Characterizing Visualization Perception with Psychological Phenomena: Uncovering the Role of *Subitizing* in Data Visualization

Arran Zeyu Wang, Ghulam Jilani Quadri, Mengyuan Zhu, Chin Tseng, and Danielle Albers Szafir

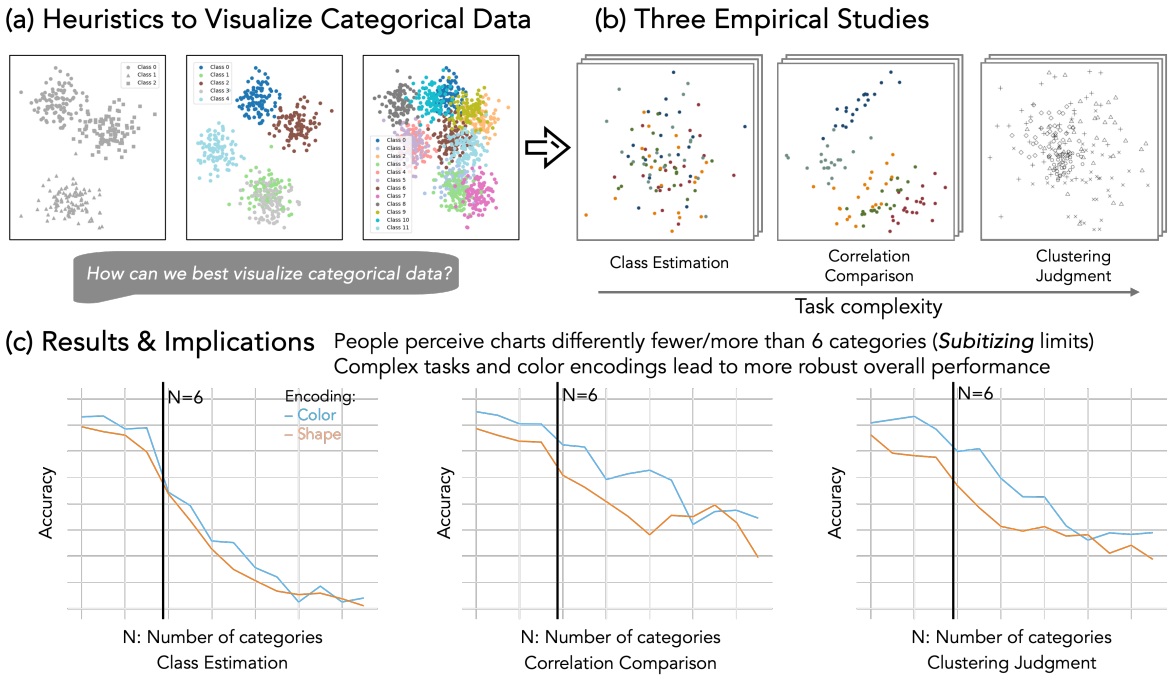

Fig. 1: We explored the potential of using psychological theories to validate visualization design heuristics by assessing a psychological phenomena called *subitizing* in categorical data visualization. (a) There are many existing heuristics when visualizing categorical data, varying from limiting designs to 5 to 12 categories. (b) We designed three empirical studies to characterize performance from 2–15 categories with color and shape encodings in visualizations with three experimental tasks: class estimation, correlation comparison, and clustering judgments. (c) Our results provide empirical evidence that people perceive charts differently before and after 6 categories with significant performance reduction. We term this threshold as the *Subitizing* limit, meaning the subitizing phenomenon can help people immediately perceive information from charts with fewer than 6 categories. More complex aggregation-based tasks and alternative encodings influence the effect of subitizing.

Abstract— Understanding how people perceive visualizations is crucial for designing effective visual data representations; however, many heuristic design guidelines are derived from specific tasks or visualization types, without considering the constraints or conditions under which those guidelines hold. In this work, we aimed to assess existing design heuristics for categorical visualization using well-established psychological knowledge. Specifically, we examine the impact of the *subitizing* phenomenon in cognitive psychology—people's ability to automatically recognize a small set of objects instantly without counting—in data visualizations. We conducted three experiments with multi-class scatterplots—between 2 and 15 classes with varying design choices—across three different tasks—class estimation, correlation comparison, and clustering judgments—to understand how performance changes as the number of classes (and therefore set size) increases. Our results indicate if the category number is smaller than six, people tend to perform well at all tasks, providing empirical evidence of *subitizing* in visualization. When category numbers increased, performance fell, with the magnitude of the performance change depending on task and encoding. Our study bridges the gap between heuristic guidelines and empirical evidence by applying well-established psychological theories, suggesting future opportunities for using psychological theories and constructs to characterize visualization perception.

Index Terms—Visualization Perception, Psychology, Subitizing, Fechner's Law, Dual-System Theory, Categorical Data, Color, Shape

1 Introduction

In visual data communication, effective visualizations enable users to interpret complex information quickly and accurately [19]. As the size and complexity of a dataset grows, designing visualizations that are both efficient and accurate becomes increasingly important, but also increasingly challenging [19,69,71]. For example, most complex datasets include both numeric and categorical data. While past studies have extensively investigated methods for representing traditional numeric

- *Arran Zeyu Wang, Mengyuan Zhu, Chin Tseng, and Danielle Albers Szafir are with the University of North Carolina at Chapel Hill. E-mail: {zeyuwang, chint, gisellez, danielle.szafir}@cs.unc.edu*
- *Ghulam Jilani Quadri is with the University of Oklahoma and the University of North Carolina at Chapel Hill. E-mail: quadri@ou.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

data [7, 59], guidelines for categorical data remain largely heuristic and relatively sparse [45, 74].

One heuristic for categorical data is to encode no more than six categories, as stated in Adobe Design Guidelines [2] and elsewhere without supporting evidence or a clear origin [28]. When the category numbers in a dataset increase, the complexity of the visualization increases as well, often leading to a significant descent of the *perception effectiveness*[1] in visualizations [30, 74]. However, modern datasets frequently contain far more than six categories, forcing designers to either reduce the granularity of categorical data by grouping related categories [4, 39], repeat encodings or facets over multiple categories [15, 32], or choose categorical encoding designs that violate this heuristic like extended color ramps or palettes [67, 86]. More recent studies suggest that, for certain tasks, people are still able to reliably reason over more than six distinctly encoded categories [74]; however, these studies also reveal a steep performance decline near six categories. In this paper, we explore the six-category heuristic over a range of visualization tasks to determine its generalizability and infer potential perceptual grounding for this common categorical design guideline.

The decline in performance after six categories is not just a matter of aesthetics [47]; it directly impacts the user's ability to make accurate and efficient judgments based on the visualized data [74]. The heuristic itself is not always employed consistently, with past work recommending a wide range of thresholds, including six [2], five to seven [48], eight [85], six to nine [56], ten [21], eleven [10], or six to twelve categories [45]. One likely source of variance in these recommendations is the lack of theoretical or empirical basis. We approach this heuristic through the lens of psychological theories and constructs. In cognitive psychology, the phenomena of *subitizing* allows people to instantly and accurately recognize and quantify small numbers of objects—typically around four to six—without counting [36]. For more than six objects, individuals must rely on counting or estimation, leading to decreased accuracy and increased cognitive load. Past work has speculated that subitizing may play a role in categorical data interpretation in visualizations, posing challenges to people's perception effectiveness [30, 74]. However, performance for categorical encodings tends to decrease non-linearly with the number of categories, potentially influencing the range of prescribed categories in past heuristics. These variations may be explained by another psychological theory, *Fechner's Law*, which states that as the *intensity* of a stimulus (in the context of visualization, the "*intensity*" corresponds to the amount of information, such as point and category number, being displayed, see Figure 2) increases, the corresponding perceived difference increases logarithmically [18]. We specifically focused on the finding that, as the number of categories grows, the perceptual difference between them diminishes logarithmically, making it harder for people to distinguish between categories [25]. This may mean that the most dramatic performance decreases happen around the subitizing threshold and level off to stable perceived differences above a secondary threshold.

In this work, we investigated the influence of subitizing in data visualization by empirically examining how increasing the number of categories affects different data analysis tasks. We conducted a series of three crowdsourced experiments that tested people's abilities to comprehend visual information across categories using multi-class scatterplots with varying numbers of categories (2 to 15), categorical encodings (color and shape), and visual complexities (point numbers) on three tasks: class estimation, correlation comparison, and clustering judgments (see Figure 1 (b)). Our results indicated that when scatterplots encode fewer than six categories, accuracy remains stable and relatively high across all tasks. This finding aligns with the concept of subitizing, suggesting that the human visual system is highly efficient at processing small sets of categories. As the number of categories increases beyond this threshold, perceptual accuracy first significantly declines and then converges to a more stable performance, aligned with Fechner's Law. However, aggregation-based tasks and color encodings remained more robust to increasing category numbers, suggesting additional visual mechanisms may be at play when engaged in these tasks. Figure 1 (c) summarizes the main implications. By grounding our results in established subitizing phenomena in cognitive science, we not only contribute to the field of data visualization but also demonstrate the potential for leveraging cognitive psychology to inform and improve visualization design. This grounding enables us to use our results to propose more generalizable and empirically-grounded guidelines that can help create more effective visualizations for categorical data.

