viewers likely perform these mathematical comparison operations with perceptual proxies. We propose and evaluate a candidate set of proxies for two visual comparison tasks.

Abstract—Perceptual tasks in visualizations often involve comparisons. Of two sets of values depicted in two charts, which set had values that were the highest overall? Which had the widest range? Prior empirical work found that the performance on different visual comparison tasks (e.g., "biggest delta", "biggest correlation") varied widely across different combinations of marks and spatial arrangements. In this paper, we expand upon these combinations in an empirical evaluation of two new comparison tasks: the "biggest mean" and "biggest range" between two sets of values. We used a staircase procedure to titrate the difficulty of the data comparison to assess which arrangements produced the most precise comparisons for each task. We find visual comparisons of biggest mean and biggest range are supported by some chart arrangements more than others, and that this pattern is substantially different from the pattern for other tasks. To synthesize these dissonant findings, we argue that we must understand which features of a visualization are actually used by the human visual system to solve a given task. We call these *perceptual proxies*. For example, when comparing the means of two bar charts, the visual system might use a "Mean length" proxy that isolates the actual lengths of the bars and then constructs a true average across these lengths. Alternatively, it might use a "Hull Area" proxy that perceives an implied hull bounded by the bars of each chart and then compares the areas of these hulls. We propose a series of potential proxies across different tasks, marks, and spatial arrangements. Simple models of these proxies can be empirically evaluated for their explanatory power by matching their performance to human performance across these marks, arrangements, and tasks. We use this process to highlight candidates for perceptual proxies that might scale more broadly to explain performance in visual comparison.

Index Terms—Graphical perception, visual perception, visual comparison, crowdsourced evaluation.

◆

---

1 INTRODUCTION

Visual comparison is a core perceptual task in data visualizations [10]. An epidemiologist might use two bar charts to assess whether, across all age groups, there is a larger overall population of women than men. An education researcher might use two bar charts to assess whether one group of students' test scores has a larger range than another. Nei-

---

- *Nicole Jardine is affiliated with Northwestern University and with the Cook County Assessor's Office in Chicago, IL, USA. E-mail: njardine@cookcountyassessor.com.*
- *Brian Ondov is with the National Institutes of Health in Bethesda, MD, USA and University of Maryland in College Park, MD, USA. E-mail: ondovb@umd.edu.*
- *Steven Franconeri is with Northwestern University in Evanston, IL, USA. E-mail: franconeri@northwestern.edu.*
- *Niklas Elmqvist is with University of Maryland in College Park, MD, USA. E-mail: elm@umd.edu.*

ther of these comparison tasks need rely on the averages of each of these sets, or identification of individual values. They simply require a judgment of which set's mean or spread is larger than the other. Although certain visual channels [3, 18] are known to convey higher-precision information (position) than other marks (hue), existing evaluations of visual comparison performance suggest that there is no single mark or spatial arrangement that optimizes visual comparison.

In prior work, Ondov et al. [20] evaluated the perceptual precision of visual comparisons of the "biggest delta between items" and "biggest correlation between sets" for different visualization marks (bar, line) and spatial arrangements (stacked, mirrored, adjacent, superposed, and animated). Precision was not optimized by a single mark type or spatial arrangement. Instead, the precision of visual comparison depended on an interaction of mark, arrangement, and task (Figure 2). The best static chart for a precise delta comparison, for example, was one that was spatially superposed, rather than juxtaposed in a stacked format (Figure 1), validating an intuitively-
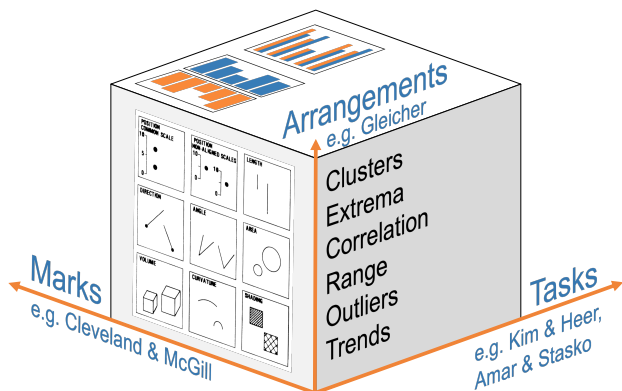
Fig. 2: Visual comparison depends not on a single dimension of mark, arrangement, or task, but of the interactions between them. These interactions can be represented as a cube. Our present goal is not to examine the full space of the cube, but rather to understand how a viewer uses visual features to serve analytic task goals depending on the marks and arrangements they see.

motivated guideline from Gleicher et al. [10]. Not predicted by current guidelines was the discovery that, to support biggest delta comparisons in data, animation provided the most precise performance. Animation, however, did not perform as well for comparisons of correlations.

We first extend this work to empirically evaluate performance across these arrangements for two new tasks—"biggest mean" and "biggest range"—and again find that performance is strongly impacted by spatial arrangement. Comparison was most precise when these two datasets were vertically stacked, and least precise when the datasets were superposed. This pattern of which arrangements were best was strikingly different than for the previous pattern for "biggest delta between items" and "biggest correlation between sets."

Why is there not a single clean emerging answer, where a given arrangement is best across various tasks? This empirical evidence for the more complex nature of visual comparison is consistent with the idea that it requires a series of visual actions at a variety of scales from one object, to multiple objects, to whole sets of objects [9]. Taxonomies of visual comparison describe multiple stages of perceptual and cognitive steps [4, 9], and vary in describing one or many types of visual comparisons, but the visual mechanisms supporting these stages are unclear. We argue that an empirical description of the precision of visual comparison across each combination of mark $\times$ arrangement $\times$ task would be valuable, but unlikely to scale to have predictive beyond its status as a lookup table. A different approach is required. We propose that instead of continuing to fill out the entries of the cube in Figure 2, it may be more productive to study *perceptual proxies* of a visualization are actually used to reason about a visual comparison task. The goal of this approach is to begin to identify the proxies for visual comparison, as opposed to merely gathering additional empirical data.

We propose several candidate visual proxies and implement them as simulations. This allows us to evaluate each proxy's objective performance in performing the same task given to the human participants. But by comparing the performance of these proxies to human performance on the same questions, we can rank proxies by which most closely mirrors human performance.

Our evaluation of perceptual proxies suggests that although these two comparison tasks show similar arrangement-driven patterns of results, these patterns are consistent with different proxies. To compare means, the visual features that best predict human performance are the ensembles of lengths and the centroids of the bars. To compare ranges, the best-predictive visual features were those that compared deltas be-

tween all items within a set or only between neighboring items.

The complex dependency of visual comparison performance on combinations of marks, arrangements, and tasks might soon be predicted by a model that accounts for such a set of perceptual proxies. We speculate that these proxies follow two broad categories of global or set-based visual comparisons and focal or item-based visual comparisons, drawing from perceptual psychology research on that division in types of visual processing.

