THE ROLE OF SYMBOL CLOSURE IN VISUAL ENCODING: FROM PERCEPTION TO VISUAL ANALYSIS OF SYNTHETIC AND REAL-WORLD DATA

by

David Burlinson

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Computing and Information Systems

Charlotte

2019

Approved by:

_____
Dr. Kalpathi Subramanian

_____
Dr. Paula Goolkasian

_____
Dr. Aidong Lu

_____
Dr. Zachary Wartell

_____
Dr. Erik Saule

ABSTRACT

DAVID BURLINSON. THE ROLE OF SYMBOL CLOSURE IN VISUAL ENCODING: FROM PERCEPTION TO VISUAL ANALYSIS OF SYNTHETIC AND REAL-WORLD DATA. (Under the direction of DR. KALPATHI SUBRAMANIAN)

Symbols and shapes are commonly employed to represent data in visualizations such as scatterplots. Practitioners, scientists, and automated visualization tools are reliant on empirical analyses of visual encoding strategies, taking into account the influence of data characteristics and visual features, to produce effective charts and graphs. In pursuit of this goal, the following questions were considered: (1) Are shapes that share bounded or unbounded structures members of a feature category that influences how they are processed and perceived? (2) Do open/closed feature categories differ in how they are processed? (3) How do shape encodings interact with characteristics of the data and types of tasks in visualization contexts? In this work I investigated the implications of perceptual categories of commonly used charting symbols sharing bounded (closed) or unbounded (open) structures using a series of experiments from low-level attentional allocation to high-level task performance in ensemble displays. Flanker and same/different tasks were used to explore the perceived similarity among open and closed symbols; participants responded to closed symbols more quickly and accurately, and discriminations within a feature category took longer than between categories, supporting the categorical distinctiveness of symbols with and without boundaries. Three relative judgment tasks (mean position, numerosity, and linear correlation) were implemented using exemplars of these

shape categories as encodings in multiclass scatterplots in order to test whether performance differences due to categorical features would subsume differences among symbols. Each task was reliably harder when marks were encoded with shapes sharing open or closed features, and conditions with closed targets received more influence from distractor features, i.e. both facilitation with different-featured distractors and inhibition with same-featured distractors. A follow-up study with a larger symbol palette and systematic variation of the level of overlap among marks in numerosity and linear correlation tasks found similar results; open target sets took significantly longer and induced significantly more errors than closed targets, regardless of overlap or distractor features. The final study incorporated more realistic displays, with data sampled from the Toxics Release Inventory, a dataset on industrial usage of toxic chemicals, and chart axes and labels. Participant performance on relative judgment tasks differed across pairs of symbols used as mark encodings, but pairs sharing open or closed bounding features always took longer than pairs differing in that feature, and displays with closed targets were always faster and less erroneous than displays with more numerous open targets, comporting with the findings from the previous studies. Overall, the categorical relationship between open and closed symbols and the perceptual preference for closed symbols was clear in all the experiments, and persisted across relative judgment tasks, when overlap among marks was systematically varied, and with palettes of symbols containing different exemplars from both feature categories. This sequence of results has implications for visualization designs in which shapes are used as categorical encodings, and also poses new questions for the vision science and visualization communities. Further studies can model the role

of shape encodings with a wider variety of data types and distributions, in tandem with more extensive tasks, and supporting more comprehensive encoding strategies involving redundant visual channels. Future work will also be required to understand the mechanisms underpinning shape perception and to explain the apparent salience of bounded symbols.

ACKNOWLEDGMENTS

I am incredibly grateful for the financial and academic support I received through the Lucille P. and Edward C. Giles Dissertation-Year Graduate Fellowship and the Graduate Assistance in Areas of National Need Fellowship, and to all the faculty and staff involved with organizing and running those fellowships at The University of North Carolina at Charlotte.

I am immeasurably thankful to Dr. Kalpathi Subramanian and Dr. Paula Goolkasian for their infinite patience, guidance and supervision in working closely with me to develop my research questions, shape my methodologies, strengthen my writing, and bolster my skills and intuitions in my academic pursuits.

I am indebted to Dr. Erik Saule, Dr. Zachary Wartell, Dr. Aidong Lu, Dr. Jamie Payton, and Dr. Min Shin for spending their time and energy giving me feedback and helping me strengthen my research and presentation skills through the Ph.D. program, and am thankful to all the other faculty at UNC Charlotte who taught me and all the students who worked with me, learned from me, and participated in my user studies over the years. I am also deeply grateful to Dr. Marguerite Doman and Dr. William Thacker for pushing me and guiding me throughout my undergraduate studies at Winthrop University, and Lisa Thompkins for encouraging my earliest fascination with computer science at Carolina Forest High School.

I cannot overstate the valuable feedback from and discussions with Dr. Steven Franconeri, Dr. Danielle Szafir, Dr. Remco Chang, Dr. Lane Harrison, Dr. Andrew Besmer, Dr. Michael Whitney, and Dr. Scott Heggen, all of whom shared wisdom

and advice on my research and career goals. There are innumerable wonderful and brilliant people who I have met at various conferences and workshops and through my higher education at Winthrop University and UNC Charlotte, all of whom have impacted my trajectory in intellectual space and through life in their own ways. I am lucky to have shared thoughts, discourse, and laughter with so many friends, colleagues, and mentors.

Finally, I am thankful to all my close friends, my parents, my siblings, my extended family, and my partner Catherine for supporting me, encouraging me, and keeping me grounded and sane throughout this excellent journey. I love and appreciate all of you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

1.1 Visualization Design

Visualization is the study and art of communicating in a visual medium. Data is represented with visual stimuli on a page or a screen so that an observer can interpret relevant characteristics of the data using the affordances of the human visual system. Sometimes those observers are business leaders, politicians, or military commanders, and the decisions they make bear weighty implications for economic prosperity, the social landscape, and human life. Scientists and domain experts need to understand complex relationships in large volumes of data, and they often rely on visual displays to present the results of their experiments, shape their interpretations, and inform future studies and analyses. In other cases, visualizations are designed to communicate information to laypeople and students, guiding their comprehension and learning in the classroom and everyday life. While all these cases entail specific considerations and nuance, each visualization used must represent its underlying data as faithfully and robustly as possible to support observers in whatever tasks they are pursuing.