The main contributions of this paper include:

- **Empirical investigation of subitizing limits in visualization and design guidelines.** We provide empirical evidence that the perceptual accuracy in visualizations significantly declines when the number of categories reaches six, corresponding to the psychological phenomenon of subitizing, and allowing for more generalizable, empirically-grounded guidance on categorical encoding based on the number of categories present in a dataset.
- **Characterizing performance patterns for categorical encodings within and beyond subitizing limits.** We find that performance remains relatively stable for fewer than six categories, consistent with subitizing, and decreases asymptotically for more than six categories, consistent with people requiring more complex cognitive processes to comprehend the visualized data and matching Fechner's Law.
- **Examining visualization perception through the lens of psychological theories and constructs.** Our results, as a whole, further show how widely established psychological theories and constructs can be used to measure the capabilities and performance for perception and cognition in data visualization from a continuum of low-level, relatively invariant to high-level, relatively variant psychological phenomena and ground the generalizability of visualization guidelines.

Providing theoretical support and empirically grounded evidence for established heuristics can be crucial for visualization research and practice [41]. Rather than establish new guidance, we aim to empirically deconstruct the underlying graphical perception process behind existing heuristics in categorical data visualization to better understand their application in practice and provide a research template for examining other design heuristics, responding to the call for better understanding the mechanisms behind how people accomplish different tasks [40].

2 Background

Research on categorical perception in visualization emphasizes the importance of visual encodings like color and shape for effective data interpretation. These studies are often tightly integrated with studies in psychology, particularly from cognitive and vision science, that highlight people's ability to recognize and perceive visual quantities. Our work builds on these foundations and provides new insights into design choices for visualizing categorical data by empirically testing how robustly people conduct high-level tasks in multiclass scatterplots.

2.1 Scatterplots for Data Communication

Scatterplots are an intuitive and widely-used visualization for bivariate quantitative data [19, 61]. They are also one of the most studied visualizations, with experiments showing how different scatterplot designs support assessing trends [14], correlation perception [31], causal inference [80], clustering [35], and detecting outliers [62]. Their visual simplicity and ability to support a range of tasks make them a common stimulus for visualization research to provide insights into general paradigms and implications of visualization design, similar to a "fruit fly" in biology [54].

Recent studies have emphasized scatterplots' effectiveness in communicating categorical information [23, 74]. For example, scatterplot designs are often used to analyze class structures in dimensionally-reduced data [77]. Categorical data visualization is a fundamental aspect of data representation, playing a crucial role in how users interpret and make decisions based on complex information [45]. Repre-

[1] In the context of this study, "perception effectiveness" is operationalized as the accuracy with which people can extract specific information about categorical data presented in multiclass scatterplots within a given time constraint.

senting categories differs from traditional numeric encoding as people must often select for and then compare patterns across different categories. As the number of categories grows, people may struggle to distinguish between them due to factors like reduced discriminability and increased clutter, leading to errors in interpretation and perceptual judgments [30, 74].

Scatterplots typically delineate categorical data using mark encodings, such as shape [11], size [33], texture [43], or color [64]. Encoding choice can significantly impact the ease with which people can interpret and analyze data. Color and shape are the most commonly used encodings for categorical data [19, 83].

Color is a common encoding choice for categorical data due to its intuitive perceptual qualities: color provides a robust perceptual cue for grouping related objects [65]. Hue, lightness, and perceptual distances have a significant effect on the efficiency of categorical color palettes in categorical visualizations [26, 74]. Color schemes specifically designed for categorical data significantly outperform sequential and diverging palettes, further emphasizing the importance of using distinct and easily distinguishable colors [75]. However, as the number of categories increases, the perceptual distinctiveness of colors diminishes, leading to reduced accuracy [74, 75]. Common heuristics for color encoding recommend using no more than six colors to represent data [2]. However, as with many common design heuristics [40], we lack empirical and mechanistic insight into these approaches to inform visualization design and to understand when, where, and how these guidelines and approaches should be applied. For example, alternative design strategies often leverage manipulations of lightness or saturation to successfully extend these palettes beyond six categories [28].

Shape is another widely-used encoding for categorical data, offering a complementary approach to color. Empirical evidence shows that the choice of shape palette also affects encoding effectiveness. For example, closed shapes may be more effective when representing two-class scatterplots compared to open shapes or mixed closed-open shape pairs [11]. Shape can be particularly useful when combined with other encodings, such as color or size, to represent multiple dimensions of data simultaneously [66]. However, Tseng et al. [76] found that when visualizing complex categorical data, shape encodings may be less effective overall than color and that shape's effectiveness may also be limited by the number of distinct shapes that can be easily recognized and differentiated by people [76]. Unlike color, we lack known heuristics for shape palette design, especially with respect to the number of potential categories.

For both color and shape, past studies show that patterns in effectiveness vary non-linearly as a function of the number of categories; however, these studies focus on averaging tasks and provide limited insight into the broader implications of this performance. This study investigates patterns in categorical data analysis in scatterplots across a wider range of tasks. Our objective is to develop more robust design guidelines and formulate theoretical hypotheses about task performance, thereby informing more effective design strategies.

2.2 Psychology in Visualization Research

Data visualization has increasingly drawn on principles from psychology to inform design practices [6, 17, 71]. Many theories, phenomena, and constructs in psychological research provide valuable insights into how people perceive and process visualized quantities that can guide more effective design practice and ground past heuristics through translational research [72]. These laws describe the relationship between the physical properties of stimuli (i.e., a visualization) and corresponding percept (i.e., an inference about the data), offering a framework for understanding the limits of human perception that, when applied to visualization, offers novel means for grounding and understanding the generalizability of different design guidance. Many cognitive or perceptual effects have been utilized by visualization research in recent years, such as the Dunning-Kruger effect [13] and inattentional blindness [9]. Those effects can help understand the perceptual and cognitive limits of visual representations and can improve the efficiency of visual data communication as a result [19, 71].

For example, psychological phenomena such as *subitizing*—our ability to rapidly process a small (magic) number of objects [36, 44]—could offer a theoretical basis for understanding why people may struggle with visualizations that involve many categories. Subitizing is typically effective for quantities of up to four or five items. Beyond this range, individuals rely on counting, which is slower and more error-prone [49]. Further, the relationship between grouping cues and subitizing, alternatively known as groupitizing [5], indicates that subitizing may play a role in categorical visualization tasks, where people draw conclusions about sets of marks rather than individual marks. The effectiveness of subitizing for less than six discrete items (or sets of items) may, in part, help account for the recommendations around encoding six categories in visualization design guidelines [2]. Past studies have speculated that subitizing may play a role in visualization interpretation such as in visual search [30], reading pictographs [29], and estimating means [74]. When people engage in subitizing versus counting or alternative comparative perceptual mechanisms may have significant implications for how we choose to construct categorical visualizations. If we are biologically adapted to reason over six or fewer categories, we may need different design approaches for datasets with smaller category numbers than for those with larger category numbers.

While subitizing offers a hypothetical threshold for understanding design, Fechner's Law [18, 31, 37] may help further explain performance patterns in larger numbers of categories. The law posits that the perceived intensity of a stimulus grows logarithmically as its physical magnitude increases. In the context of data visualization, this means that as the number of categories in a visualization increases, the perceptual difference between categories diminishes, making it harder for users to distinguish between them. Figure 2 illustrates the logarithmic differences of human perception in visualization using Fechner's Law. The difference in the number of dots in the scatterplots is the same in each pair, from 10 to 20 in (a) and from 110 to 120 in (b). However, the perceived difference in (a) is significantly higher than in (b).

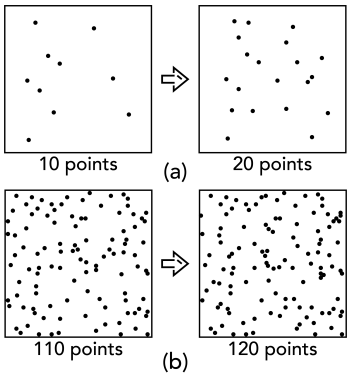

Fig. 2: Illustrating Fechner's Law with scatterplots. Charts in (a) and (b) have the same visual intensity (point number) difference, but the perceived difference in (a) is much stronger.

This principle can explain why visualizations with a large number of categories often result in decreased accuracy and increased cognitive load, but that this decrease tends to become more gradual as the number of categories increases, even if the elements of a palette remain perceptually discriminable as in many design strategies for integrating larger numbers of colors into a palette [28] or in randomly sampled palettes [75].

Our study builds on these perceptual and cognitive foundations by empirically testing how well people interpret categorical data visualizations, providing new insights into how to more effectively visualize categorical data through the theoretical lens offered by these mechanisms. The complexity of data visualizations compared to conventional perceptual studies prevents us from drawing firm causal connections between perceptual mechanisms and visualization design. We instead examine whether these mechanisms help us better model performance under different design conditions to ground actionable design guidance, understand how existing heuristics may generalize (or fail to generalize) to a range of applications, and offer translational research hypotheses for future work at the intersection of visualization and vision science.

3 Experimental Overview

We conducted three experiments to understand how well people perform several tasks when visualizing data with different numbers of categories. Our objective is to capture patterns of performance across varying numbers of categories and use these patterns to better understand effective design practices through the lens of perceptual models.

3.1 Task Selection

We assess the effect of category number on visualization task performance using scatterplots as they are one of the most representative visualization types [22, 45, 50, 54] and well-studied for categorical data [23, 61, 74]. Scatterplot visualization tasks can be categorized into three levels: object-centric, browsing, and aggregate [61]. Since object-centric tasks focus on searching for specific objects rather than reasoning over sets of objects, we focused on browsing and aggregate tasks. We selected three representative tasks—a browsing-based class estimation task, an aggregate-based correlation comparison task, and an aggregate-based clustering task— to assess the impact of categories across task complexities (see Figure 1 (b)):

Class estimation requires people to estimate the total number of classes that appeared in a scatterplot. People must browse the whole plot to identify and count the number of classes [61]. This class estimation task correlates to the long-lasting number estimation task for evaluating subitizing in psychology [55], providing a task that may heavily leverage subitizing. Tasks correlated with those from cognitive science communities, like visual search [27, 30] or centroid estimation [33], help provide translational insight into how people work with visualized data by facilitating connections between the two bodies of literature.