This work contributes (1) results on how visualization design arrangement affects two new comparison tasks with two new crowd-sourced graphical perception experiments for visual comparisons of "biggest mean" and "biggest range", (2) a framework of perceptual proxies for visual comparison, (3) implementations of these proxies for empirical evaluation, and (4) data generation procedures designed to estimate the amount of signal needed in the data to support a given visual comparison between sets of items. Our findings present a first step toward a model of human perception during visual comparison.

2 RELATED WORK

Here we review empirical research on visual comparison and suggest that visual comparisons can often be classified as being made between isolated parts (i.e., a bar distinct from other bars in its set) or whole sets (i.e., all the bars). Frequently, these correspond to analytic tasks for which a goal is identification or comparison of items, or of sets, in data. We propose that these analytic task goals correspond to proxies that determine the visual features that a viewer uses for visual comparison.

2.1 Visual Comparison

Frameworks of visual comparison are often driven by the analytic goals of the viewer; for example, Amar et al. [2] names comparison as a high-level "compound task" central to many specific analysis tasks. Gleicher et al. [9] conducted a review of the taxonomies of visual comparison with the goal of a top-down approach examining what people do when they do visual comparison. They propose that, broadly, to "compare" in a visualization involves at minimum three components: targets (which multiple "items" are being compared), the relationship between these items, and an action (a mechanism operating on the relationship between these targets). These targets may correspond to items or sets of data. Yi et al. [28] discussed adding a comparison task to their seven-task taxonomy, but ultimately decided against it because "compare is a higher-level user goal or objective." In contrast, our work here and in past work [20] frames comparison as a relatively low-level perceptual task.

2.2 Visual Comparison: Parts vs. Wholes

Visual comparison across multiple series can be comparisons of items-to-items (focal) or sets-to-sets (global). We refer to these as different visual spans of comparison. One study tested visual comparisons between smaller regions of time series charts, or between larger areas of time series charts [14]. The data were consistent with the idea that these are distinct visual actions: viewers conducting focal visual comparisons of small regions benefited from shared-space charts that overlapped in space, whereas viewers conducting visual comparisons over the entire sets were better served by separated-space charts.

Another study [20] investigated the perceptual precision of two visual tasks, and tested a series of chart arrangements to see which arrangements best supported each visual comparison. In one task, people saw two sets of 7 items and compared them to identify which of the 7 items changed the most between the first and second chart. In this item-to-item comparison, visual comparison was most precise when the charts were spatially superposed within a single chart space, or had an animated transition between them (temporally superposed). These arrangements created visually salient features that mapped to these item-to-item changes, either as overhang in superposed charts or as a salient motion cue that captured attention in the animated charts. Comparison was less precise when the sets were separated spatially such that the two chart axes were arranged horizontally, vertically, or mirroring each other. In another task, people saw two pairs of bar charts that were correlated with each other to some degree. The viewer's task

was to pick the pair with the strongest correlation. Here, performance was best when each pair of correlated charts had axes that mirrored each other. Speculatively, viewers could rely on he rapid perception of symmetry between chart items, which in this task happened to indicate correlation between charts.

These focal vs. global modes have analogues in perceptual psychology research as different perceptual modes. A viewer can opt between these modes (with a mandatory dominance for global mode for a first glance at a new image) to flexibly meet the demands of focal tasks (identify items within sets of larger items) or global tasks. In other words, when people are presented with a visual stimulus, they can flexibly choose whether to attend to the "forest" or the "trees" of that stimulus at varying spatial scales. These different attentional modes switch which scope of a visual stimulus is used for a task goal [19].

In sum, we predict that item-to-item comparisons are facilitated by animated and superposed charts that place these items in close proximity to each other, allowing a focused mode to subserve this more focused task. Set-to-set comparisons may be facilitated by arrangements that spatially separate these sets, allowing the visual system treat each set globally as its own unit, because preserving the values of specific items is not necessary. Different arrangements support different comparisons because, we propose, these arrangements offer visual proxies that people actually use to conduct visual comparisons. Biggest delta benefits from animation because a salient visual cue (motion speed of an item changing the most) naturally maps to, or is compatible with, the apprehension that the data value is also changing the most.

2.3 Interim Summary

Empirical evaluation of how human observers perceive values or relationships in a visualization suggests that people can rely on a variety of proxies, operating over the marks and arrangements and other visual properties of a visualization, to meet the demands of an analytic task.

3 HYPOTHESES

Our original hypothesis was that the two new tasks MaxMean and MaxRange would show similar patterns of performance to the MaxCorrelation and MaxDelta tasks from previous work [20]: that comparisons of means would, like comparison of correlation, benefit from mirrored charts. We were surprised to see an almost opposite pattern of performance, leading us to turn out attention toward the path of seeking perceptual proxies that might provide more explanatory power for why these strikingly different patterns emerged.

4 METHODS

Our first goal is to quantify the perceptual precision with which human observers can perform visual comparison between two charts of horizontal bars, and to measure whether that precision differs based on the spatial arrangement of those charts.

4.1 Tasks

We chose two tasks to build on previous work:

- MAXMEAN: Of two sets, which had the largest average (mean) value? Difficulty is increased by reducing the delta between the mean values, so that the difference between sets is less distinguishable. Displays were controlled so that the largest single-item in a chart was not predictive of that chart having the largest mean and so that charts in a trial were of approximately equal variance. Within-chart variance ranged from .04 to .09. Harder discriminations (smaller mean deltas) spanned the low to high variance range, whereas easier discriminations tended to be lower variance. The two extreme values that bounded data generation were directly included in a randomly selected chart, ensuring that the highest or lowest individual value did not correlate with the correct answer, and allowing us to examine whether participants used this as a proxy.

- MAXRANGE: Of two sets, which had the widest range between its min and max values? Difficulty is adjusted by varying the

delta value between the range widths of the two charts. Since range may be a less widely-understood concept, we gave our participants a detailed description with a simple example, both at the start of the trials and each time they were incorrect in training trials. See past studies [12, 22] for similar tasks.

4.2 Titer Staircase Method

Our goal was to quantify the magnitude of the difference of means for the MAXMEAN task, and the magnitude of the difference of range widths for the MAXRANGE task, for each arrangement. We dynamically titrated stimulus difficulty using a *staircase method* [20]. Briefly, this method scales task difficulty on a trial-by-trial difference. The end result quantifies a *titer*: a value between 0 and 1 that scales the magnitude of the difference between stimuli to determine the threshold at which a participant can barely perform a discrimination task (expected performance of 75%). Fig. 3 illustrates this.

In the MAXMEAN task, the titer controlled whether there was a large difference between the two chart means (large titer: easy task) or a smaller difference between the two means (smaller titer: harder task). For MAXRANGE, the titer controlled whether there was a large difference between the range widths of the two chart means or a smaller difference between the range widths of the two chart means (Fig. 1, left; orange), making this pair stand out more from the baseline pair.

