

Comparison Conundrum and the Chamber of Visualizations: An Exploration of How Language Influences Visual Design

Aimen Gaba, Vidya Setlur (Member), Arjun Srinivasan, Jane Hoffswell, and Cindy Xiong

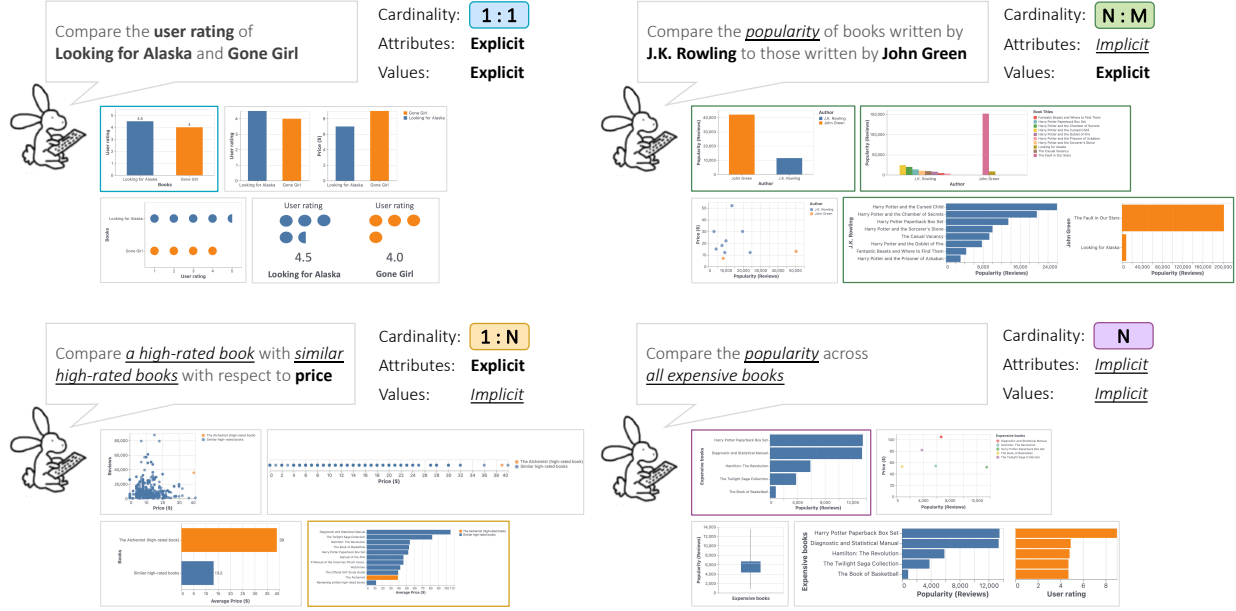


Fig. 1. Four comparison utterances from our design space with varying cardinalities for the comparison entities (1-1, 1-n, n-m, n) and different levels of concreteness (explicit and implicit). Each of these comparison utterances was included in our online survey in which participants ranked their preference for the different visualization types; the most preferred visualization(s) have a colored border.

Abstract—The language for expressing comparisons is often complex and nuanced, making supporting natural language-based visual comparison a non-trivial task. To better understand how people reason about comparisons in natural language, we explore a design space of utterances for comparing data entities. We identified different parameters of comparison utterances that indicate **what** is being compared (i.e., data variables and attributes) as well as **how** these parameters are specified (i.e., explicitly or implicitly). We conducted a user study with sixteen data visualization experts and non-experts to investigate how they designed visualizations for comparisons in our design space. Based on the rich set of visualization techniques observed, we extracted key design features from the visualizations and synthesized them into a subset of sixteen representative visualization designs. We then conducted a follow-up study to validate user preferences for the sixteen representative visualizations corresponding to utterances in our design space. Findings from these studies suggest guidelines and future directions for designing natural language interfaces and recommendation tools to better support natural language comparisons in visual analytics.

Index Terms—Comparative constructions, cardinality, explicit and implicit comparisons, natural language, intent, visual analysis.

1 INTRODUCTION

Visual comparisons are a common and critical task in analytic workflows. When people read a visualization, each comparison they make can be thought of as an analytical sentence that describes common patterns, differences, and trends in the data [63], such as “the recovery rate is overall the same for Hospitals A and B.” Even seemingly-small

design choices in a visualization can nudge viewers to see different patterns and produce different sentences [13, 69], further highlighting the complexities and nuances of comparison tasks. For example, Figure 2 shows how the juxtaposition of small multiples, either vertically or horizontally, can elicit different comparisons across the two groups.

Recommendation systems attempt to provide smart defaults to help users gain insights into their data but focus only on common operations such as filtering and sorting. They do not provide recommendations for facilitating visual comparisons [59]. Recently, visual analysis interfaces have introduced natural language (NL) queries into recommendation systems [3] that enable users to type NL utterances and receive appropriate visualization responses. Such systems make extracting insights from data more intuitive, as the process of generating utterances exposes the mental model of how people think about comparisons [14, 19, 29, 54]. However, the complexity of comparison expressions makes supporting NL-based visual comparisons a non-trivial task.

We identify two main challenges for building an effective recom-

- Aimen Gaba and Cindy Xiong are with UMass Amherst. E-mail: {agaba, yaxiong}@umass.edu.
- Vidya Setlur and Arjun Srinivasan are with Tableau Research. E-mail: {vsetlur, arjunsrinivasan}@tableau.com.
- Jane Hoffswell is with Adobe Research. E-mail: jhoffs@adobe.com.

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.xxx/TVCG.201x.xxxxxx

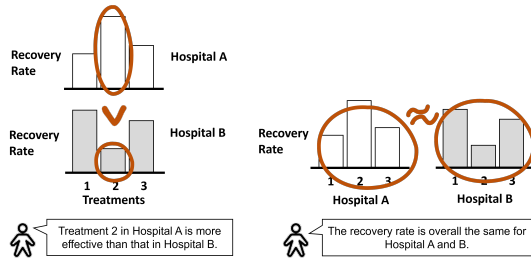


Fig. 2. Examples of two sets of spatial arrangements for visual comparisons based on findings from Xiong et al. [69]. Results found that stacking small multiples vertically tends to elicit 1-to-1 comparisons of individual values in a set, whereas aligning them horizontally tends to elicit a comparison of the two sets as two groups.

mendation system that can help users more easily make comparisons in visualized data. First, designing a visualization involves a series of decisions, each of which can afford different viewer takeaways [69, 72]. For example, past work has shown that people tend to make magnitude comparisons when viewing bar charts and trend comparisons when viewing line charts [66, 73]. However, to the best of our knowledge, prior work does not propose a comprehensive mapping between comparison intents and the best visualization for the intended comparisons.

Second, there exists semantic ambiguity and complexity in NL that makes the intent difficult to parse. These characteristics hinder the effectiveness of incorporating NL approaches to support visual comparison tasks [62]. While **explicit** comparisons such as “compare the number of trees planted in location A to the number in location B” are easy to decode, **implicit** comparisons are difficult for visualization systems to interpret precisely. For example, consider the example, “Are the number of trees planted in location A tall?”. For comparison utterances that use vague modifiers such as “tall” [33], it is difficult to determine how the user intended for “tall” to be interpreted, i.e., what constitutes something being *tall*? For comparisons that contain underspecified references to the dataset, such as referring to “performance” when “performance” is not a variable explicitly contained in the dataset, it is similarly difficult to determine how a system should behave. Figuring out how to decode comparison intents from implicit utterances and underspecified references in NL for visualization use cases still remains an open question.

Contributions: We contribute a preliminary design space of NL comparison utterances covering four cardinalities of comparisons and addressing ambiguities by differentiating implicit and explicit ways of expressing these comparisons. Through a series of interviews with 16 visualization novices and experts, we create a potential mapping between comparison utterances in our design space and different visualization techniques. We then empirically validate this mapping through an online survey with 77 participants to highlight three design implications for interpreting comparison utterances for visualization use cases. These design implications can inform the design of future NL-based recommendation systems to better support visual comparisons.

2 RELATED WORK

Designing effective visualizations for comparison utterances is an ongoing area of research and can be categorized into three main themes: (1) comparisons in visualization, (2) comparisons in computational linguistics, and (3) natural language interfaces (NLIs) for visual analysis.

2.1 Comparisons in Visualization

Representing comparisons in data visualizations is an important aspect of a user’s analytical workflow and prior work has surveyed a variety of visualization solutions to better support comparisons. For example, Tufte discussed small multiples as a way to use the same graphic to display different slices of a data set for comparison [68]. Graham and Kennedy [30] surveyed a range of visual mechanisms to compare trees, while other surveys consider methods for comparing flow fields [48]. Gleicher et al. [27] presented a broad survey with over 100 different

comparative visualization tools from information visualization domains, organized by their comparative visual designs into a general taxonomy of visual designs for comparisons. Designs were grouped into three categories: juxtaposition, superposition, and explicit encodings.

The perceptual and cognitive science communities have also considered the problem of visual comparison for decades, including the issues around change blindness [52]. Other studies focused on comparing methods for specific tasks, such as the evaluation of scalar field comparisons [46] or brain connectivity graphs [7]. Franconeri [19, 20] discussed several limitations in the mechanisms of perception that may have a direct impact on the design of comparison methods. For example, translations of an object are easy to compare, but texture, orientation, scale, space, and time may complicate comparison tasks [42, 43].