Correlation comparison requires people to choose the class with the highest correlation. Comparing correlations is a traditional task in data visualization that has been extensively studied in single-class scatterplots [31, 37, 53]. This task requires people to aggregate data to identify correlations and make comparisons between classes [61]. However, increasing the number of categories encoded in the scatterplot reduces people's correlation judgment accuracy [76]. As hypothesized in past work, we anticipate that this decrease may be in part due to a reliance on subitizing in conjunction with other visual mechanisms to perform correlation comparisons.

Clustering judgment requires people to find a class that is the most tightly clustered among all classes. Clustering judgment requires people to visually aggregate classes across the distribution [61] and then compare the resulting classes. Clustering has significant implications for designing visual quality measures [1, 35, 82]. However, people's comparison accuracy for clumpiness (higher clumpiness means more tightly clustered [84]) is significantly reduced as the number of categories increases [82], again correlating with the potential use of subitizing in selecting and characterizing individual classes.

3.2 Overall Hypotheses

Based on the three task designs and insights from prior research, we hypothesize:

H1: Accuracy will remain stable for fewer than six categories.

The first hypothesis pertains directly to how subitizing impacts people's abilities to work with certain numbers of categories [55]. Previous studies have reported preliminary insights on unusual performance drops at six categories in visual search [30] and mean judgment [74] tasks. Subitizing supports rapid, stable perception for four to six objects or sets of objects [36]. Therefore, we anticipate the limits of subitizing may be within this range for visualizations, meaning people can accurately perceive and interpret categorical visualizations for six or fewer categories and that performance will decline sharply when the number of categories reaches six as people may rely on additional perceptual mechanisms to track and enumerate larger numbers of categories.

H2: Accuracy will drop significantly for more than six categories.

Fechner's Law states that the quality of a percept changes logarithmically with an increase in cognitive load [18]. We anticipate the cognitive load required to interpret categorical visualizations increases disproportionately as the number of categories exceeds six (the anticipated limit of subitizing), leading to significant differences in distribution for accuracy and efficiency with more than six categories. More specifically, for a one-category change, Fechner's Law suggests that the decrease in performance may be more severe for category numbers near six (e.g., seven and eight) than for an equivalent step on higher numbers of categories (e.g., 12 and 13).

H3: The impact of the number of categories on perceptual accuracy varies depending on the specific visualization task.

Different tasks may require different cognitive processes [78] and analysis targets [61]. We anticipate tasks that require aggregation, such as clustering, may be more robust to increasing the number of categories than tasks that require browsing, like class estimation [12]. These aggregation-based tasks are more complex and likely rely on a larger set of perceptual mechanisms to achieve [3, 72], meaning that the ability to leverage subitizing to select for and reason over classes plays a smaller role in overall performance.

H4: Different encoding types may influence task performance as the number of categories increases.

Shape has proven a more complex and hard-to-design visual encoding channel compared to color in categorical perception [76]. These differences influence people's abilities to distinguish categories and influence overall performance as a result. Similarly, we anticipate performance will be more robust on color-coded scatterplots than on shape-coded ones.

H5: Increasing visual complexity in categorical visualizations exacerbates the decline in perceptual accuracy as the number of categories exceeds six.

More complexity in visualizations leads to poor perception performance, such as when points become overdrawn [43]. Increasing visualization complexity generally degrades overall performance. We anticipate an interaction between other aspects of visual complexity and the number of categories, with increases in both factors making it harder for people to accurately perceive and interpret data.

4 Experimental Design

We designed three experiments to test our hypotheses with each testing a different task. As these experiments leverage a number of common elements, we describe the experiments here and discuss their respective results in Section 5. The studies have been approved by [Redacted] Institutional Review Board.

4.1 Experiment One: Class Estimation

This experiment asked participants to estimate the total number of classes present in a series of scatterplots, similar to methods employed in psychology experiments studying subitizing [55]. Scatterplots varied in their encoding type (shape or color, between-participants), response duration (3 or 10 seconds, between-participants), complexity (low, medium, or high; within-participants), and number of categories ($N = 2 - -15$, within-participants). Subitizing directly assists numerosity estimation, which is the core component of this task. Therefore, we anticipate that performance in this task will be well-modeled by subitizing limits, with people reliably able to identify six or fewer categories and performance degrading significantly with more than six categories.

While people might eventually count categories accurately with unlimited time, visualizations are designed to minimize cognitive load and facilitate rapid comprehension across a range of tasks [45, 71]. Many foundational tasks in visualization rely on a gist developed very quickly, followed by a more detailed analysis of target tasks, which leads to the importance of *the speed of perception* [58], more specifically, class estimation is one of those gist tasks: people get a rough sense of the number of classes and then dig into the classes sequentially for comparison, characterization, and other analyses. While both subitizing and counting mechanisms have the potential to be highly accurate, counting takes significantly longer time [49] and subitizing operates significantly faster [63]. As a result, typical vision science studies investigate psychological phenomena like subitizing using very short time limits ranging from 500 ms to over 1500 ms [55, 79]. However, visualizations tend to operate over longer time scales and crowdsourcing platforms face potential challenges in deploying time-constrained studies, such as differences in data loading and server connection times. To account for these differences, we measure performance over two different time limits: three seconds and ten seconds. We implement the time limits as a between-subjects variable to reduce potential priming and learning effects (e.g., avoid people confusing a three-second trial for a ten-second one). These time limits were determined in piloting.

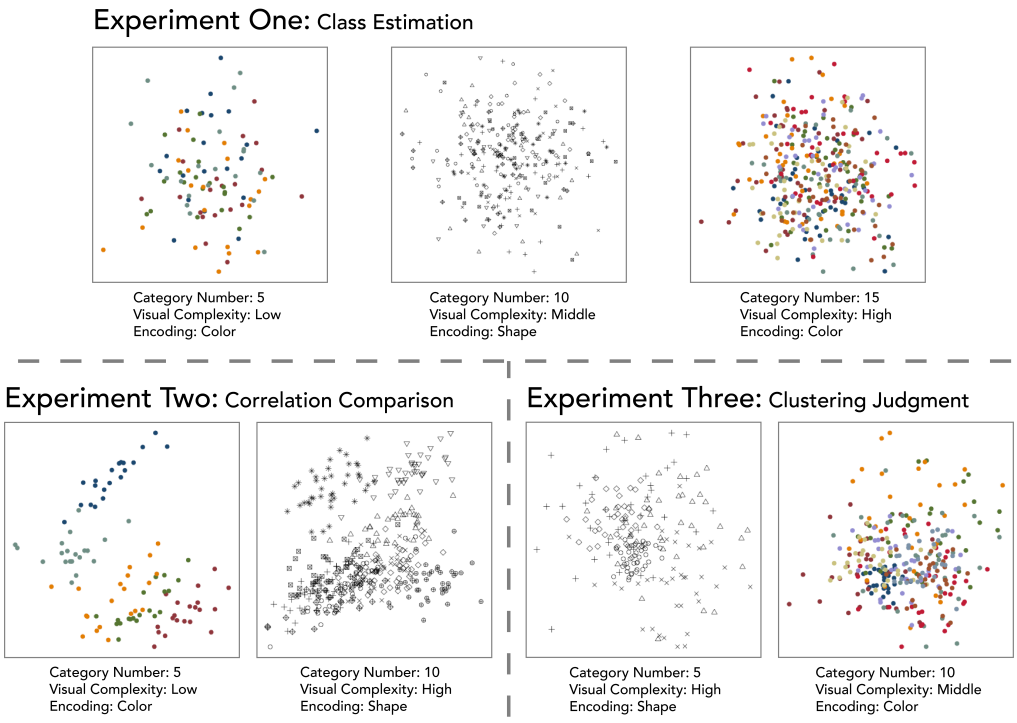

Fig. 3: Example stimuli of scatterplots applied in class estimation task (Experiment One), correlation comparison task (Experiment Two), and clustering judgment task (Experiment Three), with varying visual encodings, number of categories, and visual complexities.

We anticipate if subitizing is used to support multiclass scatterplot analysis, performance on small categories should be similar for both time limits, but will degrade with shorter exposure times for larger numbers of categories.

4.1.1 Stimuli

Scatterplots were rendered using a 400×400 pixel graph. The datasets contained between $N = 2--15$ categories. We encoded categories using two of the most widely-used visual channels for representing data in multiclass scatterplots: color and shape. To reduce the complexity of the study design, we used high-performing palettes from previous studies [74, 76], choosing the STATA S2 color palette and R shape palette to encode categories. We followed other visual representation parameters from these studies, using a filled mark with a three-pixel radius for color-encoded plots and a 6×6-pixel window for shape encodings. Colors and shapes were uniquely assigned to each category using random selection.

Class data in the employed datasets was drawn from 2D Gaussian normal distributions with variances ranging from (0, 1), with random mean x- and y-values between [0.1, 0.9]. Further, we created three different visual complexity levels for the classes by controlling the number of points from each class: low (20 points), middle (30 points), and high (40 points). Each scatterplot only contains classes with all low, middle, or high numbers of points. See Figure 3 for examples of scatterplot stimuli used in Experiment One.