Titers and stimulus datasets changed trial-by-trial depending on participant performance for the previous trial. Briefly, the initial titer value that scaled the difference-of-means or the difference-of-ranges was 0.5. Depending on whether the participant's discrimination in that trial was correct or incorrect, the next trial adjusted the titer by -0.01 or +0.03. The goal of this staircase procedure is that by the end of the trials, the titer reflects a stable magnitude of signal that allows the participant to perform with 75% accuracy for that arrangement.

4.3 Arrangements

As in previous work [20], datasets were presented in blocks of 5 different arrangements (Fig. 1):

- **Stacked**: Vertically juxtaposed small multiples (i.e. one chart is placed above the other).

- **Adjacent**: A more commonly used instance of small multiples, in which data series are horizontally juxtaposed.

- **Mirrored**: This horizontally "mirrored" variation of *adjacent* opposes the direction of the x-axis in each chart. The Gestalt nature of bilateral symmetry suggests this arrangement prompts "set" proxies rather than "item" proxies in viewers.

- **Superposed**: A combined chart depicting both data series within the same space. Past work has claimed that overlaying values, or superposition, minimizes eye movements and memory load, and may lead to efficient comparison [10].

- **Animated**: In this "arrangement," a single chart is transitioned, or morphed, from one data series to another over time, using cubic interpolation to ease the transitions [7].

In trials, the order of these 5 blocks was rotated, and each rotation reversed, for a total of 10 possible orderings, each of which was presented to 5 participants.

4.4 Task and Procedure

Before each trial began, the screen contained a centrally placed fixation dot and outlines of where the charts would appear. Participants clicked a button to start the trial. After a countdown, the visualization appeared for a short, fixed time. Static and animated charts were shown for 1.5 seconds. In contrast to previous work, at the end of the impression, both sets of data were removed from the display. Participants then clicked on the orange or blue button corresponding to the orange or blue set of bars to provide a response. For the MAXMEAN task they were instructed to "Click on the chart that had the biggest mean
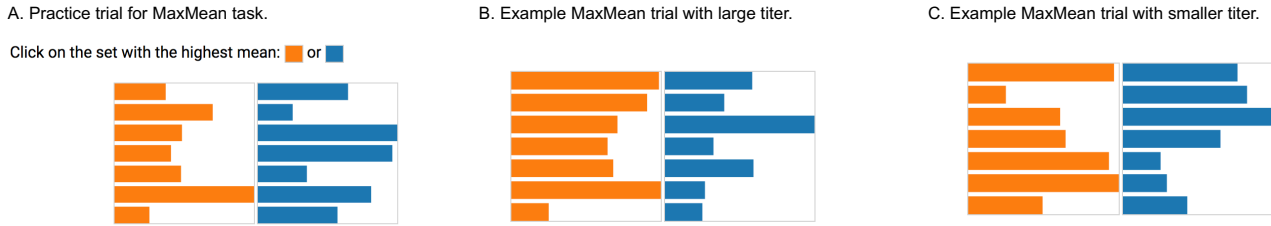
A. Practice trial for MaxMean task.

Click on the set with the highest mean: ▮ or ▮

B. Example MaxMean trial with large titer.

C. Example MaxMean trial with smaller titer.

Fig. 3: In the staircase procedure, a correct response produces a smaller difference in the subsequent trial.

values"; for MAXRANGE to "Click on the chart that had the widest range between min and max values." Participants were informed if they were correct and, if incorrect, what the correct answer was. This feedback was provided to make the task more engaging and to reinforce the goal. Between trials, the titer was adjusted based on the response. To seek 75% accuracy during trials, the titer was increased three times as much for an incorrect answer as it was decreased for a correct answer (see Figure 3). Dynamic data generation according to the titer value is described in Supplemental Materials A. Each experiment included all five arrangements. There were twenty trials for each arrangement, and arrangements were blocked. The order of the arrangement blocks was changed between participants.

4.5 Training

Before training, participants were shown examples of stimuli and the task. Before each arrangement block, participants were given a time-unconstrained version of the task, which they were required to answer correctly before proceeding (once for the MAXMEAN task, three times for MAXRANGE). Additionally, the first non-animated arrangement given to a participant followed untimed training with three timed training trials, which were identical to the real trials except that they always had the easiest (largest) titer. Data were regenerated on incorrectly answered training answers to minimize answering by elimination. Video demos are in Supplemental Materials.

4.6 Participant Recruitment

Based on previous work, we predicted $N = 50$ would provide sufficient statistical power to reliably detect the presence or absence of an effect of arrangement. We also expected that more participants would struggle to understand the MaxRange task, so we collected data from 50 MTurk workers for the MaxMean task and 54 MTurk workers for MaxRange. Participants were asked to self-select out of the study if they had color vision deficiencies. Each participant completed the staircase procedure for all 5 arrangements of one of the tasks (MAXMEAN or MAXRANGE). Worker IDs were used to ensure uniqueness of participants across all such combinations. All workers recruited for participation were adults in the United States.

5 RESULTS

We evaluated the magnitude of deltas required in the data for non-expert participants to reliably identify the set with the largest mean (Experiment 1) or the largest range (Experiment 2).

5.1 Exclusion Criterion

The dependent measure in these experiments is the average titer value from the final 10 trials of each task. We excluded datasets from participants whose average titers (averaged over spatial arrangement) was more than 2 standard deviations from the mean. This procedure eliminated 1 observer each from the MAXMEAN and MAXRANGE tasks.

We also adopted a second criterion. In a staircase procedure, the goal is to find a converged titer value for which a participant is 75% accurate. The procedure fails if a participant repeatedly reaches ceiling performance (a minimum titer value of 0.01) or floor performance (the

maximum titer value of 1.0) because at this point the stimuli cannot titrate difficulty beyond these floors and ceilings.

Because viewers performed tasks for 5 arrangements, we excluded participants for whom there were at least 5 trials of floor or ceiling titer values. These criteria excluded 0 from the MAXMEAN task, but for MAXRANGE there was 1 trial in which a participant reached ceiling performance and 109 trials who repeatedly reached the floor titer (largest delta). We excluded 7 participants for whom there were at least 5 (up to 22) trials of floor titer values (one of whom was also the participant excluded with the standard deviation procedure), leaving $N = 49$ for the MAXMEAN task and $N = 47$ for MAXRANGE.

5.2 Titer Analysis

We computed each observer's mean titer values from the final 10 trials for each arrangement. We used the final 10 trials because visual evaluation of trial-by-trial data suggested that this was approximately when the staircase procedure stabilized around a narrow range of titers, for most participants. Thus we analyze the final 10 titer values achieved for each of the five arrangements, for each subject.

5.3 Exp. 1 and 2: MaxMean and MaxRange

Figure 4 (far left) displays the mean final 10 delta values for the MAXMEAN task, and (second from the left) displays the mean values for the MAX RANGE task. These titer values correspond to the differences between the charts being compared. Means could be discriminated when they differed by approximately 5-8% of the chart axis, and range widths when they differed by approximately 14-17%. For both tasks, the precision of visual comparison was affected by arrangement.