Existing work has largely focused on how *visualization type* can influence viewer perception and decisions [45, 70, 73]. For example, bar charts plot the data as discrete objects, motivating people to compare them as two distinct units (e.g., A is larger than B), while line charts plot data as one single object, eliciting the interpretation of trends, changes over time or relations (e.g., X fluctuates up and down as time passes) [73]. Showing difference benchmarks on bar charts can not only facilitate a wider range of comparison tasks [65] but also increase the speed and accuracy of the comparisons [47]. Charts that show probabilistic outcomes as discrete objects, such as a beeswarm chart, can promote a better understanding of uncertainties [24, 32, 35, 67].

Comprehension of visual comparisons is an important aspect of determining their efficacy. Researchers have referenced knowledge from basic human perception research to generate design guidelines regarding graphical elements, such as the visualization color and shape. For example, in multi-class scatterplots, viewers can compare classes of scatterplots more effectively when they are colored differently compared to when they are plotted with different shapes [18, 28]. Shah and Freedman [63] investigated the effect of format (line vs. bar), viewers’ familiarity, and their graphical literacy skills on the comprehension of multivariate data presented in graphs. The differences between people’s perceptions of bar graphs and line graphs can be explained by differences in the visual chunks formed by the graphs based on Gestalt principles of proximity, similarity, and good continuity. Jardine et al. [34] conducted an empirical evaluation on two comparison tasks: identify the “biggest mean” and “biggest range” between two sets of values. Their work showed that visual comparisons of these tasks are most supported by vertically stacked chart arrangements, and this pattern is substantially different across different types of tasks.

Much of this prior work has focused on exploring effective *visual representation* of comparisons in data. However, with the prevalence of NLIs that support visual analysis, there is a need for such systems to be able to interpret comparison utterances and provide useful visualization responses to the user. Our work specifically focuses on better understanding the design space of mapping data comparisons expressed through *language* with useful visual representations.

2.2 Comparisons in Computational Linguistics

The syntax and semantics of comparatives have been the topic of research in computational linguistics for quite some time. Comparative expressions which establish orderings among objects according to the amount or degree to which they possess some property is a basic component of human cognition [38]. Language constructs often provide concepts to express gradable concepts such as the explicit orderings between two objects (e.g., “the price of housing in the Bay Area is higher than in Texas”) [56]. Extensive prior work in computational linguistics has explored the semantics of comparison based on gradable concepts such as “more”, “less”, “-er” [10, 12, 16, 31, 37, 40, 58].

The notion of vagueness in comparative language has previously been studied in the computational linguistics community [9, 21]. The semantics of comparatives can be vague as their interpretation depends on the context and the boundaries defining the comparative. For example, “when is it the safest time to fly?” implicitly compares flying safety across different time periods, with “safe” being a fuzzy, subjective concept. Prior work has focused on the conceptualization and representation of vague knowledge. For example, Kessler and Kuhn presented

a corpus of annotated comparison sentences from English camera product reviews. Each sentence contains the comparative predicate that expresses the comparison, the comparison type, the two entities that are being compared, and the aspect in which they are compared [39].

While linguistic vagueness has been explored for comparative expressions along with their semantic variability, little work has been done in determining how best to *visually* represent comparatives based on these variations, especially in the context of visual analysis.

2.3 Natural Language Interfaces (NLIs) for Visual Analysis

Much of the effort for supporting NLIs for visual analysis [1, 3, 4, 6, 17, 22, 59] focuses on interpreting a user's analytical intent by providing explicit visualization responses. The methods of interpreting intent typically rely on explicitly named data attributes, values, and chart types in the user's input queries. Ask Data [62] supports analytical expressions in NL such as grouping of attributes, aggregations, filters, and sorts. The system also handles impreciseness around vague numerical concepts such as "cheap" and "high" by inferring a range based on the underlying statistical properties of the data. Hearst et al. [33] explore appropriate visualization responses to vagueness by interpreting singular and plural superlatives (e.g., "highest price" and "highest prices") and numerical graded adjectives (e.g., "higher") based on the shape of the data distributions. Law et al. [44] investigated how the visual design of answers to *why* questions might influence user perceptions of a question-answering system. They found that users have a strong tendency to associate correlation with causation when systems do not provide clear explanations for the answers.

The space of analytical expression in NL is rich and much more nuanced than what these interfaces currently support. A study was conducted to assess NL input to visualization systems [62] where 75 participants were asked to write NL queries based on five underlying datasets (i.e., bird strikes, world indicators, superstore, mutual funds, and Olympic medals). Of the 578 NL queries, common ones included "Are there more strikes on takeoff or landing?", "Are certain seasons more dangerous?", and "Which country has more female medalists?" The data suggests that when participants were not restricted in the format of expression, they often chose to specify utterances with an underlying intent to either explicitly or implicitly compare values. Current NL systems, however, do not explore how utterances about comparisons ought to be interpreted even though such forms of intent are prevalent.

Our paper identifies a gap in mapping how users express comparative utterances in NLIs to appropriate visualizations. To address this problem, we explore a design space connecting language and visual representation for a range of comparison utterances, varying in what is being compared and how these comparisons are specified. Finding from our work provide implications for NLIs and recommendation tools to better interpret and support comparisons.

3 DESIGN SPACE OF COMPARISON UTTERANCES

In visual analytics, Gleicher et al. [26] define comparisons as an analytical task involving two components: a set of targets, i.e., the set of items being compared, and an action performed on the relationships among these targets, e.g., similarities and differences. We extend this definition of comparison as the task of identifying similarities and differences between two or more categories of data **values** based on one or more data **attributes** shared among those categories. For the remainder of this paper, we refer to this extended definition of comparisons.

In this work, we explore the intricacies of language constructs used to express comparisons and particularly focus on understanding the semantic variations in comparison expressions and how those variations influence the corresponding visual representations. To this end, we define a *comparison utterance* as a textual sentence, inputted by a user to an NLI, for example, that expresses the intent of performing a comparison with a given dataset. An example of a comparison utterance is "compare the sales for Washington and California," where "Washington" and "California" are the values, and sales is the attribute. Inspired by the language constructs from the computational linguistics literature [36, 64], we describe our design space for comparison expressions in terms of the **cardinality** and **concreteness** of the comparisons.

Cardinality: Comparison utterances can express relationships between individual entities, e.g., "compare the effectiveness of treatment A to treatment B", as well as relationships between individuals and larger sets containing multiple entities, e.g., "compare the effectiveness across all treatments". Similar to the process described by Xiong et al. [69], we list four cardinalities for the entities in comparison utterances in the context of visual analysis:

1-1: Compare one entity to another entity
e.g., "compare the IMDB ratings of Squid Game and Midnight Mass"

1-n: Compare one entity to another set of multiple entities
e.g., "compare the performance of Starling to other PG-13 movies"

n: Compare multiple entities within the same set
e.g., "compare the budgets across all US movies"

n-m: Compare one set of entities to another set
e.g., "compare crime shows to thriller shows in terms of box office"

Concreteness: Comparison utterances can *explicitly* mention data attributes and values to be compared; for example, the comparison expression "compare the number of silver medals won by Rebecca Adlington to all other participants in the Women's Swimming Event" explicitly refers to the values to consider ("Rebecca Adlington" and "all other participants in the Women's Swimming Event") as well as the exact data attribute to use in the comparison (silver medals). Due to language variations, we account for plurality (e.g., "book" and "books"), spelling variation (e.g., "color" and "colour"), and stemming (e.g., "costly" and "cost") for explicit references.

Comparison utterances can also be underspecified when one or more data attributes or values are *implicitly* referenced in the comparison [60]. For example, the utterance "compare the **popularity** of all movies in 2021" could be considered implicit if the dataset does not contain an attribute or value that explicitly matches the token "popularity." Rather, other attributes such as number of reviews or user rating may be more appropriate to consider when comparing the popularity of the movies. In addition, implicit comparison utterances can employ language constructs such as gradable vague modifiers like "low", "high", or "cheap" when comparing data, e.g., "compare athletes who won a **high** number of gold medals". These gradable adjectives are often mapped to an upper or lower range of data values in a data column [33, 62].

3.1 Identifying the design scope of comparisons

Both the cardinality and concreteness of the comparison utterance can impact the desired visual characteristics to perform the corresponding comparison task. Within each cardinality, there are theoretically 16 combinations of comparisons that we can identify via a truth table by permutating whether each of the two data values and data attributes are implicit or explicit ($2 \times 2 \times 2 \times 2$). However, not every combination produces a valid or natural comparison scenario. For example, the order of elements for degrees of concreteness does not fundamentally change the resulting comparison task; in other words, the comparison "compare the performance of Starling to other high rated movies" (an explicit value followed by an implicit value) is essentially the same as the comparison "compare the performance of high rated movies to Starling" (an implicit value followed by an explicit value). We can therefore reduce the space to focus on utterances that are order agnostic. Furthermore, while we initially consider a design space of two values with two corresponding attributes, comparisons that involve *different* attributes for each value are rather nonsensical like comparing "apples" and "oranges", e.g., "compare the gold medals obtained by Rebecca Adlington to the bronze medals obtained by Nathan Ghar-Jun Adrian". Finally, we chose to exclude comparisons with different degrees of concreteness for the values as many of these comparisons felt artificial compared to real-world comparison utterances, e.g., "compare a low budget TV show to Squid Game with respect to their popularity".