4.1.2 Procedure

Our experiment consisted of three or four phases, depending on the design condition: (1) informed consent, (2) color vision deficiency (CVD) screening using Ishihara plates for participants viewing color-coded scatterplots, (3) task description and tutorial, and (4) formal study. Each participant only saw either all color-coded scatterplots or all shape-coded scatterplots and were assigned to either the 3 or 10-second condition for all trials.

Participants first provided informed consent under our IRB protocol and then provided basic demographics. Participants who saw color-coded scatterplots then needed to successfully pass the CVD assessment. Afterward, participants were introduced to the class estimation task description and led to the tutorial section. They completed three tutorial questions, asking them to complete the target task with 2 to 3 categories with color or shape encodings, whichever aligned with their assigned condition. They were required to successfully answer all tutorial questions before proceeding to reduce possible ambiguities in task understanding.

During the formal study, participants completed our target task ("estimate the total number of categories in this scatterplot") for 42 stimuli presented sequentially in a random order. The 42 stimuli included one scatterplot for each combination of 14 different category numbers (2-15) and three visual complexity levels (high, middle, low). We added three simple two-class scatterplots, showing two well-separated compact classes, as engagement checks following other studies [74]. Participants had either three seconds or ten seconds to view each stimulus, depending on their assigned condition. Then the stimulus was hidden, and participants were required to provide an answer by typing in a number in an input textbox.

4.1.3 Participants

We recruited 239 participants on Amazon Mechanical Turk (MTurk) with at least a 95% approval rating and located within the US and Canada. All participants reported normal or corrected-to-normal vision. 34 participants who failed more than one engagement check were excluded, resulting in an 86% acceptance rate. Among the remaining 205 users (133 male, 72 female; 22–60 years of age), 53 participants saw color-coded scatterplots for three seconds, 54 saw color-coded scatterplots for ten seconds, 50 saw shape-coded scatterplots for three seconds, and 48 saw shape-coded for ten seconds. These studies took 4–7 minutes on average.

4.2 Experiment Two: Correlation Comparison

This experiment asked participants to pick the class that had the highest correlation. This task requires people to both select between different classes and then aggregate properties across those classes, potentially involving a larger number of perceptual mechanisms than Experiment One. Scatterplots varied in their encoding type (shape or color, between-participants), complexity (low, medium, or high; within-participants), and number of categories ($N = 2--15$, within-participants). This

study partially replicates past studies that demonstrated a potential subitizing effect [74,76] to determine whether the prior effects replicate and, if so, to evaluate the potential connection to subitizing.

4.2.1 Stimuli

The scatterplot stimulus design and complexity implementation was the same as in Experiment One; however, our data generation manipulated correlation to control task difficulty. Following the process outlined in Tseng et al. [76], we employed Pearson's correlation coefficient to control per-category correlations. The scatterplots were generated by the random multivariate normal data generation function in R [57], sampling random x- and y-mean values ranging between [0.1, 0.9].

In prior correlation studies for single-class scatterplots, the just-noticeable difference (JND) of correlations was reported to range from 0.05 to 0.15 [31,37,53]. Therefore, we set the category with the highest correlation to have a [0.8, 0.9] covariance and the second-highest category has at least a 0.1 lower correlation difference. We jittered points to avoid overlap and resampled points until the correlation coefficient values satisfied these criteria. See Figure 3 for stimuli of scatterplots used in Experiment Two.

4.2.2 Procedure

Participants followed the same general procedure as in Experiment One. Each participant completed 42 formal trials (one for each combination of category number and complexity level) and 3 engagement checks. Participants responded to the target task ("identify which category is the most correlated") by choosing the target color or shape from a set of radio buttons including all colors or shapes shown in the plot [74]. Participants were given up to 30 seconds to respond to each stimulus (duration determined in piloting).

4.2.3 Participants

We recruited 112 participants on MTurk with at least a 95% approval rating and located within the US and Canada. All participants reported normal or corrected to normal vision. 14 participants who failed more than one engagement check were excluded, resulting in an 87.5% acceptance rate. Among the remaining 98 participants (59 male, 39 female; 22–60 years of age), 49 saw color-coded scatterplots and 49 saw shape-coded scatterplots. This study took 12 minutes on average.

4.3 Experiment Three: Clustering Judgment

The third experiment asked people to choose the most tightly clustered class. People again identified different categories and drew inferences about the general shape formed by the boundaries of the class. Scatterplots varied in their encoding type (shape or color, between-participants), complexity (low, medium, or high; within-participants), and number of categories ($N = 2-15$, within-participants).

4.3.1 Stimuli

The stimuli again were generated using the same general parameters as in Experiment One. Cluster tightness was controlled using x- and y-variance. The most tightly clustered and second most tightly clustered classes differed in this variance by 0.1 along each dimension to allow perceptual differences between clusters [82], with the most tightly clustered ranging in x- and y-variance from [0.1, 0.2]. Category points were sampled using a 2D Gaussian distribution. As in previous experiments, the x- and y-mean values were randomly sampled between [0.1, 0.9]. See Figure 3 for examples of scatterplots used in Experiment Three.

4.3.2 Procedure

We employed the same general procedure as Experiments One and Two. Each participant completed the target task ("identify which category is the most clustered") for 42 formal trials and 3 engagement checks in a random sequential order. Participants selected the target category as a radio button with the corresponding shape or color. Participants had 30 seconds to respond to each stimuli (duration determined in piloting).

4.3.3 Participants

We recruited 112 participants on MTurk with at least a 95% approval rating and located within the US and Canada. All participants reported normal or corrected to normal vision. 10 participants who failed more than one engagement check were excluded, resulting in a 91% acceptance rate. Among the remaining 102 participants (70 male, 32 female; 24–64 years of age), 53 were in the color-coded group, and 49 were in the shape-coded group. This study took 12 minutes on average.

5 Results

We discuss significant results and statistical analysis using both traditional inferential measures and 95% bootstrapped confidence intervals ($\pm$ 95% CI) for fair statistical communication [16]. We used accuracy as our dependent measure, aggregated across all participants within each experiment as it's a binary response (correct or not) for each trial. We computed three-way ANOVAs for each experimental task individually to compare the impact of three different independent factors on accuracy: category number, encoding type, and visual complexity (i.e., the number of points). The data were approximately normally distributed in the results of each experimental setting. We found no significant two-way interactions and, as a result, do not discuss these here. Please see the OSF supplements for data and results including full ANOVA tables. Table 1 summarizes the results for all our experiments—*T1-3s*, *T1-10s*, *T2*, and *T3*.

Table 1: Main effects of the ANOVA for each of our experiments with three factors: category number (*CN*), encoding type (*Encoding*), and visual complexity (*VC*). Significant effects are shown in **bold**.

T1-3s	Df	Sum Sq	Mean Sq	F value	Pr(>F)
CN	1	8.34	8.34	627.37	**<.0001**
Encoding	1	0.05	0.05	3.58	0.0624
VC	2	0.24	0.12	9.00	**0.0003**
Residuals	72	0.96	0.01		
T1-10s	Df	Sum Sq	Mean Sq	F value	Pr(>F)
CN	1	7.98	7.98	593.81	**<.0001**
Encoding	1	0.02	0.02	1.32	0.2546
VC	2	0.53	0.26	19.64	**<.0001**
Residuals	72	0.97	0.01		
T2	Df	Sum Sq	Mean Sq	F value	Pr(>F)
CN	1	2.56	2.56	325.12	**<.0001**
Encoding	1	0.29	0.29	36.93	**<.0001**
VC	2	0.04	0.02	2.57	0.0834
Residuals	72	0.57	0.01		
T3	Df	Sum Sq	Mean Sq	F value	Pr(>F)
CN	1	3.32	3.32	487.00	**<.0001**
Encoding	1	0.35	0.35	52.01	**<.0001**
VC	2	0.03	0.02	2.33	0.1046
Residuals	72	0.49	0.01		

5.1 Subitizing Limit and Category Number

Our results support **H1** and partly support **H2**: for all tasks, we found that people's accuracy was relatively stable at fewer than six categories and dropped significantly after six categories (H1), but performance reduced when the category numbers became very high (H2).

Increasing category numbers significantly reduced accuracy (($p <$ .0001 for all settings). To further validate subitizing's impact across different tasks, Figure 4 shows accuracy changes across category numbers on the three experiments with four task settings (3s for class estimation, 10s for class estimation, correlation comparison, and clustering). Performance drops between five and six categories for all four task settings, as denoted by the gray dashed lines.

We found a significant difference in average accuracy between five to six categories among all tasks: 24% for estimation at 3s, 16% for 10s estimation, 12% for correlation, and 14% for clustering. This observation indicates five to six categories could correspond to potential subitizing limits in visualizations, aligning with insights from cognitive

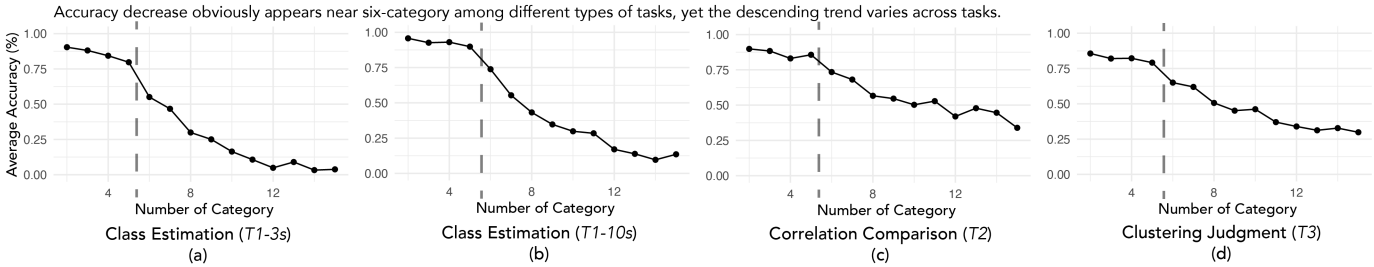

Fig. 4: Overall accuracy distributions in three experiments, separated by four different task settings. The x-axis shows the number of categories in the scatterplots, and the y-axis is the average accuracy in each task. The dashed gray lines denote the separation between five and six categories.

science [49]. Therefore, we further explored the data patterns on two category ranges: $N < 6$ and $N \geq 6$.