Titer values for the present experiment were analyzed with a mixed ANOVA to test for experiment-level and arrangement-level effects.

Titer values varied between experiments, $F(1, 94) = 9.06$, $p = .003$, $\eta_p^2 = 0.09$, but this is likely because the titer values scale to different stimulus changes between the two experiments. As such we avoid a meaningful comparison between differing titer values.

More meaningful is that there was a significant effect of arrangement on precision, $F(3.09, 290.7) = 8.17$, $p < .0001$, $\eta_p^2 = 0.08$, without evidence for an interaction between arrangement and experiment, $F(3.09, 290.7) = .34$, $p = .85$, $\eta_p^2 = 0.004$, both Greenhouse-Geisser corrected. This suggests that arrangement produces largely similar effects on the precision of visual comparisons of means and of ranges.

We conducted pairwise comparisons within each experiment. Stacked charts were the most precise arrangement overall for both tasks. Precision was better in stacked relative to adjacent charts for the MAXRANGE task, $t(49) = 2.73$, $p = .009$, although not significantly better in the MAXRANGE task, $t(47) = 1.70, p = .09$. Superposed charts resulted in the lowest precision for both tasks. Note that these patterns are strikingly different compared to prior evaluation of visual comparisons of items, which were best supported by animated and superposed charts.

5.4 Accuracy

The goal of a staircase procedure is to titrate the task's difficulty so difficulty might change across arrangements, but that accuracy is equiv-
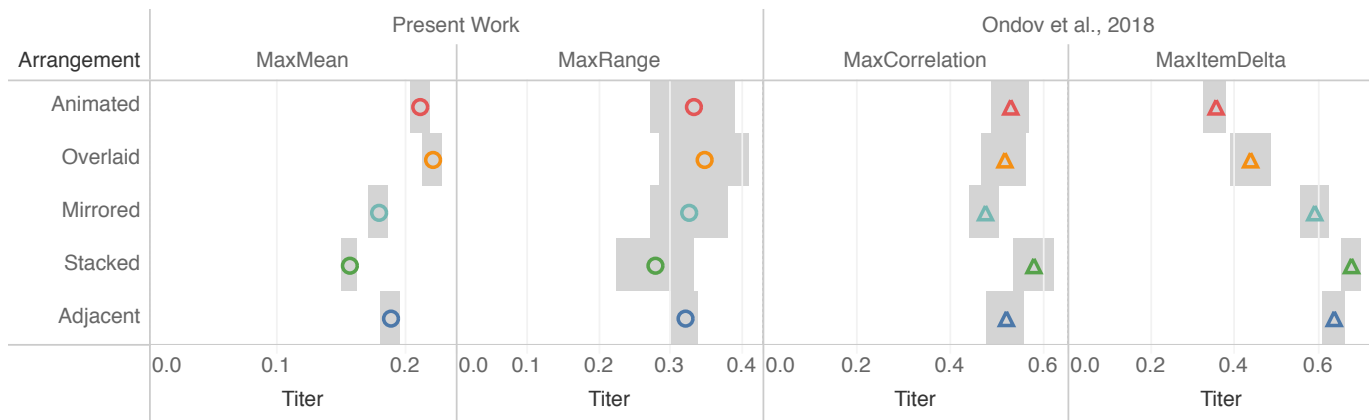
Fig. 4: Means of averaged final titer values across participants performing the MAXMEAN and MAXRANGE tasks. Smaller titers correspond to more precise differences between means (range widths). The precision of both the MAXMEAN and MAXRANGE tasks was affected by chart arrangements. Also presented are titer values from previously published empirical evaluations of the precision of other comparison tasks. Note that different chart arrangements support different visual comparisons. Gray bars represent 95% confidence intervals.

alent between arrangements. Mean accuracy in the MAXMEAN task for each arrangement ranged from 76.4% (stacked) to 79.9% (mirror), with no evidence that accuracy was different between arrangements. This suggests the staircase procedure reliably converged for this task.

Mean accuracy in the MAXRANGE task ranged from 75.2% (superposed) to 84.7% (stacked), and a repeated measures ANOVA found that accuracy consistently differed between arrangements, $F(4, 184) = 4.34$, $p = .002$. The staircase procedure did not reliably converge for all arrangements in the MAXRANGE task due to large effects of arrangements on people's ability to perceive range widths. Stacked charts allowed for higher accuracy and high precision than other arrangements. In a pilot version of this experiment with fewer participants, we tested a larger initial titer value so that participants unfamiliar with statistical ranges could use a very large signal in this task, but found the same pattern: superposed charts simply outperform stacked charts regardless of initial task difficulty. Because of these strong effects of chart arrangement, future work using this method might better titrate task difficulty by relying on a greater number trials, or a different kind of adaptive titration procedure.

6 DISCUSSION

The precision of visual comparison of MAXMEAN and MAXRANGE tasks were best supported by vertically stacked charts, and least supported by superimposed charts. This is in contrast to previous findings for item-item comparisons, which were best supported by animated and superimposed charts. These differential findings are consistent with the principle that item-item and set-set comparisons of data are supported by arrangements that enable focal and global visual feature comparison, respectively. That vertically stacked charts were best may be unsurprising because marks are horizontally extending bars. Viewers can simply slice downward to extract lengths between charts.

We hypothesize that the reason these arrangements best supported these tasks is that visual system can, somewhat flexibly, adopt a series of focal or global "perceptual proxies" that operate on a visualization. To understand visual comparison we must understand not only what visual features exist in a given visualization, but when they are used for which tasks. But these are post-hoc explanations. Are people actually using that proxy for both tasks? Or are people using different visual features to support MAXMEAN and MAXRANGE? This question cannot be answered by evaluating the precision of visual comparisons. Next we study whether the same patterns of precision arising from the same arrangements for different tasks arise because of the different visual features that people use for comparison.

7 PERCEPTUAL PROXIES

Instead of mathematical operations, people more likely rely on *heuristic perceptual proxies* to extract data values and patterns from data visualizations. Heuristics are shortcuts that rely on a simplified metric—a *proxy* metric—to convey the desired information. Perceptual heuristics are easily computable features that (at least) correlate with the right answer, allowing viewers of visualizations to use visual features as accurate or inaccurate proxies for the data those features represent.

One example is the perception of correlation in scatterplots. The perceptual process does not appear to calculate the true mathematical correlation, and there are instead proposals for multiple proxies that might underlie correlation perception [12, 22, 27], including the aspect ratio of the bounding box surrounding the points [27]. This proxy can be efficient because it relies on a rapid perceptual process of inspecting a shape boundary around the points. This proxy is also fairly accurate [27]: the width of the bounding box of the visualization corresponds strongly to the correlation in the data.