The final four combinations of concreteness in our design space are as follows: *explicit* data values paired with *explicit* data attributes (EV-EA); *explicit* values and *implicit* attributes (EV-IA); *implicit* values and *explicit* attributes (IV-EA); and *implicit* values paired with *implicit* attributes (IV-IA). Our final design space, therefore, includes a total

of 16 utterances across four cardinalities (**1-1**, **1-n**, **n**, **n-m**) and four combinations of concreteness (EV-EA, EV-IA, IV-EA, IV-IA). The final design space along with example queries for both datasets can be found in our supplemental material.¹

4 STUDY 1: DESIGNING VISUALIZATIONS FOR COMPARISONS

We conducted user studies with visualization experts and non-experts to better understand how they visualize data for comparison utterances.

4.1 Participants

We recruited 16 participants (six female, ten male). All participants were fluent in English and self-reported being comfortable designing visualizations. Eight participants (four female, four male) self-reported as data visualization experts with an average of ~11 years of visual analytics experience and were recruited from a visual analytics company. The data visualization experts had a variety of job backgrounds including three data visualization consultants, two user experience designers, one business strategist, one visualization lead expert, and one solution engineer. The other eight (two female, six male) self-reported as data visualization non-experts and were recruited from an academic institution. All of the non-expert participants had a computer science or engineering background, with six being graduate students and two being undergraduates. Participants were recruited via a company mailing list, Slack, and using snowball sampling [25]. The students were offered to enter a \$30 gift card raffle as an incentive to participate, while the participants recruited from the visual analytics company were not offered any financial incentive due to company policy.

4.2 Procedure

Each study session lasted 60 minutes and was conducted remotely via Zoom. Two researchers supported each session: one facilitator and one note-taker. All sessions were video recorded. Field notes were expanded to a video log after the study through partial transcription of the videos. The video log (and raw video for reference) was then qualitatively coded for high-level themes.

Participants recruited from the academic institution were asked to fill out a consent form to indicate their interest in receiving a gift card. All participants were asked for their consent for the session to be recorded, and participants were then asked to share their screens in order to make the session more interactive between the participant and the facilitator.

During each session, participants were first introduced to the study and asked about their visualization background and current role. Participants were then given instructions for the study and were asked to draw one or more visualizations for four comparison queries. The main study prompt was as follows: “Imagine that you are a visualization recommendation tool that generates visualization(s) when prompted by a user. Assume that the user imported a [Netflix or Olympics] dataset, and inputted some utterances describing what they want to see from the data. Your task is to generate a visualization or multiple visualizations (if you think that it can be visualized in multiple ways) as a response to the utterance and present it back to the user. Please sketch your visualization response that you think will best respond to their request.”

Participants were first provided a weblink to a Google Sheet containing the dataset along with a metadata summary sheet. They then read four comparison utterances one at a time and used Google Jamboard [5] to sketch out the visualizations. Participants were encouraged to think aloud during the study, and we employed a question-asking protocol to elicit qualitative feedback. Responses to common questions were documented ahead of time so that the facilitator could provide consistent responses to every participant.

4.3 Experimental Design

This study employed a 4×4 Graeco-Latin square design that covered four different cardinalities (**1-1**, **1-n**, **n**, and **n-m**) and two levels of concreteness (I: implicit and E: explicit) for the data values (V) and data attributes (A). Each participant came across all four cardinalities and

¹ Additional details regarding the experimental protocol, datasets, and all the comparison utterances can be found in supplemental material.

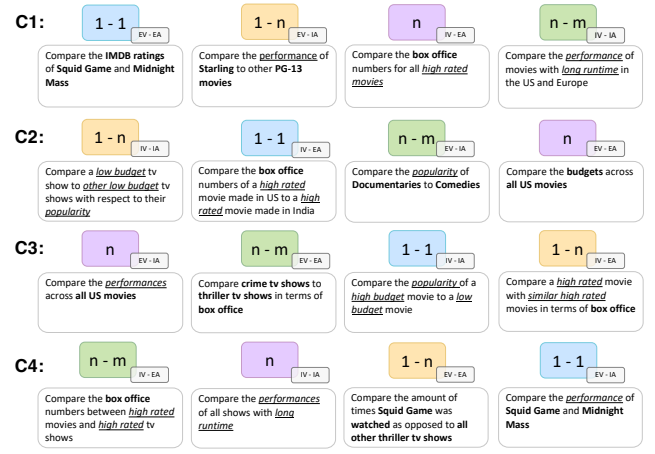


Fig. 3. A Graeco-Latin square with example utterances from the Netflix dataset and four participants came across each condition.

all four combinations of concreteness (EV-EA, EV-IA, IV-EA, IV-IA); the 4×4 Graeco-Latin square design balances the order in which the participant comes across the utterances. The conditions are shown in Figure 3 and described as follows:

- **Condition 1:** 1-1 EV-EA, 1-n EV-IA, n IV-EA, n-m IV-IA
- **Condition 2:** 1-n IV-IA, 1-1 IV-EA, n-m EV-IA, n EV-EA
- **Condition 3:** n EV-IA, n-m EV-EA, 1-1 IV-IA, 1-n IV-EA
- **Condition 4:** n-m IV-EA, n IV-IA, 1-n EV-EA, 1-1 EV-IA

4.4 Experimental Dataset

In each condition, participants interacted with one of two datasets: a Netflix dataset [11] and an Olympics medals dataset [53]. The Netflix dataset contains 13 attributes including seven categorical (e.g., Title, Genre), six numeric (e.g., IMDB Rating, Box office), and one temporal (Release year). The Olympics dataset also contains 13 attributes composed of six categorical (e.g., City, Event), six numeric (e.g., Height, Gold), and one temporal attribute (Year). We chose these datasets as they contain a combination of continuous and categorical variables that can support a variety of comparison utterances. The datasets were also familiar to a wide set of participants and hence could be easily interpreted. We used a subset of 60-70 rows for each of the two datasets to prevent participants from becoming overwhelmed with the amount of data when performing the tasks.

4.5 Pilot Studies

We conducted nine pilot studies to test the study protocol and stimuli. Given the remote experimental setup using a combination of online tools, the goal of the pilots was to streamline the user study process and resolve any points of confusion, particularly around the instructions and utterances. We conducted, transcribed, and analyzed ~540 minutes of pilot interviews before conducting the final user study.

Based on feedback from the pilot participants, we updated the instructions for the study and refined the utterances and datasets. To better orient the participant about the dataset and its attributes and values, we included a metadata summary. We also incorporated a short tutorial in the study protocol to familiarize participants with the various drawing options in Jamboard and data manipulation features in Google Sheets.

4.6 Study 1: Qualitative Analysis

In total, we gathered ~960 minutes of video and audio recordings, transcribed them using Amazon AWS Transcribe [41], and then placed each unique statement into our analysis software, a shared spreadsheet. We extracted the following details from the recordings:

Visualization design: We extracted details about the visualization type (e.g., bar chart, scatter plot) and encoding (e.g., color, size), which included details about the mapping between data values and visual

properties. We also considered high-level properties of the visualization layout such as the *orientation* (e.g., vertical or horizontal), *arrangement* [69] (e.g., adjacent, overlaid), and whether or not the visualization used a *small multiples* design [8].

Visualization details: We also extracted additional narrative characteristics of the visualization such as the use of *highlighting* to emphasize certain features and the use of *annotations* (e.g., labels, reference lines, legends). We also recorded whether or not the visualization was *sorted*.

Visualization or data transformations: An important characteristic of this study was how participants *interpreted* implicit data values and attributes; if participants identified multiple interpretations, we also recorded which of them the participant preferred. As part of the design process, we recorded how the data was *filtered* before drawing; for instance, for the utterance “*compare the performance of Starling with other PG-13 movies*,” some participants chose only the “top 3” movies to compare with Starling, whereas others chose 5, and so on. Finally, we extracted the desired *interactive* features of the visualization described by participants, such as the inclusion of a tooltip or data label on hover, or other ways in which the visualization could change on demand.

Participant details & preferences: In addition to *demographic* characteristics like the participant's occupation, and expertise, we extracted the *number* of visualizations created during the study and in cases where multiple visualizations were drawn for an utterance, we also extracted which visualizations the participant *preferred* and why.

4.6.1 Interpreting ambiguity of implicit values and attributes

One of the main goals of this study was to understand how participants interpret the concreteness of the data values and data attributes. Participants came across multiple implicit data **values** such as “high rated movie”, “high budget movie”, “tall athlete”, “successful wrestler”, and “high achieving”, as well as implicit data **attributes** such as “performance”, “popularity”, “physique”, “achievements,” and “long runtime”. Additional details regarding the datasets and their metadata summaries can be found in the supplementary material.

