The Arrangement of Marks Impacts Afforded Messages: Ordering, Partitioning, Spacing, and Coloring in Bar Charts

Racquel Fygenson, Steven Franconeri, and Enrico Bertini

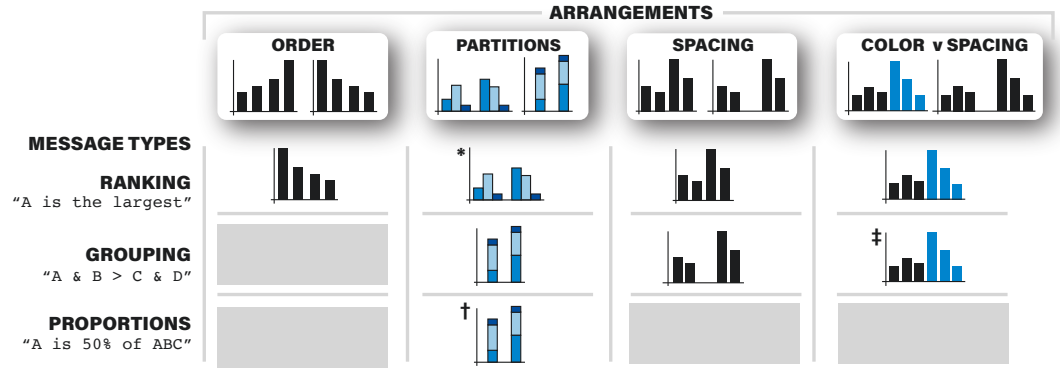

Fig. 1: A sample of presented findings: each cell shows the best option we found. Column headers show the options we compared, and row headers show tested message types. Gray rectangles indicate that the corresponding message-arrangement was not studied. Of tested arrangements, $*$ is best when comparing individual parts and † is best when 3 colors are present – see Fig. 4. ‡ is best when 2 colors are present – see Fig. 6.

Abstract—Data visualizations present a massive number of potential messages to an observer. One might notice that one group's average is larger than another's, or that a difference in values is smaller than a difference between two others, or any of a combinatorial explosion of other possibilities. The message that a viewer tends to notice – the message that a visualization 'affords' – is strongly affected by how values are arranged in a chart, e.g., how the values are colored or positioned. Although understanding the mapping between a chart's arrangement and what viewers tend to notice is critical for creating guidelines and recommendation systems, current empirical work is insufficient to lay out clear rules. We present a set of empirical evaluations of how different messages–including ranking, grouping, and part-to-whole relationships–are afforded by variations in ordering, partitioning, spacing, and coloring of values, within the ubiquitous case study of bar graphs. In doing so, we introduce a quantitative method that is easily scalable, reviewable, and replicable, laying groundwork for further investigation of the effects of arrangement on message affordances across other visualizations and tasks. Pre-registration and all supplemental materials are available at https://osf.io/np3q7 and https://osf.io/bvy95, respectively.

Index Terms—Perception & cognition, Methodologies, Human-subjects qualitative studies, Human-subjects quantitative studies, Charts, diagrams and plots, General public

1 Introduction

Visualization evaluation and design are often guided by a ranking of visual variables developed on precision-based criteria (e.g., response time, exactness of read values) [19, 22, 26, 27, 51]. Other visualization guidance is based off intuition [43], or extrapolated from cognitive psychology experiments that use far simpler stimuli (e.g., sets of shapes) and different participant tasks [2, 18, 28, 29, 33, 52].

Precision-based evaluation provides limited guidance in designing an effective visualization. While precision can ensure quantitative data is estimated accurately from graphical depictions, it is not sufficient to guarantee efficacy. Visualization designs can convey data in precise ways, yet not make an intended message obvious, and imprecise designs can still make intended messages obvious and intuitive [3]. Similarly, two visualizations can show the same data with equal precision, but communicate significantly different messages. Consider the pair of graphs in Figure 2, taken from Alberto Cairo's popular blog "The Functional Art." Both graphs encode number of COVID-19 cases using the height of aligned bars, but differences in ordering and spatial proximity of their bars convey markedly different trends in case numbers. The top graph sorts bars in descending order regardless of time, implying a consistently decreasing trend, while the bottom communicates that the numbers of cases by county increase before they decline. Thus, it is possible for simple changes in the arrangement of parts of a chart to impact the message that a viewer is likely to grasp. More generally, as existing research posits, data visualizations' design can afford potential takeaways [37, 53, 54, 60]. In practice, past research has investigated afforded messages by examining how differences in visualizations can compel viewers to reason differently [53], alter the type of comparisons they make [54], and most commonly vary their description of underlying information [37, 54, 60].

In this paper, we explore a novel metric for evaluating afforded takeaways: Do some arrangements of marks (i.e., visual objects in a graph) make messages more obvious than others? To answer this question, we need to 1) enumerate a possible set of mark arrangements and a possible set of subsequently afforded messages and 2) investigate if these arrangements impact the obviousness of the identified messages.

- *Racquel Fygenson and Enrico Bertini are with Northeastern University. E-mail: fygenson.r | e.bertini @northeastern.edu*
- *Steven Franconeri is with Northwestern University. E-mail: franconeri@northwestern.edu.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

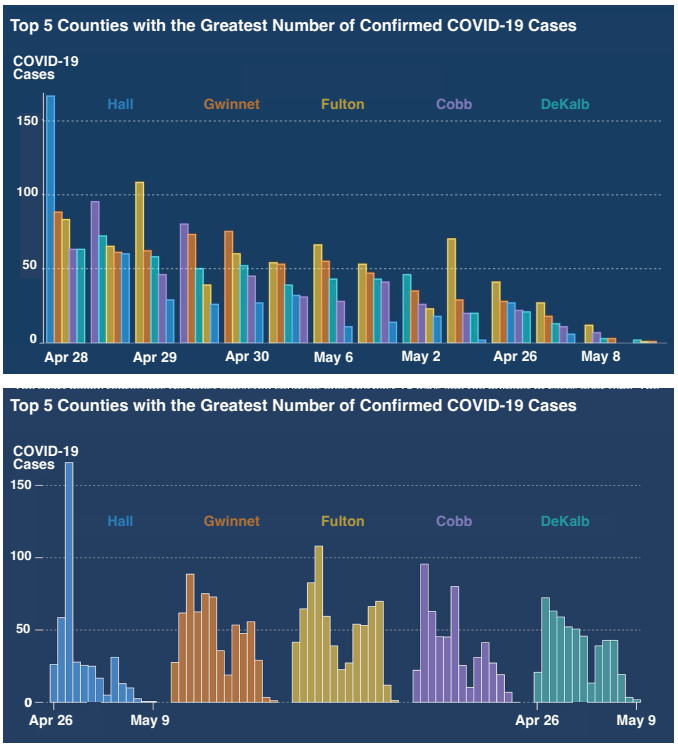

Fig. 2: These charts show the same information with differently ordered and grouped bars. What patterns are most obviously communicated by the top chart? Are these same patterns obvious in the bottom? Recreated from [10].

We present four experiments that investigate how four arrangements of data marks (ordering, partitioning, spacing, coloring) affect the subjective match of visualization designs to a set of messages (ranking judgments, group comparisons, and part-to-whole relationship judgments).

2 Related Work

While precision-centric evaluation methods remain extremely popular [19], alternate methodologies have been advocated in a collection of evaluation-focused papers [3, 4, 27, 32, 48], and motivated the long-running IEEE VIS workshop *BEyond time and errors: novel evaLuation methods for Information Visualization* (BELIV). In alignment with growing consideration for new metrics of evaluation, studies have explored visual metaphors [61], memorability [1, 5, 6], deeper insights [27], implicit takeaways [54], and afforded reasoning [53].

2.1 Current Methods for Exploring Visualization Affordances

In the late 1990s, Zacks and Tversky, and Shah et al. explored differences in reader takeaways incurred by bar and line charts, finding that bars imbue a sense of discreteness, while line charts imply continuous relationships [37, 60]. In both of these seminal works, researchers employed qualitative methodologies, showing graphs to participants and then hand-coding their open-ended descriptions. This method provides strong benefits by allowing findings to arise organically, without the need for pre-declared hypotheses. The inverse of this method–in which researchers describe a relationship between data points and ask participants to draw a corresponding graph–offers similarly beneficial evidence [16, 49, 60].

Hand-coding qualitative survey responses continues to be used to study visualization affordances, by asking participants to type out or voice their takeaways. [8, 23, 53, 54]. But this methodology is time-consuming and labor-intensive. Even more problematic, as diligently reported by Xiong et.al, this hand-coding is often unable to resolve the natural syntactic and semantic ambiguities in sentences that people type [54]. As a simple example, imagine a bar graph showing the sizes of two birds and two squirrels. If a viewer says, "the birds are bigger than the squirrels", they could mean that any bird is bigger than any squirrel, or that the birds are bigger than squirrels on average. Open-ended methods are powerful exploratory tools, but require, the sometimes impossible, resolving of ambiguities to effectively study visualization affordances. Thus, we present a complementary methodology to such approaches. We employ a confirmatory design that restricts the space of tested stimuli and messages, but provides efficient, replicable data to verify the impact of visual arrangements on afforded messages.

