# Open vs. Closed Shapes: New Perceptual Categories?

David Burlinson, Kalpathi Subramanian, Paula Goolkasian

**Abstract**— Effective communication using visualization relies in part on the use of viable encoding strategies. For example, a viewer's ability to rapidly and accurately discern between two or more categorical variables in a chart or figure is contingent upon the distinctiveness of the encodings applied to each variable. Research in perception suggests that color is a more salient visual feature when compared to shape and although that finding is supported by visualization studies, characteristics of shape also yield meaningful differences in distinctiveness. We propose that open or closed shapes (that is, whether shapes are composed of line segments that are bounded across a region of space or not) represent a salient characteristic that influences perceptual processing. Three experiments were performed to test the reliability of the open/closed category; the first two from the perspective of attentional allocation, and the third experiment in the context of multi-class scatterplot displays. In the first, a flanker paradigm was used to test whether perceptual load and open/closed feature category would modulate the effect of the flanker on target processing. Results showed an influence of both variables. The second experiment used a Same/Different reaction time task to replicate and extend those findings. Results from both show that responses are faster and more accurate when closed rather than open shapes are processed as targets, and there is more processing interference when two competing shapes come from the same rather than different open or closed feature categories. The third experiment employed three commonly used visual analytic tasks - perception of average value, numerosity, and linear relationships with both single and dual displays of open and closed symbols. Our findings show that for numerosity and trend judgments, in particular, that different symbols from the same open or closed feature category cause more perceptual interference when they are presented together in a plot than symbols from different categories. Moreover, the extent of the interference appears to depend upon whether the participant is focused on processing open or closed symbols.

## 1 Introduction

Scatterplots are some of the most common tools for exploring multivariate data, and are widely-used graphing idioms in statistical and charting software. The usefulness of a scatterplot, as with any visualization strategy, is contingent upon effectiveness and expressiveness [18]. In short, we ask how accurately and fairly the scatterplot reflects the underlying data, and how well we can derive insights from it. Central to this question is the choice of symbols used to map the data to visual attributes; this is especially important as the number of data points and categorical classes increase. The literature is rich with endeavors rooted in visual perception, to optimize such displays [5, 10, 24].

Scatterplots mapping multiple categorical classes of variables require careful selection of visual primitives and attributes in order to facilitate discriminability of each variable. This discriminability directly impacts the ease with which an individual can attend to multiple classes of points and make quick, accurate inferences about the nature of the data they represent. If the encoding is suboptimal, a user's task performance will suffer, with errors in higher level judgements (numerosity, density, clustering), or be misled by the prominence of a class of points, or face difficulties discriminating between classes of points.

A number of researchers have focused on understanding symbol characteristics based on basic features of visual perception. Research by Cleveland and McGill [4] indicate an ordering in efficacy for mapping visual attributes to categorical variables. Attributes from the identity channel are most effective, with spatial position the superior option, followed by hue, direction of motion, and finally, differences in shape. Given that scatterplots already utilize spatial position for each point, an additional encoding is required to identify each class. Encoding strategies for scatterplots have largely focused on discriminability between symbols, by exploring symbol contrast and size [16], or by building perceptual distance kernels of plotting symbols [11]. An additional body of work focused on experiments using higher level tasks, such as the perception of correlation [20], average value [10], and discriminability between symbols based on their topological characteristics [1–3].

Our focus is on symbol shapes, and more specifically, two classes of shapes that we term *open* and *closed*; open shapes are shapes that are composed of line segments that are not bounded (plus, asterisk, etc.), while closed shapes bound a finite region of space (square, circle, etc.). We describe three experiments to study these two classes of shapes to understand if they constitute categories of perceptual awareness. In this regard, our work builds on the work by Chen et al. [1] by providing a better understanding of the distinction between these classes of shapes.

**Contributions.** We provide evidence through findings from three experiments, utilizing a variety of perceptual and visualization tasks, that shapes having open or closed features make a difference in the effectiveness of how they are processed. This warrants consideration for their use in visualization tasks. Open/closed feature categories for shapes are found to be particularly important with cluttered rather than sparse displays and with heterogeneous rather than homogeneous items.

## 2 Related work

### 2.1 Perceptual Organization

The psychology literature includes a mature body of work on the lowest level of visual perception. A number of theories suggest that characteristics of items in the visual field are segmented and processed well before the influence of attentional focus [12, 22, 26], then re-organized into meaningful units of perception [25]. The list of 'pre-attentive' features includes characteristics such as hue, motion, curvature, and line endings. Large enough differences in these features can be discerned effectively and instantaneously in the visual field regardless of the number of 'distractor' elements.

Recognition of shape, even simple and two-dimensional, seems to occur further in the pipeline, as shapes themselves are comprised of pre-attentive features. In particular, the number of line endings, angles between and number of line segments, and curvature, all differ among triangles, squares, asterisks, plus signs, and other commonly used charting symbols. Despite the maturity of the literature on pre-attentive characteristics, it is not fully understood how these elements combine to create visual salience in composite objects, or how the features of these composite shapes influence perception across larger displays and in abstract cognitive tasks.

- *David Burlinson and Kalpathi Subramanian are with the Department of Computer Science at the The University of North Carolina at Charlotte. E-mail: dburlins@uncc.edu, krs@uncc.edu.*
- *Paula Goolkasian is with the Department of Psychology at The University of North Carolina at Charlotte. E-mail: pagoolka@uncc.edu.*

*Manuscript received 31 Mar. 2017; accepted 1 Aug. 2017.*
*Date of publication 28 Aug. 2017; date of current version 1 Oct. 2017.*
*For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.*
*Digital Object Identifier no. 10.1109/TVCG.2017.2745086*

Chen et al. [1–3] have argued for a different interpretation of early vision and perception. They suggest a more global perceptual organization based on topological characteristics (invariants) of symbol shapes. Their model proposes a *global-to-local* topological model in perception of shapes, whereby 'wholes are coded prior to perceptual analysis of their separable properties or parts'. Our study of open vs. closed shapes follows this idea to some extent, and suggests its use as a new visual channel that can carry information.

### 2.2 Symbol Discrimination

Research on symbol discrimination is situated between the psychology literature on early vision and work in visualization and statistics. Lewandowsky and Spence [15] investigated visual encoding strategies involving shape, color, amount of fill, letters, and oriented lines using an experimental paradigm in which subjects judged relative degrees of correlation between multiple categorical strata in scatterplot displays. Their results provided mixed support for Cleveland and McGill's ordering of retinal variables from the identity channel; they found that hue was the most useful encoding strategy followed by shape, amount of fill, then confusable letters, although certain discriminable letters introduced similar performance to that of shapes. They also underscored the importance of examining response latency as well as error rates when studying performance with statistical graphs. Tremmel [23] used numerosity judgment tasks to explore symbol differences, with more of a focus on elementary visual features and visual perception theory. Results suggested that differences in fill superseded differences in shape, contrasting the findings from Lewandowsky and Spence; the combination of fill and shape provided even greater discriminability between encodings. In addition, circular symbols were shown to separate well from shapes with multiple line terminators, corroborating findings from theories of pre-attentive vision.

Further examination of plotting symbols in scatterplot displays was undertaken by Li et. al. [17], who studied the influence of rotational symmetry and polygonal and asterisk-based shapes. An internal separation space was computed using Multidimensional scaling (MDS), showing pairwise perceptual distances between the shape and size encodings under investigation. Difference in size was found to introduce a more dominant effect than differences in shape, and shape differences within the broad categories of shape were negligible when compared to differences across those two categories. This is a finding which our results also seem to support.

More recently, Demiralp et. al. [5] used multiple subjective, crowdsourced measures to elicit more direct pairwise perceptual distances between some common shapes, colors, sizes, and their combinations. After applying MDS to the results, a few separate clusters of shapes were exposed - shapes composed of line segments, triangular shapes of various orientations, square and diamond shapes, and circles - each formed clusters separate from each other, reflecting the perceptual distances that arose from their featural differences.