Scatterplots are commonly used to show correlation data, but not all links between visualization and analytic task are so strong. Furthermore, an analyst might not always know what kind of visualization they will see. Finally, the selection of a proxy will be strongly affected by the visual features that are available in a visualization.

Different proxies may afford not only different data patterns, but different conceptual associations of what those values might mean. The same two data points graphed as two bars or as two endpoints of a line chart can evoke different visual actions taken on visual features of the visualization. Zacks and Tversky [29] presented simple line or bar charts to participants for open description. Participants' descriptions of bar charts overwhelmingly tended to involve discretizing words, such as "Y is higher than Z," and descriptions of line charts entirely used continuous relations, such as "as X increases, Y decreases." This bar-line message correspondence seems to occur because the type of mark is associated with metaphors of bars being containers or groups, in contrast to lines, which are continuous entities.

7.1 Candidate Proxies

A visualization contains any number of visual features potentially available as a proxy for a given task, such as the lengths of the topmost items of each set, or the perceived symmetry of each set. Different visual features might be better proxies than others for different visual comparisons. Here we explore which visual features appear to be most similar to participant performance (making the same decision), when used as a proxy for MAXMEAN or MAXRANGE. We developed two broad categories of candidate features, informed by research in both visualization and perceptual psychology.
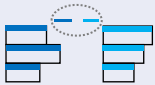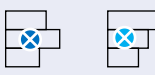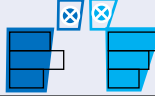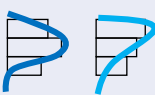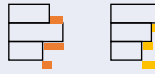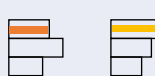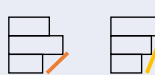
Possible proxies		Description	Visual cognition principle
Global Proxies	Mean*	Extracts lengths of bars of each set, computes ensembles, chooses chart with longer ensemble.	People can extract the mean size of a set of items [25], though that mean is likely through some proxy.
	Centroid	Picks chart with largest centroid of the bar areas (along just relevant x axis).	Eye movements rapidly deploy to centroids of groups, but those centroids appear to be computed across the bounding hull of the objects, not their true center of gravity [17].
	Hull Area	Calculates convex hull of chart, picks chart with bigger hull area.	
	Hull Centroid	Calculates convex hulls, picks larger centroid (along relevant dimension only).	
	Trap Area	Draws trapezoids between each chart's top and bottom bars, picks bigger area.	A shape's external boundaries can be more visually salient than internal boundaries [6], which could produce overweighting of the first and last bars when judging the boundary contour.
	Trap Centroid	Draws trapezoids between each chart's top and bottom bars, picks trapezoid with larger centroid.	
	Symmetry Bias	Calculates skew (i.e. symmetry) of each set, chooses which chart is less skewed (i.e. more symmetric).	People are sensitive to symmetry [24] and are biased to select symmetric objects even when the task is not a symmetry judgment [15].
Focal Proxies	Range*	Extracts all pairwise deltas within a set, chooses set with the longest pairwise length difference.	Length differentiation has high acuity [16].
	Biggest Mover Pair (Abs)	Between charts, finds largest delta between item pairs (a1-b1, a2-b2…), picks chart with largest positive delta.	
	Biggest Mover Pair (Rel)	Same as Biggest Mover Pairwise, but scaled relative to the smaller item within the pair.	
	Biggest First Item	Compare top items, picks chart with larger top item.	When faced with multiple objects, people are biased to attend to the top object [26].
	Biggest Middle Item	Compare middle items, picks chart with larger middle item.	People might select a group's center item [17] and then select the larger of the two [16].
	Slope Min to Max	Finds each chart's min/max, computes slope between them. Picks the chart with the least-vertical slope.	People might select min/max outliers [13], then calculate offset [16].

Fig. 5: A set of candidate perceptual proxies that might be used in visual comparison of means and ranges (and possibly other tasks). The proxies are arranged by their correspondence with hypothesized distinctions between global and local visual scopes.

7.1.1 Global Features

Global-level features describe properties aggregated over a visual set of items, rather than comparing two focal items. Viewers can rapidly compute global statistics such as the mean of a collection of items [1, 11, 23, 25], though from present work it is unclear if this ability is mediated by a proxy. One high-precision proxy is that the lengths of bars in a set are veridically averaged together and the chart with the largest ensemble length is chosen as the answer for the task.

The mean length feature tests this veridical averaging. Viewers might also perceptually organize the bars into a coherent object, such that what they perceive is the convex hull of the bounded object that includes the heights of the bars and the white space between bars, and then compare the centroids or areas of these two hulls. These object boundary proxies might be subject to perceptual biases, such as over-weighting outer edges in contour judgments [6]. Empirical research on human attention suggests that the allocation of attention through-

out visual displays is preceded by the organization of the scene into objects and groups [5], and that the center-of-area of those objects can be rapidly computed [17]. The hull area and hull centroid proxies test whether this visual feature is consistent with participant responses and consistent with differences in the data. Note that for superposed charts, the two hulls are overlapping, such that this particular visual feature may be harder for people to see because it involves filtering using color rather than space (as with the stacked, mirrored, and vertical arrangements). Finally, people are highly sensitive to symmetry in displays [24] and are biased to select symmetric over asymmetric information [15]. One possible heuristic is that people use symmetry as a proxy for range, such that any chart that is less symmetric is selected as the one having the bigger range.

We suspected that visual features describing global, rather than focal, characteristics of the visualization would be better predictors of human decisions in the MAXMEAN task, because this task involves set-level comparisons. Conversely, because precision for a MAXRANGE task requires the isolation of individual item lengths for comparison between sets, we predicted visual features describing focal characteristics would be better predictors of human decisions.

7.1.2 Focal Features

Focal features describe pairwise differences between two items. People can discriminate small differences in line segment lengths [16]. Chart viewers might be sensitive to the deltas, either between charts (Biggest Mover Pair) or within a chart (Neighbor Delta). In addition, focal attention can be biased to attend to the topmost item in a collection [26], so one possible proxy is that people compare only the lengths of the topmost items of the two sets (Biggest First Item).

7.1.3 Implementations

We implemented these global and focal perceptual proxies for all charts (Figure 5 and pseudocode in Appendix B in the supplemental materials). Some visual features may be salient [13] to human observers, but not useful for an analytic task (uncorrelated with the answer). For example, the delta between adjacent bars (i.e., the amount of overhang) might be a salient and useful indicator for an analytic task involving comparing items, but if the viewer's goal is to compare means, relying on this feature should impair task performance.

To evaluate these proxies, we simulated what would happen if each proxy was tested on every data series combination that each observer actually saw in the two experiments. Each proxy was used to make a decision about a visual comparison (e.g., Hull Area generated a convex hull around each of the two charts, calculated their areas, and evaluated the pixel difference in their areas), and provided an "answer" to the task (i.e., larger area is used as a proxy for mean or for range).