Similar approaches include asking participants to report their opinion, often using Likert scales [40], on how much different visualizations support semantic variables (e.g., "stable", "rigid", "complete") [62], on the trustworthiness or bias of visualizations [21, 30], on their agreement with provided statements [20, 21], or on amount of risk to themselves or others before and after seeing visualizations showing pandemic information [31].

We present a comparably empirical approach, but focus on general reader takeaways, a subject matter that–to the best of our knowledge–has only been quantitatively evaluated once before, in Xiong et al.'s Experiment 2, and with a much smaller (n=45 experts vs our $n \geq 130$ general public) sample [54]. For a more detailed comparison of our work to Xiong et. al, see SM7 in Supplemental Materials.

2.2 Bar Chart Research

Bar charts, one of the most prevalent types of visualization [24], are a common subject of visualization evaluations, and produce study results that have been generalised to other graph types [26]. Foundational research in visualization, including the widely cited Cleveland & McGill, Zacks & Tversky, Shah et al, Bateman et al, and Heer & Bostock papers seek to evaluate fundamental paradigms of visualizations and their communication by studying bar charts [1, 11, 17, 41, 59, 60]. Takeaways from these papers establish core tenets of bar chart interpretation, including how accurately one can discern bar chart lengths given different placement and heights of the bars [11, 17, 41, 59], and that arranging bars in groups using irregular spacing leads to readers "visual[ly] chunking" their takeaways accordingly [37]. For the most part, best practices using bar charts can be informed via the effectiveness ranking of channels [26].

In this work, we explore the effect of bar chart arrangements beyond classic manipulated variables (e.g., x-axis alignment, height differences) and the classic dependent variable of precision (e.g., response time, read accuracy). To further the understanding of visualization design, we include conditions that have been tested (e.g., bar alignments vs misalignment), and those that have yet to be investigated (e.g., spacing vs coloring to convey grouping of bars), and focus on the impact of these conditions on message obviousness. Thus we present novel results on how bar chart arrangements' obviousness of messages can align, or fail to align, with precision-based design decisions. See Figures 4, 5, 6, and 7 for tested messages, bar charts, and results.

3 Study Rationale

We start by identifying arrangements of marks that may have an influence on afforded messages (e.g., formats that compel viewers to compare data, focus on trends, or guide a specific reading sequence). We prioritize arrangements that can generalize to multiple graphical representations and therefore have an impact beyond bar charts. We avoid studying arrangements inherent to a specific visual mark (e.g., density/texture in choropleths which does not have an equivalent in line charts). We also aim to limit use of arrangements that require multiple visual channels to encode information. In this study, we test **Ordering**, **Partitioning**, **Spacing**, and **Coloring**.

Ordering describes the sequence in which objects can be arranged in a visualization. Examples of ordering in different visualizations include the sequence of bars in a bar chart, segments in a pie chart, columns and rows in a matrix visualization, and plots in a small multiple visualization.

Partitioning pertains to the division of a graphical mark into sub-parts, and those sub-parts' corresponding placement in a visualization. Partitioning can be found in pie charts, treemaps, and stacked bar charts.

Spacing, in this context, addresses the use of spatial proximity to organize visual objects into groups. Spacing is used in grouped bar charts, exploded pie charts, and grouping within Sankey diagrams or alluvial plots.

Coloring, in this context, describes the use of different hues or levels of saturation to group elements and/or distinguish between elements of different types. Examples of coloring used in this way include colored dots in scatter plots, colored bars in bars charts, colored rectangles in treemaps, and colored lines in multi-line charts.

We hypothesize that these four variations of arranging marks influence the strength of different messages extracted from visualizations. Specifically, we identify the following types of afforded messages and their influencing arrangements: **Ranking** (Ordering), **Grouping** (Spacing, Partitioning, Coloring), and **Part-to-whole relationships** (Partitioning). We provide details of how these versions are compared and measured experimentally in the Materials & Methods section.

3.1 Experimental Approach

The following is a high-level overview of the motivation behind our experimental approach. For experimental design details, see the Materials & Methods section.

Our main experimental goal is to explore how the arrangements outlined in the previous subsection influence readers' interpretation of visualizations.

As detailed in Section 2.1, we seek to complement popular open-ended methods of studying visualization affordances with the following, more structured approach:

1. Identify message types that may be afforded by visualizations.
2. For each message type (e.g., ranking, grouping), develop example messages to test
3. Using previous research, knowledge, or open-ended experimental methods, hypothesize which variations in mark arrangements may strengthen or weaken the affordance of test messages.
4. Design visualizations that contain the arrangements hypothesized to be a good match for each message type, along with visualizations that are hypothesized to not match well.
5. Showing one message at a time, ask participants to select which of the visualization designs best matches the message. We do so in practice by asking participants which visualization makes the message "the most obvious" to them.
6. Quantify the strength of the arrangement-to-message fit according to the frequency with which participants select tested visualizations.
7. Determine support, or lack thereof, for hypotheses by evaluating the proportion of participants that report each visualization as making the tested message the most obvious.

This approach has complementary advantages and disadvantages to the qualitative approach of collecting readers' interpretations through open-ended questions. Our approach makes message matching more quantifiable, less noisy and more accessible to reviewing and replication. On the other hand, it is solely confirmatory, hinging upon researchers' choice of tested messages and graphic variations, and is potentially influenced by nuances in message wording.

Fortunately, partnering our approach with an open-ended exploratory method (see Sec. 2.1) can mitigate the first limitation. The second limitation can be addressed, as in Experiment 2, by testing messages with the same meaning but slight re-wordings to investigate if nuanced wording is confounding results. In this study, we have found rewording to have no effect.

4 Materials & Methods

In this work, we examine the arrangements of marks in bar charts, and how they impact the obviousness of afforded messages. We conduct four separate, confirmatory, within-subjects studies in which we study the effects of ordering, partitioning, coloring, and irregularly spacing bars (see top row of Figs. 3 to 6). In contrast with the majority of current research on visualization affordances [16, 37, 44, 56, 60], we employ a quantitative methodology, in which research participants select one of four shown graphs that makes a given message the most obvious to them. In doing so, we reduce the uncertainty around confirming and replicating qualitative experiments, but also reduce the investigative scope of our experiment; our experimental conclusions, and thus our hypotheses, are context-specific. Accordingly, any hypothesis that message M will be made the most obvious by graph G1, must be qualified with the context that M is only made more obvious by G1 than G2, G3, or G4. Our pre-registered hypotheses, experimental design, and analysis plan are available at `https://osf.io/np3q7`.

4.1 Investigative Questions & Hypotheses

4.1.1 Experiment 1 - Order

In Experiment 1, we investigate the effect of sorting bars on messages concerning rank (see Figure 3). While some work has established written language influences mental ordering schema [12, 44], and other research has hypothesized about the cognitive effort required to identify extrema given variously ordered bar charts [34], we are unaware of any studies that explore differences in ascending and descending graphs' interpretation.

We hypothesize that **(H1A)** bar charts arranged in descending order from left to right (Fig. 3, C) will make messages about first-, second-, and third-largest bars (Fig. 3, Ordering.1, .2, .3) the most obvious. Conversely, bar charts arranged in ascending order from left to right (Fig. 3, D) will make messages about first-, second- and third- smallest bars (Fig. 3, Ordering.4, .5, .6) the most obvious.

This hypothesis stems from previous cognitive psychological research that shows left-to-right visual scanning associations stemming from left-to-right languages influence mental ordering schema [12, 44]. Because all of our participants speak English fluently and currently reside in the United States, we hypothesize that they are pre-disposed to reading bar charts from left-to-right, and thus any ordering-specific messages would be made most obvious by chart arrangements in which the extreme associated with the ordering in question is further towards the left.

4.1.2 Experiment 2 - Partitions

Experiment 2 investigates the arrangement of bars that encode part-to-whole data. Specifically, this experiment studies how bars that are placed side-by-side afford proportion-specific and comparison messages more or less strongly than those that are stacked vertically (see Figure 4). Previous studies establish that axis-aligned bars will be more precisely interpreted than those that are not [11, 17, 41]. Talbot et al. further establish that vertically aligned, adjacent bars are more likely to be mis-estimated to add up to 100%–and thus considered a whole–than vertically aligned, spatially separated bars [41]. Building on Talbot et al.'s discovery, we investigate stacked and un-stacked bars' effect on the obviousness of messages concerning comparison between parts, comparison between wholes, and the existence of proportional relationships, in an effort to see if affordances of these messages align with previously accepted paradigms of stacked bar charts.

We hypothesize **(H2A)** stacked bar charts (Fig. 4, C & D) will make messages about the the whole of parts (e.g., comparison among summed parts, Fig. 4, Partitions.1, .2) more obvious than side-by-side bar charts (Fig. 4, A & B). We also hypothesize **(H2B, H2D)** stacked bar charts (Fig. 4, C & D) will make messages about individual parts as a proportion of a whole (e.g., clip sales in the West make up 50% of all clip sales, Fig. 4, Partitions.7, .8, .9, .10) more obvious than side-by-side bar charts (Fig. 4, A & B). These hypotheses are informed by theory on physical-visual metaphors [50], and past research that indicates that stacked bars imply wholeness and a part-to-whole relationship more strongly than side-by-side layouts [41]. We also hypothesize **(H2C)** side-by-side bar chart arrangements (Fig. 4, A & B) will make messages specific to the identification and comparison of individual parts in a part-to-whole visualization (e.g., clip sales in the West vs clips sales in the East, (Fig. 4, Partitions.3, .4, .5, .6) more obvious, via the same, albeit inverse, reasoning as our previous two hypotheses.