Design decisions in this space include the visual primitives, such as the shapes, marks, and colors used; the visual metaphors, from simple graphs like bar charts, pie chart, and scatterplots, to more esoteric stream graphs and chord graphs; and more complex analytics systems coordinating multiple dynamically linked views. A

great number of books and papers have been written on the creation of information-rich charts and diagrams. Bertin's seminal Semiology of Graphics [9] was one of the earliest methodological approaches to mapping data into visual forms, and Tufte's Envisioning Information [122] and Munzner's Visualization Analysis and Design [88] are also indispensable resources for visualization students and practitioners, organizing key conceptual components and design factors to take into account. Rensink's On the Prospects for a Science of Visualization [95] makes a compelling argument for a rigorous, quantitative science of visualization, backed by perception and vision science paradigms and operating in parallel to a more qualitative, design-focused approach.

Two general metrics are used to assess a visualization: effectiveness – how well the capabilities of the display medium and human visual and intellectual capabilities are harnessed – and expressiveness – how well the underlying information is represented with graphical primitives and their combinations [84]. Much work has been done to bolster decision-making in creating and analyzing effective visualization strategies on the back of these metrics. Cleveland et. al. [28] demonstrated the influence of axis scale on perceptual inference of correlation in scatterplot displays, showing just how easily a chart can mislead an observer, hampering both expressiveness and effectiveness. Brath [16] presented a variety of metrics related to designing effective visualizations in a 3D context, many of which are relevant to visualizations in general, such as number of data points and dimensionality, identifiability and occlusion of points, and cognitive overhead.

If brevity is the soul of wit, then perhaps simple, functional charts are the heart of visual communication; indeed many experts have decried artistic embellishments,

or 'chart junk' [123]. However, evidence suggests that people may find these adornments make charts more appealing and memorable in the long run [7]. Clearly there is no one-size-fits-all solution in such a large design space, so how can researchers and practitioners investigate better and worse decisions? In some cases, new visual idioms are required. When presented with the challenge of visualizing large quantities of hierarchical information, like directory structures in a file system, Johnson and Schneiderman [59] created Tree-maps to take better advantage of spatial judgments and make maximal use of space. In an effort to visualize relationships among variables in an otherwise unintuitive high-dimensional space, Inselberg [57] popularized parallel coordinates. Tree-maps and parallel coordinates have now become widely used charting idioms. In other cases, it is valuable to return to more common chart styles and run careful experiments to understand how people view and interpret them. Skau and Kosara [109] examined the humble pie chart and systematically isolated arc length, center angle, and slice area, finding that, contrary to all expectations, internal angle is their least important feature. Whether it is appropriate to revisit and study existing charts or design exciting new ones, the deeper one digs the more critical it becomes to consider the foundations of human visual capacities and cognitive faculties to understand how people see a chart and reason about the information it displays.

1.2 Vision

The human visual system is a complex structure with numerous functional subsystems that all combine to produce the phenomenological experience of sight. At the lowest level, light signals with wavelengths between 370 and 730 nanometers are fo-

cused onto the eye's retina by a series of precise contractions of muscle groups around each eye's lens. These signals are transduced into neuronal action potentials and fed into the visual cortex, which is roughly divided into topologically organized functional areas, each of which synthesizes more and more exhaustive representations of a visual scene. Local spatial frequencies and orientations cause activation in retinotopically mapped neurons' receptive fields, whose signals are combined to produce inference of edge boundaries spanning larger regions of the visual field. Information from the primary visual areas are streamed to brain regions responsible for object and categorical recognition, and spatial awareness and movement tracking.

The fovea, the high-acuity region at the center of the visual field, is associated with an outsized proportion of neurons in early vision areas in the brain, and vision researchers often design experiments so as to keep relevant stimuli within foveal vision when measuring response latency as a dependent variable. In more natural viewing conditions, the foveal region is shifted rapidly and automatically between salient locations in the visual field. This saccadic motion is made several times per second, and visual information is suppressed for its duration. Peripheral vision, the rest of the visual field outside foveal fixation, is marked by significantly reduced clarity and susceptibility to crowding effects; research on peripheral vision [101, 100, 113] aims to tune more general models of vision to account for these factors.

1.2.1 Perceptual Organization

Many theoretical models of vision [60, 134, 119, 118, 127] suggest that features of objects in the visual field are segmented and processed well before the influence of

attentional focus then re-organized into meaningful units of perception. Such 'pre-attentive' features include basic 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.

Perceptual organization of visual information from low-level preattentive features into topological figures and grounds is mediated in part by closure, or at least perceived closure, of detected edges [65]. Some models of vision treat closure itself as a feature [119], while alternate theories have purported that line segments, crossings, and endings are the more meaningful signals [60, 61]. In either case, it is well established that detection and recognition of objects, which are themselves composed of various low-level features, rely on the relative salience of the items and the expectations and top-down influences of the observer [134].

Chen [24, 26, 25] has argued for an opposing interpretation of early vision and perceptual organization involving topological characteristics such as connectedness and closedness. His models suggest a global-to-local topological model in perception of shapes, whereby 'wholes are coded prior to perceptual analysis of their separable properties or parts.' Evidence from his studies support potential explanations for the relative distinctiveness of simple shapes and symbols, which can differ in terms of these basic topological properties.

Not only do the bottom-up and topological characteristics of the visual field influence what is seen at a given point in time, but expectations and goals exert a top-down influence to refine the preferential selection of visual characteristics in any given moment [134].

1.2.2 Ensemble Coding

In addition to rapid processing of objects and their distinctive features, the visual system is tuned to quickly assess statistical summaries of certain types of information across spatially distributed objects, including lower-level features such as average size and orientation, and higher-level features such as the mean emotion of a crowd of faces or the distributional structure of a set of data points in a visualization [3, 49, 129, 55, 115, 124]. Human reasoning and inference are underpinned by assumptions regarding the statistical regularities of the world around us, so it is no surprise that our visual system has developed powerful mechanisms to exploit that regularity to extract meaning quickly and with a surprising degree of accuracy. Szafir and colleagues [115] enumerate four areas where ensemble coding is relevant for tasks in visualization displays, including identification, summarization, segmentation, and structure estimation, but they are quick to note that this is an area ripe for investigation, and crucially, collaboration among visualization and visual perception researchers.

This coarse overview outlines important peaks on the landscape of visual perception literature, but elides many nuances that can be taken into account when thinking about the design of effective visual displays. Visual illusions such as change blindness, dominance of particular visual channels like hue, the role of task and cognitive load on attentional allocation, and rapid extraction of statistical properties in foveal vision, in the periphery, or across the entire visual field all play a role in a holistic understanding of human vision and its application in visualization contexts.