Netflix Dataset: The following list describes the implicit data values and data attributes, their implicit type (i.e., vague modifier or underspecification), and examples of how they were commonly interpreted by participants for the Netflix dataset:

- **performance** (*vague modifier*): [Watched] OR [IMDB rating] OR [Rotten tomatoes rating] OR ROI/Profit [Box office – Budget] OR [Box office]
- **popularity** (*vague modifier*): [Box office – Budget] OR [Watched]
- **high rated** (*underspecified*): [Rotten tomatoes rating] > 80/100 OR [Rotten tomatoes rating] > AVG(Rotten tomatoes rating) OR [IMDB rating] > AVG(IMDB rating) OR [IMDB rating] > 8/10
- **high budget** (*underspecified*): [Budget] > AVG(Budget) OR [Budget] > 95th percentile
- **low budget** (*underspecified*): [Budget] < AVG(Budget) OR [Budget] < 5th percentile OR lowest budget value
- **long runtime** (*underspecified*): [Duration] > 100 minutes OR [Duration] > 80th percentile

Olympics Dataset: The following list describes some of the implicit values and attributes, their implicit type, and examples of how they were interpreted for the Olympics dataset:

- **achievements** (*vague modifier*): SUM(Gold, Silver, Bronze)
- **performance** (*vague modifier*): SUM(Gold, Silver, Bronze) OR weighted metric of [Gold, Silver, Bronze]
- **physique** (*vague modifier*): [Height] OR [Weight] OR metric of [Weight, Age, Height]

- **tall athlete** (*underspecified*): [Height] > AVG(Height) OR [Height] > MEDIAN(Height) OR [Height] ≥ 180cm
- **short athlete** (*underspecified*): [Height] < AVG(Height) OR [Height] < MEDIAN(Height) OR [Height] < 180cm
- **strong physique** (*underspecified*): [Weight] > AVG(Weight) OR [Weight] > weighted sum of [Weight, Height, Age]
- **successful player** (*underspecified*): SUM(Gold, Silver, Bronze) > 0
- **high achieving** (*underspecified*): [Gold] > 0
- **young athlete** (*underspecified*): [Age] < MEDIAN(Age) OR [Age] < 20
- **top-winning** (*underspecified*): [Gold] > 0 OR weighted sum of [Gold, Silver, Bronze] > 0
- **senior athlete** (*underspecified*): [Age] > AVG(Age)

While a few of the participants interpreted the implicit data values or attributes in only a single context from the list given above, most of them interpreted them in multiple contexts. In the case of underspecified values, we notice a general trend of participants interpreting values as being above or below average. In the case of vague modifiers, participants usually picked all the attributes from the dataset related to the implicit value or attribute and mapped all of them to it. For example, P9 explained their approach to resolving the ambiguity as follows: “*performance is kind of a vague thing. So let's just assume number of medals. That would be the best thing to assume. So performance is how many accolades they got.*”. Participants also tended to use some form of a weighted average, algorithmic, or parametric approach to interpret the implicit terms; for example, P4 interpreted “high rated” as “*I would statistically identify the high ratings more of like in a statistical or algorithmic way.*”

4.6.2 Developing clear charts for the intended comparison task

Some participants found the decision process for developing the visualization designs particularly difficult when trying to handle aggregation or other forms of data transformation. Reflecting on their role as “a visualization recommendation tool,” one participant explained that “*I would prefer for a recommendation engine to say don't look at the mean, look at the median. But if it's a normal distribution, I'd prefer it to say, let's look at the mean... you can trust us and that we are automatically providing the best results.*” [P4].

A few participants suggested leveraging appropriate design decisions, such as highlighting certain parts of the visualizations or adding data transformations, to make the comparison easier for a viewer to comprehend. P9 shared insights on how design decisions can support comparison tasks: “*So even something as simple as like a dot on the left to draw your eye to Rebecca because she is central to the question... or I would do the same chart, but put her at the very top, so she's floated to the top and then beneath her we're still in descending order.*”

Another participant argued that the initial visualizations should be rather simple when describing their 2-bar bar chart: “*I don't want to start with more than they asked. The two bars next to each other, I feel like they're good if there are not too many things that you're comparing. So if it's just two movies, I feel like this works really well.*” [P4].

4.7 Study 1: Quantitative Analysis

We analyzed the time spent designing visualizations based on the four cardinalities and level of concreteness, as well as the number of visualizations that the participants sketched. Additional details can be found in the supplementary materials.

Time Spent: Two-way ANOVAs comparing the time people spent for each of the 16 queries from both datasets showed no difference between cardinalities, concreteness, nor datasets ($p > 0.05$). On average, participants spent 9.75 min (SD = 0.52) per query.

Number of Visualizations: Two-way ANOVA comparing the number of visualizations (including the scratched out ones) that participants

Table 1. The percentage of visualizations that participants described as *interactive* or *annotated* for each cardinality and concreteness.

Percent of visualizations with interaction							
Cardinality				Concreteness			
1-1	1-n	n	n-m	EV-EA	EV-IA	IV-EA	IV-IA
23.3	25.7	45.2	24.4	13.2	28.6	37.0	36.1

Percent of visualizations with annotation							
Cardinality				Concreteness			
1-1	1-n	n	n-m	EV-EA	EV-IA	IV-EA	IV-IA
36.7	42.9	45.2	43.9	34.2	40.5	51.9	38.9

sketched for each query in both datasets suggests that there is no significant effect of concreteness, cardinality, or their interaction. On average, participants drew 2.25 visualizations (SD = 0.15) per query.

We also recorded the number of visualizations with *interactive* features and *annotations* for each dimension (Table 1). Based on the four cardinalities, we see that cardinality **n** had the highest percentage of visualizations with interactive features, followed by cardinality **1-n**. In terms of concreteness, IV-EA had the highest percentage of visualizations with interactive features, closely followed by IV-IA. When P11 was asked why they added interactivity to their sketch, they commented, “I don’t know anyone who would really want to look at gold and silver, but I would look at gold medals versus all medals seems like a reasonable question... being able to do things like add events at years, filter by medals, they seem to be the adjacent questions to me in this.”

Regarding the percentage of annotated visualizations, **n** had the highest, followed by **n-m** and **1-n**. In terms of concreteness, IV-EA had the highest percentage of annotated visualizations, followed by EV-IA, and then IV-IA. Many participants emphasized the importance of annotations and highlighting to make the comparisons clearer. P2 explained that “you could highlight a specific show... so it stands out and then all the other ones would be a different color.” This strategy was common across participant sketches and informed the design of the representative visualizations in Study 2 (Section 5).

4.8 Identifying Common Visualizations per Cardinality

Based on the participants’ sketches, we created illustrations of the visualizations using Notability software [2]; a sample set shown in Figure 4. We clustered the visualizations based on the four cardinalities. Two authors went through the visual clusters to find commonalities between the charts. We initially thought that the top three visualizations from each cardinality would be sufficient to represent the common visual charts. However, after some extensive examination of the visual clusters, the top four visualizations for each cardinality were chosen.

The top four visualizations were chosen based on the occurrence of the chart type, arrangement, and orientation of the chart. We then looked at whether and how these charts were annotated in order to include those annotations in our generalized drawings. We also looked in detail into how the implicit values and attributes were interpreted by the participants. We commonly noticed the introduction of interaction, grouped/small multiples/stacked, just mapping of one attribute or value from the dataset in visualizations that were sketched for levels of concreteness involving implicitness. As the focus of this paper is not to look into interactions in visualizations, we decided to map multiple or single attribute(s)/value(s) in the levels of concreteness that involve ambiguity. In the following paragraphs, we describe in detail the commonalities in the sketches by the participants and the process by which 16 visualizations were extracted, i.e., the top four for each cardinality.

Cardinality 1-1: From a total of 30 visualizations, there were 15 bar charts, 8 unit charts, 2 scatter plots, 2 pie charts, 1 line chart, and 1 box plot. 10 of the charts had a vertical orientation and 7 had a horizontal orientation. For the bar charts, seven were grouped bar charts and the rest were simple two-bar bar charts. We noticed a trend of both the values being in a different color. The simple bar charts were annotated with the value of the attribute on top of the bar. There were three

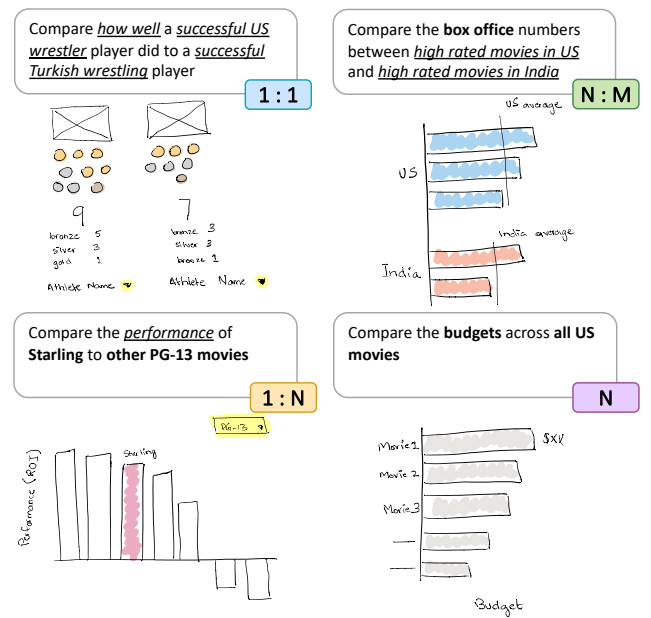


Fig. 4. Sketches elicited from Study 1. Charts are recreated for enhanced clarity based on the original sketches and interview script.

adjacent labeled unit charts where the data values were mentioned in a bigger font below the circles, and there were four other unit charts. There was a trend of values included in the charts as legends or labels. For **1-1**, we created (A) a simple bar chart, (B) an adjacent unit chart, (C) a horizontal unit chart, and (D) a grouped bar chart.