For fewer than six categories, performance drops were small for increasing category numbers, at only 3-4% for each added category across all three experiments. Average accuracy remained relatively high. For example, even with restricted three-second viewing time limits, people still achieved nearly 80% accuracy in judging the number of classes in scatterplots with five categories (Figure 4 (a)). This insight aligns with previous research where people achieved robust performance (nearly 90%) at five categories in mean judgment tasks [74].

However, for six or more categories, the performance distribution significantly changed. We observed the strongest performance drops when increasing from five to eight categories, showing an average accuracy reduction per added category at around 16% for Figure 4 class estimation tasks and around 8% for correlation and clustering. However, as category numbers continued to increase, the effects of adding more categories on accuracy were reduced despite class encodings remaining discriminable. This effect was especially notable for category numbers greater than 12, where performance appears to essentially plateau. The observed two-segment performance descending distribution (decreasing sharply after six and slowing to near stable performance after 12) could be at least in part modeled by Fechner's Law. However, we cannot find a single inflection point across all tasks that can separate the fast-descending and flat-descending ranges of category numbers to indicate a consistent logarithmic model. This finding indicates the strength of the hypothesized effects may be different across tasks: in our case, it is clearer in browsing-based tasks (which more closely resemble subitizing tasks in psychology) than in aggregate-based tasks (which likely involve a larger set of mechanisms [72]).

Overall, these results show that the distribution of accuracy changes notably for more than six categories, lending empirical support to traditional design heuristics around the limits of categorical encodings. Before reaching this limit, accuracy remains relatively stable and high, indicating that the perceptual system can manage up to six categories effectively. Beyond this point, the decrease in accuracy reduces at a rate correlated with Fechner's Law. However, the rate of decline varies across tasks, suggesting that other mechanisms may be at play for different tasks and the performance of these mechanisms degrades at different rates than those predicted by subitizing.

5.2 Higher Task Complexity is More Robust to Increasing Categories

Our results support **H3**: we found the impact of category numbers on perceptual accuracy varied depending on the specific visualization task.

To explore the impact of task settings, we conducted simple slope comparisons with Bonferroni correction as a post-hoc analysis. In addition to the general results shown in Figure 4, there was a significant difference between *T1-10s* and *T2* ($t = -8.515$, $p < .0001$), *T1-10s* and *T3* ($t = -6.961$, $p < .0001$), *T1-3s* and *T2* ($t = -8.949$, $p < .0001$), and *T1-3s* and *T3* ($t = -7.396$, $p < .0001$). People's judgment accuracy on *T1-3s* or *T1-10s* have a significantly steeper slope than both *T2* and *T3*. In other words, the same increase in category numbers will result in stronger performance drops in the class estimation task than both correlation comparison and clustering tasks.

This pattern suggests that aggregation (T2 and T3) may actually be more robust to increases in the number of categories compared to browsing tasks like class estimation. This may also align with cognitive insights that visual aggregation usually relies more upon complex cognitive processes like ensemble coding [3, 52, 72]. We note that the data distributions differed across experiments to control task difficulty, creating a potential confounding effect between task and stimulus design. While our focus on relative performance within each task mitigates the potential impact on our findings, future work should consider examining different tasks using the same data distributions.

5.3 Color is More Effective for Aggregation Tasks

Our results partly support **H4**: color encodings were more robust than shape encodings to increasing numbers of categories in aggregate-based tasks (clustering and correlation comparison). We found a significant effect of category encoding types on accuracy in tasks *T2* ($F(1,71) = 36.93, p < .0001$) and *T3* ($F(1,71) = 52.01, p < .0001$), but did find any significant effect in tasks *T1-3s* and *T1-10s*.

Figure 5 illustrates the accuracy results and Figure 6 (a) shows the post-hoc analysis results of two category encodings with specific task types. These results show that the accuracy of color encodings is generally higher than shape encodings among all tasks. However, we also observed that for browsing-based class estimation tasks (see Figure 5 (a) and (b)), the accuracy distributions of color and shape encodings cannot be distinguished clearly from each other. For aggregation-based correlation comparison and clustering tasks (see Figure 5 (c) and (d)), the accuracy differences between color and shape encodings are significant at $t = 6.077$, $p < .0001$ (*T2*) and $t = 7.212$, $p < .0001$ (*T3*) respectively. This suggests that color could be a more perceptually effective encoding than shape when rendering categorical data, but only for certain kinds of tasks.

5.4 Reducing Point Numbers May Not Significantly Improve Performance

Our results do not support **H5**: lower visual complexity failed to significantly improve accuracy when modeled as varying point number. We found a significant effect of visual complexity on accuracy in tasks *T1-3s* ($F(1,71) = 9.00, p = .0003$) and *T1-10s* ($F(1,71) = 19.64, p < .0001$); however, that effect did not match our hypothesized outcomes.

Figure 6 (b) shows the post-hoc analysis results between complexity and task. The general pattern for aggregation-based tasks (*T2* and *T3* in Figure 6 (b)) aligns with our assumption that accuracy for lower visual complexities is slightly higher than higher complexities; however, the differences are not significant. For browsing-based class estimation tasks (*T1-3s* and *T1-10s* in Figure 6 (b)), low complexity led to lower accuracy compared to middle and high complexity.

Neither of these results support the hypothesis that reducing visual complexity leads to improved perceptual accuracy. This effect was especially notable for browsing-based tasks, where the low points number may cause their overall distributions to be too sparse for people to make immediate judgments. The limited correlation between complexity and performance matches observations from Tseng et al. [74]. Complexity factors beyond point number, such as patterns in distribution density, may lead to different outcomes. Future studies should better explore this impact using visual quality measures that align with human perception [1, 35, 82] to guide visual complexity control.

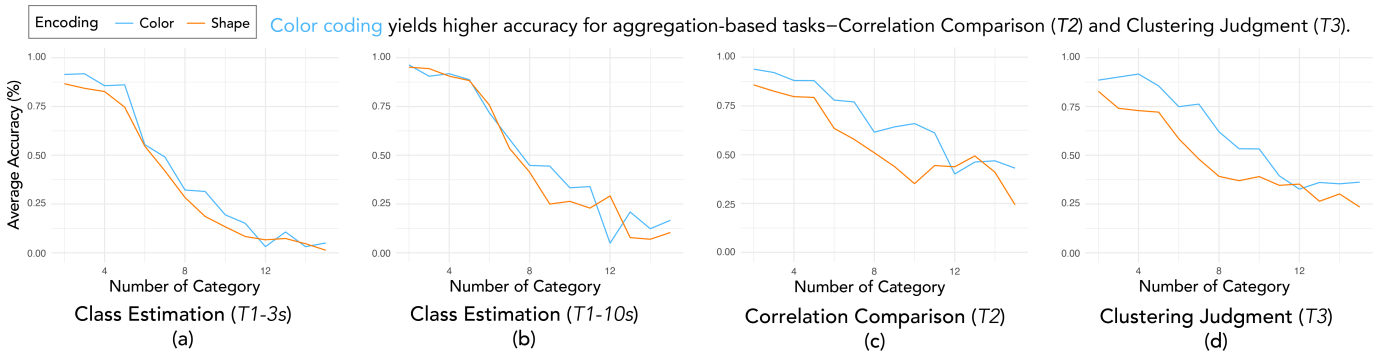

Fig. 5: Accuracy comparisons with color and shape encodings on four different task settings. (a) and (b): class estimation tasks with 3s and 10s time limits. (c): correlation comparison task. (d): clustering judgment task.

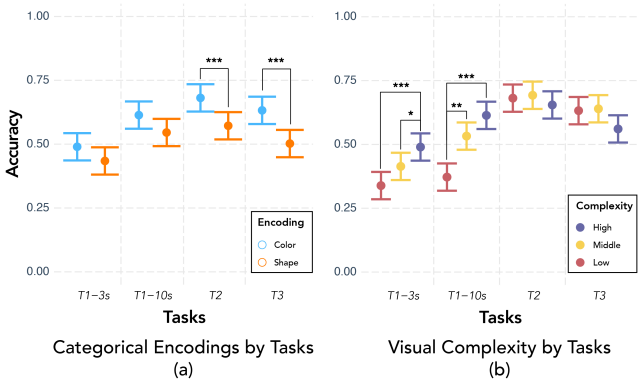

Fig. 6: (a) shows the post-hoc results comparison with 95% confidence intervals (CI) per encodings with tasks. (b) plots the post-hoc results comparison with 95% CI per visual complexities with tasks. Significant differences are denoted as * ($p < .05$), ** ($p < .01$), and *** ($p < .0001$).

6 Discussion

We explored how varying category numbers influence accuracy on three multiclass scatterplot tasks, with an emphasis on how two relevant psychological phenomena and theory—subitizing and Fechner's Law—may explain patterns in categorical data interpretation. Our results provide further evidence of the role of subitizing in data visualization, offer new perspectives reflecting on prior heuristics and research findings, and suggest actionable design guidelines and future research opportunities.