1.2.3 Marks and Channels

One area of overlap between visual perception research and visualization design is the study of marks and channels, the visual primitives available for use in representing information. Marks include basic graphical elements like points, lines, areas, and volumes, and channels describe how marks can vary, such as position, size, shape, orientation, color [88]. In one example, Cleveland and McGill [29] recommended an ordering of efficacy, later validated in crowdsourcing studies by Heer and Bostock [56], for encoding categorical relationships: spatial region (related to Gestalt grouping [127]) is the most effective, and hue, shared direction of motion, and shape (all related to Gestalt similarity [127]) follow with less and less effectiveness. Making good design choices with respect to categorical attributes allows effective grouping of marks in a visual display so that a viewer can selectively attend to one class or another, and make relative judgments among classes in aggregate across the display. This is concretized in multi-class scatterplots, where marks from multiple categorical classes are differentiated by careful selection of an identity channel encoding, and high-level summary statistics, such as variance and clusters, can be computed and compared with little effort from an observer.

Channels differ in terms of how discriminable exemplars are, and how many different exemplars a viewer can readily distinguish [88]. Channels can also be utilized in tandem to varying degrees of success, roughly characterized in terms of their separability. Channels such as hue and position are reasonably independent, but pairings of size with shape or color have an influence above and beyond differences in either one

of the channels [110, 114]; the relationship among these channels have been the focus of recent scrutiny. Redundant encoding, combinations of channels such as shape and hue, was sanctioned with some caveats by Tufte [122], and recent studies [90] support the notion that it can be useful in segmenting and grouping visual items.

1.3 Computational Modeling

Building upon the theories of salience and allocation of attention, researchers have constructed computational models of vision attempting to predict fixations and saccades in natural scenes [58, 50, 51]. Computational models of gaze prediction have also been examined for graphical and statistical displays. Harrison and colleagues [52] demonstrated that current models of bottom-up salience can predict fixations at marginally above-chance rates in statistical graphs, but that incorporating top-down factors, while difficult, is crucial for increasing accuracy.

Recent work by Tsotsos and colleagues [133, 121] and Peters and Itti [94] attempted to model temporal sequences of fixations in a way that incorporates top-down influences, including task goals and previous fixations, moving beyond saliency maps and bottom-up features alone. It is not yet clear how these types systems will fare in artificial contexts such as visualizations. Harrison et. al. point out that models trained on natural scenes tend to give significant weight to chart labels and text rather than the marks and channels encoding the data in the chart, suggesting that the spatial frequencies and requirements differ significantly between natural and artificial displays [52]. In addition, the top-down influences brought to bear in visual analysis contexts are complex, and may be hard to capture parsimoniously.

1.4 Automated Visualization Design and Encoding

A long-term goal among visualization researchers and practitioners is to automate larger portions of the pipeline from raw data to graphical representations in order to incrementally encode state-of-the-art wisdom for combinations of tasks, visual channels, and data types [84, 85]. Bertini et. al. [12] and Tatu et. al. [116] described the difficulty in selecting, automatically or otherwise, the most appropriate two-dimensional projections when dealing with high-dimensional data, and explored techniques and metrics to pare down that decision space using perceptual judgments of scatterplot clusters. Tools such as Pixnostics [106], Voyager [135], and SeekAView [66] aim to streamline the process of selecting relevant subsets, projections, and visual encodings of data by permuting the parameter space and automatically suggesting the most relevant views. The increasing intrigue in data science and visualization has propelled the rise of web-based toolkits such as D3 [15] and Vega [105], and statistical packages like R's ggplot2 [131], which facilitate the creation of charts and graphs with a bit of coding expertise, and other commercial software like Tableau and PowerBI support flexible data exploration and creation of visualizations through nice interfaces. The utility of these tools is contingent upon user expertise and automated heuristics to produce effective and expressive visual displays.

Research supporting the automation of visualization design aims to incorporate tasks and models of visual encoding strategies to refine the heuristics underpinning such systems and provide general guidelines for practitioners. For example, Amar and colleagues [2] constructed a taxonomy of low-level analysis tasks for arbitrary data,

and Sarikaya and Gleicher [104] reviewed the literature on scatterplots and built a framework of tasks and designs to standardize the language and explore tradeoffs in that particular type of chart. Saket and colleagues [103] examined how well a set of simple visualization types, including line, bar, and pie charts, and scatterplots, supported a set of tasks, including cluster and correlation detection, characterizing a distribution, assessing outliers and extremum, and so on. In similar fashion, Kim and Heer [64] explored four low-level analysis tasks and a variety of data distributions and encoding strategies, demonstrating an interplay of visual channels such as color and size, and a significant role of data density or overplotting in determining appropriate visualization approaches. Szafir and Smart built models to begin accounting for the interaction among visual channels in visualization displays, such as the effect of chart area on color perception [114] and the asymmetric interaction of shape, size, and color channels [110]. Modeling these perceptual factors, and the relationship between visual characteristics and task needs, are crucial pieces of the larger puzzle of automated visualization design, and work is underway to bridge from high-level analysis goals to low-level sub-tasks and encodings [67].

Perception of correlation is a specific area that has received much thought, with multiple studies assessing correlation judgments across visual encodings [53, 62, 75, 28], and particular focus by Rensink on modeling perception of correlation using psychophysical laws [97] and positing the nature of this regularity in terms of ensemble judgments of entropy [96]. Rensink has been a leading proponent of studying simple visualizations in controlled environments to shed light on the underlying perceptual and cognitive mechanisms we use to interpret them, in addition to his more general

call for a rigorous science of visualization mentioned earlier.

Automated visualization design and encoding have already inspired powerful tools to explore and present data, but a great deal of work needs to be done to fully take into account the effects and interactions of separable and integral channels, redundant encoding of marks, the influence of data and various distributional effects, and tasks and high-level analysis goals. It is my goal to help inform this area of inquiry and contribute meaningfully to the growing body of literature on perceptual factors in visualization design.

1.4.1 Overplotting

Directly visualizing large datasets becomes challenging as the number of data points increases, and is next to impossible when the data size approaches and exceeds the number of pixels in the display. Vis designers will run into issues even more quickly when multiple categorical variables are required, because distinct marks will generally require multiple pixels. One particular issue that arises in these circumstances is overdraw or overplotting, in which data points are drawn closely enough together that some are obscured or partially hidden [16], potentially leading to biases in analysis tasks [31].

A number of techniques have been proposed to address this issue, including filtering, reducing, or binning the data, augmenting charts with marginal distributions along the primary axes, and creating new charts by aggregating data into contour plots or splatterplots. Sarikaya and Gleicher [104] provide a framework for mapping various tasks and data characteristics into a space of scatterplot design options. Ellis and Dix

[36] enumerate a taxonomy of clutter reduction techniques, Cottam et. al. [30] provide an efficient implementation of a variety of approaches for combatting overplotting, Chen et. al. [23] have preliminary work on utilizing animation to combat overdraw, and Keim et. al. [63] present generalized scatterplots as a solution to the overdraw problem.