### 2.3 Visual Analytic Tasks

Ultimately, our goal is to get a deeper understanding of relationships between symbols that are commonly used in typical visualization displays in order to contribute to automatic visualization design. To this effect, our work needs to combine metrics of perceptual differences with measures from more abstract visual analytics tasks. The types of tasks relevant to our investigation generally rely on *ensemble* coding: extraction of meaningful higher-level information from distributed elements in the visual field. Szafir et al. [21] discuss ensemble coding in the context of data visualization, and provide a categorization of task types including similarity judgements, numerosity [9], clustering, summary statistics (mean, average value [10]), and trend judgements, such as correlation [20]. Extending simpler cognitive tasks and perceptual differences to higher level analytic tasks is frequently more complicated in practice, as we have ourselves found and describe in the following sections.

## 3 Methods

Our overall goal is to investigate whether open and closed shapes represent categories of perceptual awareness in the pipeline of object perception. In other words, we are studying whether exemplars are rapidly and automatically clustered into one of these two classes and if these extend into typical scatterplot displays. Three experiments were conducted to test if the open vs. closed shapes constitute meaningful perceptual categories. Experiments 1 and 2 use basic perceptual tasks (Flanker task, Same/Difference task) to test whether exemplars representing open and closed categories have different influences on target processing as measured by reaction times (RTs). The Flanker task varied perceptual load and measured the influence of open/closed items on attentional selection, while the Same/Difference task measured perceptual responses to paired items (same symbol, different symbol/same category, different symbol/different category) to measure the speed of processing for items within and between categories of symbols. Experiment 3 tests whether the findings can be extended to scatterplot displays in typical visual analytic tasks (average value, numerosity, and trend judgement).

### 3.1 Experiment 1: The Flanker Task

Flanker tasks [8] involve identification of a target, presented at one of a number of locations in a circular display, while a flanker appears 3 degrees to the right or left of the target. The display locations of the target are all equidistant from the fixation point and within the focus of attention, while the flanker, which is a distractor item, is positioned just outside the attentional focus. Reaction time (RT) responses to the target are typically found to be influenced by the flanker compatibility [6, 7].

For each block of trials in our study, two shapes comprised the set of possible targets and participants were required to discriminate between them. The flanker was a shape that varied in compatibility with the target in one of the following ways. Compatible flankers used the same shape as the target, incompatible flankers used the other shape in the target set, and neutral flankers were a shape that was unrelated to the target item. By measuring the flanker compatibility effect, the difference in target RTs when presented together with compatible and incompatible flankers, we can assess participants' ability to selectively attend to the target shape and ignore the flanker. Compatible flankers should facilitate target responses, while incompatible flankers should interfere due to the flanker shape's importance in the attentional set. Neutral flankers should neither facilitate nor hinder the speeded response, as they are not part of the attentional set of potential targets.

To extend our results to visualization displays, we incorporated perceptual load as a variable, similar to Normand et. al. [19], in order to study selective attention with sparse and cluttered displays. Low load displays included only the target and a flanker, while high load displays filled the remaining locations in the circular array of possible target locations with random non-target shapes to signify a cluttered display, as illustrated in Fig. 1. Compatibility effects are found to be much stronger with low load than high load displays [8, 13, 14].

A comparison of open and closed shapes was studied by varying across each block of trials, whether the target pair consisted of exemplars of the same or different category of open/closed shapes. Each block used two particular shapes as potential targets, and primed a participant's attentional set to favor these symbols. Same feature pairs for the open dimension were star and asterisk, and, for the closed dimension, square and triangle. Different feature pairs include one item from each of the two categories. We hypothesized differences in the compatibility effect as a function of same/different feature pairs and open/closed target/flanker pairs. If the open/closed features represent a relevant perceptual category, then target RTs should vary in response to open and closed shapes and this variable should interact with flanker compatibility and/or load.

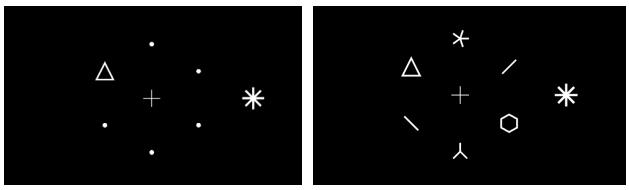
#### 3.1.1 Participants, Stimulus Materials

For all three experiments, we recruited student volunteers from UNC Charlotte, which awards class credit for participating in approved research studies. The inclusion criteria required all participants to be over the age of 18, with 20/20 (or corrected to 20/20) vision and no history

of visual impairment. In Experiment 1, 43 (7 males and 34 females) students were used.

The four shapes we assigned as targets were square, triangle, asterisk, and plus sign, two open and two closed shapes. For a given trial, the flanker that appeared with the target could be compatible, incompatible, or neutral. In the compatible condition both the target and the flanker were the same (either square, triangle, asterisk, or plus-sign); while in the incompatible condition the flanker was the other member of the target set (i.e., square target with triangle, asterisk, or plus-sign flanker; triangle target with square, asterisk, or plus-sign flanker). Neutral flankers incorporated one of the two feature categories but with a shape not used as a target (i.e., square target with circle or $\times$ flanker, or plus sign target with circle or $\times$ flanker).

Other than instructional material, the two forms of visual stimuli utilized in each trial were fixation and target displays. The fixation displays had a black background with a white fixation dot at the center to orient the participant's gaze, and appeared for 500, 600, 700, 800, 900, or 1000 milliseconds (See Fig. 1). The target display featured six positions, marked by dots, spaced equally around the center of the screen within foveal vision (1.5°), and a flanker position placed 3° to the left or right of the fixation point, just outside the focus of attention. A target shape was placed in one of the six target locations, and a flanker shape appeared in one of the two flanker positions. Display locations were based on previous research with this paradigm [7].



(a) (b)

Fig. 1. Displays for the flanker test trials. (a) low-load condition (b) high-load condition. Participants were shown a fixation display for 500 to 1000ms, the target display for 100ms, then a post-stimulus fixation display until a keypress response was made.

In the low load condition, the target was presented in one of 6 locations on a circular array and the other locations were marked by a decimal point. However, in the high load conditions, the non-target locations were filled with a mixture of shapes from the open and closed dimension: diamonds, pentagons, and hexagons for the closed shapes, and symbols with three or five equidistant radial line segments at various rotational positions for the open shapes. None of the filler shapes were used for the target set of items.

Target and flanker pairs were presented in blocks of trials, and there were 6 blocks of trials – two same-feature blocks (triangle/square, asterisk/plus) and 4 different-feature blocks (square/asterisk, square/plus, triangle/asterisk, and triangle/plus). Each block contained 120 experimental trials, split evenly among high and low-load trials. Each set of 60 high and low-load trials in each block included 20 compatible, incompatible, and neutral trials, with targets, flankers, and filler distractors distributed as evenly as possible between all possible locations. Within each block of trials the target and flanker locations were random, but they appeared an equal number of times at each of the possible locations. As shown in Fig. 1, target, flanker, and filler shapes were presented as white against a black background.

In total, there were 6 blocks of 120 trials for a total of 720. We used a Latin Square to balance the presentation order of the six blocks to account for effects of presentation order and sensitization; participants were randomly assigned to one of the 6 orders.

In all three of the experiments described in this report, the visual stimuli were presented on a Macintosh G4 computer with a 15" flat screen monitor. Stimulus presentation and data collection were controlled by SuperLab 4.0.

### 3.1.2 Procedure

Participants were run individually in 40-minute sessions. After filling out an informed consent sheet, they were positioned 60 cm from a computer screen in a well-lit room. They participated in 6 blocks of trials in one of 6 presentation orders.

For each trial, participants were sequentially presented with a fixation display, a target display, and a post-stimulus fixation display. The first fixation display was shown until each trial began, randomly for 500, 600, 700, 800, 900, or 1000 milliseconds, then the target display was presented for 100 ms, followed by the post-stimulus display which was terminated by the participant's keypress.