Cardinality 1-n: From a total of 35 visualizations, there were 21 bar charts, 5 scatterplots, 4 dot plots, 2 unit charts, 2 line charts, and 1 pictograph. 12 of the charts had a horizontal orientation. For the bar charts, 2 were grouped bar charts and 2 stacked. Generally, there were multi-bar bar charts and two-bar bars charts where one bar depicts a value and the other depicts the average value of others. We also noticed a general trend of coloring the value mentioned in the query in a different color as compared to others. The simple bar charts were annotated with the value of the attribute on top of the bar. There was another trend of data values included in the charts as legends or labels. For **1-n**, we created (E) a horizontal multi-bar chart, (F) a horizontal simple bar chart, (G) a dot plot, and (H) a scatterplot.

Cardinality n: From a total of 31 visualizations, there were 17 bar charts, 5 scatterplots, 3 box plots, 1 line chart, 1 parallel coordinates chart, and 1 sentiment chart. 10 of the charts had a horizontal orientation and 9 had a horizontal orientation. For bar charts, there were four small multiples, five stacked, and two grouped. We also noticed a general trend of all the values being in the same color representing a single attribute in the case of bar charts. All the values were shown in different colors or depicted in a legend in the case of scatter plots. For **n**, we created (I) a horizontal multi-bar chart, (J) a small multiple bar chart, (K) a scatterplot, and (L) a box-plot.

Cardinality n-m: From a total of 41 visualizations, there were 21 bar charts, 5 scatterplots, 6 dot plots, 2 box plots, 2 unit charts, 4 line charts, and 1 pie chart. 15 of the charts had a vertical orientation and 10 had a horizontal orientation. For bar charts, six were grouped bar charts, five small multiples, and two stacked. We also noticed a general trend of all the values of n and all the values of m in a different color. All the values were in a different color in the case of scatterplots and grouped bar charts. For **n-m**, we created (M) a grouped bar chart, (N) a small multiples bar chart, (O) a simple vertical bar chart, and (P) a scatterplot.

5 STUDY 2: IDENTIFYING VISUALIZATION PREFERENCES

We conducted an online crowdsourced experiment to cross-validate user preferences for the 16 representative visualizations identified in Section 4.8 for our design space. We focus on non-expert user preferences because the designs from visualization experts are often meant

for communicating key patterns to naive viewers to help them make comparisons. The results of this study can inform the design choices for visualization recommendation systems focused on comparisons.

5.1 Participants

Based on a pilot study with 10 participants using only **1-1** queries, we conducted a power analysis to determine the number of participants required for the experiment to find an overall difference in preference rankings. With a medium effect size of ~ 0.49 , our analysis suggests that a target sample of 76 would yield 95% power to detect an overall difference between preference rankings for the four visualizations at an alpha level of 0.05. We recruited 79 participants via Prolific.com [49] to complete an online survey through Qualtrics [51]. They were compensated at 10.55 USD per hour. In order to participate in our study, the workers had to be based in the United States and fluent in English. After excluding participants who failed attention checks (e.g., failing to select a specific answer in a multiple-choice question) or entered illegible/nonsensical response, we ended up with 77 participants, with 58 that identified as women ($M_{age} = 39.34$, $SD_{age} = 16.18$), 18 as men ($M_{age} = 40.94$, $SD_{age} = 13.90$), and one chose to not disclose.

The participants completed a subjective graph literacy report [23] and reported an average value of 3.87 out of 6 ($SD = 0.82$, 1 = low self-reported literacy, 6 = high self-reported literacy), suggesting that most participants were comfortable with visualizations but did not identify as visualization experts. Only 5 people reported that they create visualizations often for work or as a hobby, and 17 people reported that they rarely interact with visualizations in their day-to-day life.

We asked our participants how much effort they put in completing our study; they were told that the answer to this question would not affect their compensation and were encouraged to answer honestly. 56 participants reported having put in a lot of effort into our study, carefully thinking through their answers before responding. 21 participants reported having put in some effort, having answered the questions without thinking too deeply about anything.

5.2 Stimulus and Design

Similar to Study 1, we considered four cardinalities and four levels of concreteness, resulting in 16 comparison queries. We used the Amazon Books dataset [55] and generated queries similar to those from Study 1. The dataset had seven attributes including three categorical (e.g., Book Title, Author), three numeric (e.g., User rating, Reviews), and one temporal attribute (Year). We created visualizations following the representative designs extracted from Study 1 (Section 4.8). We varied the data attributes and values used to generate the visualization to match the content of the queries for each cardinality. For the 16 queries, we made 64 visualizations using Vega-Lite [57]; 16 visualizations for each cardinality that included four visualizations for each level of concreteness (query). We briefly describe how the visualizations looked for each level of concreteness in Study 2. The full set of comparison queries and example visualizations can be found in Figure 5.

EV-EA: For **1-1** (charts A, B, C), **1-n** (E, F, G), **n** (I, L), and **n-m** (M, N, O) the visualizations were straightforward with explicit data values compared to an explicit data attribute (User rating) in case of **1-1** and the data attribute (Price) in case of **1-n**, **n**, and **n-m**.

EV-IA: For **1-1** (charts A, B, and C), **1-n** (E, F, and G), **n** (I, and L), and **n-m** (M, N, and O) the visualizations were straightforward with explicit data values compared to an implicit data attribute ("Popularity"), which is a vague modifier and is mapped to the attribute (Reviews).

IV-EA: For **1-1** (charts A, B, and C), **1-n** (E, F, and G), **n** (I, and L), and **n-m** (M, N, and O) the visualizations were straightforward with implicit data values compared to an explicit data attribute (User rating) in case of **1-1** and the attribute (Price) in case of **1-n**, **n**, and **n-m**. The titles of implicit values such as "a bestseller book in 2012" in case of **1-1**, "high rated books" in case of **1-n** and **n**, and "high rated fiction books" in case of **n-m** are shown in the charts as labels or legends.

IV-IA: For **1-1** (charts A, B, and C), **1-n** (E, F, and G), **n** (I, and L), and **n-m** (M, N, and O) the visualizations show implicit data values

compared to an implicit data attribute ("Popularity") which is mapped to the attribute (Reviews). The titles of implicit values such as "a cheap book" in case of **1-1** and **1-n**, "expensive books" in case of **n**, and "high rated non fiction books" in case of **n-m** are shown in the charts as labels or legends.

5.3 Procedure

Participants were given a link to our survey. After consenting to participate, participants were given a metadata sheet describing the variables in the Amazon Books dataset. Participants had access to this metadata sheet throughout the experiment. After a brief introduction to the survey, participants viewed all sixteen comparison queries in random order on separate pages. For each comparison query, participants were given the four corresponding visualizations based on the cardinality of the query and were instructed to rank them in terms of how well they enable a viewer to make the intended comparison. Participants were told to assume that the viewer was not familiar with the dataset and there were no correct answers. At the end of the survey, participants reported demographic information, completed the self-report visual literacy test, and reported the amount of effort they put in completing the study.

5.4 Results for Each Comparison Utterance

We conducted a Friedman Rank Sum test for each comparison query using the PMCMRplus R package [50] to compare preference rankings for the four visualizations, with post-hoc pair-wise comparisons via Conover's test with Bonferroni's correction to determine the specific ranking differences (Table with complete analysis can be found in the supplementary material). Figure 5 shows the summary results.

Cardinality 1-1: Accounting for all combinations of concreteness, participants generally preferred (A) **simple bar charts** or (B) **adjacent unit charts** to help them make **1-1** comparisons. They preferred (C) horizontal unit charts and (D) grouped bar charts significantly less. There is a significant difference in user preferences for the EV-EA query (Friedman $\chi^2 = 57.28$, $p < 0.001$), the EV-IA query (Friedman $\chi^2 = 21.85$, $p < 0.001$), the IV-EA query (Friedman $\chi^2 = 80.97$, $p < 0.001$), as well as the IV-IA query (Friedman $\chi^2 = 18.29$, $p < 0.001$).

Cardinality 1-n: Accounting for all combinations of implicit and explicit data variables and data attributes, participants generally preferred (E) **multi bar charts** to help them make **1-n** comparisons. Participants preferred (G) dot plots and (H) scatterplots significantly less. There is a significant difference in user preferences for the EV-EA query (Friedman $\chi^2 = 71.42$, $p < 0.001$), the EV-IA query (Friedman $\chi^2 = 60.45$, $p < 0.001$), the IV-EA query (Friedman $\chi^2 = 91.51$, $p < 0.001$), as well as the IV-IA query (Friedman $\chi^2 = 88.51$, $p < 0.001$).

Cardinality n: Accounting for all combinations of implicit and explicit data variables and data attributes, participants generally preferred (I) **multi bar charts** to help them make **n** comparisons, similar to the **1-n** cardinality. Participants preferred (L) box plots the least, with (J) small multiples bar charts and (K) scatterplots in the middle of the pack. There is a significant difference in user preferences for the EV-EA query (Friedman $\chi^2 = 123.34$, $p < 0.001$), the EV-IA query (Friedman $\chi^2 = 123.34$, $p < 0.001$), as well as both the IV-EA and IV-IA queries (Friedman $\chi^2 = 113.83$, $p < 0.001$ for both queries).