6.1 Viewing Data Visualization via the Lens of Psychology

Our results identify a potential role of subitizing in visualization—supporting categorical reasoning—to help predict how category number affects people's abilities to use visualizations effectively. Prior research in cognitive psychology has established that subitizing is effective for small numbers of items, beyond which individuals must rely on other, often less efficient, cognitive processes such as counting [36, 49]. Our study extends these findings by demonstrating that these limits directly affect task accuracy in categorical visualizations. More specifically, our results confirm that when the number of categories reaches six, people start to experience significant performance drops for a range of categorical tasks, providing empirical and theoretical support for past design heuristics and highlighting the importance of considering these cognitive boundaries for visualizations.

This result aligns with findings from related studies that found sudden performance drops for tasks such as mean judgments [74,75], visual search [30], isotype visualizations [29], and correlation judgments [76]. When viewed as an outcome of subitizing, such performance drops imply a more general visualization phenomenon described by known constraints of subitizing for $N < 6$ and by Fechner's Law for $N \geq 6$. That is, people are able to quickly and efficiently identify and reason across six or fewer encoded categories (*subitizing*) and, after a certain threshold (approximately 12 in our results), additional categories have little to no effect on analysis across categories (Fechner's Law).

These results, as a whole, can further connect to *Dual-System Theory* in psychology [20], which suggest individuals have two different sets of decision-making processes, System 1—impulsive, fast, and acts without thinking—and System 2—a more cognitive, deliberate, thinking process. Subitizing, the rapid and accurate perception of small quantities, aligns with System 1 processing, characterized by its automatic and effortless nature. When quantities exceed the subitizing limit, processing shifts to the more effortful and conscious System 2, involving deliberate counting or estimation.

Our results also help explain tensions in prior studies. Tseng et al. [74] failed to confirm the guideline proposed by Gleicher et al. [23], which reported adding additional distractor classes does not significantly impact mean judgment performance in multiclass scatterplots. Considering visualization through the lens of subitizing explains this contradiction: Gleicher et al.'s study only tests category numbers within the subitizing range ($N \leq 3$), so their results confirmed the subitizing ability of near-instant categorical perception. Alternatively, Tseng et al. [74] explored beyond the subitizing range ($N \leq 10$) and observed the performance decrease that may be explained by Fechner's Law under more complex cognitive processes such as counting or aggregation.

6.2 In Dialogue with Existing Heuristic Design Guidelines

Many design heuristics suggest limiting the number of categories to avoid overwhelming the viewer. However, empirical evidence supporting these guidelines has been sparse, and both research and commercial tools offer strategies for pushing encodings beyond these limits [28]. By evaluating these heuristics using psychological models, we can apply an empirical lens to better understand if and when this and other heuristics hold in practice.

Our study provides both empirical evidence and a theoretical basis for the limit of six. We found that accuracy in categorical analysis tasks remained stable for smaller categories but declined sharply with more than six categories, providing empirical support for the up-to-six claims from prior heuristics [2, 28]. The six-category baseline, when viewed in relation to subitizing, is likely to generalize well across most visualization tasks that require selecting and comparing features of different groups. Further, characterizing performance through subitizing and Fechner's Law enables designers to make more informed predictions about how effectiveness will change as they increase the number of categories they choose to encode. While encoding more than six categories leads to lower overall performance, people still are capable of analyzing data at these scales, and the performance degradation from adding more categories quickly tapers off. This fall-off implies that people can make sense of larger numbers of categories (if less efficiently) and that visualizations are unlikely to benefit from reducing category count unless the data can be reduced to six or fewer categories.

While the six category rule is among the most common design heuristics employing a prescriptive quantity for design, other heuristics also raise specific design thresholds that subitizing and Fechner's Law may in part explain. For example, in palette design, an up-to-10 suggestion

states "*If you add too many colors to the palette, it'll be difficult to comprehend the chart.*" [21]. Others note that it may be difficult to find more than eight distinctive colors for categorical palettes [85]. This guidance contradicts the existence of at least thirteen readily-namable, readily distinguished colors [8] as well as our ability to sample far more than eight distinguishable colors from a limited numerical color space [70]; however, this heuristic does align with subitizing limits. Reasoning across more than eight color-coded categories is likely more difficult due to an inability to efficiently leverage subitizing than to an inability to distinguish between the colors themselves.

Other heuristics for design beyond categorical data emphasize the potential negative impact of excess visual features in visualizations, such as how text color, gridlines, and backgrounds may interfere with chart perception [45]. In past studies from psychology, adding additional distracting elements [24] or additional encodings [73] may slow subitizing, meaning distracting visual information may also interfere with categorical analysis. Certain encoding types may use visual features that are well-suited to the cognitive processes used to accomplish different tasks. Understanding the connections between features and task-relevant processes can help guide more effective task-driven design guidelines [3, 38, 60]. For example, we found that more complex aggregation tasks that likely leverage other processes in addition to subitizing performed better with color than with shape, which may imply colors are efficiently processed by these processes. As people may process these channels differently [3], future work should explore if the perceptual mechanisms we use to process particular visual features make them well-correlated to certain tasks.

6.3 Design and Research Implications

Our results provide recommendations to promote future visualization design and research in general. Given our stimuli, we primarily focus on visual data communication to inform static graphs.

When possible, limit visualizations to six categories: Aligned with several existing heuristic guidelines, this recommendation reflects likely subitizing limits in data visualization. Above six categories, people will be less accurate and slower in analyzing data as they are likely to need to engage less efficient perceptual mechanisms to reason over categories. When the number of categories surpasses this threshold, people experience a significant decline in perceptual accuracy, leading to more potential data misinterpretation.

Higher category numbers are more robust with aggregation-based tasks: Our findings suggest that category number's impact on perceptual accuracy varies depending on specific tasks at hand when exceeding the subitizing limit. For aggregate-based tasks like clustering, people may be able to process a higher number of categories with relatively high accuracy assuming a well-designed visual encoding. This is likely due, at least in part, to additional mechanisms that are not subject to the same limits as subitizing playing a larger role in these tasks.

Prioritize color encodings: We confirm past findings [76] that color is more easily distinguishable than shape when encoding categorical data. While color encodings may have accessibility limitations, if these limitations can be minimized or avoided, color provides a more robust and effective categorical encoding channel across a variety of tasks.

Psychology can provide theoretical grounding for design heuristics: Integrating advances in psychology into data visualization research offers a promising interdisciplinary way to improve the field [6, 17]. Our study further confirms the value of using these grounded theories as guidance to inform design guidelines, which aligns with previous visualization research that investigates the application of psychological theories in visualization design and interpretation, like Weber's Law [31, 68] and confirmation bias [42, 81]. These theories help *ground, refine, and generalize heuristic guidance* by bridging design practice with the processes by which people interpret visualizations. Connecting visualization interpretation to perceptual processes offers new hypotheses for vision science, such as opportunities to understand how different mechanisms work together to influence different kinds of data interpretation (e.g., different tasks), therefore building up real empirical foundations [40]. However, our results stress the importance of translational empirical research that confirms the effects these theories have in practice within visualizations. For example, subitizing and Fechner's Law help model performance across tasks, but are insufficient to fully explain performance differences as a function of category number.

Our work reveals that understanding the underlying psychological mechanisms behind visualization heuristics offers more than just empirical validation; it provides genuine explanatory and predictive value. By studying these mechanisms, researchers can determine the boundaries, edge cases, and potential conflicts between heuristics, leading to more robust and generalizable visualization design guidelines. We hope our work can promote future interdisciplinary efforts to foster a better understanding of perception in the context of data visualization to advance both visualization and vision science research.

6.4 Limitations & Future Work

Visualization types other than scatterplots, such as bar charts, pie charts, or heatmaps, may present different challenges and opportunities for graphical perception [51] and may offer different insights into the role of various perceptual laws and functions. Further, we only investigated a limited set of categorical encodings. Even though these palettes were chosen based on results in previous studies, they do not encompass the full range of possible encoding options available to designers. For instance, other palettes and encoding channels, such as textures [34], size [66], and position [33], could also benefit perception efficiency.

Redundant encoding—using multiple visual channels to represent the same factors in data—can enhance accuracy for visualization tasks [23, 46]. Our study did not investigate the potential benefits of redundant encodings, such as using both color and shape simultaneously to encode one category [66]. This approach may benefit task performance, particularly in scenarios involving higher category numbers, and may offer additional features that the visual system can reason over simultaneously.

While our study focused on three specific tasks, these represent only a subset of the tasks that people commonly perform when interacting with categorical visualizations [61]. These tasks specifically require distinguishing different classes of points and analyzing data across these classes. Other tasks, such as identifying outliers, comparing skewness, or performing time-series analysis, may place different demands on people's perceptual and cognitive resources and may require using class information in different ways. Further, these tasks may help provide insight into whether there is a causal or correlative relationship between subitizing and multiclass analysis, as well as other mechanisms that might be at play. Future work should explore a wider range of tasks to better understand subitizing and the types of tasks it helps model. We aimed to cover a large range of category numbers, which reduced the trials for participants at each number, resulting in a lack of insights into individual differences which should be studied in the future.

7 Conclusion

This study deconstructs design heuristics and visualization perception with psychological phenomena. More specifically, our study provides a better understanding of data visualization in relation to *subitizing* phenomena. By empirically investigating the effects of category number, visual encoding, and task complexity, we have identified critical thresholds and patterns in behavior that suggest the use of subitizing in interpreting categorical visualizations. Performance patterns indicate that people's abilities to make sense of multiclass data change at six categories, beyond which perceptual accuracy declines significantly and correlates with behaviors predicted by Fechner's Law. Our results culminate in a set of empirically grounded design guidelines that can inform more effective categorical visualizations. These design guidelines and implications offer a roadmap for creating more effective visualizations for categorical data. By adhering to these principles, designers can ensure that their visualizations align with the cognitive capabilities of their users, thereby maximizing both accuracy and usability. Our work also demonstrates the significance of connecting theories from cognitive psychology and practices in data visualization, offering practical guidelines that enhance the effectiveness of visualizations and insight into how and why guidelines may generalize.