Prior to participation in each block of trials, participants were shown the two shapes in the target set and the two response keys used to indicate the target shape. There were 20 practice trials to familiarize the participants with the key press responses that were associated with each of the shapes in the target set. They were instructed to keep their index fingers over the two response keys and to indicate as quickly and accurately as they could which of the two target shapes appeared on each trial. They were also told to ignore all other shapes in the display. After the practice trials, the participants began the experimental trials for that block. Each block was followed by a brief break.

### 3.1.3 Analysis

In all three of the experiments, RTs were trimmed if they exceeded 2.5 standard deviations from each individual's mean, and data from participants were removed prior to the analysis if there were error rates in excess of 50 percent in at least two conditions. For the remaining participants, mean correct RTs were computed across the trials in each of the experimental conditions. The mean trimmed correct RTs and proportion of incorrect responses were analyzed with separate repeated measures analysis of variances (ANOVAs). A significance level of 0.05 was used for all statistical tests, and the Greenhouse–Geisser correction was made to the p-value where appropriate to protect against possible violations of assumptions of sphericity. When appropriate, the analysis on the RTs also included a between-subjects effect of counterbalanced group. Follow-up Bonferroni comparisons (at the $p < .05$ level of significance) were also used when main effects were found to be significant. [1]

For this experiment, two percent of the trials were trimmed and data from 6 participants were removed prior to the analysis. The ANOVAs for the remaining 37 participants were averaged across the 20 trials within each condition to test for load (high, low), compatibility (compatible, incompatible, neutral), and block effects (2 same feature 4 different feature).

**Reaction Times.** As expected load and compatibility main effects were consistent with past research. We found target identification took longer with high load or cluttered displays than with low load or sparse displays ($F(1,31) = 136.267$, $p < 0.001$, $\eta_p^2 = 0.815$). Means were 756 ms and 660 ms, respectively. We also found a significant effect of flanker compatibility ($F(2, 62) = 47.372$, $p < 0.001$, $\eta_p^2 = 0.604$). Follow-up Bonferroni comparisons (at the $p < .05$) significance level) showed incompatible trial response times greater on average (727 ms) than those of compatible (697 ms) or neutral (700 ms) trials. Importantly, the interaction between load and flanker compatibility was also significant ($F(2,62) = 4.377$, $p = 0.017$, $\eta_p^2 = 0.124$).

Our analysis also found a significant effect of block ($F(5, 155) = 13.503$, $p < 0.001$, $\eta_p^2 = 0.303$); Table 1 presents the average RTs and error rates for the target pairs that were used in each block. Follow-up Bonferroni comparisons ($p<.05$) showed that RTs in the same feature block with the two closed targets were significantly shorter in comparison to all other blocks and RTs were significantly longer in the same feature block with the two open targets in comparison to all other blocks. RTs to the different feature blocks were in between the two same feature conditions with some minor differences. Block 5 differed from 2 and 4, and the difference between blocks 3 and 4 was also

[1] Counterbalanced group was not found to be significant, nor did it interact with any of the variables of experimental interest. The analyses (which include group) for all three of the experiments are provided as supplementary material.

significant, otherwise there were no significant RT differences among the 4 blocks of different feature target pairs.

Block also interacted with load ($F(5,155) = 12.281$, $p < 0.001$, $\eta_p^2 = 0.284$) and in a significant three-way interaction with load, flanker, and block ($F(10,310) = 2.870$, $p = 0.013$, $\eta_p^2 = 0.085$). To understand these complex effects, we calculated the compatibility effect, the difference between incompatible and compatible trials for each combination of load and block, and reanalyzed the data by looking at the effect of block and load. The analysis on the compatibility effect showed that block was not significant ($p = 0.121$), however load ($F(1,36) = 6.843$, $p = 0.013$, $\eta_p^2 = 0.160$) and block by load interactions ($F(5, 180) = 4.702$, $p = 0.002$, $\eta_p^2 = 0.116$) were significant.

Fig 2 presents the compatibility effect for each of the experimental conditions together with 95% confidence intervals. *What is compelling about these data is the fact that the compatibility effect for the two same feature blocks do not show the same load effects as the other blocks with the exception of square and plus. In three of the four different feature blocks, the strong compatibility effects evident in the low load conditions are diminished under high load or cluttered displays.* With the two same feature displays, however, compatibility effects are similar under the two load conditions.

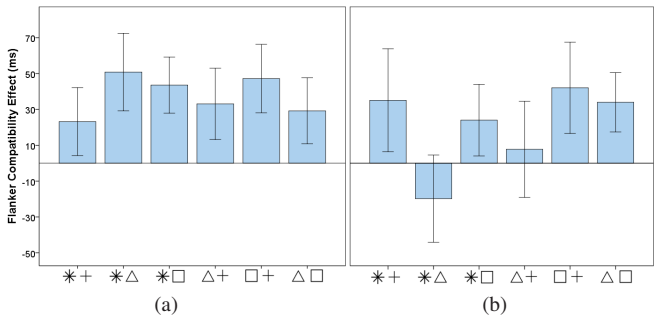



Fig. 2. Flanker compatibility (difference between mean incompatible and compatible response times) for each block in (a) low load and (b) high load conditions. Error bars are 95% confidence intervals.

**Errors.** The average proportion of errors was moderately low in the experimental conditions, varying from 2.9% to 17.297% with a mean of 7.38%. The ANOVA on average error proportions yielded a significant effect of flanker compatibility ($F(2,62) = 8.438$, $p = 0.001$, $\eta_p^2 = 0.214$), with mean error proportions in the incompatible condition at 8.67%, and the compatible and neutral conditions slightly lower at 6.39% and 6.72% respectively. Follow-up Bonferroni comparisons showed that the mean error proportion for the incompatible condition was significantly different from both the compatible ($p = 0.010$) and neutral ($p = 0.007$) conditions, but the two latter conditions did not differ significantly from each other ($p = 1.000$).

We found a significant effect of load ($F(1,31) = 56.12$, $p < 0.001$, $\eta_p^2 = 0.644$); mean error in high load trials was 9.5% compared to 5.02% in low load trials. Load effects were also found to vary by block ($F(5, 155) = 4.638$, $p = 0.001$, $\eta_p^2 = 0.130$). It was greatest in same feature open condition, smaller in the different feature blocks, and nonexistent in the same feature closed condition.

We also found a significant effect of block on error proportions ($F(5,155) = 13.320$, $p < 0.001$, $\eta_p^2 = 0.301$) and the mean error rates, shown in Table 1, display a pattern among the blocks that is similar to the RT data. Both the same-feature blocks were outliers, where the same-feature block with open shapes had the highest error rate while the same-feature closed shapes yielded the lowest error rate.

#### 3.1.4 Discussion

These findings show that when the target set included shapes from different feature categories, interference from response incompatible flankers presented outside the focus of attention was much stronger in low load or uncluttered displays than when high load or cluttered displays were used. However, when the target set included shapes from the same feature category, response interference was consistent irrespective of perceptual load. With this first experiment, we successfully replicated the influence of load and flanker compatibility effects, but the extent of these effects depended on open and closed feature categories and whether target sets included same or different feature categories.

| Block | Feature/ Shapes | RT (ms) | SD (ms) | Error Rate | Error SD |
|---|---|---|---|---|---|
| 1 | Same(∗/+) | 752 | 20 | 11.19% | 1.00% |
| 2 | Different(∗/△) | 718 | 18 | 8.35% | 1.01% |
| 3 | Different(∗/□) | 699 | 22 | 6.48% | 0.78% |
| 4 | Different(△/+) | 738 | 22 | 6.96% | 0.75% |
| 5 | Different(□/+) | 685 | 17 | 6.35% | 0.87% |
| 6 | Same(△/□) | 655 | 17 | 4.24% | 0.45% |

Table 1. Breakdown of response time and error rate differences across the blocks.

Response time differences between compatible trials and neutral trials were not statistically significant whereas differences between incompatible and neutral were. These data and the significant effect of flanker compatibility suggests that, rather than having compatible flankers facilitate response times, incompatible flankers seem to cause response competition, in accordance with Forster and Lavie [8] and Normand et. al. [19].