Aside from circumstances in which aggregations and new charts are required, there are still better and worse design decisions for combatting overdraw. Conventional wisdom and various charting software packages suggest a few techniques – avoid filled shapes because they occupy more space (default in Tableau and R's ggplot2 package), use alpha blending to mitigate occlusion [87], reduce the size of the marks, and so on. These approaches can be effective but do have their downsides, and more work is required to fully model the design space for designers and automated visualization tools. Smart recently demonstrated how size and color are asymmetrically influenced by symbol choice, and built models to predict their perceptibility in scatterplot displays, but crucially, as with much other work in this domain, did not account for overlap among symbols [110].

Bertini and Santucci developed models for quantifying visibility in scatterplots with large sets of points and improving visibility by sampling points while preserving overall densities and distributions [10, 11]. Their published methods are designed for single-pixel encodings though, and they point out a number of potential difficulties in extending results to shape encodings. Urribarri and Castro built upon those models by defining a visibility index based on the percentage of encoded data symbols not fully covered by other symbols, and explored semiautomatic recommendations for

glyph size [125]. However, they only utilized filled square glyph areas, leaving aside differences in features and space consumed by symbols within that square bounding region in order to focus on the general question of relative size differences.

Few [41] recommends using shapes that are not shaped like containers, such as X instead of a circle. Most other discussions on overplotting only cursorily consider the shapes or glyphs involved, generally dismissing them as uninformative for the broader question of overplotting in very large datasets. Some data, task, and display constraints can certainly preclude the use of individual point encodings, particularly very large numbers of points and increasing numbers of classes [104]. Point encodings can be useful for many other combinations of these constraints though, and the work in this document, particularly in later chapters, is designed to explore the influence of overplotting on symbols with open and closed features.

1.5 Color Space

There are few concepts more derided among visualization researchers than the rainbow color map [14, 99, 80]. Color is one of the most dominant visual channels [29], and many visualizations use color to represent continuous or categorical data, so what is so wrong with the rainbow mapping? In short, the rainbow color map does not comport with a linear perception of the colors it contains. In other words, two points an arbitrary distance apart on the rainbow map are not guaranteed to seem as similar or dissimilar perceptually as any other two equidistant points; this can be misleading in a visualization that claims to represent continuous data.

Multiple alternatives have been built to support visualization designers [54, 1], and

an entire color space – CIELAB – was developed to account for perceptual differences in lightness, green/red, and blue/yellow values [83]. Unlike the rainbow color map, colors sampled from uniform steps along each of those three dimensions or from arbitrary locations in CIELAB's continuous color space will have predictable perceptual qualities, allowing visualization designers to map differences in data values to perceived color differences with a high degree of accuracy. Stone and colleagues [112] modeled CIELAB color differences as a function of size for uniform display swatches and found that smaller sizes require larger differences in color to achieve the same level of perceptual similarity. Szafir [114] extended this model to account for noticeable color differences across various marks common to visualizations, such as bars, lines, and points, and worked with Smart [110] to measure the interaction with multiple categories of filled and unfilled symbols. This type of modeling has immediate and wide-reaching importance for visualization designers and automated visual encoding systems, as the continuous nature of these color spaces and their relative stability across viewers (not accounting for color deficiencies, which can affect a non-trivial proportion of individuals) support computationally inexpensive measurements of perceptual color similarity and difference.

1.6 Shape Space

In contrast to the brief discussion of color perception above, reasoning about symbols and shapes is not nearly as straightforward. Nobody has contributed a 'shape space' comparable to CIELAB color space, nor does there appear to be a nice continuous representation of symbols from which to sample. Some attempts to quantitatively

analyze general shapes and symbols have focused on mathematical and gestalt representations, [4, 32, 81] and there is a degree of overlap with the study of the complexity of symbols in language [93, 22, 128]. One of the common findings across these investigations is that there is no single straightforward measure; instead symbol perception is influenced by a number of complementary, perhaps overlapping mechanisms at different stages of visual processing. Research from the vision sciences community has uncovered an illusory size effect, in which simple objects missing parts of their boundaries are reliably seen as larger than the same fully bounded shapes [86]. Further, summary statistic models of peripheral vision predict sensitivity to shapes with open and closed features, with differing degrees of underestimation of the numerosity of points for both shape types in comparison to normal dot displays [6].

Another interesting domain is that of shape skeletons, which are theorized to underpin biological form perception and similarity judgments in two- and three-dimensional contexts [117]. Feldman and Singh [40] proposed a probabilistic bayesian approach to compute a shape's skeletons, sidestepping some of the issues in previous deterministic methods, which were sensitive to local perturbations. Firestone and Scholl [42] demonstrated the primacy of shape skeletons with an interesting crowdsourced approach - people were instructed to tap a shape anywhere they liked, and the distribution of touches aligned with the medial-axis skeleton in a variety of shapes.

Despite the complexities involved in representing their complexity and similarity, symbols have long been utilized in visualizations, particularly as visual encodings for marks in scatterplot displays, supporting discrimination and comparison of categorical relationships. Some researchers discount the utility of shapes for visualizations on

the basis that they are only useful for certain multivariate glyph-based approaches or categorical encodings, where they are outperformed by differences in hue [64, 91, 46]. Some notable exceptions include attempts to understand perceptual orderings of shapes [27], similarity metrics from different shapes [33, 77] and in sets of glyphs [73], and incorporation of shape and size into models of optimal color differences [114]. Commonalities among the palettes of shapes have tended toward the use of basic geometric and radially symmetric elements such as circles, squares, plus signs, and other simplistic arrangements of line segments. More complex combinations of these primitives have been used when larger collections of symbols have been desirable. The chosen shapes generally don't confer meaning in the semiotic sense, but rather serve as distinctive sets of categorical encodings so viewers can readily distinguish among points related to particular variables of interest. Glyph-based visualization is an interesting tangential domain in which symbols and their attributes are more closely tied to particular features and meaning of the underlying data. See Borgo et. al. [13] for a state of the art report on glyphs, and Legg et. al. [72], who proposed a quasi-Hamming distance to quantify the perceptual similarity of glyph sets and explored methods for creating and validating these measures as well as their application for icons in file systems.

Work in the statistical charting community built upon emerging models of vision such as Texton theory [60] and Feature Integration Theory [118, 119] to evaluate charting symbols with differing curvature, fill, and line endings [120]. Lewandowsky and Spence [74] investigated visual encoding strategies involving shape, color, amount of fill, letters, and oriented lines using relative correlation judgments between multiple

categorical strata in scatterplot displays. 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.