Cardinality n-m: Accounting for all combinations of concreteness, participants equally preferred (M) **grouped bar charts**, (N) **small multiples bar charts**, and (O) **simple bar charts** to help them make **n-m** comparisons. Participants generally preferred (P) scatterplots the least. There is a significant difference in user preferences for the EV-EA query (Friedman $\chi^2 = 58.24$, $p < 0.001$), the EV-IA query (Friedman $\chi^2 = 69.36$, $p < 0.001$), the IV-EA query (Friedman $\chi^2 = 76.78$, $p < 0.001$), as well as the IV-IA query (Friedman $\chi^2 = 92.38$, $p < 0.001$).

5.5 Summary Findings

Results from Study 2 provide insights into which visualization arrangement may be preferred for a given type of comparison utterance from our design space. Overall, we found that the preference ranking of

1-1 comparisons

Across all levels of concreteness, the simple bar chart (A) was the most preferred chart type.

Value	Attribute	Comparison Query
Explicit	Explicit	Compare the <u>user rating</u> of <u>Looking for Alaska</u> and <u>Gone Girl</u>
Explicit	Implicit	Compare the <u>popularity</u> of <u>Looking for Alaska</u> and <u>The Litigators</u>
Implicit	Explicit	Compare the <u>user rating</u> of <u>a bestseller book in 2012</u> to <u>a bestseller book in 2019</u>
Implicit	Implicit	Compare the <u>popularity</u> of <u>a cheap book</u> to <u>an expensive book</u>

1-n comparisons

Across all levels of concreteness, the multi bar chart (E) was the most preferred chart type.

Value	Attribute	Comparison Query
Explicit	Explicit	Compare the <u>price</u> of <u>The Alchemist</u> to <u>all other fiction books</u>
Explicit	Implicit	Compare how <u>popular</u> <u>The Alchemist</u> is to <u>other fiction books</u>
Implicit	Explicit	Compare <u>a high-rated book</u> with <u>similar high-rated books</u> with respect to <u>price</u>
Implicit	Implicit	Compare <u>a cheap book</u> to <u>other cheap books</u> with respect to its <u>popularity</u>

n comparisons

Similar to the 1-n cardinality, the multi bar chart (I) was the most preferred chart type.

Value	Attribute	Comparison Query
Explicit	Explicit	Compare the <u>average price</u> across <u>all of J.K. Rowling's books</u>
Explicit	Implicit	Compare the <u>popularity</u> across <u>all J.K. Rowling's books</u>
Implicit	Explicit	Compare the <u>average price</u> across <u>all high-rated books in 2016</u>
Implicit	Implicit	Compare the <u>popularity</u> across <u>all expensive books</u>

n-m comparisons

All three of the bar charts were equally preferred: (M) grouped, (N) small multiples, and (O) simple.

Value	Attribute	Comparison Query
Explicit	Explicit	Compare the <u>price</u> of <u>all books written by J.K. Rowling</u> to those of <u>John Green</u>
Explicit	Implicit	Compare the <u>popularity</u> of <u>books written by J.K. Rowling</u> to those written by <u>John Green</u>
Implicit	Explicit	Compare the <u>price</u> of <u>high-rated fiction books</u> and <u>high-rated non-fiction books in 2016</u>
Implicit	Implicit	Compare the <u>popularity</u> of <u>high-rated fiction books</u> and <u>high-rated non-fiction books in 2016</u>

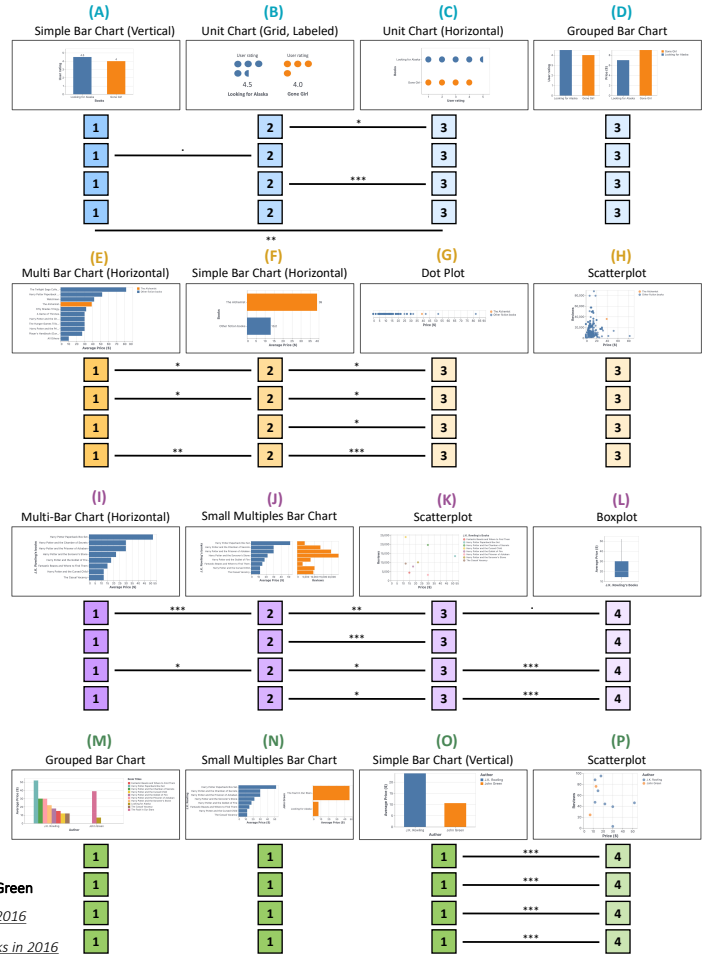


Fig. 5. Study 2 preference rankings for visualizations extracted based on Study 1 (Section 4.8). Each row represents a query with varying cardinalities and levels of concreteness (**explicit** or **implicit**). The boxed number represents the average preference ranks assigned to each visualization, based on a Friedman Rank Sum test. Connecting lines represent significant differences based on post-hoc comparisons. Three asterisks (***) indicate that the comparison p-value < 0.001. Two asterisks (**) represent p < 0.01. One asterisk (*) represents p < 0.05. One dot (.) represents p < 0.1.

charts within the four cardinalities (i.e., **1-1**, **1-n**, **n**, and **n-m**) was consistent, even taking into account the combinations of concreteness (Figure 5). We use the notation [W#] when referring to participants in these studies. We now discuss specific preferences and participant feedback for the various cardinalities of comparison utterances.

1-1 Comparisons: Viewers preferred the (A) **simple bar charts** without the need for unit charts or grouped bar charts: “The bar graphs are easier to compare than the bubbles.” [W18] and “The simple bar plot was easiest to understand and provided a perfect conveyance of the data.” [W33]. Other viewers had trouble interpreting unit charts; for example, “I have no idea what [the unit chart] is trying to say.” [W12].

1-n Comparisons: Viewers preferred (E) **multi-bar charts** sorted in descending order with the singleton value in the comparison utterance highlighted for easier relative judgment tasks. As a second option, people indicated a preference for (F) **a simple horizontal bar chart** comparing the singleton value to an aggregated bar of other entities: “The first one [multi-bar chart] is the best visual as it gives the most accurate ranking system, while the other one [simple horizontal bar chart] only shows averages, which could cause misrepresentation as it lumps all the books.” [W68]. (G) Dot plots and (H) scatterplots were both equally ranked lower in the rankings. W65 summarized that “The bar chart is easy to compare due to The Alchemist having a brighter color that stands out on the chart. The scatterplot is somewhat confusing due to the cluster of circles bunched together.”

n Comparisons: The top two user preferences were (I) **a multi-bar chart** and (J) **small multiples of simple bar charts** showing the entities

sorted by a common property such as the price. Participants ranked the (K) scatterplot lower than the bar charts with the (L) box plot ranked last. W5 commented that “The first two graphs [bar graphs] are just as easy to understand and the third graph [scatterplot] is a little less easy to understand and the last one [box plot] is objectively pretty terrible at illustrating the popularity across all of JK Rowling’s books.”

n-m Comparisons: There was a tie in preferences among the three types of bar charts: (M) **grouped**, (N) **small multiples**, and (O) **a simple bar chart**. (P) Scatterplots were consistently ranked last. Viewers liked to see the breakdown of the entities being compared rather than viewing aggregated data: “The first two choices breakdown the price by book, while the last two choices don’t have that breakdown.” [W2]. Several participants expressed difficulty interpreting the scatterplot for a comparison task; for example, W63 noted that “Scatterplots are difficult to read. The first option is cleanest and easiest to read to compare the cost of books by the author.” and W25 explained in detail that “My first choice showed at a quick glance the exact comparison between data sets - and number two was relatively easy as well as three. But number four, again... those dots!”

6 DISCUSSION

To summarize, the findings from both studies are consistent with that from prior work, where bar charts are generally conducive to visual comparisons [69]. However, we found interesting insights regarding the type of bar charts that were preferred. Feedback from participants showed that bars were easily comparable when visually aligned and spatially proximate. In terms of language, we observed a trend towards

requiring interactivity and annotations where implicit values/attributes were involved. Insightful trends were found regarding the general interpretation of vague modifiers and underspecified implicit values by the participants. Observations from these studies provide four key design implications for recommendation tools and NLIs to support comparisons during visual analysis.