4.1.3 Experiment 3 - Spacing

Experiments 3 and 4 investigate the affordance of grouping messages, given bars with varied ordering, spatial proximity, and coloring. Experiment 3 tests uniform vs irregular spacing to determine if, as is currently maintained in visualization [26,37], visual perception [47], and psychology [7,52] literature, increasing the space between bars–and therefore the proximity of some bars to others–affords grouping.

We hypothesize that **(H3A)** bar charts with uniform spacing (Fig. 5, A & B) will make messages about overall extrema more obvious than bar charts with grouping implied via irregular spacing (Fig. 5, C & D). This hypothesis stems from pilot study results and, while logical given gestalt principles [47], was not immediately obvious to us before collecting pilot data. We also hypothesize that **(H3B)** bar charts with irregular spacing defining elements in groups (Fig. 5, C & D) will make messages that discuss those groups more obvious than bar charts with uniform spacing (Fig. 5, A & B).

4.1.4 Experiment 4 - Color vs Spacing

Experiment 4 compares the strength of spatial grouping to color grouping in bar charts (see Figure 6). Research on the hierarchy of visual grouping mechanisms has found that proximity conveys grouping more strongly that similar coloring [2,7,15,26,33]. Interestingly, the majority of studies that investigate visual grouping do not examine bar charts, instead, focusing on dot lattices [2,7,18,28,29,33].

We hypothesize that **(H4A)** bar charts with color groups and regular spacing (Fig. 6, A & B) will make messages about overall extrema more obvious than those with groups defined by spacing (Fig. 6, C & D), because the communication of grouping will be less strong in the color-grouped charts and thus easier to ignore when evaluating extrema over multiple groups. Informed by the same known hierarchy, we also hypothesize that **(H4B)** bar charts with groups defined by spacing (Fig. 6, C & D) will make messages that discuss those groups more obvious than charts with groups defined by color (Fig. 6, A & B).

4.1.5 Methodological Checks

Lastly, to check the methodological rigor of our survey, we include the following message-graph questions and their pre-registered hypotheses. To confirm that participants are not swayed by familiarity bias, thus always reporting that height-ordered bar charts (Fig. 3, C & D) make all messages most obvious, we hypothesize that **(H1C)** charts with bars sorted in a specific order (Fig. 3, A) will make messages comparing bars grouped in that order (Fig. 3, Order.7) most obvious due to the proximity of the bars in questions.

In Experiment 2, we also test some pairs of messages with alternate wordings and sentence structures to examine the effect of the style of messages on our results. We hypothesize that **(H2E)** messages communicating the same concept, despite rewording, (Fig. 4, Partitions.3 and .4, .7 and .8, .9 and .10) will not produce different results.

4.2 Stimuli Design

Experiments 1-4 investigate variations in bar chart arrangements (see Figs. 3 to 6). Unlike much prior affordance work in visualization, we focus on varying arrangements within bar charts as opposed to visualization encoding types (e.g. bar charts, line charts, pie charts) in the interest of evaluating design decisions that are less commonly investigated in visualization literature, education, and recommendation systems. This decision also reinforces the validity of our survey design; asking participants to select between multiple types of visualizations increases the risk of familiarity bias (i.e., participants always selecting visualizations that they have seen more often) clouding participant judgement. Due to the lack of visual variance, we believe the familiarity differential of bar charts with varying arrangements (e.g., ordered ascending vs descending) is much smaller than that plausible of different visualization types (e.g., pie chart vs tree-map).

4.2.1 Experiment 1 - Order

Experiment 1 investigates ordering of bars and is motivated by a lack of research into the effects of such design decisions. The majority of research on ordering of bars generally fails to investigate higher-level, participant-reported takeaways in favor of response time, precision, eye-tracking, and cognitive effort models [12,14,25,34].

In all four experiments we test four different visualization conditions for the sake of methodological consistency. Experiment 1 consists of the same bar chart in ascending, descending, and alphabetical order, as well as a fourth, "wildcard" ordering in which tallest bars are centered forming a $\wedge$ shape (see Fig. 3). This last arrangement was motivated by the desire to test four conditions, and by an interest in how the Gestalt Law of Symmetry, which states that people tend to perceive symmetrical shapes and prefer visual symmetry [47,52], might have an unexpected effect on message obviousness (spoiler alert: it didn't).

4.2.2 Experiment 2 - Partitions

Experiment 2 investigates the representation of part-to-whole bar charts. The primary motivation behind the development of its visualization conditions was to explore the obviousness of part-to-whole relationships given different partitioning. The hierarchy of visual encoding channels (as discussed in the Related Work) is universal in informing effective visualization design [26,51], and can be used to justify the replacement of all part-to-whole visualizations (i.e., pie charts, stacked bars) with side-by-side bar charts. This replacement prioritizes precision but has not been shown to better facilitate the communication of relationships or other non-precision messages. In fact, previous work investigating the efficacy of pie and bar charts challenges the effectiveness hierarchy when completing certain tasks [38]. Experiment 2 seeks to investigate how the hierarchy of effectiveness compares to the affordance of part-to-whole messages in side-by-side (aligned) and stacked (unaligned) bar charts. Thus, Experiment 2's visualization space consists of one dataset split into two groups of three bars, both side-by-side and stacked (Fig. 4, B & D), and the same data split into three groups of two bars both side-by-side and stacked (Fig. 4, A & C).

To determine color scheme, we selected three colors from a widely used categorical color palette from Tableau[1], a popular software for making visualizations. We selected these colors by avoiding hues that are strongly associated with warning (i.e., red, orange, yellow). Next we used Color Oracle[2], free software that simulates common forms of Color Vision Deficiency (CVD) [46], to evaluate and slightly alter the luminance of our chosen colors so as to increase their distinction for viewers with CVD.

4.2.3 Experiment 3 - Spacing

Experiments 3 and 4 seek to investigate how color and spatial arrangements of marks afford grouping. Experiment 3 is designed to replicate previous findings that irregular spacing strongly implies groups among bar charts [9]. Thus Experiment 3's stimuli design consists of two different orderings of bars, each regularly spaced (Fig. 5, A & B), and then irregularly spaced into groups (a condition with two groups of three bars, and a condition with three groups of two bars (Fig. 5, C & D).

4.2.4 Experiment 4 - Color vs Spacing

Experiment 4 shares much of the same motivation as Experiment 3, but investigates a less strongly supported theory on grouping in bar charts. While generally proximity is agreed to imply grouping more strongly than similar coloring [2,7,15,26,33], this has not been directly measured in bar charts. Thus, Experiment 4 presents a novel investigation into the hierarchy of afforded grouping in bar charts. To do so, Experiment 4 replicates proximity grouped conditions from Experiment 3 (Fig. 6, C & D), and compares them to equivalent bar charts that use color grouping instead (Fig. 6, A & B). We reuse the CVD-friendly color scheme from Experiment 2 in this experiment as well.

4.3 Procedure

All four of our within-subjects experiments were implemented through a Qualtrics[3] survey. After reading and approving a consent form, par-

[1] https://help.tableau.com/current/pro/desktop/en-us/viewparts_marks_markproperties_color.htm

[2] https://colororacle.org/index.html

[3] https://www.qualtrics.com/

ticipants were given the option to self-report their education level and if they had CVD (mentioned by name and colloquialized as "color-blindness" in our survey). Participants were then instructed to make their browser window as large as possible and primed on the types of graphs they would see (see SM1 in Supplementary Materials for language used). They were then shown a page comprised of four charts with varied arrangements, a short sentence describing the content of the charts, and, as an attention check, a message with a fill-in-the-blank drop-down consisting of two possible answers, one of which correctly described the data depicted in all four chart conditions. Participants were instructed to *1. Use the charts below to fill in the blank.* and *2. Then select the chart that makes the statement below most obvious to you.* See SM2 in Supplemental Materials for an example survey question. Participants were shown between 6 and 10 of these questions, depending on the number of tested messages in the experiment.

This methodology, which presents quantitative and easily replicable evidence of subjective takes, stands in contrast to many similarly motivated investigations, which show participants visualization stimuli and ask them to describe it [37, 54, 56, 60], or show participants a description and ask them to represent the information with a visual creation [16, 44, 60]. As mentioned in the Section 2.1, to the best of our knowledge, the only experiment with similar methodology to ours is Xiong et al.'s Experiment 2 (see SM7 in Supplemental Materials for a more detailed comparison) [54].

Both the order in which participants viewed questions, and the order charts were presented in the quadrant of every question were randomized using the Qualtrics "randomization" functionality. The order of the drop-down answers for the fill-in-the-blank was not randomized, but held consistent with terms like "smaller," "less," and "least" appearing above terms like "larger," "more," and "most," so as not to confuse participants or lead to incorrect selection despite correct comprehension.

We added the fill-in-the-blank question as both an attention check, and to compel participants to actually read and consider the message when reporting the graph that made it the most obvious. Without this experimental design detail, we would have little way of knowing if participants actually read and reported their opinions on the message, because all four conditions show the same data and are therefore technically "correct" answers. While our pre-registered analysis plan[4] dictates excluding a participant's chart selection if they incorrectly answer the corresponding drop-down question, we find very little inconsistency between reported obviousness of charts from participants who correctly and incorrectly answer the drop-down. See SM4 in Supplemental Materials for a comparison of results with and without this exclusion criteria.