Note that this procedure necessarily shows the proxies different stimuli depending on arrangement: because the stimuli have been titrated to respond to viewer accuracy, the charts "shown" for stacked stimuli will have different properties than the charts "shown" for superposed stimuli. Because the data in the charts "shown" to the proxies is arrangement-specific, proxies were implemented to be arrangement-invariant. The proxies were calculated using raw data values, the length of each mark, and the relative location of each mark (e.g., the first datum in a chart was at the "top" location), not as visual features extracted from an image-based representation. Future work should also test proxy performance using image-based implementations.

We computed two outputs for each of these proxies: which chart would the proxy have chosen, and was this choice correct? Although we excluded some participants from the titer analysis for low accuracy, we included their data in the simulation to allow for the future possibility of testing whether their poorer task performance is consistent with using different perceptual proxies than other viewers with higher-precision visual comparison.

Files that contain trial-by-trial data for properties of the stimuli, human responses, the pixel information used by each perceptual proxy to inform a heuristic about a chart decision, and each proxy's decision, for all combinations of arrangement and task, are posted at `https://osf.io/uenzd/`.

7.2 Perceptual Proxy Results

The goal of this proxy approach is to evaluate which visual features are consistent with human performance, and which are actually useful for the task. As such we evaluate the "decisions" of each proxy against two baselines. On what proportion of trials did the proxy agree with the participant's response? And on what proportion of trials did the proxy agree with the true answer of the stimulus? We treat all of the following results as initial speculations, and make no claims of their statistical reliability. These values are depicted in Fig. 6. A visual feature can be considered useful if a decision using the differences in that visual feature is consistent with the task-dependent differences in the data. The dots in Fig. 6 to the right of 50% show features that give above-chance performance at the task. We highlight a few patterns.

First, the most useful visual feature depends on comparison task. Broadly speaking, global visual features such as centroids are better candidates for the MAXMEAN comparisons, and focal visual features such as Neighbor Delta are better candidates for MAXRANGE comparisons. For the MAXMEAN task, visual features of the Mean lengths (global), Bar Centroids (global), and Biggest Mover Pair (focal) were the most predictive of the difference in the means. It was unexpected that the Biggest Mover Pair, which computes pairwise differences between chart items, predicted the difference of means at above-chance levels. It suggests that in the data, the largest between-item change (neighbor delta in superposed charts, motion in animated charts) was predictive of the chart means, moreso than other global features. For MAXRANGE, the Range proxy (which computed all pairwise distances between items) was most useful, closely followed by pairwise differences only between neighboring items (Neighbor Delta).

Second, people tend to make decisions consistent with using the most useful visual features: the bars that show agreement between proxy responses and human responses tend to follow the dots that shows the most task-relevant useful features in Figure 6.

Third, we note the absence of a symmetry bias. The Symmetry proxy, which uses stimulus symmetry as a proxy on which to make MAXMEAN and MAXRANGE decisions, was predictive neither of actual differences in means or range, nor of human responses.

Fourth, there is weak evidence of a bias for people to perform the MAXRANGE task with the global proxies of Hull Centroid and/or Area Trapezoid Centroid, to a higher degree than is actually useful in the task: note where in Figure 6 the human behavior bars are to the right of the proxy dots.

We speculate that these findings are broadly consistent with the idea that global visual features are useful for set-level visual comparisons, and local visual features are useful for item-level visual comparisons. MAXMEAN and MAXRANGE tasks benefit from the same chart arrangements, but use different emergent visual features in these chart arrangements for visual comparison. Visual comparison is afforded by more than precision of marks and their arrangements. The "visual" component of visual comparison may rely on a flexible suite of visual proxies that viewers can rely on to accomplish a given task, depending on what visual features are present. The slight bias to erroneously use global features for the MAXRANGE task raises the speculative possibility that, in some tasks and arrangements, viewers use global shape-based proxies even when these proxies are not useful.

7.3 Caveats, Limitations, and Future Directions

We enumerate several important caveats below and how they suggest possible future avenues of research.

1. **Whither the cube?** We began this work with the intent of filling out more cells of the mark × arrangement × task "cube" in Figure 2. While we quickly found that this model does not seem to scale, it is possible that more data (varying marks, arrangements, and tasks) might show that it can scale with some modifications. For example, dividing the tasks axis into "local" and "global" might allow a useful level of predictive validity, even if the underlying perceptual explanations are less satisfying.

2. **More data on the cells of the cube:** Existing work has only now empirically evaluated four visual comparison tasks, and
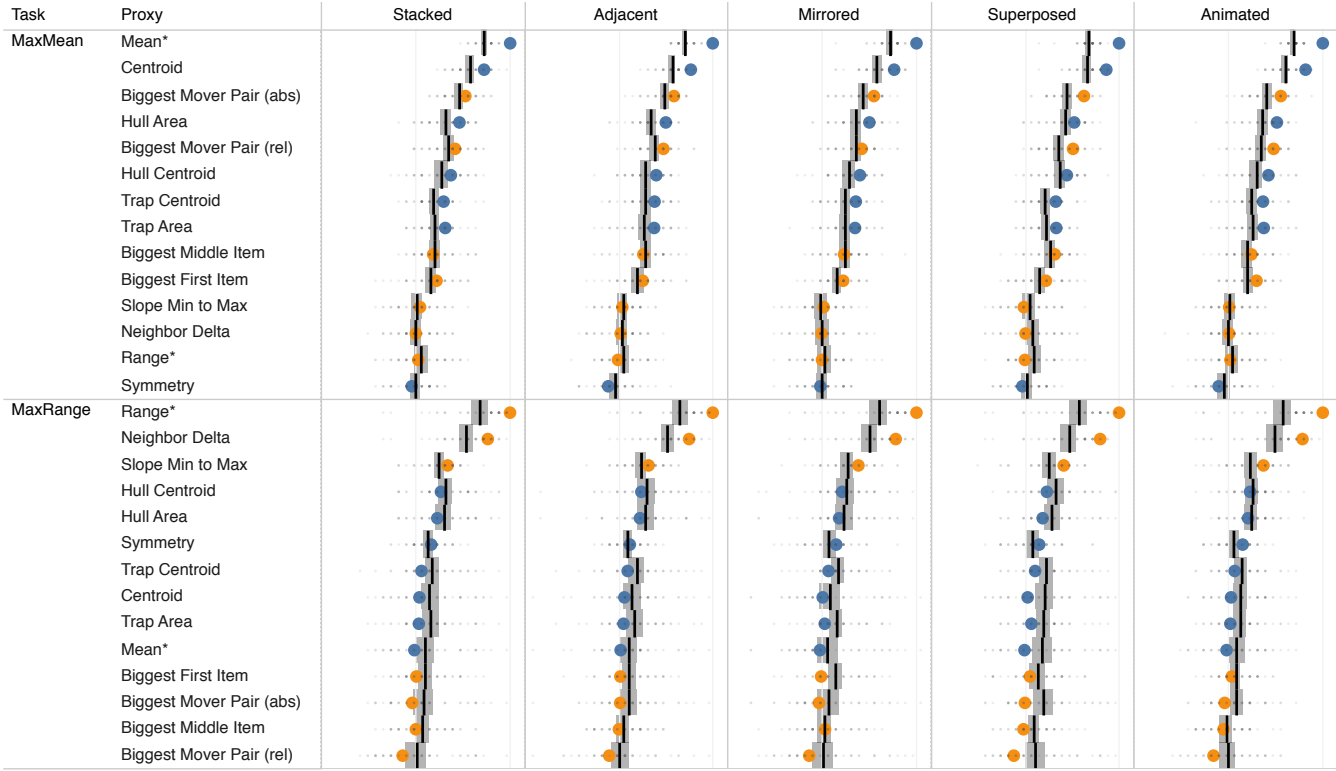
Fig. 6: Results of the two analyses of visual feature performance, split by task: MAXMEAN and MAXRANGE. The x-axis is the percentage of trials for which the visual proxy was predictive, for human behavior (vertical bars), and for true answer for the comparison (colored dots). The small dots show individual subjects, and light gray around the black lines shows 95% confidence interval. True answer dots are color-coded to show whether we informally coded them as a global proxy feature (blue) or focal proxy feature (orange). The true answer dots indicates that some features are more useful than others for a given visual comparison.