Experiments by Demiralp et. al. [33] and Li et. al. [77, 76] focused in on the relative discriminability of simple shapes. The former constructed normalized kernels of pairwise similarity among common symbols, sizes, and colors using subjective tasks including Likert scales, triplet matching, and spatial alignment. The latter gathered quantitative data from a series of tasks and modeled an internal separation space using a modified Power Law [111] and multidimensional scaling; they also provided evidence for the distinctiveness of bounded geometric shapes and shapes composed of radially symmetric line segments.

For visualization practitioners, it may well be sufficient to provide a black box, where one can ask for a set of categorical symbols, a range of ordinal shapes, or a set of shapes x units apart in some perceptual space. For vision scientists, and for the more fundamentally curious, I suspect that black box would be too simple even if it were to exist. In any case, the wide difference between the approaches to the same fundamental questions underscores the complexity of this topic, and there is more work to be done to synthesize knowledge on shapes as such and as encoding strategies.

1.7 Contributions

There are better and worse ways to create scatterplots, and practitioners are better served relying on well-supported recommendations rather than artistic intuitions or unverified norms. The visual attributes used to encode marks vary in their effectiveness, particularly in contexts where the amount of information does not preclude a one-to-one mapping from data to marks and where a goal of the plot involves attending to specific marks, as well as supporting summary judgments about the overall data features and distribution. Research in this domain illustrates the primacy of color encodings. While prior work addresses additional encodings such as shape and size, there remain a number of unanswered questions regarding the influence and interaction of these elements, especially in conjunction with analysis tasks. The shapes commonly used in scatterplot visualizations tend to be simple, symmetric, and relatively distinguishable from each other. Studies indicate that such shapes tend to be clustered into filled or unfilled bounded geometric shapes and unbounded collections of line segments, and that this categorization both reflects their distinctiveness in terms of low-level visual features and provides a useful starting point for encoding categorical variables in practice.

The overarching goal of the work detailed in this document has been to find empirical evidence for design decisions related to the shapes and glyphs used to carry information and convey categorical distinctions in visualization contexts. The studies described mesh paradigms from the vision science community and tasks and displays from the visualization community in an attempt to further that aim and expand the

current state of knowledge of categorical shape encoding, support more robust automated visual encoding systems, and help practitioners develop more effective and expressive charts.

My investigations have explored the role of the topological or gestalt closure for shape encodings in multiclass scatterplot displays, and the findings I will expound throughout the rest of this document comport with many of the previously mentioned findings, such as the categorical distinctiveness of polygonal and texton-like shapes [19, 77, 76, 120] and the seeming perceptual preference for processing symbols with closed boundaries, which have been shown to be discriminated [35] and recognized [45] more easily than their open counterparts in other contexts. I also show that pairwise measures of similarity among symbols do not fully predict the performance of those symbols in relative judgment tasks in ensemble displays, and that symbols do not make a difference on task performance when judgments are made in side-by-side plots.

The successive chapters follow the temporal and investigative sequence undertaken during my studies at UNC Charlotte.

In chapters 2, 3, and 4, I describe methods from psychological sciences for examining the influence of symbol discriminability from the perspective of perceptual awareness and attentional allocation, including spatial cuing, and Flanker and Same-Different tasks.

Chapter 2 details a preliminary attempt to investigate symbols using liminal perception of spatial cues. Color cues had been shown to produce benefits when sharing a primed color and appearing in the same location as a target singleton, and produce

costs when differing in that color [69]. I was not successful in replicating the results of that earlier work though, so was unable to explore whether the same results would hold true for shape cues.

Chapter 3 describes an alternate approach to the questions from the liminal perception study, with a flanker task presenting target shapes alongside distractor shapes with varying featural similarities and differences. Compatibility effects arose, with differences in performance between compatible trials (in which distractors and targets were the same shape) and incompatible trials (in which targets and distractors differed), and these effects varied based on the features of symbols primed in the attentional set for each block of trials. Blocks in which both primed targets shared boundary closure (or both lacked it) were not mediated by the systematic varying of cognitive load, supporting the categorical distinctiveness of shapes with a bounding edge (closed shapes) and unbounded shapes (open shapes). In addition, closed shapes were processed more quickly and accurately, and discriminations within a category took longer than between categories.

Chapter 4 discusses a same/different task, a more straightforward procedure used to extend the symbol palette and test the results from the Flanker study. Participants indicated whether the symbols present in each display were the same or not, with the delay in reaching that determination varying depending on the relative similarity among the symbols and their features. Results in this experiment supported the findings from the Flanker study, with closed shapes outperforming open ones, and differences between open and closed symbols significantly outperforming differences within either feature category.

Chapter 5 presents a first pass at extending the results from low level paradigms into visualization displays and analysis tasks involving higher-level judgments. Participants performed tasks involving average value, numerosity, and linear relationship judgments among multiple categorical variables in synthetic displays. The results lend further credence to the relative effects of using open and closed shapes to encode marks, and find that closed symbols are more influenced by distractor features. Performance also varied significantly among the tasks: judgments of linear trends and numerosity were much more automatic than judgments of the average position of sets of symbols.

Chapter 6 represents a continuation of the visual summary task approach from chapter 5 with an increased symbol palette and a particular focus on varying amounts of overlap among symbols in synthetic displays. Continued support was found for the categorical nature of bounded and unbounded symbols and for the processing preference of closed symbols. Response latency increased linearly and error rates increased quadratically with the proportion of symbols overlapping other symbols.

Chapter 7 presents the final experiment, in which a real-world dataset of toxic chemical usage was sampled to produce realistic displays with chart axes and a larger number of data points than the studies in the preceding chapters. Participants made relative numerosity judgments supported by various shape encodings from a simplified symbol palette. Support was found for the processing preference of closed symbols, and mixed support was found for the role of bounded and unbounded feature categories. The variation in performance between pairs of symbol encodings has implications for the limits of pairwise similarity measures. Some symbol pairs

comported with subjective measures of pairwise similarity in the literature [33], but other pairs did not adhere to differences predicted in that work, suggesting that ensemble mechanisms and task-based constraints exert further influence on perceptual judgments.

Chapter 8 summarizes the findings from all of the experiments in more detail and addresses methodological considerations among the studies. Avenues for future refinement of this work and areas where additional research will be valuable are discussed.