4.4 Participants

Participants were recruited via the online platform Prolific[5]. Prolific connects scientific researchers with eligible human studies participants, and offers a number of services to facilitate high-quality, ethical human-subjects research, including enacting specified inclusion and exclusion criteria, encouraging fair pay rates for participants, and facilitating compensation directly. Using Prolific, we recruited participants who were over the age of 18, fluent in English, current residents of the United States, and had high ($\geq 98\%$) approval rates on the platform, and constructed a study population that was roughly balanced on reported sex, as stated in our pre-registered study plan[6]. Also via Prolific, we compensated all participants 1.60USD for their participation, given an anticipated participation time of 8 minute, for an estimated rate of 12.00USD/hour.

5 Results

5.1 Participants

A total of 610 participants were recruited via Prolific. Of these, 591 (Exp. 1 n = 147, Exp. 2 n = 166, Exp. 3 n = 140, Exp. 4 n = 138) completed the full survey with no higher than a 30% error rate, passing the universal exclusion criteria, and were included in our final data analysis. For a breakdown of participants' reported sex, education and color vision deficiency for each experiment, see Table 1. Exact sample size per message varies based on number of participants who selected the corresponding drop-down correctly, although all sample sizes are equal to or more than our minimum pre-registered sample size of 128. For a breakdown of sample size per tested message see SM3 in Supplementary Materials.

Table 1: Demographics of Participants per Experiment

	Exp. 1	Exp. 2	Exp. 3	Exp. 4
Female	75	84	72	70
Male	72	82	68	68
Some high school	2	0	0	2
High school/GED	32	48	35	27
Tech/community college, associates degree	26	30	26	29
Undergraduate degree	63	55	51	61
Graduate degree	21	24	24	16
Doctoral degree	3	8	4	3
Does not have CVD	141	163	139	135
Has CVD	5	1	140	0
Did not answer	1	2	0	3
n	**144**	**166**	**140**	**138**

Initially, Experiment 1 included multiple un-piloted messages that resulted in very high ($> 25\%$) error rates. We hypothesized that these errors were most likely due to ambiguous or overly convoluted messages. We re-wrote these messages to be more straightforward [7], and re-ran the entire experiment, drastically decreasing error rates to $\leq 12\%$. We report the results from the final Experiment 1 in this paper.

5.2 Analysis

We exclude all participants who answered $> 30\%$ of all drop-down answers incorrectly.

Participant responses are analyzed using the Sison-Glaz procedure for estimating multinomial proportion confidence intervals [39], as implemented by the Python library statsmodels.stats [36]. Due to the multinomial nature of this procedure, no correction for family-wise error rate is necessary. Using the worst-case multinomial proportion table (Table 1) from Steven K. Thompson's "Sample Size for Estimating Multinomial Proportions" [42], we determine minimum sample size for a 95% confidence interval within a maximum specified distance from the true proportion, d, of 0.1 to be 128 participants per experiment. We elect to only conduct a visual analysis of the confidence intervals, avoiding null hypothesis significance testing and its common pitfalls (e.g. type II statistical errors) [13]. We present and discuss results of all four experiments using language and best practices of statistical analysis for Human Computer Interaction [13].

In Figures 3 to 6, we visualize the actual proportions and 95% confidence interval for all four conditions given each message tested. We highly encourage all readers to view and determine strength of results for themselves, but will summarize visual findings using hedged language as advised by [13].

5.3 Experiment 1 Results

Experiment 1 results are visualized in Figure 3 and present strong support for hypotheses **H1A**, and **H1C** (H1B was rendered irrelevant in Exp. 1's re-run and thus dropped). For an overview of all hypotheses see Section 4.1.

The data in Figure 3, Maximum-Centric provide a consistent, visually distinct signal that bar charts formatted in descending order (condition C) make messages concerning the largest, second-, and third-largest bars more obvious than those formatted in ascending, alphabetical, or centrally-peaked order (conditions D, A, B). It is worth

[4] https://osf.io/np3q7
[5] prolific.co
[6] https://osf.io/np3q7
[7] for differences in the preliminary and final run of Experiment 1, compare the pre-registered design (https://osf.io/np3q7) with the design reported in this paper

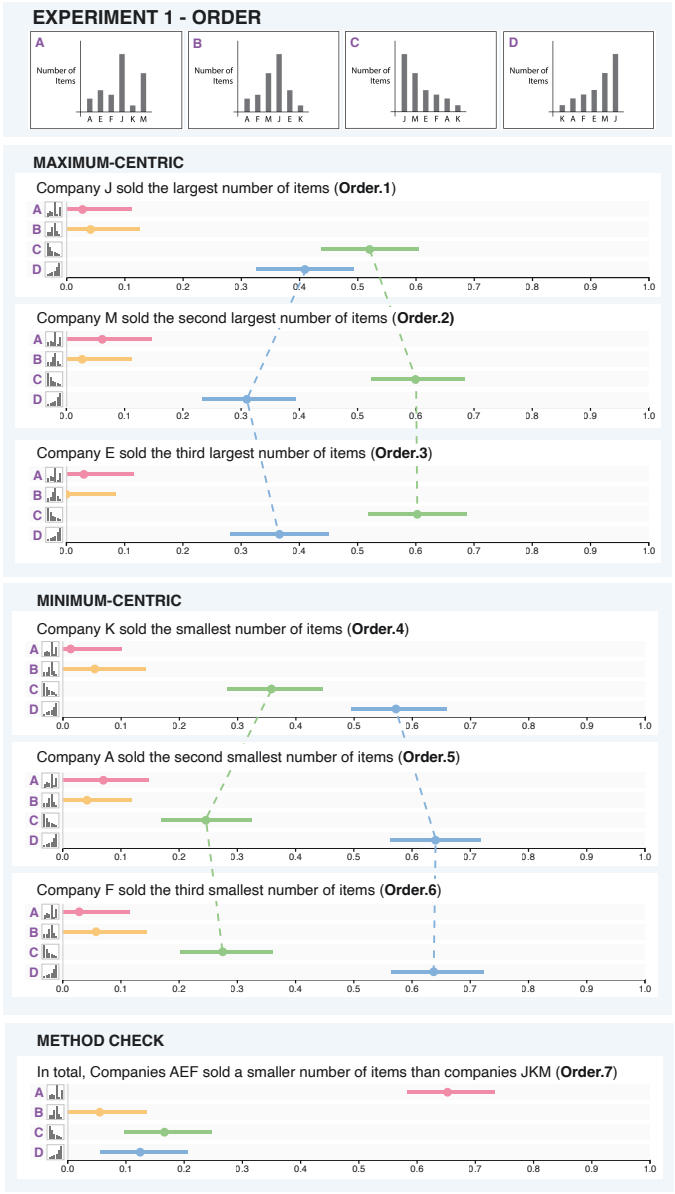

Fig. 3: Experiment 1 Results. Tested conditions are shown across the top of the figure. Below, lines encode 95% CIs for the proportion of respondents that report each condition makes the given message most obvious. Circles encode the actual proportion observed from the experiment.

noting that this signal is much weaker for message Order.1, which concerns the largest value in the bar chart. Observed alone, there is not a sizeable difference in CIs to support condition C affording Order.1 more than condition D. Yet, when taken into context with messages Order.2 and Order.3, a more consistent signal of interest emerges.

The data in Figure 3, Minimum-Centric provide a similar series of visually distinct signals that bar charts formatted in ascending order (condition D) make messages concerning the smallest, second-, and third-smallest bars more obvious than those formatted in descending, alphabetical, or centrally-peaked order (C, A, B). This signal can be seen to grow stronger (i.e., the distance between CIs for ascending and descending conditions increases) as the messages concern more convoluted (e.g., second- and third- order) rankings. This increase in signal suggests that readers do not simply report increased obviousness due to marks of interest being immediately proximate to the left side of a chart, and that ascending and descending conditions still have an impact on obviousness of messages concerning bars that are more centrally located (e.g., third-largest and -smallest bars).

Finally, the data shown in Figure 3, Method Check support **H1C** with a strong signal that bar charts formatted in a particular order (condition A) make messages comparing companies grouped in that order (Order.7) more obvious than any other tested bar charts. This result is supported by cognitive psychology research that credits proximity with the ability to suggest grouping [7, 26, 28, 47, 50, 52].

5.4 Experiment 2 Results

Experiment 2 results are visualized in Figure 4. Supporting **H2A**, the data in Figure 4, Whole Comparisons presents a consistent, visually distinct signal that, in part-to-whole charts, stacking bars (conditions C, D) make messages concerning comparison of the whole more obvious than arranging them side-by-side(conditions A, B). Inversely, the data in Figure 4, Part Comparisons support **H2C** by presenting a consistent, visually distinct signal that bars arranged side-by-side (conditions A, B) make messages regarding the comparison of single parts more obvious than their stacked equivalents (conditions C, D).