Acknowledgments

We thank the reviewers for their insightful comments. This work was supported by NSF IIS #2046725 and by NSF CNS #2127309 to the Computing Research Association for the CI-Fellows Project.

References

[1] M. Abbas, M. Aupetit, M. Sedlmair, and H. Bensmail. Clustme: A visual quality measure for ranking monochrome scatterplots based on cluster patterns. *Computer Graphics Forum*, 38(3):225–236, 2019. doi: 10.1111/cgf.13684 4, 7

[2] Adobe. Color for data visualization, 2024. Accessed: 2024-08-15. 2, 3, 8

[3] D. Albers, M. Correll, and M. Gleicher. Task-driven evaluation of aggregation in time series visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 551–560. ACM, 2014. doi: 10.1145/2556288.2557200 4, 7, 9

[4] B. Alsallakh, L. Micallef, W. Aigner, H. Hauser, S. Miksch, and P. Rodgers. The state-of-the-art of set visualization. *Computer Graphics Forum*, 35(1):234–260, 2016. doi: 10.1111/cgf.12722 2

[5] G. Anobile, E. Castaldi, P. A. M. Moscoso, D. C. Burr, and R. Arrighi. "groupitizing": a strategy for numerosity estimation. *Scientific Reports*, 10(1):13436, 2020. doi: 10.1038/s41598-020-68111-1 3

[6] S. S. Bae, K. Cave, C. Görg, P. Rosen, D. A. Szafir, and C. X. Bearfield. Bridging network science and vision science: Mapping perceptual mechanisms to network visualization tasks. *IEEE Transactions on Visualization and Computer Graphics*, PrePrints:1–16, 2025. doi: 10.1109/TVCG.2025.3541571 3, 9

[7] M. Behrisch, M. Blumenschein, N. W. Kim, L. Shao, M. El-Assady, J. Fuchs, D. Seebacher, A. Diehl, U. Brandes, H. Pfister, et al. Quality metrics for information visualization. *Computer Graphics Forum*, 37(3):625–662, 2018. doi: 10.1111/cgf.13446 2

[8] B. Berlin and P. Kay. *Basic color terms: Their universality and evolution*. University of California Press, 1991. doi: 10.2307/2798490 9

[9] T. Boger, S. B. Most, and S. L. Franconeri. Jurassic mark: Inattentional blindness for a datasaurus reveals that visualizations are explored, not seen. In *2021 IEEE Visualization Conference (VIS)*, pp. 71–75. IEEE, 2021. doi: 10.1109/VIS49827.2021.9623273 3

[10] R. M. Boynton. Eleven colors that are almost never confused. In *Human Vision, Visual Processing, and Digital Display*, vol. 1077, pp. 322–332. SPIE, 1989. doi: 10.1117/12.952730 2

[11] D. Burlinson, K. Subramanian, and P. Goolkasian. Open vs. closed shapes: New perceptual categories? *IEEE Transactions on Visualization and Computer Graphics*, 24(1):574–583, 2017. doi: 10.1109/TVCG.2017.2745086 3

[12] P. Chandler and J. Sweller. Cognitive load theory and the format of instruction. *Cognition and Instruction*, 8(4):293–332, 1991. doi: 10.1207/s1532690xci0804_2 4

[13] M. Chen, Y. Liu, and E. Wall. Unmasking dunning-kruger effect in visual reasoning & judgment. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):743–753, 2025. doi: 10.1109/TVCG.2024.3456326 3

[14] M. Correll and J. Heer. Regression by eye: Estimating trends in bivariate visualizations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 1387–1396. ACM, 2017. doi: 10.1145/3025453.302592 2

[15] D. Deng, W. Cui, X. Meng, M. Xu, Y. Liao, H. Zhang, and Y. Wu. Revisiting the design patterns of composite visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 29(12):5406–5421, 2022. doi: 10.1109/TVCG.2022.3213565 2

[16] P. Dragicevic. Fair statistical communication in hci. In *Modern Statistical Methods for HCI*, pp. 291–330. Springer, 2016. doi: 10.1007/978-3-319-26633-6_13 6

[17] M. A. Elliott, C. Nothelfer, C. Xiong, and D. A. Szafir. A design space of vision science methods for visualization research. *IEEE Transactions on Visualization and Computer Graphics*, 27(8):3504–3522, 2021. doi: 10.1109/TVCG.2020.3029413 3, 9

[18] G. T. Fechner. *Elements of psychophysics, 1860*. Appleton-Century-Crofts, 1948. 2, 3, 4

[19] 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. doi: 10.1177/15291006211051956 1, 2, 3

[20] K. Frankish. Dual-process and dual-system theories of reasoning. *Philosophy Compass*, 5(10):914–926, 2010. doi: 10.1111/j.1747-9991.2010.00330.x 8

[21] FusionCharts. Colors for charts: How to use them effectively, 2024. Accessed: 2024-08-15. 2, 9

[22] K. Gadhave, J. Görtler, Z. Cutler, C. Nobre, O. Deussen, M. Meyer, J. M. Phillips, and A. Lex. Predicting intent behind selections in scatterplot visualizations. *Information Visualization*, 20(4):207–228, 2021. doi: 10.1177/14738716211038604 4

[23] 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. doi: 10.1109/TVCG.2013.183 2, 4, 8, 9

[24] L. Goldfarb and S. Levy. Counting within the subitizing range: The effect of number of distractors on the perception of subset items. *PLoS ONE*, 8(9):e74152, 2013. doi: 10.1371/journal.pone.0074152 9

[25] R. L. Goldstone and A. T. Hendrickson. Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(1):69–78, 2010. doi: 10.1002/wcs.26 2

[26] C. C. Gramazio, D. H. Laidlaw, and K. B. Schloss. Colorgorical: Creating discriminable and preferable color palettes for information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):521–530, 2016. doi: 10.1109/TVCG.2016.2598918 3

[27] C. C. Gramazio, K. B. Schloss, and D. H. Laidlaw. The relation between visualization size, grouping, and user performance. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1953–1962, 2014. doi: 10.1109/TVCG.2014.2346983 4

[28] M. Graze and J. Schwabish. Building color palettes in your data visualization style guides. *Journal of the American Medical Informatics Association*, 31(2):488–498, 2024. doi: 10.1093/jamia/ocad084 2, 3, 8

[29] S. Haroz, R. Kosara, and S. L. Franconeri. Isotype visualization: Working memory, performance, and engagement with pictographs. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 1191–1200. ACM, 2015. doi: 10.1145/2702123.2702275 3, 8

[30] S. Haroz and D. Whitney. How capacity limits of attention influence information visualization effectiveness. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2422–2431, 2012. doi: 10.1109/TVCG.2012.233 2, 3, 4, 8

[31] L. Harrison, F. Yang, S. Franconeri, and R. Chang. Ranking visualizations of correlation using weber's law. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1943–1952, 2014. doi: 10.1109/TVCG.2014.2346979 2, 3, 4, 6, 9

[32] J. Heer, N. Kong, and M. Agrawala. Sizing the horizon: the effects of chart size and layering on the graphical perception of time series visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2009. doi: 10.1145/1518701.1518897 2

[33] M.-H. Hong, J. K. Witt, and D. A. Szafir. The weighted average illusion: Biases in perceived mean position in scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):987–997, 2021. doi: 10.1109/TVCG.2021.3114783 3, 4, 9

[34] V. Interrante. Harnessing natural textures for multivariate visualization. *IEEE Computer Graphics and Applications*, 20(6):6–11, 2000. doi: 10.1109/mcg.2000.888001 9

[35] H. Jeon, G. J. Quadri, H. Lee, P. Rosen, D. A. Szafir, and J. Seo. Clams: A cluster ambiguity measure for estimating perceptual variability in visual clustering. *IEEE Transactions on Visualization and Computer Graphics*, 2023. doi: 10.1109/TVCG.2023.3327201 2, 4, 7

[36] E. L. Kaufman, M. W. Lord, T. W. Reese, and J. Volkmann. The discrimination of visual number. *The American Journal of Psychology*, 62(4):498–525, 1949. doi: 10.2307/1418556 2, 3, 4, 8

[37] M. Kay and J. Heer. Beyond weber's law: A second look at ranking visualizations of correlation. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):469–478, 2015. doi: 10.1109/TVCG.2015.2467671 3, 4, 6

[38] Y. Kim and J. Heer. Assessing effects of task and data distribution on the effectiveness of visual encodings. *Computer Graphics Forum*, 37(3):157–167, 2018. doi: 10.1111/cgf.13409 9

[39] J. Kogan, C. Nicholas, and M. Teboulle. *Grouping multidimensional data*. Springer, 2006. 2

[40] R. Kosara. An empire built on sand: Reexamining what we think we know about visualization. In *Proceedings of the Workshop on BEyond time and errors: novel evaLuation methods for Information Visualization (BELIV)*. IEEE, 2016. doi: 10.1145/2993901.2993909 2, 3, 9

[41] B. Lee, K. Isaacs, D. A. Szafir, G. E. Marai, C. Turkay, M. Tory, S. Carpendale, and A. Endert. Broadening intellectual diversity in visualization research papers. *IEEE Computer Graphics and Applications*, 39(4):78–85, 2019. doi: 10.1109/MCG.2019.2914844 2