Important evidence in support of our hypothesis of open/closed feature categories was the significant differences in the speed and accuracy of performance across the 6 blocks of trials. When both items in the target set were closed shapes, responses were faster and more accurate than all other blocks. When both were open shapes, attentional selection took the longest and was most prone to errors. However, when the target set included one item from each of the two feature categories, performance in the attentional selection task fell in between the two same feature target pairs.

These findings have two important implications for visualization tasks. There are differences in processing open and closed shapes when used as symbols and these differences are particularly evident with cluttered rather than sparse or low load displays. Secondly, it is easier to focus attention on a target shape and ignore other distractor shapes when the target is from a different open or closed feature category than the other shapes in the display.

The results of this first experiment support the existence of the hypothesized open/closed perceptual categories. *Participants' ability to focus on the target (and ignore the flanker) varied with open and closed shapes, and also varied with blocks of same and different feature combinations.* Some shapes and combinations were harder to ignore than others. Still under question, however, is whether these findings can be replicated with other perceptual tasks and when the stimulus set is expanded to include other exemplars of the open and closed categories. It is possible that low-level feature differences among the 4 exemplars used as targets could have contributed to the finding of differences in RTs between open and closed shapes. However, replication of the findings with more varied exemplars of open/closed shapes and another perceptual paradigm would help to refute that interpretation.

### 3.2 Experiment 2: Same-Different Task

In order to determine whether our findings were valid or whether they were tied specifically to the task or shapes in the first experiment, Experiment 2 used a *Same-Different* paradigm. The open/closed shape categories are directly examined by testing more varied examples of open and closed shapes, and pairing them with all possible same and different combinations.

In a Same-Difference task, each trial features a set of shapes organized at the center of the screen, and participants indicate with a keypress whether all the shapes in the display are the same or different. RTs reflect the degree of difficulty participants face in discerning the

homogeneity or heterogeneity of the presented shapes. We expected fastest performance for trials in which all shapes were the same, as the human visual system rapidly computes summary statistics across the field of vision prior to attentional allocation; however, based on our previous findings, we expected that closed shapes would be associated with faster RTs than open shapes. Moreover, additional support for our hypothesized feature categories would be obtained if there were differences between exemplars of the two feature categories but no differences between exemplars within a feature category.

The trials with two different shapes belonged to one of two conditions. Different items same feature presented two shapes that were from the same feature category (such as a circle paired with a square), while different items different feature category had two different items one from each of the open and closed categories (such as plus sign with square). Because discrimination between perceptual categories is quicker than discrimination within a category, support for our hypothesis would be found if trials with different stimulus elements took longer when the shapes still shared the same open or closed category as compared to stimuli from different feature categories.

#### 3.2.1 Stimulus Materials

Each trial consisted of fixation and target displays. We inverted the background and foreground colors from the first experiment to use black targets on white backgrounds, simulating a common characteristic of visualization displays. Fixation displays had a white background with a black fixation cross at the center to orient the participant's gaze. Target displays featured either 2 or 3 shapes positioned along the central horizontal axis and spaced equally around the fixation point. Two-shape trials featured shapes 1.25° to either side of the fixation cross, and three-shape trials featured a shape at the center and 2.5° to each side so all shapes had uniform spacing across 2- and 3-shape trials. See supplemental materials for sample stimulus displays.

We expanded the shapes representing each of the feature categories from Experiment 1 to include the following six shapes: circle, square, and triangle for the closed shapes, and asterisk, ×, and plus sign for the open shapes. For any given trial, the shapes were either exactly the same, or two different shapes were selected. For trials with three elements in the different-shape condition, the middle shape was always the position differing from the other two.

Each block contained 192 experimental trials split evenly among same-shape and different-shape trials. Different-shape trials were split evenly between different-feature and same-feature trials. Each block contained either only 2-shape trials or only 3-shape trials, as pilot tests suggested mixing the number of elements caused confusion within a block. The trials distributed the shape combinations and locations as evenly as possible. In order to maintain an even number of trials between shape conditions, same-shape trials were oversampled with 96 trials per block while different feature category and different item same feature category each had 48 trials.

In total, there were 4 blocks of 192 trials for a total of 768 trials. We counterbalanced the presentation order of the blocks so that some participants began with 2-shape blocks and others began with 3- shape blocks. The blocks alternated between two and three elements until the participant finished all the trials.

#### 3.2.2 Procedure

Forty-two participants (31 female) were recruited from the same participant pool as the first experiment with the same inclusion criteria. Participants were tested individually in 40-minute sessions. They were given an informed consent form and then positioned 60 cm from the computer screen in a well-lit room. Each participant completed all 4 blocks of trials.

Within each trial, participants were sequentially presented with a fixation display and target display. The fixation display was shown for 500, 600, 700, 800, 900, or 1000 milliseconds, then the target display was presented until the participant responded with a keypress. We decided to leave the stimuli on the screen rather than flashing them on the screen as in Experiment 1, both to simulate more realistic visualization scenarios and so that participants could take as much time as necessary to respond.

Each participant began with 20 practice trials to familiarize themselves with the response keys and the association with 'same' and 'different' responses. They were instructed to use the 'f' and the 'j' key to indicate whether the shapes on the screen in each trial were the same or different shapes. A note at the bottom of the screen reminded the participants of the keys associated with each of the responses. As in Experiment 1, participants were told to respond as quickly and accurately as possible. Experimental trials for the first block followed the practice trials and each block was followed by a brief break before the next block began.

#### 3.2.3 Analysis

On average, 2.6% of trials were removed for each participant, and the largest trim proportion was 4%. Data from 4 participants were removed from the final analysis due to error rates in excess of 50 percent in at least two conditions. For the remaining 38 participants, the ANOVAs tested for feature category (open or closed shapes), condition (same-shape, different-shape/same-feature, and different-shape/different-feature), and block effects (2 or 3 shapes). Since we did not find any significant effect ($p = 0.376$) when 2 or 3 shapes were used and the number of shapes was not found to interact with the other variables of interest, we combined the 2-and 3-shape trials and tested for differences across feature category and conditions.

**Reaction Times.** Our analysis found a strong main effect of same/different condition ($F(2,74) = 55.234$, $p < 0.001$, $\eta_p^2 = 0.599$). Same-shape trials had the fastest mean RTs (651 ms), different-item/different-feature trials were the second fastest (675 ms), and different-item/same-feature trials took the longest (709 ms). Follow-up Bonferroni comparisons ($p < .05$) showed that each was significantly different from the other.

Consistent with the findings from Experiment 1, feature category was also significant ($F(1, 37) = 40.099$, $p < 0.001$, $\eta_p^2 = 0.520$), with faster RTs ( M= 668 ms) for closed shapes than for open shapes (M=689 ms). The interaction between condition and feature presented in Fig. 3 was significant ($F(2,74) = 6.902$, $p = 0.004$, $\eta_p^2 = 0.157$), with closed shapes faster than open shapes, except for the different-shape/different feature trials (which was not significant). Since different shape/different feature trials included items from both feature categories it was arbitrary which feature category the paired items were assigned to. When this category was excluded from the analysis, however, the two-way interaction between same/different condition and feature category was not significant ($p = .077$).

To further understand the feature differences for individual shapes and pairwise relationships between the shapes, we examined the trials for each condition separately and compared RTs for each of the individual shapes. [2]

In the same-shape condition, we found significant differences in RTs among the 6 shapes ($F(5, 185) = 20.836$, $p < .001$, $\eta_p^2 = 0.360$); but follow-up Bonferroni comparisons (at $p < .05$ level of significance) showed that all three of the open items had significantly longer RTs than the three closed items. Moreover, there were no differences in RTs to the three items within either the open or the closed category.

For the different-shape/same-feature condition, we again found significant RT differences among the 6 pairs of shapes ($F(5, 185) = 14.136$, $p < 0.001$, $\eta_p^2 = 0.276$) but the findings were not as clear as in the previous condition. Trials with × and plus took significantly longer (760 ms) than all other shape combinations. And circle/triangle trials were significantly faster (630 ms) than the other combinations of closed shapes (700 ms for circle/square and 716 ms for square/triangle). Asterisk/plus-sign trials were also significantly faster (683 ms) than square/triangle trials (716 ms).