CHAPTER 2: LIMINAL PERCEPTION STUDY

Prior research on the limits of unconscious processing has convincingly shown that subliminal primes are capable of affecting responses to subsequent targets in timed visual search exercises. Lamy et al. [69, 68] showed that there is a dissociation between attention and conscious perception, and found that capture of spatial attention is largely independent of conscious perception of a subliminal prime. Their studies indicated that subliminal color singletons strongly captured attention when their color matched the target-defining feature, providing evidence for same-location benefit when the cue and target appeared in the same location on the screen. On the other hand, when the subliminal cue's color differed from the target, it did not capture attention, yet still incurred a same-location cost when appearing in the same location as the target. They concluded that this reflects the temporal cost of updating an object's episodic representation in visual memory, suggesting that conscious perception of a visual object may be required in order to create an object file for that object.

The plan for this first study was to validate Lamy and colleagues' results by performing a similar set of trials, and then extend their research on updating object files and same-location cost and benefit by exploring some simple shape features. I set out to test whether cues and targets comprised of bounded and unbounded shapes would mirror the temporal cost or benefits found with color singletons. For exam-

ple, whether circles would provide effective cues for colocated target circles, whether squares would incur same-location costs when cuing plus signs, and how effects such as these would interact with participants' conscious awareness.

The hypotheses for this study were as follows:

H1 *Color cues appearing in the same position as a target will produce same-location costs when differing from the cued target color and same-location benefits when sharing the target color*

H2 *Shape cues will follow the same trend as colors; shapes appearing in the same position as a target will produce same-location costs when differing in structure from the cued target symbol, and same-location benefits when sharing structure*

2.1 Methodology

2.1.1 Participants

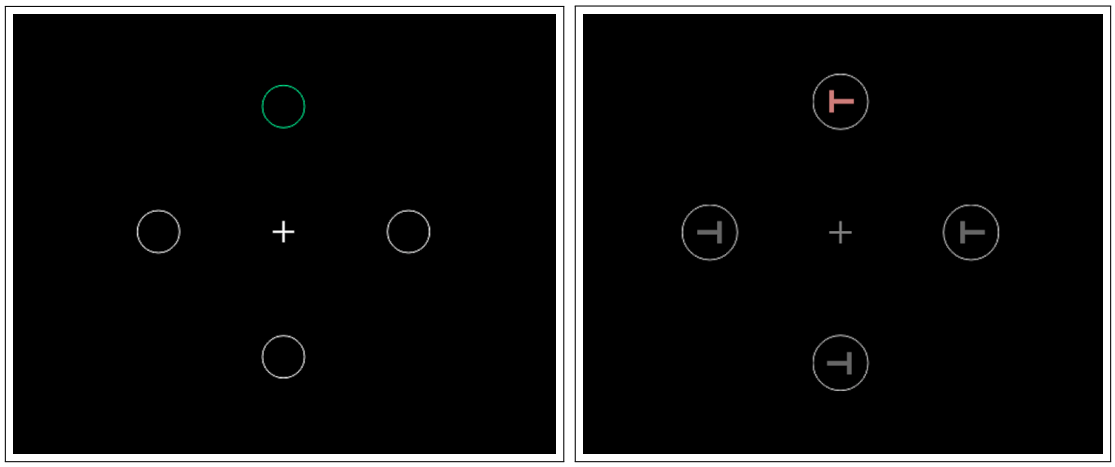
Thirty-five participants were recruited from the UNCC SONA system, and awarded one research credit for 35-45 minutes of their time. Participants were at least 18 years old with 20/20 or corrected to 20/20 vision and no history of visual impairment.

2.1.2 Stimulus Materials

The visual stimuli were presented on an iMac computer with a 17" flat screen LCD monitor. Stimuli were created using Javascript and SVG on the same computer to guarantee uniform spacing, positioning, and color effects. Stimulus presentation and data collection were controlled by SuperLab 4.0.

Within each trial, participants were sequentially presented with a fixation display,

a cue display, the same fixation display, and then a target display. The first fixation display was shown until each trial began, the cue was displayed for 40 ms, the inter-stimulus fixation display was shown for 110 ms, then the target display was presented for 150 ms. The fixation display had a black background with a white fixation cross at the center to orient the participant's gaze (See figure 2.1).



(a) (b)

Figure 2.1: (a) Cue display with green color cue and (b) target display with red target from the Liminal Perception study. For a more comprehensive set of stimuli, refer to Appendix B.1 and B.2

In the first set of trials, the cue and target displays featured unfilled white circles on a black background at one of four equidistant locations around the center of the visual field. For each of the displays, one of the circles changed border thickness and color. For the second set of trials, the cue and target displays featured white dots at the equidistant locations around the center of the visual field. For each of the displays, one of the dot locations was replaced with either a filled or unfilled shape. The shapes chosen for this study were diamonds, triangles, squares, and circles.

In part one, half of the color group had either red or green, and the other half had

either yellow or blue as the possible colors for cues and targets in a given trial. For part two, the shape group was to be divided evenly between the four shapes for their cues and targets so that each person would have only one shape to look for in all trials to determine if it is open or closed.

The locations of the cues and targets were distributed as evenly as possible between the four possible locations. Cues were present in 80% of trials and targets were present in 100% of trials.

Each participant's 400 trials were split evenly among the four experimental conditions yielding 100 trials within each of the following conditions: 1) Same cue/target color - same target/cue location 2) Same cue/target color - different target/cue location 3) Different cue/target color - same target/cue location 4) Different cue/target color - different target/cue location

2.1.3 Procedure

Participants were positioned 60 cm from the computer screen in a well-lit room. Each participant read and signed an informed consent document and learned which color possibilities they would be presented with (Red/Green, Yellow/Blue). Participants went through 40 practice trials at the beginning of the study to become familiar with the procedure and the stimuli. Each participant then performed 400 trials. Participants pressed either the f or j key to indicate whether the target was red or green, or yellow or blue, respectively. They were instructed to respond as quickly and accurately as possible.

Once the target response was logged for each trial, participants pressed the *space*

key to indicate if they were aware of having seen the color cue. Once they completed each trial, they pressed the *enter* key to begin the next trial.

2.2 Analysis and Results

Overall, eight participants were excluded from the analyses due to invalid responses, high false positive rates, or high cue awareness. Participants reported not seeing the cue on 85% of cue- absent trials and on 42% of cue-present trials. Trials with errors (9.4% of all trials) and RT outliers (3.1% of all trials with correct responses) were excluded from all analyses. Preliminary analyses showed no effect involving target color, and the data were therefore collapsed across target-color conditions. The results conformed to our predictions: the three-way interaction among cue awareness, cue color (same as target vs. different from target), and target location (same as cue vs. different from cue) was significant, ($F(4, 45) = 13.74$, $p < .0001$).