6.1 Design Implications

Basic charts are reasonable responses for a variety of comparisons. The preference ranking consistently showed that bar charts are preferred for their simplicity in representing comparisons. Our findings suggest that for **1-1**, vertical bar charts and unit charts were preferred. For **1-n**, preference was towards horizontal bar charts. Horizontal and small multiple bar charts were popular for **n**; whereas for **n-m**, grouped, small multiple or simple bar charts were preferred. Many tools are already capable of creating basic charts such as sorted horizontal and vertical bar charts with support for highlighting a subset of bars for easier comparison. By extending the language parser in these tools to recognize comparison expressions, they can support a repertoire of comparison types without the need to generate bespoke chart types.

Include necessary information useful for the comparison task. Viewers ranked charts that contained only information relevant to the comparison higher than those that had extra or unnecessary information, an observation synergistic with previous NLI research for visual analysis tasks [33, 61]. When generating visualization responses, tools should ensure that they contain only information that is indicated in the comparison utterances. Participants explained their rationale for ranking charts with superfluous information lower than others: *"The bottom chart [small multiples bar chart] contains extraneous information that is not needed for the comparison sentence (price)."* [W6]. Similarly, for comparisons involving specific values, viewers preferred charts showing unaggregated data: *"Preferred graph gives comparative information. All of the others compare one book with an aggregate."* [W31].

Indicate how implicit entities are interpreted. Emulating previous NL systems [33, 59, 60], we map vague modifiers to more concrete representations. Viewers appreciate the importance of exposing the provenance of how implicit entities such as "cheap" and "best-selling" are mapped to specific data attributes in the visualization. W72 remarked that *"The first option [small multiples bar chart] allowed readers to gauge both price and review at the same time, while the other graphs did not show them at all. 'Expensive' books are subjective"*. In general, for vague modifiers, participants usually picked all the attributes from the dataset that somewhat matched the implicit value or attribute and mapped all of them to it (e.g., 'performance' as a combination of box-office revenue and IMDB ratings). Underspecified values were often interpreted as being above or below average.

Support user interaction and text with the visualizations. We found that users preferred interactivity (e.g., filter controls, changing the attribute, or modifying the data values), particularly for comparison utterances concerning implicit concepts. In addition, users preferred the inclusion of text describing how these implicit concepts were interpreted in the visualization responses. As shown in Table 1, IV-EA had the highest percentage of visualizations with interactive features, closely followed by IV-IA. IV-EA also had the highest percentage of visualizations with annotations, closely followed by EV-IA and IV-IA. In terms of cardinality, **n** had the highest percentage of visualizations with interactive features, followed by **1-n**. **n** also had the highest number of visualizations with annotations, followed by **n-m**.

6.2 Limitations and Future Directions

Our work specifically explores the interplay between language and visual representations for comparisons. However, we discuss some limitations of our work and identify promising future directions in this important area of research.

Explore additional design considerations for comparisons. The visualizations used in Study 2 varied across several design dimensions (e.g., arrangement, encodings, chart type, annotation), unlike prior

studies [69, 71]), where one design element (such as visual arrangement) was isolated and tested comprehensively. We considered the current approach because past work has shed light on how vast and complex the experimental space can be. To avoid a combinatorial explosion of experimental conditions, we hence adopted a design-and-then-validate approach to cover a wide range of visualization designs for multiple types of comparison utterances. The approach enabled us to produce immediately actionable guidelines for visualization recommender and NL systems. Future work should examine how interactivity, such as brushing, linking, dynamic filtering and scaling, and annotation techniques might facilitate different types of comparisons.

Explore the role of ethics for interpreting comparison utterances. With the impreciseness of language, we acknowledge that people may interpret implicit attributes and vague concepts in unique, subjective ways. There is a need to further explore the ethical issues surrounding how tools implementing comparison utterances can be transparent about their underlying assumptions and decision-making processes. Future work should specifically evaluate the ethical considerations for NLIs and recommendation systems [15, 74] for responsibly supporting analytical inquiry for comparisons.

Leverage additional metrics beyond user preferences. The visualization responses for comparison utterances were informed by a series of interviews that were subsequently validated by a user preference study. Future research should evaluate the effectiveness of these preferred visualizations by examining the speed and accuracy with which users can make the intended comparisons, or test the types of comparisons they afford via qualitative responses. Other research directions should explore how NLIs can take into account *implicit* feedback by users through refinement and repair of system choices for visual comparisons as well as query reformulations [62]. In addition, telemetry data collected from production systems and deployed visual analysis tools can be used as a training set for machine learning models to learn and improve user expectations over time. Other metrics involve *explicit* feedback through surveys and a like/dislike functionality built into NLIs could provide additional insight into the utility of the visualization response in supporting a comparison intent.

Extend comparisons to support exploratory data analysis. Our paper focuses on a question-answering form of interaction for supporting comparison intents. However, future research should explore how users express comparisons to support exploratory data analysis, focusing specifically on language pragmatics for follow-up inquiry during an analytical conversation. Such insights can also be useful to help inform the design of recommender tools for suggesting comparison utterances and views as part of a larger analytical workflow.

7 CONCLUSION

Visual comparisons are an important form of analytical workflows as people reason about and make sense of data. With the growing popularity of NLIs and recommendation systems, interpreting and visualizing the semantic nuances of comparison utterances can be challenging. In this paper, we explore a preliminary design space of NL comparison utterances covering four cardinalities of comparisons and addressing varying degrees of concreteness in these comparison intents. Through a series of interviews with 16 visualization novices and experts, we create a potential mapping between comparison utterances in our design space and different visualization techniques. By empirically validating this mapping in a preference ranking study with 77 participants, we provide design implications for interpreting the comparison language for a range of scenarios. We hope that our findings are a step towards addressing the complex interplay between language and visualization.

ACKNOWLEDGMENTS

The authors wish to thank Shubham Mishra and Nisarga Patil from UMass Amherst for helping with the coding process, the study participants for their feedback, and the reviewers for their thoughtful comments and suggestions.