When we analyzed the 9 different-shape/different-feature conditions, we found a significant main effect ($F(8, 296) = 2.836$, $p = 0.012$, $\eta_p^2 =$

[2]Figures showing the means with 95% confidence intervals in each of the 3 conditions are provided as supplementary material.

0.071), but follow-up Bonferroni tests showed no significant differences among the items.

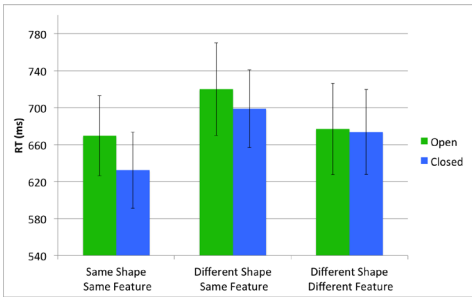



Fig. 3. The interaction between feature and condition on RTs. Same shape was significantly easier, different-shape/different-feature were close across both features, and different-shape/same-feature trials took the longest. Error bars show 95% confidence intervals.

**Errors.** The average proportion of errors was moderately low in the experimental conditions, varying from 0 to 15% with a mean of 3.1%. The ANOVA on average error proportions showed a significant effect of condition ($F(2, 74) = 6.488$, $p = 0.006$, $\eta_p^2 = 0.149$) with different shape/different feature lower (2%) than the other two conditions–same shape (3.3%) and different shape/same feature (3.7%). We also found significance for feature errors ($F(1, 37) = 10.936$, $p = 0.002$, $\eta_p^2 = 0.228$). Closed shapes had significantly fewer errors (2.7%) compared to open shapes (3.4%).

#### 3.2.4 Discussion

As hypothesized, same-shape trials yielded the fastest RTs, and different-item/same-feature trials took the longest. Participants took longer deciding that the shapes were heterogeneous when the different shapes shared the open or closed feature category than when the shapes were taken from both categories. These findings are consistent with those of Experiment 1 and provide additional evidence for the importance of open/closed categories in perceptual awareness.

Similarly to Experiment 1, participants reliably had faster and more accurate RTs to closed shapes than to the open shapes and the effect was consistent for both same shape and different items same category conditions. Interestingly, in the same shape condition, closed shaped items were responded to more quickly than open shaped items and there were no significant RT differences among the items within the feature categories. This provides additional evidence that feature category differences reflect differences in the way open and closed shapes are perceived rather than a result of low-level differences in the shapes. Items within each of the two categories differed in similar ways in terms of low-level features such as differences in line elements and angles; if these factors were the basis of the category difference there would have been more differences between items within a category. For example, the plus sign had fewer elements than the asterisk and the triangle had fewer angles than the square yet differences were not observed when these shapes were presented in the same shape condition.

The analysis on the error proportions also lends credence to our hypotheses. *Different-item/different-feature trials had the fewest errors, reflecting the ease with which participants discerned between open and closed shapes. Different-item/same-feature trials had the highest errors due to participants' relative difficulty in discriminating between different shapes sharing the open or closed dimension.*

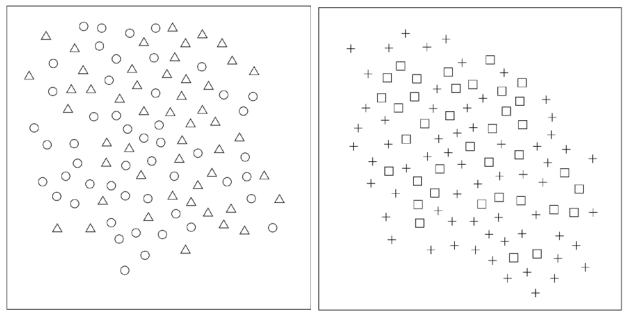
### 3.3 Experiment 3: High Level Tasks

Experiment 3 investigated the discrimination within and between open and closed shapes in visualization displays. It was of interest to assess how the deployment of such encoding strategies influences participants' abilities to perform the types of tasks commonly used in scatterplot displays. A key feature of scatterplot displays is that they extract generalized, higher-order information from large sets of elements in the visual field with ensemble coding. We chose 3 visual analytic tasks with relative judgments to test our findings from the first two experiments:

1. *Average Value:* Determine which of the two sets of shapes has a higher position on the y-axis,
2. *Numerosity:* Determine which of the two sets of shapes contains more elements, and,
3. *Trend Judgements.* Determine which set of shapes exemplifies a linear relationship.

For each of the tasks we hypothesized that if open/closed features represented an important perceptual category, then there should be some difference in task performance when open rather than closed symbols are used in the scatterplot displays. Based on previous findings, we expected (1) that visualization tasks involving closed symbols would be associated with faster RTs than open symbols, and (2) that when two symbols are used together for visualization tasks requiring discrimination between symbols in a single display, symbols from different open/closed categories would be more easily distinguishable and lead to faster RTs than symbols from the same open or closed category.

With each task two kinds of displays were tested: separate-plot displays, which appeared side by side and required participants to select the plot with the higher average value, higher numerosity, or the one showing a linear relationship; and single-plot displays that paired two shapes within one plot and required participants to determine the one that depicted the higher average value, numerosity, or linear trend. The separate-plots contained only homogeneous shapes in each display and were used to get baseline data on the participants' ability to perform the visualization task while the single plot used two symbols within the same display and required discrimination between the two symbols to make a relative judgment. Performance in the single-plot displays provided a direct test of the hypothesis that it is easier to discriminate between two symbols and perform a visualization task when the symbols are from different open/closed categories.



(a) (b)

Fig. 4. Medium-difficulty single-plot displays for the scatterplot analysis trials. (a) Average Value Task (b) Numerosity Task. Participants were shown a fixation display for 500 to 1000ms, then the stimulus display until a keypress response was made. Linear Relationship Task figure is included in supplemental material.

#### 3.3.1 Stimulus Materials

We used the same shape palette as in Experiment 2: circle, square, and triangle for the closed shapes, and asterisk, ×, and plus-sign for the open shapes. Fig. 4 has examples of the single-plot displays for two of the tasks. All of the displays contained 100 items, which were split evenly between two sets of shapes in the single-plot displays and split evenly between left and right displays in separate-plot displays. The only deviation from that even split was within numerosity trials, which necessitated a difference between the number in each set.

We used the concept of 'delta' for each task to introduce different levels of difficulty, similar to Gleicher et. al. [10], who found difficulty to correlate with task performance. Delta represented the difference in pixels between the average position on the y-axis for two sets of shapes in average value judgments, the difference in number of shapes between the two sets in numerosity tasks, and the degree of correlation displayed by the set of shapes with the linear relationship in the third task. We used pilot testing to obtain reasonable delta values for easy,

medium, and hard conditions within each task. For average value judgments, the deltas were 50, 35, and 20 pixels, respectively. For numerosity judgments, deltas were 36, 26, and 16 shapes. For linear relationship judgments, correlations were within 0.05 of 0.8, 0.6, and 0.4 as measured by the Pearson product-moment correlation coefficient.

The stimuli were black on a white background, with display regions of 500 by 500 pixels, and all shapes were rendered at 15 by 15 pixels within a circular area with a diameter of 30 pixels to prevent overlap and introduce a minimum distance between elements. For separate-plot displays, two display regions of the same size were placed side by side in the center of the screen.

For the single-plot target displays in average value and numerosity judgments, we adapted the algorithm from Gleicher et. al. [10]. First, we randomly selected the center point of the entire set at a location in the middle third of the display. Then we utilized a dart-throwing approach to maintain spatial distance between shape positions and best-candidate sampling to prefer positions providing the desired mean, alternating between the two desired shapes to intersperse the categorical sets. For the average value displays, we made small vertical adjustments to the resultant sets of points to reach the desired pixel delta for the given difficulty level. We also de-correlated the top and bottom shapes from the actual higher and lower sets to counter the response heuristic relying on these extremes. To maintain particular delta values in the numerosity tasks, we alternated between shapes until the desired maximum number for the smaller set, then drew the shape from the larger set in the rest of the generated positions.