Follow-up analyses on location effects showed that on same-color trials, attentional capture was as large when the cue had been consciously perceived (M = 104 ms, SE = 11 ms), ($F(1, 45) = 75.66$, $p < .0001$, Cohen's $d = 2.55$), as when it had been invisible (M = 104 ms, SE = 12 ms), ($F(1, 45) = 84.25$, $p < .0001$, Cohen's $d = 2.41$). There was no difference between these two location effects ($F < 1$). On different-color trials, there was a significant same location benefit when the cue was invisible (M = 43 ms, SE = 13 ms), ($F(1, 45) = 9.95$, $p < .003$, Cohen's $d = 0.90$), but not when the cue was consciously perceived (M = 8 ms, SE = 13 ms), ($F < 1$, Cohen's $d = 0.17$). The difference between these two effects approached significance, ($F(1, 45) = 3.92$, $p < .055$). Similar analyses on error rates showed only a significant main effect

of location, ($F(1, 12) = 28.30$, $p < .0001$), with higher accuracy on same- than on different-location trials. No other effect approached significance.

RTs to the target were again slower following visible cues (M = 793 ms, SE = 34 ms) than following invisible cues (M = 723ms, SE = 34ms), ($F(1,12) = 11.29$, $p < .0001$), but they were more accurate (visible cues: M = 88.5% correct, SE = 1.8%; invisible cues: M = 85.86% correct, SE = 1.6%), ($F(1, 11) = 5.01$, $p < .05$), which suggests that there was a speed/accuracy trade-off.

2.3 Discussion

Overall, support was found for Lamy et. al.'s conclusions about same location benefits for same color singletons and costs for different color singletons, but some significant effects for the theorized interactions of attentional capture did not replicate. Participants also displayed high variance in their reports of conscious awareness of the cues. In addition, I encountered difficulty in reliably rendering the color cue display at liminal rates; in some pilot tests I found the cues to be perfectly superliminal at the fastest possible exposure speeds (even pushing down to the refresh rate of the monitor, 17ms for 60Hz).

The best option for continuing this investigation was to seek alternative paradigms less reliant on rendering barely liminal stimuli and supporting a more direct route to exploring the open vs closed shape relationship.

CHAPTER 3: FLANKER STUDY

In order to address the same fundamental questions as the liminal shape study – how features of bounded and unbounded shapes interact with visual attention – while sidestepping some of the methodological pitfalls reported in chapter 2, I adopted the flanker task paradigm. The experiment remained centered on the relationship between open and closed shapes. Basic charting symbols have been shown to differ in their discriminability and affect performance in prior research [120, 77, 76], and symbols with open and closed features appeared to be more distinct from each other. I wanted to determine whether exemplars of these classes of shapes would be rapidly and automatically segmented in the pipeline of object perception. This would lend support to some of the reports from Tremmel [120] and Li et. al. [77, 76], and help explain an important mechanism underlying visual perception of shapes in general.

Flanker tasks involve timed identification of a target presented briefly at one of a number of specific locations in a display while a distractor appears elsewhere in the display [43]. One common approach is to use a circular array of target positions to provide uniform distance between any given target position and the fixation point at the center of the display, keeping all targets within foveal vision. The flanker, which is the distractor item, is positioned just outside attentional focus, generally to the left or right of the central array of target locations. Reaction time (RT) responses to the target are typically found to be influenced by the flanker compatibility – the

relationship between the visual forms of the targets and flanking distractors [38, 39].

For each block of trials in this 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, participants' ability to selectively attend to the target shape and ignore the flanker can be assessed. 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 begin extending these results to visualization displays, I incorporated perceptual load as a variable, similar to Normand et al. [89], in order to study selective attention with sparse and cluttered displays. Compatibility effects are found to be much stronger with low load than high load contexts, as high perceptual load mitigates the interference introduced by distractors [43, 71, 70]. In this experiment, 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.

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 category were star and asterisk, and, for the closed category, square and triangle. Different feature pairs include one item from each of the two categories.

I hypothesized that differences in the compatibility effect would arise 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.

The specific hypotheses for this study were as follows:

H1 *Compatible flankers will produce shorter RTs and incompatible flankers will lengthen RTs*

H2 *High load displays will weaken compatibility effects and low load displays will strengthen them*

H3 *Open vs closed features will drive differences in the compatibility effect; targets and flankers sharing either feature will cause larger compatibility effects*

3.1 Methodology

3.1.1 Participants

Forty-three (seven male and thirty-four female) student volunteers were recruited from UNC Charlotte, and awarded class credit for participating in approved research studies where relevant. 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.

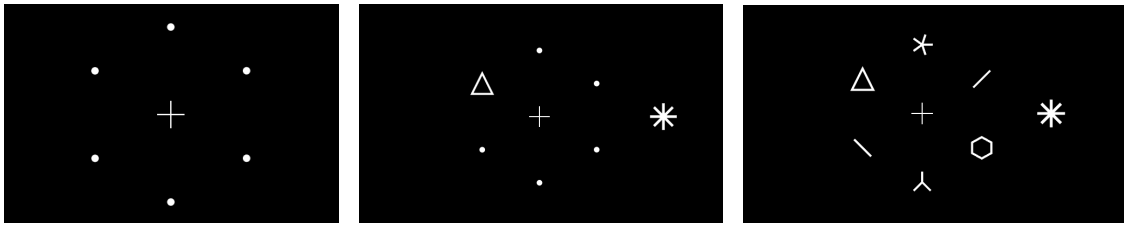
3.1.2 Stimulus Materials

The visual stimuli were presented on an iMac computer with a 17" flat screen LCD monitor. Stimulus presentation and data collection were controlled by SuperLab 4.0. All Stimuli were created using Javascript and SVG on the same computer to guarantee uniform display properties.

The four shapes 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. 3.1). 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 [39].



(a) (b) (c)

Figure 3.1: Displays for the flanker test trials. (a) fixation display (b) low-load condition (c) high-load condition. Participants were shown a fixation display for 500 to 1000 ms, the target display for 100 ms, then a post-stimulus fixation display until a keypress response was made. For a more comprehensive set of stimuli, refer to Appendix B.3 and B.4

In the low load condition, the target was presented in one of six 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 feature categories: 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 six blocks of trials – two same-feature blocks (triangle/square, asterisk/plus) and four 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 sixty high and low-load trials in each block included

twenty 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. 3.1, target, flanker, and filler shapes were presented as white against a black background.

In total, there were six blocks of 120 trials for a total of 720. A Latin Square was used 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 six orders.

3.1.3 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 six blocks of trials in one of six 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 twenty 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.