The data in Figure 4, Proportions provide fairly strong evidence to support **(H2B)** stacking bars (condition D) makes messages regarding a single part as a percentage of a three-part whole (message Partitions.7, .8) more obvious than a side-by-side arrangement (condition B). At the same time, the visualized CIs in Figure 4, Proportions provide no evidence to support **(H2D)** the same difference in signal when messages regard a single part as a percentage of a two-part whole (messages Partitions.9, .10). This difference could be explained by a visual processing capacity limit of two colors at once [35, 55]. For further discussion, see Section 6.1.

Finally, Experiment 2 renders very similar CI results when testing re-wordings of the same messages (see red annotations in Fig. 4). This similarity supports **H2E** and the methodological validity of the survey by addressing concerns of potential confounding due to phrasing variations.

5.5 Experiment 3 Results

Experiment 3 results are visualized in Figure 5. The data in Figure 5, Ranking support **H3A** by depicting a consistent, visually distinct signal that bar charts without irregular spacing (conditions A, B) make messages concerning overall extrema more obvious than bar charts with irregular spacing (conditions C, D). The data in Figure 5, 3 Groups and Figure 5, 2 Groups support **H3B** by displaying a consistent, visually distinct signal that bar charts grouped via irregular spacing (conditions C, D) make messages concerning those groups more obvious than bar charts with identical ordering but uniform spacing (conditions A, B).

5.6 Experiment 4 Results

Experiment 4 results are visualized in Figure 6. The data in Figure 6, Ranking slightly support **H4A** with a consistent signal that bar charts with color grouping (conditions A, B) make messages concerning overall extrema more obvious than bar charts with proximity grouping (conditions C, D). The difference in signal between conditions A-B and C-D appear to be significant but are not as widely spread as hypothesis **H4A** postured. The data in Figure 6, 2 Groups do not support **H4B**. Instead, they show a consistent, visually distinct signal that bar charts with color grouping (condition A) make messages concerning two groups of three bars more obvious than bar charts with spatial grouping (condition C), which is the inverse of our hypothesized hierarchy. The data in Figure 6, 3 Groups display visually approximate confidence intervals for conditions B and D, which–while different from results in Figure 6, 2 Group–still do not support **H4B**. We speculate the reason for this difference could be a visual processing capacity limit of two colors [35, 55], as discussed in Experiment 2, as well. For further discussion, see Section 6.1.

6 Discussion

In this paper, we present four experiments investigating differences in visual arrangements' afforded messages. We do so through an empirical methodology that evaluates which arrangements of marks increase the obviousness of potential takeaways.

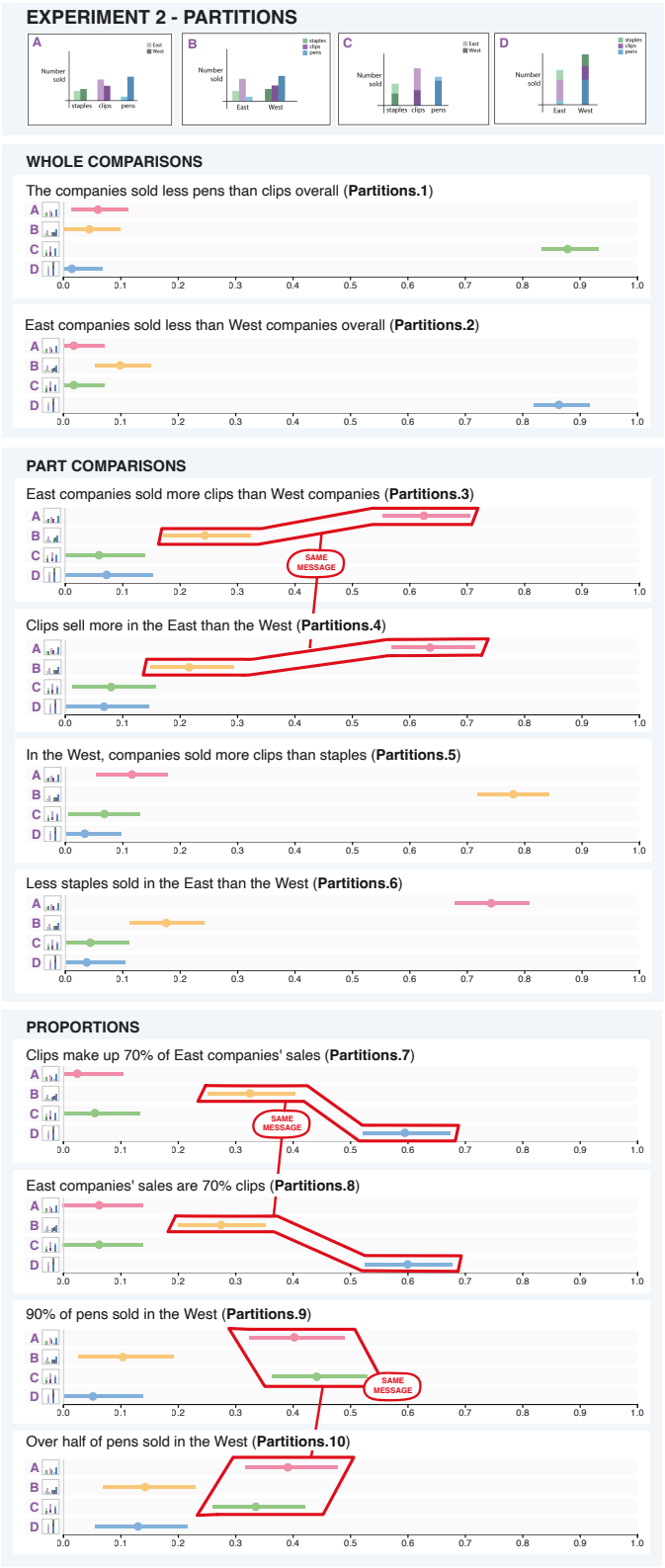

Fig. 4: Experiment 2 Results. Tested conditions are shown across the top of the figure. Below, lines encode 95% CIs for the proportion of respondents that report each condition makes the given message most obvious. Circles encode the actual proportion observed from the experiment.

6.1 Main Takeaways

We summarize the following outcomes from the Results section:

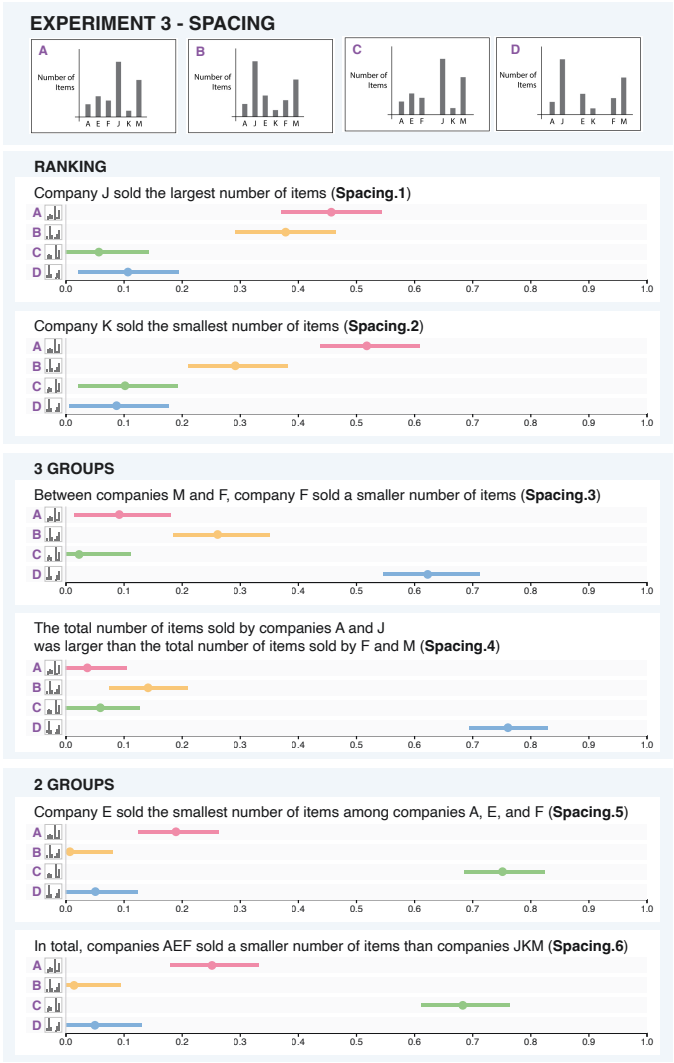

Fig. 5: Experiment 3 Results. Tested conditions are shown across the top of the figure. Below, lines encode 95% CIs for the proportion of respondents that report each condition makes the given message most obvious. Circles encode the actual proportion observed from the experiment.

1. Messages concerning largest, second-, and third- largest bars are made the most obvious by bars sorted in descending order from left to right (Fig. 3, Maximum-Centric).
2. Messages concerning smallest, second-, and third- smallest bars are made the most obvious by bars sorted in ascending order from left to right (Fig. 3, Minimum-Centric).

Takeaways 1 and 2 advise researchers and designers alike that the ordering of marks in a bar chart affects the affordance of messages about ranking. These takeaways exist within the context of the tested charts in Experiment 1 and the English-speaking nature of our participants. Still, these findings help bolster empirical evidence surrounding the impact of sorting bars, much of which has been conflicting. For example, Tversky et al. found similar evidence in their cognitive psychology study in 1991, discovering that children that speak directionally-ordered languages (e.g. left-to-right for English) associate ordering schema accordingly [44]. At the same time, newer perceptual effort models, supported by eye-tracking experiments, suggest the opposite; ascending bars require more effort to extract the minimum value than descending bars [34]. Regardless, this paper presents actionable recommendations for visualization designers who aim to draw attention to ranking-related messages.