there are many more to test. The present experiment focused only on bar charts, due to their ubiquity in real visualizations, and combination of position and length encodings. It will clearly be important to fill out the "cube" by testing other marks, including lines, orientations, saturations, etc., both to test the robustness of the cube model and to provide more data for the enterprise of searching for candidate perceptual proxies for visualization tasks. The bar charts in the present work and that of Ondov et al. [20] were horizontally extended, an increasingly common design [8]. Other variants even of bar charts might reveal the use of different proxies for comparison.

3. **Different dependent measures:** The current experiments titrate the size of the compared difference, instead of other potential difficulty manipulations, such as the time allowed to make the judgment, or the number of objects in each set. But some visual comparisons need not be precise, and future work should test whether the same patterns of results hold for these alternative dependant measures. We would not be surprised by substantial differences in those results, as the perceptual proxies that help make precise judgments could differ substantially from those that allow coarser judgments, or judgments over larger sets of values.

4. **Artificial artificial artificial intelligence:** We measured the performance of workers on Mechanical Turk (whose slogan is "artificial artificial intelligence," because human workers take on tasks that are often automated) on making visual comparisons, and then matched their trial-by-trial performance to the predictions of each candidate proxy. Another way to test the match

of the proxies is to create "bots" that perform the same experiments, simulating Mechanical Turk workers that exclusively use only one proxy. Large numbers of these simulated participants could run through the real experiment, with difficulty titrated according to their performance across trials, in a way that produces mean titer levels for each "proxy bot," allowing another type of comparison of the proxy to human performance.

5. **Proxies merely fit performance:** This human-proxy agreement metric can only reveal features that are *consistent* with human performance. They cannot confidently reveal what features people actually use. As in science more broadly, we can rarely be sure of an answer, but we can be sure which of many generated potential answers is most consistent with the data.

6. **Proxy overlap:** The tested visual features are highly intercorrelated. As such, we cautiously refrain from strong conclusions about which best predict performance. As above, future work should use more sophisticated modeling that accounts for this shared variance. It could also rely on datasets that are intentionally designed to maximally differentiate among the predictions of the candidate proxies. For example, our data generation for the MAXMEAN task was specifically designed so that the largest item was not predictive of the largest set mean. As such a largest-item proxy could not be useful for the current MAXMEAN data sets, but could be for other data sets.

7. **A proxy for proxies:** The proxies implemented here did not use a computer visual system to "look at" pixels of a chart's visual

features and parse those pixels into values. We used the actual data values to generate models of these perceptual proxies. The value of this approach is that if we can determine the properties of the data and arrangements that lend themselves to particular proxies for comparison, then a potential application of this approach is that an automated visualization system would only need know the data values and the designer's desired comparison to construct the mark and arrangement to support that comparison. In other words, these proxies do not directly take into account limitations in perceptual visual acuity, or the capacity limitations of attention and memory.

8. **How can one generate new candidate proxies?** Future work should generate more, and more sophisticated, proxies (including combinations of proxies, and eventually, predictions for who will use which, and when). We generated proxies with a combination of intuition and consultation with the perceptual psychology literature, including a strong influence of the literature on focal vs. global processing modes in vision. Our list is by no means exhaustive, and identifying new candidates will be a creative process that, like hypothesis generation across the rest of science, relies on engaging a diverse group of people with different types of background knowledge across both the perception and data visualization communities. A brute force approach would be to generate the full space of mathematically possible pairwise and set-wise proxies. Another route could be based on interviews with viewers engaged in a particular task, to see which aspects of their proxies might be consciously verbalized.

7.4 Implications for Visualization

The study of the visual features used for visual comparison points to fruitful paths forward for both perceptual psychology as well as for inspiring guidelines for effective visual comparison. For practitioners, because our results indicate that people use different visual features for different tasks, a visualization designer could use these rules to optimize their visualizations and arrangements depending on the perceived task. For example, a visualization where the key task is perceiving a range, such as trend over time in, e.g., a stock market visualization, should clearly optimize for focal visual comparisons. In visualizations where understanding the biggest mean is central, value arrangements should favor global visual features.

Our study here was confined to bar charts, but it is clear from the richness of our results that this limitation did not restrict the complexity of the performance results. Bar charts are clearly flexible visual representations in that they support both global and focal visual comparison. Nevertheless, another clear next step is to expand this work to other visual mark types.

In our empirical study of perceptual proxies, the features that yielded the most similar responses suggests that some proxies are more likely than others to explain human behavior. This is not to say such a correspondence would prove that people use these extremely simplified proxies exactly, or alone, but instead point a path forward to possible mechanisms that can be empirically evaluated.

7.5 How Might Viewers Choose Proxies?

The proxy approach has been a fruitful one for the study of value estimation and comparison of correlation in scatterplots [12, 22, 27]. Scatterplots are an ideal "Petri dish" with which to test perceptual proxies for visualization for a number of reasons. Scatterplots are also used often to communicate a single statistic (correlation) of a set for which precision is important. A viewer seeing a scatterplot will likely develop the analytic goal of perceiving correlation, which should be more likely to trigger analysis of the proxies available in the scatterplot visualization to calculate correlation [21]. Designers are likely to be implicitly aware of some of these perceptual proxies for global and focal tasks. It is possible that the ubiquity of bar chart histograms, or multi-item bar charts in general, is because they allow for both global and focal comparisons.