Single-plot target displays for the linear relationship judgment tasks were generated in a fashion similar to the description given by Rensink et. al. [20]. We first selected a linear equation from a predetermined set of candidate lines with slopes ranging from -1 to 1 and y intercepts within the central two-thirds of the display. From there, we alternated between the two sets of shapes. For elements in the set of linearly associated shapes, we randomly selected x-coordinates and generated y-coordinates for each point within a constrained distance of the associated y-coordinate from the linear equation depending on the correlation delta. For elements in the set without linear relationship, we used a pseudo-random number generator for both x- and y-coordinates. For shapes of either set, we made small adjustments to prevent overlaps and maintain spacing between shapes, and re-randomized the positions if a satisfactory position could not be achieved with minor adjustments. See Fig. 4 for examples of each stimulus display.

For the separate-plot target displays for each of the three analysis tasks, we followed the same sequence of steps as the single-plot display generation, but drew alternate shapes in two separate regions of the display rather than the same region.

The three tasks were arranged into blocks of 108 trials, separated into sub-blocks of 36 separate-plot trials and 72 single-plot trials. Single-plot target displays were split evenly among easy, medium, and hard trials and all 6 of the shapes were used an equal number of times as both target and distractors. Separate-plot target displays were also split evenly among the 3 difficulty levels, and contained an even number of instances when the target display was on either the left or right side. For the linear relationship task trials, we maintained an even number of positive and negative correlations. Within any given sub-block of trials, there was a random arrangement of difficulty levels and shapes.

The presentation order of the 3 task blocks was arranged into a Latin Square order and, as often as was possible, an equal number of participants were randomly assigned to one of the three orders. Within each block, participants always began with the separate-plot trials as these involved easier binary decisions and some baseline data, followed by the single-plot trials.

### 3.3.2 Procedure

We recruited 26 student volunteers (19 were male and 7 female) as participants. Participants were tested individually in 40-minute sessions. They were given an informed consent form and then positioned 60 cm from the computer screen in a well-lit room. Each participant completed the three blocks of trials.

For each trial, participants were sequentially shown a fixation display followed by the target display. The fixation display was shown for 500, 600, 700, 800, 900, or 1000 milliseconds, then the target display was presented until the participant responded with a keypress or 30 seconds elapsed. Participants began each of the three blocks with 12 practice trials to familiarize themselves with the separate-plot task and the associated key responses for the left/right decision. They were instructed to respond with the 'f' key to indicate left and the 'j' key to indicate right as quickly as possible without sacrificing accuracy. In the average value task, participants were told to identify which of the side by side graphs had the items with the higher average Y value. For numerosity, the task was to identify which plot had the greater number of symbols and for the trend task, the participants identified which plot had the linear relationship. Thirty-six experimental trials followed the practice trials.

A second set of 24 practice trials was used to learn the key associations for the shape responses in the single-plot displays. For these displays, participants were instructed to identify which of the two shapes that appeared in the heterogeneous display indicated the higher average Y value, or the greater number of symbols, or the linear relationship. Key responses for the six shapes were mapped to six easily-accessible keys in the center of the keyboard (sdf, jkl). The shape/key mappings remained accessible to participants throughout the duration of the study with a note at the bottom of the monitor. We reordered the keypress mappings for the six shapes for every other participant so that open and closed feature shapes were mapped to right/left finger responses an equal number of times across participants to account for any handedness bias. Seventy-two experimental trials followed with the single-plot displays. After a brief break, participants moved on to the second and third block of trials following the same procedure.

### 3.3.3 Analysis

On average 2% of the trials were trimmed and the data from 6 participants were removed prior to the analysis due to error rates in excess of 50 percent in at least two conditions. For the remaining 20 participants, mean correct RTs were computed across the 6 trials in each of the experimental conditions. The ANOVAs tested for task (average value, numerosity, linear relationship), difficulty (easy, medium, hard), target feature (open, closed), and distractor feature (same, different). Follow-up Bonferroni comparisons (at the $p < .05$ level of significance) were also conducted to explore the significant main effects of task and difficulty level.

### 3.3.4 Separate-Plot Displays

**Response Times.** The analysis on the responses to separate-plot displays showed considerable difference in RTs to the three tasks ($F(2, 38) = 9.5$, $p = 0.006$, $\eta_p^2 = .333$); average value took significantly longer on average (M =1602, SD =1492 ) than numerosity (M= 662, SD = 205 ) and linear relationship (M=632, SD=149), which did not differ significantly from each other. Unexpectedly, we did not find a significant effect of target feature ($F < 1$, $p = .664$), nor did target feature interact with any other variables of interest: target feature by task ($F(2, 38) = 2.23$, $p = .12$), target feature by difficulty $F(2, 38) = 1.44$, $p = .250$), target feature by task by difficulty ($F(2, 38) = 2.73$, $p = .089$).

Task difficulty had the strongest effect on the response times ($F(2, 38) = 21.314$, $p < 0.001$, $\eta_p^2 = .529$), and follow-up Bonferroni tests ($p < .05$) showed as expected that easy trials (790) were significantly faster than all others, hard trials were the longest (1156) with medium difficulty trials (950) in between, mirroring accounts from the literature. Difficulty also interacted with task ($F(4, 76) = 6.09$, $p = .004$, $\eta_p^2 = .243$), as shown in Fig. 5.

Because of the variability in RTs among the three tasks, we reanalyzed the data separately for each of the tasks looking for effects of target feature and task difficulty. We found a significant effect of task difficulty for all three tasks. For average value ($F(2, 38) = 10.81$, $p < 0.001$ , $\eta_p^2 = .363$), follow-up Bonferroni tests ($p < .05$) showed that the easy trials were significantly different from the hard trials. For numerosity ($F(2, 38) = 22.26$, $p < 0.001$, $\eta_p^2 = .540$), the main effect resulted from a significant difference among all three levels of difficulty; and for linear relationship ($F(2, 38) = 22.25$, $p < 0.001$ , $\eta_p^2 =$

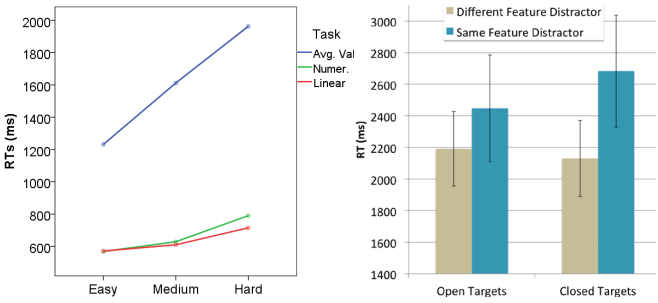



Fig. 5. (a) Difficulty by task interaction for the side-by-side plots. (b) Target by distractor interaction for the single plot numerosity tasks. Error bars express 95% confidence intervals.

.539), the significant main effect was due to the longer RTs for the hard trials in comparison to the other difficulty levels. However in all three of the analyses, there were no significant effects of target feature, nor any significant interactions between target feature and task difficulty.

**Errors.** Performance across all of the conditions was high with average error rates ranging from 0 to 13% of the trials. The analysis on the proportion of errors was similar to the RTs in showing a main effect of task ($F(2, 38) = 13.754$, $p = 0.001$, $\eta_p^2 = .420$), with more errors on the average value task ($M = .053$, $SD = .08$ ) in comparison to the numerosity ($M = .004$, $SD = .01$ ) and linear relationship ($M = .011$, $SD = .03$ ) tasks. As in the previous RT analysis, there is more variability associated with performance in the average value task than in the other tasks. Difficulty remained a significant factor ($F(2, 38) = 5.016$, $p = .012$, $\eta_p^2 = .209$), although none of the difficulty levels were significantly different from each other (means of .014, .017, and .038 for easy, medium, and hard, respectively). Surprisingly, target feature was also significant ($F(1, 19) = 28.023$, $p < .001$, $\eta_p^2 = .596$), as were its interactions with task ($F(2, 38) = 5.635$ $p = .007$, $\eta_p^2 = .229$) and three-way interaction with task and difficulty ($F(4, 76) = 2.907$, $p = .044$, $\eta_p^2 = .133$). These interactions may have resulted from a floor effect with negligible error rates in the numerosity and trend tasks in comparison to some low error rates in response to closed targets in the average value task.