3. In bar charts depicting part-to-whole data, messages concerning

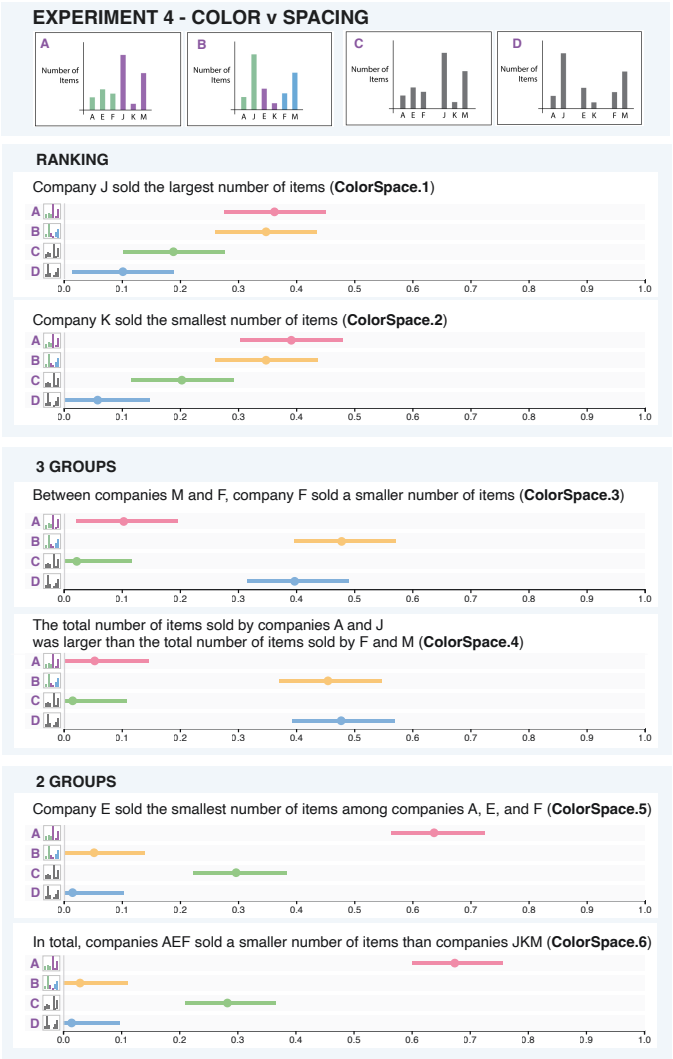

Fig. 6: Experiment 4 Results. Tested conditions are shown across the top of the figure. Below, lines encode 95% CIs for the proportion of respondents that report each condition makes the given message most obvious. Circles encode the actual proportion observed from the experiment.

the whole(s) are made more obvious by stacking than by side-by-side arrangements (Fig. 4, Whole Comparisons).

4. In bar charts depicting part-to-whole data, messages concerning the parts(s) are made more obvious by side-by-side arrangements than by stacking (Fig. 4, Part Comparisons).

5. In bar charts depicting part-to-whole data, messages concerning parts as percentages of the whole are sometimes made more obvious by stacking (Fig. 4, Proportions).

Takeaway 3 is hardly surprising since comparing visualized sums is easier than trying to mentally sum parts and then compare. The same, but inverse logic holds for Takeaway 4. More interestingly, Takeaway 5 finds messages about parts as a percentage of a whole are made equally, if not more, obvious by stacked bars over side-by-side bars. This holds true, even when an side-by-side arrangement would, from a precision standpoint, more effectively facilitate said comparison over its stacked counterpart [17,41]. Thus, we present initial evidence that precision and affordance (at least in the way it is operationalized as "obviousness" in our experiments) can diverge. In other words, a graph may lead to more precise comparisons and still be worse from an arrangement-message matching standpoint.

6. Spacing bars such that groups are formed by proximate bars makes messages concerning groupings of those bars more obvious than spacing bars uniformly. (Fig. 5, 3 Groups, 2 Groups)

7. Uniformly spaced bar charts make messages concerning overall extrema more obvious than charts with irregular spacing. (Fig. 5, Ranking)

Takeaway 6 is to be expected, as it is supported by decades of research on perceptual grouping [7,26,28,29,50,52]. Thus, Takeaway 6 provides for a methodological sanity check. More interestingly, Takeaway 7 is reiterative of the same perception research, but provides evidence that perceptual grouping can *hinder* the affording of messages concerning groups. This result is also an interesting case in which precision-centric visualization guidelines do not align with affordances. All arrangements in Experiment 3 support the same level of precision when identifying a maximum. Yet, our studies find bar charts with regular spacing make messages about maxima more obvious.

8. Uniformly spaced, color-grouped bar charts obstruct the obviousness of messages concerning extrema less than uncolored charts with spatial grouping (Fig. 6, Ranking).

9. Uniformly spaced, color-grouped bar charts depicting two groups of three bars make messages concerning those groups more obvious than corresponding uncolored charts with spatial grouping Fig. 6, 2 Groups). Though this difference in obviousness is not apparent when the bar charts depict three groups of two (Fig. 6, 3 Groups).

Takeaway 8 aligns with Takeaway 7 and the current hierarchy of perceptual grouping techniques, which maintains that proximity more strongly indicates grouping than color [2,7,26,57,58]. As we see in Takeaway 7 from Experiment 3, grouping implied via proximity can hinder the obviousness of messages concerning extrema across groups. Thus, it is expected that color grouping would inhibit the obviousness of such messages slightly less than spatial grouping.

Takeaways 5 and 9 were not predicted, and were surprising at first. However, after considering these takeaways in relation to recent accounts of the mechanisms underlying color grouping from perceptual psychology literature, we are excited that these findings might have a clear explanation, and with further testing, could result in clear and novel design guidelines.

We first note that the human visual system can encounter a powerful capacity limitation when required to process multiple colors at the same time; many tasks, like grouping sets of objects by color, may even be a strictly serial process in which only one group is conceived in any one perceptual moment [57,58]. In tasks that require people to temporarily associate a color with a label or other meaning (e.g., associating a color to a legend), capacity appears to be limited to two colors [35,55].

The results associated with Takeaways 5 and 9 might be explained by this reluctance to process more than two colors at once. Recall that in the Partitions experiment, 'Proportions' comparisons involving pens in the West vs. the East (messages Partitions.9, .10) were rated as equally obvious for stacked and side-by-side bars. Note that this comparison should be between two bars, requiring inspection of two colors (light blue and dark blue). But comparisons of clips to two other products (messages Partitions.7, .8) require juggling three colors (that are also categorically different hues: purple, green, and blue), which may prove more aversive. In this three-color condition, participants suddenly strongly report that a stacked bar makes proportional messages the most obvious. In this case, the stacked bar might allow viewers to more easily select the relevant hue (purple) as a percentage of the whole bar, leading to a preference for arrangements that make information more clear when color capacity is reached.

Similarly, in Experiment 4, participants surprisingly preferred making comparisons between two groups of bars when those bars were defined by two colors instead of by two spatially separated regions. But when those groups were defined by three colors, participants were equivocal in their preference between color and spatial grouping. This finding could also be explained by an aversion to processing more than two colors at once.

This explanation is speculative, and requires additional empirical support. Currently our studies confound number of colors, number of objects to be compared, and number of total groups. Additionally, some conditions use saturation differences (e.g., light purple and dark purple) while others use hue differences. While we doubt that these factors drive asymmetries in our results, our understanding could be better supported by experiments that are specifically designed to isolate the effect of number of color hues. Still, we remain excited about this speculative account because, while surprising, it is consistent with new models of color grouping and processing capacity [35,55]. If this speculative reasoning holds, it would produce a clear design guideline: use color to distinguish among two groups, use either for three categories, and use space to distinguish among four or more.

6.2 Limitations

While the method with which we study visualization affordances presents many positive features (see Section 3.1), its confirmatory nature also restricts the scope of possible findings. As noted in the Material & Methods section, our findings must be digested with their restricted scope in mind. For example, Takeaway 2 (*Messages concerning smallest... bars are made the most obvious by bars sorted in ascending order from left to right.*) holds in comparison to the other three orders tested, but it may not do so when compared to other bar chart arrangements. Fortunately, this limitation can be mitigated in part by pairing our experimental design with an exploratory method, as detailed in the Related Work section.

Similar contextual restrictions surround our study population. We recruit participants who are fluent in English, over the age of 18, and currently reside in the US. Ordered language conventions could very likely influence findings [44], and the replication of our work with other populations is prudent before generalizing results on a global scale. Fortunately, due to the easily replicable and modifiable nature of our method, such experiments could be run affordably.

Additionally, the results we present only consist of responses from participants who correctly filled in the drop-down of a given message. If a participant incorrectly filled in message A, their response to which chart made message A the most obvious was discarded, though all of their other responses were included. This exclusion has the potential to bias results towards an audience with a high graphic literacy. But to maintain a high quality of data, such removal is necessary to ensure that analyzed participants are actually answering survey questions with care. Due to both of these considerations, we provide a comparison of all results with and without this exclusion in SM4 in the Supplemental Materials – no large differences are apparent between the two.