One possible avenue of future work is a survey of the literature of what kinds of visualizations are used for different kinds of visual comparisons, to test the validity of the global-versus-focal distinction. A possible eventual extension of this line of research is the potential automation of constructing marks and arrangements, based on characteristics of the data and on the designer's desired comparison they wish to enable in the viewer. This line of work would invert the approach used here. In the current work, we assessed performance of proxies and of observers with a known task. The experimental task may have guided the proxy that observers used. But what if a viewer does not have a specific task before seeing a visualization? One possibility is that the proxy that yields the most salient comparison compels the viewer to perform a task guided with that proxy. For example, in a visualization with two charts with one large item outlier, the single item outlier might capture attention provoke a focal comparison. Future work could manipulate the salience of different proxies in a stimulus and evaluate which task observers perform based on the salience of those proxies, and verify this approach with interviews.

8 CONCLUSION

In this work, we found evidence that visual comparisons of biggest mean and biggest range are supported by some chart arrangements more than others, and that this pattern is substantially different from the pattern for other tasks. We proposed a series of potential proxies across different tasks, marks, and spatial arrangements. Simple models of these proxies can be empirically evaluated for their explanatory power by comparing their performance to human performance for these marks, arrangements, and tasks. We used this process to highlight candidates of perceptual proxies that might scale more broadly to explain performance in visual comparison.

REFERENCES

[1] G. A. Alvarez. Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences*, 15(3):122–131, 2011. doi: 10.1016/j.tics.2011.01.003

[2] R. Amar, J. Eagan, and J. Stasko. Low-level components of analytic activity in information visualization. In *Proceedings of the IEEE Symposium on Information Visualization*, pp. 111–117. IEEE, Piscataway, NJ, USA, 2005. doi: 10.1109/INFOVIS.2005.24

[3] J. Bertin. *Semiology of Graphics*. University of Wisconsin Press, Madison, Wisconsin, 1983.

[4] M. Brehmer, J. Ng, K. Tate, and T. Munzner. Matches, mismatches, and methods: Multiple-view workflows for energy portfolio analysis. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):449–458, 2016. doi: 10.1109/TVCG.2015.2466971

[5] Z. Chen. Object-based attention: A tutorial review. *Attention, Perception, & Psychophysics*, 74(5):784–802, 2012. doi: 10.3758/s13414-012-0322 -z

[6] H. Choo, B. R. Levinthal, and S. L. Franconeri. Average orientation is more accessible through object boundaries than surface features. *Journal of Experimental Psychology: Human Perception and Performance*, 38(3):585, 2012.

[7] P. Dragicevic, A. Bezerianos, W. Javed, N. Elmqvist, and J.-D. Fekete. Temporal distortion for animated transitions. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pp. 2009–2018. ACM, New York, NY, USA, 2011. doi: 10.1145/1978942.1979233

[8] S. Few. *Now You See It: Simple Visualization Techniques for Quantitative Analysis*. Analytics Press, Oakland, CA, USA, 2009.

[9] M. Gleicher. Considerations for visualizing comparison. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):413–423, 2017. doi: 10.1109/TVCG.2017.2744199

[10] M. Gleicher, D. Albers, R. Walker, I. Jusufi, C. D. Hansen, and J. C. Roberts. Visual comparison for information visualization. *Information Visualization*, 10(4):289–309, Oct. 2011. doi: 10.1177/1473871611416549

[11] J. Haberman and D. Whitney. Ensemble perception: Summarizing the scene and broadening the limits of visual processing. *From Perception to Consciousness: Searching with Anne Treisman*, pp. 339–349, 2012. doi: 10.1093/acprof:osobl/9780199734337.003.0030

[12] L. Harrison, F. Yang, S. Franconeri, and R. Chang. Ranking visualizations of correlation using Weber's law. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1943–1952, 2014. doi: 10.1109/TVCG.2014.2346979

[13] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (11):1254–1259, 1998.

[14] W. Javed, B. McDonnel, and N. Elmqvist. Graphical perception of multiple time series. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):927–934, 2010. doi: 10.1109/TVCG.2010.162

[15] M. King, G. E. Meyer, J. Tangney, and I. Biederman. Shape constancy and a perceptual bias towards symmetry. *Perception & Psychophysics*, 19(2):129–136, 1976. doi: 10.3758/BF03204219

[16] D. M. Levi and S. A. Klein. Vernier acuity, crowding and amblyopia. *Vision Research*, 25(7):979–991, 1985. doi: 10.1016/0042-6989(85)90208 -1

[17] D. Melcher and E. Kowler. Shapes, surfaces and saccades. *Vision Research*, 39(17):2929–2946, 1999. doi: 10.1016/S0042-6989(99)00029-2

[18] T. Munzner. *Visualization Analysis and Design*. A.K. Peters Visualization Series. CRC Press, Boca Raton, FL, USA, 2014.

[19] D. Navon. Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, 9(3):353–383, 1977. doi: 10.1016/0010-0285(77)90012-3

[20] B. D. Ondov, N. Jardine, N. Elmqvist, and S. Franconeri. Face to face: Evaluating visual comparison. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):861–871, 2019. doi: 10.1109/TVCG.2018.2864884

[21] S. Pinker. A theory of graph comprehension. In R. Freedle, ed., *Artificial Intelligence and the Future of Testing*, pp. 73–126. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, USA, 1990.

[22] R. A. Rensink and G. Baldridge. The perception of correlation in scatterplots. *Computer Graphics Forum*, 29(3):1203–1210, 2010. doi: 10.1111/j.1467-8659.2009.01694.x

[23] D. A. Szafir, S. Haroz, M. Gleicher, and S. Franconeri. Four types of ensemble coding in data visualizations. *Journal of Vision*, 16(5):11–11, 2016. doi: 10.1167/16.5.11

[24] J. Wagemans. Characteristics and models of human symmetry detection. *Trends in Cognitive Sciences*, 1(9):346–352, 1997. doi: 10.1016/S1364 -6613(97)01105-4

[25] D. Whitney and A. Yamanashi Leib. Ensemble perception. *Annual Review of Psychology*, 69:105–129, 2018. doi: 10.1146/annurev-psych -010416-044232

[26] Y. Xu and S. L. Franconeri. Capacity for visual features in mental rotation. *Psychological Science*, 26(8):1241–1251, 2015. doi: 10.1177/0956797615585002

[27] F. Yang, L. Harrison, R. A. Rensink, S. Franconeri, and R. Chang. Correlation judgment and visualization features: A comparative study. *IEEE Transactions on Visualization and Computer Graphics*, 25(3):1474–1488, Mar. 2018. doi: 10.1109/TVCG.2018.2810918

[28] J. S. Yi, Y. ah Kang, J. T. Stasko, and J. A. Jacko. Toward a deeper understanding of the role of interaction in information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1224–1231, 2007. doi: 10.1109/TVCG.2007.70515

[29] J. Zacks and B. Tversky. Bars and lines: A study of graphic communication. *Memory & Cognition*, 27(6):1073–1079, 1999. doi: 10.3758/BF03201236