Reanalyzing the data separately for the three tasks exposed significant effects of target feature ($F(1, 19) = 18.424$, $p < .001$ , $\eta_p^2 = .492$) in average value tasks and difficulty in average value tasks ($F(2, 38) = 7.535$, $p = .006$, $\eta_p^2 = .284$) and linear relationship tasks ($F(2, 38) = 4.147$, $p = .05$, $\eta_p^2 = .179$). Numerosity tasks received so few errors across the conditions that none of the effects achieved significance.

Taken together, the RT and error data from the side by side displays show that although there were differences across the tasks, participants could perform all three of the visualization tasks with a high degree of accuracy. Detecting numerosity and linear relationships were accomplished more quickly than determining average value but in all of the tasks performance for the most part was above 90% correct.

These results also show that for each of the three tasks, there was no difference in task performance with the open and closed category of shapes, other than a slight increase in errors with closed shapes in the average value task. In contrast to the findings from the perception tasks, however, there are no observable differences in task performance when either open or closed shapes are used as symbols in homogeneous scatterplot displays.

### 3.3.5 Single-Plot Displays

**Response Times.** When participants were asked to identify which of the two shapes presented in a single display met the task requirements, we again found a strong, significant influence of task on RTs ($F(2, 38) = 15.67$, $p < 0.001$, $\eta_p^2 = .452$), with follow-up test showing that all three task means differing significantly from each other ($M = 3226$, $SD = 1417$; $M = 2363$, $SD = 404$; $M = 1787$, $SD = 347$ for average value, numerosity, and linear relationship respectively). Task was also found to interact with distractor feature ($F(2, 38) = 6.135$, $p = .01$, $\eta_p^2 = .244$), and target feature ($F(2, 38) = 5.03$, $p = .012$, $\eta_p^2 = .209$).

As with the side by side displays, there was a main effect of difficulty level ($F(2, 38) = 16.05$, $p < 0.001$, $\eta_p^2 = .458$); however, Bonferroni comparisons showed that easy trials (2261) differed significantly from medium (2512) and hard (2603), and the latter two did not differ significantly from each other.

Importantly, distractors from the same feature category as the targets lengthened RTs relative to distractors from a different feature category ($F(1, 19) = 52.595$, $p < 0.001$, $\eta_p^2 = .735$), and this variable interacted with target feature ($F(1, 19) = 25.11$, $p < .001$, $\eta_p^2 = .569$), and in a three-way interaction with target feature and difficulty ($F(2, 38) = 5.051$, $p = .011$, $\eta_p^2 = .210$). There was no main effect of target feature ($p = .552$), but there was an additional interaction of this variable with difficulty ($F(2, 38) = 3.401$, $p = .044$, $\eta_p^2 = .152$).

To understand the influence of target and distractor features on each task, the data were reanalyzed separately for each of the tasks. In the analysis on the average value task, the subjects displayed a great deal of variability in response latency. Although the trends appeared to be moving in the right directions for the hypothesized effects of difficulty and distractor feature, we found no significant main or interaction effects among any of the experimental conditions in the task.

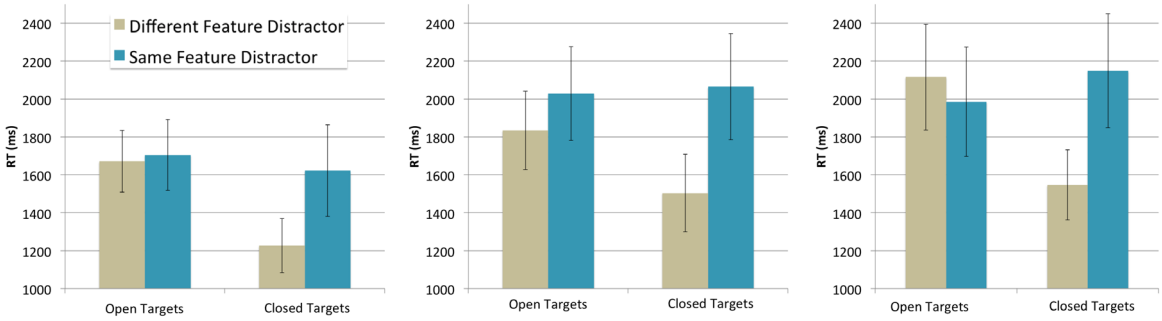
For the numerosity task, however, we found a number of significant effects. Distractor feature was significant ($F(1, 19) = 48.16$, $p < .001$, $\eta_p^2 = .717$); same-feature distractors took 400 milliseconds longer on average. Difficulty level was also significant ($F(2, 38) = 8.87$, $p = .002$, $\eta_p^2 = .318$); easy trials (2142) differed significantly from both medium (2436.97) and hard (2511), but the latter two did not differ significantly from each other. A significant interaction effect of target feature and distractor feature ($F(1, 19) = 8.70$, $p = .008$, $\eta_p^2 = .314$) indicated that same-feature distractors reliably caused longer reaction times but the effect was modulated by target features Fig. 5(b).

It was the linear relationship task, however, where we found strong effects on target RTs from all of the manipulated variables. As with the previous task, there was an effect of level of difficulty ($F(2, 38) = 24.38$, $p < .001$, $\eta_p^2 = .562$), and easy trials (1556) were significantly different from medium (1857) and hard trials (1949); medium and hard did not differ significantly.

Closed target features led to quicker RTs than open targets ($F(1, 19) = 20.95$, $p < .001$, $\eta_p^2 = .524$), and same featured distractors lengthened RTs relative to distractors from different categories than the target ($F(1, 19) = 25.33$, $p < .001$, $\eta_p^2 = .571$). Additionally, these two variable interacted with each other ($F(1, 19) = 26.27$, $p < .001$, $\eta_p^2 = .58$) as well as in a three-way interaction with level of difficulty ($F(2, 38) = 3.15$, $p < .054$, $\eta_p^2 = .142$). Fig. 6 presents the 3-way effect.

These results show that *when making judgments of numerosity and linear relationships from displays with heterogeneous items, the feature category of both the target and distractor shapes are important. When the distractors are from a different open/closed category than the target, RTs are faster, but the effect is particularly evident when processing closed targets.*

**Errors.** The mean proportion of errors varied considerably across the experimental conditions (0% to 37%) showing effects largely consistent with the RTs data. The analysis of error proportions in single display trials yielded a number of significant effects and interactions. Task ($F(2, 38) = 31.920$, $p < 0.001$, $\eta_p^2 = .627$) and difficulty ($F(2, 38) = 45.075$, $p < .001$ , $\eta_p^2 = .703$) both showed strong significant effects. Error rates in all three tasks were significantly different from each other (average value: $M = .272$, $SD = .225$; numerosity: $M = .118$, $SD = .13$, and linear relationship: $M = .024$, $SD = .06$). At an error rate of .201, hard trials introduced significantly more errors than medium (.119) and easy (.093) conditions. Task and difficulty also showed a significant interaction ($F(4, 76) = 10.607$, $p < .001$, $\eta_p^2 = .358$). Distractor feature ($F(1, 19) = 20.083$, $p < .001$, $\eta_p^2 = .514$), its interaction with task ($F(2, 38) = 3.677$, $p = .035$, $\eta_p^2 = .162$), and its three-way interaction with task and difficulty ($F(4, 76) = 4.002$, $p = .016$, $\eta_p^2 = .174$) were all




(a) (b) (c)

Fig. 6. Three-way interaction for single plot linear tasks with 95% confidence intervals. (a) easy, (b) medium, and (c) hard conditions.

significant effects in the error proportion analysis.