[42] S. Li, T. J. Davidson, C. X. Bearfield, and E. Wall. Confirmation bias: The double-edged sword of data facts in visual data communication. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–16, 2025. doi: 10.1145/3706598.3713831 9

[43] A. Mayorga and M. Gleicher. Splatterplots: Overcoming overdraw in scatter plots. *IEEE Transactions on Visualization and Computer Graphics*, 19(9):1526–1538, 2013. doi: 10.1109/TVCG.2013.65 3, 4

[44] G. A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2):81–97, 1956. doi: 10.1525/9780520318267-011 3

[45] T. Munzner. *Visualization analysis and design*. CRC Press, 2014. 2, 4, 9

[46] C. Nothelfer, M. Gleicher, and S. Franconeri. Redundant encoding strengthens segmentation and grouping in visual displays of data. *Journal of Experimental Psychology: Human Perception and Performance*, 43(9):1667–1672, 2017. doi: 10.1037/xhp0000314 9

[47] S. E. Palmer, K. B. Schloss, and J. Sammartino. Visual aesthetics and human preference. *Annual Review of Psychology*, 64:77–107, 2013. doi: 10.1146/annurev-psych-120710-100514 2

[48] D. J. Peterson and M. E. Berryhill. The gestalt principle of similarity benefits visual working memory. *Psychonomic Bulletin & Review*, 20:1282–1289, 2013. doi: 10.3758/s13423-013-0460-x 2

[49] M. Piazza, A. Mechelli, B. Butterworth, and C. J. Price. Are subitizing and counting implemented as separate or functionally overlapping processes? *NeuroImage*, 15(2):435–446, 2002. doi: 10.1006/nimg.2001.0980 3, 4, 7, 8

[50] G. J. Quadri and P. Rosen. A survey of perception-based visualization studies by task. *IEEE Transactions on Visualization and Computer Graphics*, 28(12):4893–4911, 2022. doi: 10.1109/TVCG.2021.3098240 4

[51] G. J. Quadri, A. Z. Wang, Z. Wang, J. Adorno, P. Rosen, and D. A. Szafir. Do you see what i see? a qualitative study eliciting high-level visualization comprehension. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–26. ACM, 2024. doi: 10.1145/3613904.3642813 9

[52] R. M. Ratwani, J. G. Trafton, and D. A. Boehm-Davis. Thinking graphically: Connecting vision and cognition during graph comprehension. *Journal of Experimental Psychology: Applied*, 14(1):36–49, 2008. doi: 10.1037/1076-898X.14.1.36 7

[53] R. Rensink and G. Baldridge. The perception of correlation in scatterplots. *Computer Graphics Forum*, 29(3):1203–1210, 2010. doi: 10.1111/j.1467-8659.2009.01694.x 4, 6

[54] R. A. Rensink. Information visualization and the study of visual perception. *Journal of Vision*, 18(10):1350, 2018. doi: 10.1167/18.10.1350 2, 4

[55] S. K. Revkin, M. Piazza, V. Izard, L. Cohen, and S. Dehaene. Does subitizing reflect numerical estimation? *Psychological Science*, 19(6):607–614, 2008. doi: 10.1111/j.1467-9280.2008.02130.x 4

[56] A. Richner. Using colors for data visualization with large categories, 2024. Accessed: 2024-08-15. 2

[57] B. Ripley, B. Venables, D. M. Bates, K. Hornik, A. Gebhardt, and D. Firth. *MASS: Support Functions and Datasets for Venables and Ripley's MASS*, 2013. R package version 7.3-58.4. 6

[58] T. A. Ryan and C. B. Schwartz. Speed of perception as a function of mode of representation. *The American journal of psychology*, 69(1):60–69, 1956. doi: 10.2307/1418115 4

[59] B. Saket, A. Endert, and Ç. Demiralp. Task-based effectiveness of basic visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 25(7):2505–2512, 2018. doi: 10.1109/TVCG.2018.2829750 2

[60] B. Saket, A. Srinivasan, E. Ragan, and A. Endert. Evaluating interactive graphical encodings for data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 24(12):3040–3051, 2018. doi: 10.1109/TVCG.2017.2680452 9

[61] A. Sarikaya and M. Gleicher. Scatterplots: Tasks, data, and designs. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):402–412, 2018. doi: 10.1109/TVCG.2017.2744184 2, 4, 9

[62] A. Sarikaya, M. Gleicher, and D. A. Szafir. Design factors for summary visualization in visual analytics. *Computer Graphics Forum*, 37(3):145–156, 2018. doi: 10.1111/cgf.13408 2

[63] P. Schleifer and K. Landerl. Subitizing and counting in typical and atypical development. *Developmental Science*, 14(2):280–291, 2011. doi: 10.1111/j.1467-7687.2010.00976.x 4

[64] K. B. Schloss, L. Lessard, C. S. Walmsley, and K. Foley. Color inference in visual communication: the meaning of colors in recycling. *Cognitive Research: Principles and Implications*, 3(1):1–17, 2018. doi: 10.1186/s41235-018-0090-y 3

[65] M. F. Schulz and T. Sanocki. Time course of perceptual grouping by color. *Psychological Science*, 14(1):26–30, 2003. doi: 10.1167/1.3.385 3

[66] S. Smart and D. A. Szafir. Measuring the separability of shape, size, and color in scatterplots. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, article no. 669. ACM, 2019. doi: 10.1145/3290605.3300899 3, 9

[67] S. Smart, K. Wu, and D. A. Szafir. Color crafting: Automating the construction of designer quality color ramps. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):1215–1225, 2019. doi: 10.1109/TVCG.2019.2934284 2

[68] U. Soni, Y. Lu, B. Hansen, H. C. Purchase, S. Kobourov, and R. Maciejewski. The perception of graph properties in graph layouts. *Computer Graphics Forum*, 37(3):169–181, 2018. doi: 10.1111/cgf.13410 9

[69] 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. doi: 10.1145/3231772 1

[70] D. A. Szafir. Modeling color difference for visualization design. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):392–401, 2018. doi: 10.1109/TVCG.2017.2744359 9

[71] D. A. Szafir, R. Borgo, M. Chen, D. J. Edwards, B. Fisher, and L. Padilla. *Visualization Psychology*. Springer Nature, 2023. doi: 10.1007/978-3-031-28469-8 1, 3, 4

[72] D. A. Szafir, S. Haroz, M. Gleicher, and S. Franconeri. Four types of ensemble coding in data visualizations. *Journal of Vision*, 16(5):11, 2016. doi: 10.1167/16.5.11 3, 4, 7

[73] L. M. Trick. More than superstition: Differential effects of featural heterogeneity and change on subitizing and counting. *Perception & Psychophysics*, 70(5):743–760, 2008. doi: 10.3758/pp.70.5.743 9

[74] C. Tseng, G. J. Quadri, Z. Wang, and D. A. Szafir. Measuring categorical perception in color-coded scatterplots. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, 2023. doi: 10.1145/3544548.3581416 2, 3, 4, 5, 6, 7, 8

[75] C. Tseng, A. Z. Wang, G. J. Quadri, and D. A. Szafir. Revisiting categorical color perception in scatterplots: Sequential, diverging, and categorical palettes. In *Eurographics/IEEE VGTC Symposium on Visualization (EuroVis) - Short Papers*. The Eurographics Association, 2024. doi: 10.2312/evs.20241073 3, 8

[76] C. Tseng, A. Z. Wang, G. J. Quadri, and D. A. Szafir. Shape it up: An empirically grounded approach for designing shape palettes. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):349–359, 2025. doi: 10.1109/TVCG.2024.3456385 3, 4, 5, 6, 8, 9

[77] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11):2579–2605, 2008. 2

[78] T. Van Gog, L. Kester, and F. Paas. Effects of concurrent monitoring on cognitive load and performance as a function of task complexity. *Applied Cognitive Psychology*, 25(4):584–587, 2011. doi: 10.1002/acp.1726 4

[79] E. Vuokko, M. Niemivirta, and P. Helenius. Cortical activation patterns during subitizing and counting. *Brain Research*, 1497:40–52, 2013. doi: 10.1016/j.brainres.2012.12.019 4

[80] A. Z. Wang, D. Borland, and D. Gotz. An empirical study of counterfactual visualization to support visual causal inference. *Information Visualization*, 23(2):197–214, 2024. doi: 10.1177/14738716241229437 2

[81] A. Z. Wang, D. Borland, T. Peck, W. Wang, and D. Gotz. Causal priors and their influence on judgements of causality in visualized data. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):765–775, 2025. doi: 10.1109/TVCG.2024.3456381 9

[82] Y. Wang, Z. Wang, T. Liu, M. Correll, Z. Cheng, O. Deussen, and M. Sedlmair. Improving the robustness of scagnostics. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):759–769, 2019. doi: 10.1109/TVCG.2019.2934796 4, 6, 7

[83] C. Ware. *Information visualization: perception for design*. Elsevier, 3rd ed., 2012. 3

[84] L. Wilkinson, A. Anand, and R. Grossman. Graph-theoretic scagnostics. In *IEEE Symposium on Information Visualization (InfoVis)*, pp. 157–164. IEEE, Oct 2005. doi: 10.1109/INFVIS.2005.1532142 4

[85] B. Wong. Points of view: Color coding. *Nature Methods*, 7(8):573, 2010. doi: 10.1038/nmeth0810-573 2, 9

[86] J. Yuan, S. Xiang, J. Xia, L. Yu, and S. Liu. Evaluation of sampling methods for scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1720–1730, 2020. doi: 10.1109/TVCG.2020.3030432 2