Lastly, while we posit that affordance is an important metric in evaluating visualizations, the line between affordance and effectiveness is blurry. Can a graph make a desired message obvious but be ineffective? Or can a graph be effective but not make a desired message obvious? These are questions that need to be clarified, but are difficult due to a lack of agreed upon definition for effectiveness in visualization (see Related Work for a summary of metrics). Our current intuition, advised by our presented findings, is that affordance should correlate with increased graph comprehension, reduced reading effort, and general viewer preference. Future work is warranted to investigate if strong arrangement-message matches lead to increased efficacy of a graph, perhaps through the use of response time and precision as Vessey's cognitive fit model suggests [45] or via other metrics like cognitive effort and memorability.

6.3 Implications & Future Work

The studies we present firmly suggest that visual arrangements can directly impact the messages people perceive from a graph. That is, the various arrangements of identical marks in a graph can alter the strength of perceived messages. While our experiments cover a limited set of visual arrangements and messages, they point to a number of implications, and compel the expansion of this work.

To continue to build out academic and practical understanding of the effect of arrangements on afforded messages, our work can be extended as follows:

- *Study different arrangement variations*. Future works may maximize their impact by investigating properties that are generic enough to apply to a wide variety of visual representations. Candidate arrangements include: orientation, rotation, styling of negative marks, and visual linking through outlines or edges.
- *Study different message types*. We cover a small subset of potential messages afforded by visualizations. Further exploration of other messages could drastically expand our understanding of visualizations and what they communicate.
- *Study different visualization types*. Future work could also examine the arrangements studied in this paper (or an extension of them) with new types of graphs. Candidates include: spacing or ordering in pie charts or tree maps, color grouping in scatter plots or choropleths, and ordering in Sankey diagrams.
- *Study the relative strength of arrangements' affordance of messages*. Our methodology provides continuous, as opposed to binary, output allowing researchers to investigate both whether arrangements afford a message, and also possible hierarchies of arrangements' affordance (e.g., spacing affords grouping $>$ color affords grouping $>$ shape affords grouping), as demonstrated in Experiment 4.

Lastly, the work presented in this paper has relevant implications for practitioners. This work provides infrastructure to build a "library" of visual arrangements and their afforded message,s which designers could use to inform and evaluate their visualizations. Practically, a designer could begin either with a desired message to communicate, or with a set of visualizations they want to narrow down, and use our framework, or a repository of results from our framework, to better understand the implications behind their designs.

The same affordance library could be used as an evaluation tool to review existing visualizations. Existing designs could be evaluated so as to confirm that intended messages are conveyed strongly and, equally paramount, that unintended messages are not strongly communicated.

7 Conclusion

In this paper, we investigate how four different arrangements of marks – ordering, partitioning, spacing, and coloring – in bar charts afford messages on ranking, part-to-whole relationships, and grouping. We present an replicable, scalable, modifiable, confirmatory methodology for investigating arrangements of marks within visualizations and their relative impact on afforded messages.

In our Related Work, we establish current methods of investigating visualization affordances and current understanding of bar charts to provide context for our findings. In our Discussion, we summarize our findings into nine key takeaways which provide insight for visualization designers, researchers, and educators on the affordance of messages when considering spatial and color arrangements of marks. We then contextualize said findings, comparing them to the closest existing research.

In summary, we provide two useful contributions: 1) four experiments resulting in nine takeaways on how bar chart arrangements afford various messages and 2) the tools to continue this work through an easily scalable and modifiable method for evaluating visualization arrangements' impact on their afforded messages.

Supplemental Materials

All supplemental materials are available on OSF at `https://osf.io/bvy95/files/osfstorage`. In particular, they include (1-2) screenshots of the Qualtrics survey for posterity, (3) a table showing sample size for each tested message, (4) a side-by-side visual comparison of results excluding and including participants who answered a specific question incorrectly (does not apply to participants fully excluded from studies), (5) raw data files and a runnable jupyter notebook with all analysis, (6) .csv files used in the visualized CIs for Figures 3 to 6, and (7) a comparison of our work to [54].

Figure Credits

Figure 2 is a partial recreation of figures that appear in [10].

Acknowledgments

The authors wish to thank Laura South, Myrl Marmarelis, and Sydney Purdue for their advice on statistical analysis. This work was supported in part by a grant from the National Science Foundation (Award #2236644)

References

[1] S. Bateman, R. L. Mandryk, C. Gutwin, A. Genest, D. McDine, and C. Brooks. Useful junk? the effects of visual embellishment on comprehension and memorability of charts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, p. 2573–2582. Association for Computing Machinery, New York, NY, USA, 2010. doi: 10.1145/1753326.1753716 2

[2] M. B. Ben-Av and D. Sagi. Perceptual grouping by similarity and proximity: Experimental results can be predicted by intensity autocorrelations. *Vision Research*, 35(6):853–866, 1995. doi: 10.1016/0042-6989(94)00173-J 1, 4, 8

[3] E. Bertini, M. Correll, and S. Franconeri. Why shouldn't all charts be scatter plots? beyond precision-driven visualizations, 2020. doi: 10.48550/ARXIV.2008.11310 1, 2

[4] E. Bertini, A. Perer, C. Plaisant, and G. Santucci. Beliv'08: Beyond time and errors: Novel evaluation methods for information visualization. In *CHI '08 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '08, p. 3913–3916. Association for Computing Machinery, New York, NY, USA, 2008. doi: 10.1145/1358628.1358955 2

[5] R. Borgo, A. Abdul-Rahman, F. Mohamed, P. W. Grant, I. Reppa, L. Floridi, and M. Chen. An empirical study on using visual embellishments in visualization. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2759–2768, 2012. doi: 10.1109/TVCG.2012.197 2

[6] M. A. Borkin, A. A. Vo, Z. Bylinskii, P. Isola, S. Sunkavalli, A. Oliva, and H. Pfister. What makes a visualization memorable? *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2306–2315, 2013. doi: 10.1109/TVCG.2013.234 2

[7] J. L. Brooks. Traditional and new principles of perceptual grouping. In J. Wagemans, ed., *Oxford Handbook of Perceptual Organization*. Oxford Handbook of Perceptual Organization: Oxford University Press, Oxford, UK, September 2015. 4, 6, 8

[8] A. Burns, C. Lee, T. On, C. Xiong, E. Peck, and N. Mahyar. From invisible to visible: Impacts of metadata in communicative data visualization. *IEEE Transactions on Visualization I& Computer Graphics*, pp. 1–16, dec 5555. doi: 10.1109/TVCG.2022.3231716 2

[9] R. Burns, S. Carberry, and S. Elzer Schwartz. An automated approach for the recognition of intended messages in grouped bar charts. *Computational Intelligence*, 35(4):955–1002, 2019. doi: 10.1111/coin.12227 4

[10] A. Cairo. About that weird georgia chart, 2020. 2, 10

[11] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):531–554, 1984. 2, 3

[12] C. Dickinson and H. Intraub. Spatial asymmetries in viewing and remembering scenes: Consequences of an attentional bias? *Attention, perception I& psychophysics*, 71:1251–62, 09 2009. doi: 10.3758/APP.71.6.1251 3, 4

[13] P. Dragicevic. *Fair Statistical Communication in HCI*, pp. 291–330. Springer International Publishing, 2016. doi: 10.1007/978-3-319-26633-6_13 5

[14] A. Feeney and L. Webber. Analogical representation and graph comprehension. In A. Butz, A. Krüger, and P. Olivier, eds., *Smart Graphics*, pp. 212–221. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003. 4

[15] S. L. Franconeri, L. M. Padilla, P. Shah, J. M. Zacks, and J. Hullman. The science of visual data communication: What works. *Psychological Science in the Public Interest*, 22(3):110–161, 2021. PMID: 34907835. doi: 10.1177/15291006211051956 4

[16] A. Gaba, V. Setlur, A. Srinivasan, J. Hoffswell, and C. Xiong. Comparison conundrum and the chamber of visualizations: An exploration of how language influences visual design. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):1211–1221, 2023. doi: 10.1109/TVCG.2022.3209456 2, 3, 5

[17] J. Heer and M. Bostock. Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, p. 203–212. Association for Computing Machinery, New York, NY, USA, 2010. doi: 10.1145/1753326.1753357 2, 3, 8

[18] J. Hochberg and A. Silverstein. A quantitative index of stimulus-similarity proximity vs. differences in brightness. *The American Journal of Psychology*, 69(3):456–458, 1956. 1, 4

[19] T. Isenberg, P. Isenberg, J. Chen, M. Sedlmair, and T. Möller. A systematic review on the practice of evaluating visualization. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2818–2827, 2013. doi: 10.1109/TVCG.2013.126 1, 2

[20] H.-K. Kong, Z. Liu, and K. Karahalios. Frames and slants in titles of visualizations on controversial topics. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, p. 1–12. Association for Computing Machinery, New York, NY, USA, 2018. doi: 10.1145/3173574.3174012 2

[21] H.-K. Kong, Z. Liu, and K. Karahalios. Trust and recall of information across varying degrees of title-visualization misalignment. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, p. 1–13. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3290605.3300576 2

[22] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. Empirical studies in information visualization: Seven scenarios. *IEEE Transactions on Visualization and Computer Graphics*, 18(9):1520–1536, 2012. doi: 10.1109/TVCG.2011.279 1