Separate reanalysis of each task yielded a significant effect of difficulty in average value tasks ($F(2, 38) = 17.263$, $p < .001$, $\eta_p^2 = .476$) and numerosity tasks ($F(2, 38) = 41.148$, $p < .001$, $\eta_p^2 = .684$). In the former, hard trials (.365) involved significantly more errors than medium (.244) and easy (.206) trials. The same relationship was observed for numerosity trials (easy, medium, and hard trials had .052, .092, and .210 error proportions respectively). No other effects were significant in average value tasks. In numerosity tasks however, we found a significant effect of distractor feature ($F(1, 19) = 13.612$, $p = .002$, $\eta_p^2 = .417$) and its interaction with target feature ($F(1, 19) = 6.050$, $p = .024$, $\eta_p^2 = .242$).

As with RT results, the error rates indicate same-feature distractors caused more errors than different-feature distractors, but closed shapes were more sensitive than open shapes to facilitation and inhibition effects. For linear relationships tasks, the only effect to achieve significance was the three-way interaction between target feature, distractor feature, and difficulty ($F(2, 38) = 4.546$ , $p = .022$, $\eta_p^2 = .193$). However, since the error rates in this task were low, ranging from 1% to 6% of the responses in any given condition, it appears that this complex effect may have resulted from a floor effect in many of the conditions.

### 3.3.6 Discussion

Support for the hypothesized open/closed feature categories was found in two of the three visualizations tasks (numerosity and average value) and only in the single-plot displays with heterogeneous items. An effect of feature category was not evident in the baseline task when homogenous items filled side by side displays. Because visualization tasks require integration of information through ensemble coding, they may not be as sensitive to feature category differences as perceptual tasks when homogeneous items fill the display. Feature category differences, however, were more evident in the single-plot displays because participants were required to discriminate among heterogeneous items; when dealing with visual clutter, same and different feature categories had important influences on numerosity and average value judgments.

Also notable in our findings is the fact that the average value task took much longer and exhibited far more variance than the other two tasks. The disparity in individual differences in response times when added together with the length of the RTs and the error rates could suggest that the perception of average value required a lot of processing time, while the numerosity and perception of linear relationship tasks were more automatic.

After reanalyzing the data for each task, we found a great deal of variability in response latency in average value tasks. Although the trends in this task appeared to be moving in the right directions for the hypothesized difficulty and distractor feature effects, we found no significant main or interaction effects among the experimental conditions.

Numerosity and linear relationship tasks were more interesting: consistent with our hypothesis, same-feature distractor shapes lengthened RTs relative to different-feature distractors. The effect of distractor feature was modulated by target features; closed-feature targets were impacted more drastically – both facilitation by different-feature and interference by same-feature distractors – than open-feature counterparts (see Fig. 5(b)). In particular, closed target shapes with closed-feature distractors exhibited a great deal of interference. Specifically in linear relationship tasks, closed targets were responded to more quickly than open targets for both same and different distractor features in easy trials. However, when the level of difficulty increased to medium the pattern of the interaction changed, and RTs to the closed targets were facilitated when distractors were from a different category rather than the same category as the target. With the hardest difficulty level, the effect of distractor feature was found only with the closed targets (see Fig. 6).

## 4 Conclusions

Taken together, the results of the three studies provide support for the importance of open and closed features as a new category to consider when using symbols in visualization displays. Although experiments 1 and 2 show that exemplars of these categories are processed more quickly in basic perceptual tasks, results from the side-by-side visualization displays seemed to indicate that there was no difference in processing these symbols when presented alone. It was only when the symbols were presented together with other symbols in the same display that differences became evident. With these displays, the presence of distractor symbols from the same feature category interfered more than distractors from different feature categories when participants were focused on processing a given symbol. Interestingly however, the effect of distractor feature category is more evident when processing closed targets than open targets. Also, the effects occurred with numerosity and trend tasks rather than average value. The processing advantage for closed symbols was observed in the easiest task (linear trend) under the easy level of difficulty. With higher levels of difficulty however, that processing advantage depended upon the distractor feature category.

Earlier work on symbol discrimination [5, 16] had focused on computing the perceptual distances between symbols that included the shapes considered in this work, and showed the separation between open and closed shapes. However, our experiments illustrate the complexities and nuances involved in their direct application to visualization design. Symbol selection in visualization design also needs to take into account the underlying data distribution (clutter, overplotting). The results from experiment 3 showed that the symbol features were not a factor in performance on separate-plot displays; these results should extend to plots with well separated clusters in single-plot displays.

More future research about the new open/closed feature category should help unravel some of the complex effects that were shown in our experiments. With a better understanding of the underlying data, this should help us set guidelines for a more automated symbol selection in complex visualization displays.

## Acknowledgments

The authors wish to thank the thoughtful comments of the reviewers. This work was supported in part by a DOE GAANN fellowship and an award from the National Science Foundation (DUE-1245841).

## REFERENCES


[1] L. Chen. Topological structure in visual perception. *Science*, 218(4573):699–700, 1982.

[2] L. Chen. Perceptual organization: To reverse back the inverted (upside-down) question of feature binding. *Visual Cognition*, 8(3-5):287–303, 2001.

[3] L. Chen. The topological approach to perceptual organization. *Visual Cognition*, 12(4):553–637, 2005.

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

[5] Ç. Demiralp, M. S. Bernstein, and J. Heer. Learning perceptual kernels for visualization design. *IEEE transactions on visualization and computer graphics*, 20(12):1933–1942, 2014.

[6] B. A. Eriksen and C. W. Eriksen. Effects of noise letters upon the identification of a target letter in a nonsearch task. *Attention, Perception, & Psychophysics*, 16(1):143–149, 1974.

[7] C. W. Eriksen. The flankers task and response competition: A useful tool for investigating a variety of cognitive problems. *Visual Cognition*, 2(2-3):101–118, 1995.

[8] S. Forster and N. Lavie. High perceptual load makes everybody equal eliminating individual differences in distractibility with load. *Psychological science*, 18(5):377–381, 2007.

[9] S. Franconeri, D. Bemis, and G. Alvarez. Number estimation relies on a set of segmented objects. *Cognition*, 113(1):1–13, 2009.

[10] 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.

[11] J. Heer and M. Bostock. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 203–212. ACM, 2010.

[12] B. Julesz. A brief outline of the texton theory of human vision. *Trends in Neurosciences*, 7(2):41–45, 1984.

[13] N. Lavie. Distracted and confused?: Selective attention under load. *Trends in cognitive sciences*, 9(2):75–82, 2005.

[14] N. Lavie. Attention, distraction, and cognitive control under load. *Current Directions in Psychological Science*, 19(3):143–148, 2010.

[15] S. Lewandowsky and I. Spence. Discriminating strata in scatterplots. *Journal of the American Statistical Association*, 84(407):682–688, 1989.

[16] J. Li, J.-B. Martens, and J. J. van Wijk. A model of symbol size discrimination in scatterplots. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2553–2562. ACM, 2010.

[17] J. Li, J. J. van Wijk, and J.-B. Martens. Evaluation of symbol contrast in scatterplots. In *2009 IEEE Pacific Visualization Symposium*, pp. 97–104. IEEE, 2009.

[18] J. Mackinlay. Automating the design of graphical presentations of relational information. *Acm Transactions On Graphics (Tog)*, 5(2):110–141, 1986.

[19] A. Normand, F. Autin, and J.-C. Croizet. Evaluative pressure overcomes perceptual load effects. *Psychonomic bulletin & review*, 22(3):737–742, 2015.

[20] R. A. Rensink and G. Baldridge. The perception of correlation in scatterplots. In *Computer Graphics Forum*, vol. 29, pp. 1203–1210. Wiley Online Library, 2010.

[21] D. A. Szafir, S. Haroz, M. Gleicher, and S. Franconeri. Four types of ensemble coding in data visualizations. *Journal of vision*, 16(5):11–11, 2016.

[22] A. M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980.

[23] L. Tremmel. The visual separability of plotting symbols in scatterplots. *Journal of Computational and Graphical Statistics*, 4(2):101–112, 1995.

[24] D. Urribarri and S. M. Castro. Prediction of data visibility in two-dimensional scatterplots. *Information Visualization*, 16(2):113–125, 2017.

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

[26] J. M. Wolfe. Guided search 2.0 a revised model of visual search. *Psychonomic bulletin & review*, 1(2):202–238, 1994.