REFERENCES

- [1] Microsoft Q & A. <https://powerbi.microsoft.com/en-us/documentation/powerbi-service-q-and-a/>. Accessed: 2022-08-04.
- [2] Notability. <https://notability.com/>. Accessed: 2022-03-30.
- [3] Tableau's Ask Data. <https://www.tableau.com/products/new-features/ask-data>. Accessed: 2022-08-04.
- [4] ThoughtSpot. <http://www.thoughtspot.com/>. Accessed: 2022-08-04.
- [5] Google Jamboard. <https://support.google.com/jamboard>, 2022. Accessed: 2022-03-30.
- [6] IBM Watson Analytics. <http://www.ibm.com/analytics/watson-analytics/>, 2022. Accessed: 2022-03-30.
- [7] B. Alper, B. Bach, N. Henry Riche, T. Isenberg, and J.-D. Fekete. Weighted graph comparison techniques for brain connectivity analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, p. 483–492. Association for Computing Machinery, New York, NY, USA, 2013. doi: 10.1145/2470654.2470724
- [8] D. Archambault, H. Purchase, and B. Pinaud. Animation, small multiples, and the effect of mental map preservation in dynamic graphs. *IEEE Transactions on Visualization and Computer Graphics*, 17(4):539–552, 2011. doi: 10.1109/TVCG.2010.78
- [9] J. Bacsakai-Atkari. The syntax of comparative constructions : operators, ellipsis phenomena and functional left peripheries. 2014.
- [10] O. Bakhshandeh and J. Allen. Semantic framework for comparison structures in natural language. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 993–1002. Association for Computational Linguistics, Lisbon, Portugal, Sept. 2015. doi: 10.18653/v1/D15-1115
- [11] S. Bansal. Netflix movies and tv shows. <https://www.kaggle.com/shivamb/netflix-shows>, Sep 2021. Accessed: 2022-08-04.
- [12] M. Bierwisch. The semantics of gradation. 1989.
- [13] A. Burns, C. Xiong, S. Franconeri, A. Cairo, and N. Mahyar. How to evaluate data visualizations across different levels of understanding. In *2020 IEEE Workshop on Evaluation and Beyond-Methodological Approaches to Visualization (BELIV)*, pp. 19–28. IEEE, 2020.
- [14] H. H. Clark and W. G. Chase. On the process of comparing sentences against pictures. *Cognitive psychology*, 3(3):472–517, 1972.
- [15] M. Correll. Ethical dimensions of visualization research. 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: 10.1145/3290605.3300418
- [16] M. CRESSWELL. The semantics of degree. In B. H. PARTEE, ed., *Montague Grammar*, pp. 261–292. Academic Press, 1976. doi: 10.1016/B978-0-12-545850-4.50015-7
- [17] K. Dhamdhere, K. S. McCurley, R. Nahmias, M. Sundararajan, and Q. Yan. Analyza: Exploring data with conversation. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, IUI 2017, pp. 493–504, 2017.
- [18] J. Duncan and G. W. Humphreys. Visual search and stimulus similarity. *Psychological review*, 96(3):433, 1989.
- [19] S. Franconeri. *The Nature and Status of Visual Resources*. Oxford Library of Psychology. Oxford University Press, 2013.
- [20] S. Franconeri. Three perceptual tools for seeing and understanding visualized data. *Current Directions in Psychological Science*, 30:096372142110095, 09 2021. doi: 10.1177/09637214211009512
- [21] C. Friedman. A general computational treatment of the comparative. In *27th Annual Meeting of the Association for Computational Linguistics*, pp. 161–168. Association for Computational Linguistics, Vancouver, British Columbia, Canada, June 1989. doi: 10.3115/981623.981643
- [22] T. Gao, M. Dontcheva, E. Adar, Z. Liu, and K. G. Karahalios. Datatone: Managing ambiguity in natural language interfaces for data visualization. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software Technology*, UIST 2015, pp. 489–500. ACM, New York, NY, USA, 2015.
- [23] R. Garcia-Retamero and E. T. Cokely. Communicating health risks with visual aids. *Current Directions in Psychological Science*, 22(5):392–399, 2013.
- [24] R. Garcia-Retamero and M. Galesic. Communicating treatment risk reduction to people with low numeracy skills: a cross-cultural comparison. *American journal of public health*, 99(12):2196–2202, 2009.
- [25] O. C. G. Gelo, D. Braakmann, and G. Benetka. Quantitative and qualitative research: Beyond the debate. *Integrative Psychological and Behavioral Science*, 42:266–290, 2008.
- [26] M. Gleicher. Considerations for visualizing comparison. *IEEE Transactions on Visualization and Computer Graphics*, 24:413–423, 2018.
- [27] 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
- [28] M. Gleicher, M. Correll, C. Nothelfer, and S. Franconeri. Perception of average value in multiclass scatterplots. *IEEE transactions on visualization and computer graphics*, 19(12):2316–2325, 2013.
- [29] L. R. Gleitman, D. January, R. Nappa, and J. C. Trueswell. On the give and take between event apprehension and utterance formulation. *Journal of memory and language*, 57(4):544–569, 2007.
- [30] M. Graham and J. Kennedy. Exploring multiple trees through dag representations. *IEEE transactions on visualization and computer graphics*, 13:1294–301, 11 2007. doi: 10.1109/TVCG.2007.70556
- [31] C. Hamann, I. Heim, D. Lewis, P. Seuren, and W. Sternefeld. Comparing semantic theories of comparison. *Journal of Semantics*, 2005.
- [32] S. T. Hawley, B. Zikmund-Fisher, P. Ubel, A. Jancovic, T. Lucas, and A. Fagerlin. The impact of the format of graphical presentation on health-related knowledge and treatment choices. *Patient education and counseling*, 73(3):448–455, 2008.
- [33] M. Hearst, M. Tory, and V. Setlur. Toward interface defaults for vague modifiers in natural language interfaces for visual analysis. In *2019 IEEE Visualization Conference (VIS)*, pp. 21–25, 2019. doi: 10.1109/VISUAL.2019.8933569
- [34] N. Jardine, B. D. Ondov, N. Elmqvist, and S. Franconeri. The perceptual proxies of visual comparison. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):1012–1021, 2020. doi: 10.1109/TVCG.2019.2934786
- [35] M. Kay, T. Kola, J. R. Hullman, and S. A. Munson. When (ish) is my bus? user-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 2016 CHI*, pp. 5092–5103, 2016.
- [36] R. Keefe. Vagueness: Supervaluationism. *Philosophy Compass*, 3(2):315–324, 2008. doi: 10.1111/j.1747-9991.2008.00124.x
- [37] C. Kennedy. Projecting the adjective: The syntax and semantics of gradability and comparison. 01 1997.
- [38] C. Kennedy. Comparatives, semantics of. *Encyclopedia of Language & Linguistics*, 08 2004. doi: 10.1016/B0-08-044854-2/01028-2
- [39] W. Kessler and J. Kuhn. A corpus of comparisons in product reviews. 05 2014.
- [40] E. Klein. A semantics for positive and comparative adjectives. *Linguistics and Philosophy*, 4(1):1–45, 1980. doi: 10.1007/BF00351812
- [41] G. H. Kranz. Transcribe, 1986.
- [42] A. Larsen and C. Bundesen. Size scaling in visual pattern recognition. *Journal of experimental psychology. Human perception and performance*, 4 1:1–20, 1978.
- [43] A. Larsen and C. Bundesen. Effects of spatial separation in visual pattern matching: Evidence on the role of mental translation. *Journal of experimental psychology. Human perception and performance*, 24:719–31, 07 1998. doi: 10.1037/0096-1523.24.3.719
- [44] P. Law, L. Y.-H. Lo, A. Endert, J. Stasko, and H. Qu. Causal perception in question-answering systems. *ArXiv*, abs/2012.14477, 2020.
- [45] 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, 2016.
- [46] M. A. Livingston and J. W. Decker. Evaluation of trend localization with multi-variate visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2053–2062, 2011. doi: 10.1109/TVCG.2011.194
- [47] C. Nothelfer and S. Franconeri. Measures of the benefit of direct encoding of data deltas for data pair relation perception. *IEEE transactions on visualization and computer graphics*, 26(1):311–320, 2019.
- [48] H.-G. Pagendarm and F. Post. Comparative visualization - approaches and examples. 01 1995.
- [49] S. Palan and C. Schitter. Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27, 2018.
- [50] T. Pohlert and M. T. Pohlert. Package ‘pmcmrplus’. *R Foundation for Statistical Computing*, Vienna, Austria, 2018.
- [51] I. Qualtrics. Qualtrics. Provo, UT, USA, 2013.
- [52] R. A. Rensink. Change detection. *Annual Review of Psychology*, 53(1):245–277, 2002. PMID: 11752486. doi: 10.1146/annurev.psych.53.100901.135125

- [53] Rgriffin. 120 years of olympic history: Athletes and results. <https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results>, Jun 2018. Accessed: 2022-08-04.
- [54] J. Roth and S. Franconeri. Asymmetric coding of categorical spatial relations in both language and vision. *Frontiers in psychology*, 3:464, 2012.
- [55] S. Saalu. Amazon top 50 bestselling books 2009-2019. <https://www.kaggle.com/sootersaalu/amazon-top-50-bestselling-books-2009-2019>, Oct 2020. Accessed: 2022-08-04.
- [56] E. Sapir. Grading, a study in semantics. *Philosophy of Science*, 11:93–116, 1944.
- [57] A. Satyanarayan, D. Moritz, K. Wongsuphasawat, and J. Heer. Vega-lite: A grammar of interactive graphics. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):341–350, jan 2017. doi: 10.1109/TVCG.2016.2599030
- [58] R. Schwarzschild and K. Wilkinson. Quantifiers in comparatives: A semantics of degree based on intervals. *Natural Language Semantics*, 10:1–41, 2002.
- [59] V. Setlur, S. E. Battersby, M. Tory, R. Gossweiler, and A. X. Chang. Eviza: A natural language interface for visual analysis. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, UIST '16, p. 365–377. Association for Computing Machinery, New York, NY, USA, 2016. doi: 10.1145/2984511.2984588
- [60] V. Setlur and A. Kumar. Sentifiers: Interpreting vague intent modifiers in visual analysis using word co-occurrence and sentiment analysis. In *2020 IEEE Visualization Conference (VIS)*, pp. 216–220, 2020. doi: 10.1109/VIS47514.2020.00050
- [61] V. Setlur and M. Tory. How do you converse with an analytical chatbot? revisiting gricean maxims for designing analytical conversational behavior. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM, 2022.
- [62] V. Setlur, M. Tory, and A. Djalali. Inferencing underspecified natural language utterances in visual analysis. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, p. 40–51. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3301275.3302270
- [63] P. Shah and E. G. Freedman. Bar and line graph comprehension: An interaction of top-down and bottom-up processes. *Topics in cognitive science*, 3(3):560–578, 2011.
- [64] S. Shapiro. *Vagueness in Context*. Oxford University Press, 2006.
- [65] A. Srinivasan, M. Brehmer, B. Lee, and S. M. Drucker. What's the difference? evaluating variations of multi-series bar charts for visual comparison tasks. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2018.
- [66] D. A. Szafir. The good, the bad, and the biased: five ways visualizations can mislead (and how to fix them). *interactions*, 25(4):26–33, 2018.
- [67] A. R. Tait, T. Voepel-Lewis, B. J. Zikmund-Fisher, and A. Fagerlin. Presenting research risks and benefits to parents: does format matter? *Anesthesia and analgesia*, 111(3):718, 2010.
- [68] E. R. Tufte. *Envisioning information*. Graphics Press, Cheshire, Conn., 1990.
- [69] 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 and Computer Graphics*, 28(1):955–965, 2021.
- [70] C. Xiong, J. Shapiro, J. Hullman, and S. Franconeri. Illusion of causality in visualized data. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):853–862, 2019.
- [71] C. Xiong, C. Stokes, and S. Franconeri. Visual salience and grouping cues guide relation perception in visual data displays.
- [72] C. Y. Y. Xiong. *Perceptual and Cognitive Affordances of Data Visualizations*. PhD thesis, Northwestern University, 2021.
- [73] J. Zacks and B. Tversky. Bars and lines: A study of graphic communication. *Memory & Cognition*, 27(6):1073–1079, 1999.
- [74] Z. Zeng, P. Moh, F. Du, J. Hoffswell, T. Y. Lee, S. Malik, E. Koh, and L. Battle. An evaluation-focused framework for visualization recommendation algorithms. *IEEE Transactions on Visualization and Computer Graphics*, 28:346–356, 2022.