[23] S. Lee, S.-H. Kim, Y.-H. Hung, H. Lam, Y.-A. Kang, and J. S. Yi. How do people make sense of unfamiliar visualizations?: A grounded model of novice's information visualization sensemaking. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):499–508, 2016. doi: 10.1109/TVCG.2015.2467195 2

[24] S. Lee, S.-H. Kim, and B. C. Kwon. Vlat: Development of a visualization literacy assessment test. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):551–560, 2017. doi: 10.1109/TVCG.2016.2598920 2

[25] A. Michal, D. Uttal, P. Shah, and S. Franconeri. Visual routines for extracting magnitude relations. *Psychonomic Bulletin I& Review*, 23, 05 2016. doi: 10.3758/s13423-016-1047-0 4

[26] T. Munzner. *Visualization Analysis and Design*. AK Peters Visualization Series. CRC Press, 2015. 1, 2, 4, 6, 8

[27] C. North. Toward measuring visualization insight. *IEEE Computer Graphics and Applications*, 26(3):6–9, 2006. doi: 10.1109/MCG.2006.70 1, 2

[28] T. Oyama. Perceptual grouping as a function of proximity. *Perceptual and Motor Skills*, 13(3):305–306, 1961. doi: 10.2466/pms.1961.13.3.305 1, 4, 6, 8

[29] T. Oyama, M. Simizu, and J. Tozawa. Effects of similarity on apparent motion and perceptual grouping. *Perception*, 28(6):739–748, 1999. PMID: 10664768. doi: 10.1068/p2799 1, 4, 8

[30] L. Padilla, R. Fygenson, S. C. Castro, and E. Bertini. Multiple forecast visualizations (mfvs): Trade-offs in trust and performance in multiple covid-19 forecast visualizations. *IEEE Transactions on Visualization I& Computer Graphics*, 29(01):12–22, jan 2023. doi: 10.1109/TVCG.2022.3209457 2

[31] L. Padilla, H. Hosseinpour, R. Fygenson, J. Howell, R. Chunara, and E. Bertini. Impact of covid-19 forecast visualizations on pandemic risk perceptions. *Scientific reports*, 12(1):1–14, 2022. 2

[32] C. Plaisant. The challenge of information visualization evaluation. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, AVI '04, p. 109–116. Association for Computing Machinery, New York, NY, USA, 2004. doi: 10.1145/989863.989880 2

[33] P. T. Quinlan and R. N. Wilton. Grouping by proximity or similarity? competition between the gestalt principles in vision. *Perception*, 27(4):417–430, 1998. PMID: 9797920. doi: 10.1068/p270417 1, 4

[34] S. Schwartz, N. Green, S. Carberry, and J. Hoffman. A model of perceptual task effort for bar charts and its role in recognizing intention. *User Modeling and User-Adapted Interaction*, 16:1–30, 01 2006. doi: 10.1007/s11257-006-9002-9 3, 4, 7

[35] J. Scimeca and S. Franconeri. Selecting and tracking multiple objects. *Wiley Interdisciplinary Reviews: Cognitive Science*, 6, 12 2014. doi: 10.1002/wcs.1328 6, 8, 9

[36] S. Seabold and J. Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010. 5

[37] P. M. Shah, R. E. Hegarty, and Mary. Graphs as aids to knowledge construction: Signaling techniques for guiding the process of graph comprehension. *Journal of Educational Psychology*, 91(4):690–702, 12 1999.

doi: 10.1037/0022-0663.91.4.690 1, 2, 3, 4, 5

[38] D. K. Simkin and R. Hastie. An information-processing analysis of graph perception. *Journal of the American Statistical Association*, 82:454–465, 1987. 4

[39] C. P. Sison and J. Glaz. Simultaneous confidence intervals and sample size determination for multinomial proportions. *Journal of the American Statistical Association*, 90(429):366–369, 1995. doi: 10.1080/01621459.1995.10476521 5

[40] L. South, D. Saffo, O. Vitek, C. Dunne, and M. A. Borkin. Effective use of likert scales in visualization evaluations: A systematic review. *Computer Graphics Forum*, 41(3):43–55, 2022. doi: 10.1111/cgf.14521 2

[41] J. Talbot, V. Setlur, and A. Anand. Four experiments on the perception of bar charts. *IEEE Transactions on Visualization I& Computer Graphics*, 20(12):2152–2160, dec 2014. doi: 10.1109/TVCG.2014.2346320 2, 3, 8

[42] S. K. Thompson. Sample size for estimating multinomial proportions. *The American Statistician*, 41(1):42–46, 1987. doi: 10.1080/00031305.1987.10475440 5

[43] E. Tufte. *The visual display of quantitative informations 2nd ed*. Graphics Press, Cheshire, Conn., 2001. 1

[44] B. Tversky, S. Kugelmass, and A. Winter. Cross-cultural and developmental trends in graphic productions. *Cognitive Psychology*, 23(4):515–557, 1991. doi: 10.1016/0010-0285(91)90005-9 3, 5, 7, 9

[45] I. Vessey. Cognitive fit: A theory-based analysis of the graphs versus tables literature*. *Decision Sciences*, 22(2):219–240, 1991. doi: 10.1111/j.1540-5915.1991.tb00344.x 9

[46] F. Viénot, H. Brettel, and J. D. Mollon. Digital video colourmaps for checking the legibility of displays by dichromats. *Color Research I& Application*, 24(4):243–252, 1999. doi: 10.1002/(SICI)1520-6378(199908)24:4<243::AID-COL5>3.0.CO;2-3 4

[47] J. Wagemans, J. H. Elder, M. Kubovy, S. E. Palmer, M. A. Peterson, M. Singh, and R. von der Heydt. A century of gestalt psychology in visual perception: I. perceptual grouping and figure–ground organization. *Psychological Bulletin*, 138(6):1172–1217, (2012). doi: 10.1037/a0029333 4, 6

[48] E. Wall, C. Xiong, and Y.-S. Kim. Vishikers' guide to evaluation: Competing considerations in study design. *IEEE Computer Graphics and Applications*, 42(3):29–38, 2022. doi: 10.1109/MCG.2022.3152676 2

[49] J. Walny, S. Huron, and S. Carpendale. An exploratory study of data sketching for visual representation. *Computer Graphics Forum*, 34(3):231–240, 2015. doi: 10.1111/cgf.12635 2

[50] C. Ware. *Visual Thinking: For Design*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2008. 3, 6, 8

[51] C. Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann Series in Interactive Technologies. Morgan Kaufmann, Amsterdam, 3 ed., 2012. 1, 4

[52] M. Wertheimer. Untersuchungen zur lehre von der gestalt. ii. *Psychol. Forsch.*, 4(1):301–350, 1923. 1, 4, 6, 8

[53] C. Xiong, E. Lee-Robbins, I. Zhang, A. Gaba, and S. Franconeri. Reasoning affordances with tables and bar charts. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–13, 2022. doi: 10.1109/TVCG.2022.3232959 1, 2

[54] C. Xiong, V. Setlur, B. Bach, E. Koh, K. Lin, and S. Franconeri. Visual arrangements of bar charts influence comparisons in viewer takeaways. *IEEE Transactions on Visualization I& Computer Graphics*, 28(01):955–965, jan 2022. doi: 10.1109/TVCG.2021.3114823 1, 2, 5, 9

[55] Y. Xu and S. L. Franconeri. Capacity for visual features in mental rotation. *Psychological Science*, 26(8):1241–1251, 2015. 6, 8, 9

[56] L. Yang, C. Xiong, J. K. Wong, A. Wu, and H. Qu. Explaining with examples: Lessons learned from crowdsourced introductory description of information visualizations. *IEEE Transactions on Visualization I& Computer Graphics*, 29(03):1638–1650, mar 2023. doi: 10.1109/TVCG.2021.3128157 3, 5

[57] D. Yu, D. Tam, and S. L. Franconeri. Gestalt similarity groupings are not constructed in parallel. *Cognition*, 182:8–13, 2019. doi: 10.1016/j.cognition.2018.08.006 8

[58] D. Yu, X. Xiao, D. Bemis, and S. Franconeri. Similarity grouping as feature-based selection. *Psychological Science*, 30:095679761882279, 01 2019. doi: 10.1177/0956797618822798 8

[59] J. Zacks, E. Levy, B. Tversky, and D. Schiano. Reading bar graphs: Effects of extraneous depth cues and graphical context. *Journal of Experimental Psychology: Applied*, 4:119–138, 06 1998. doi: 10.1037/1076-898X.4.2.119 2

[60] J. Zacks and B. Tversky. Bars and lines: A study of graphic communication. *Memory I& cognition*, 27:1073–9, 12 1999. doi: 10.3758/BF03201236 1, 2, 3, 5

[61] C. Ziemkiewicz and R. Kosara. The shaping of information by visual metaphors. *IEEE transactions on visualization and computer graphics*, 14:1269–76, 11 2008. doi: 10.1109/TVCG.2008.171 2

[62] C. Ziemkiewicz and R. Kosara. Implied dynamics in information visualization. In *Proceedings of the International Conference on Advanced Visual Interfaces*, AVI '10, p. 215–222. Association for Computing Machinery, New York, NY, USA, 2010. doi: 10.1145/1842993.1843031 2