3.2 Analysis

Response times 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% in at least two conditions. For the remaining thirty seven 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. Counterbalanced group was not found to be significant, nor did it interact with any of the variables of experimental interest.

For this experiment, two% of the trials were trimmed on average, and data from six participants were removed prior to the analysis because they responded incorrectly to more than half of the trials in at least two conditions. The ANOVAs for the remaining

thirty-seven participants were averaged across the twenty trials within each condition to test for load (high, low), compatibility (compatible, incompatible, neutral), and block effects (two same-feature four different-feature).

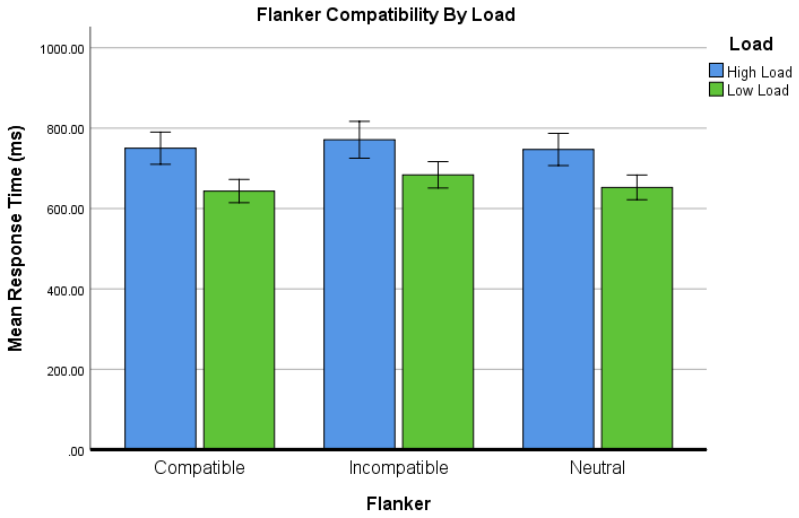


Figure 3.2: Mean correct response times for flanker compatibility by load. RTs from compatible and neutral trials were similar within each load condition. Incompatible flankers lengthened RTs, demonstrating response competition from distractor shapes when they were part of participants' attentional sets.

3.2.1 Reaction Times

As expected, load and compatibility main effects were consistent with past research. 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. A significant effect of flanker compatibility ($F(2, 62) = 47.372$, $p < 0.001$, $\eta_p^2 = 0.604$) also arose. Follow-up Bonferroni comparisons (at the $p < .05$ significance level) showed incompatible trial response times greater on average (M = 727 ms) than those of compatible (M = 697 ms) or neutral (M = 700 ms) trials. Importantly, the interaction between load and flanker compatibility was

also significant ($F(2,62) = 4.377$, $p = 0.017$, $\eta_p^2 = 0.124$).

This analysis also found a significant effect of block ($F(5, 155) = 13.503$, $p < 0.001$, $\eta_p^2 = 0.303$); Table 3.1 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 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, I 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 3.3 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*

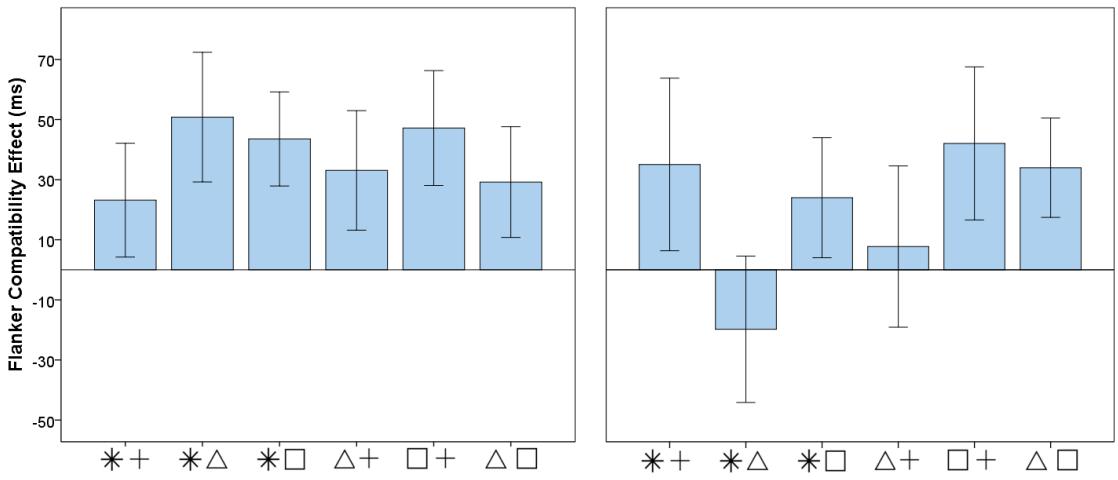


Figure 3.3: Flanker compatibility (the difference between mean incompatible and compatible response times) for each block in (a) low load and (b) high load conditions. Error bars show 95% confidence intervals.

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

3.2.2 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 condi-

Table 3.1: Breakdown of response time and error rate differences across the blocks. Same-feature pairs with closed features had the fastest RTs and fewest errors, and same-feature pairs with open features took the longest and induced the most errors. Different-feature pairs were stratified in between.

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%

tions did not differ significantly from each other (p = 1.000).

Load exerted a significant effect on error proportions ($F(1,31) = 56.12$, $p < 0.001$, $\eta_p^2 = 0.644$); 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 the same-feature, open condition, smaller in the different-feature blocks, and nonexistent in the same-feature, closed condition.

I 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 3.1, display a pattern among the blocks that is similar to the RT data. Both the same-feature blocks were outliers; the same-feature block with open shapes had the highest error rate while the closed, same-feature shapes yielded the lowest error rates.

3.3 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. In other words, blocks with shapes sharing open or closed features caused participants difficulty even under the conditions that usually mitigate those difficulties. The influence of load and flanker compatibility effects were successfully replicated, but the extent of these effects depended on open and closed feature categories and whether target sets included same or different feature categories.

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 [43] and Normand et al. [89].

Important evidence in support of the 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. 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.

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, was whether these findings could be replicated with other perceptual tasks and when the stimulus set was expanded to include other exemplars of the open and closed categories. It is possible that low-level feature differences among the four 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.

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.

CHAPTER 4: SAME-DIFFERENT STUDY

In order to test the validity of the findings from the flanker study in the previous chapter and examine whether they were tied specifically to the task or shapes used therein, a *Same-Different* paradigm was introduced. Focus was maintained on the open/closed shape feature categories by testing more varied examples of shapes from both categories and pairing them with a larger number of same and different combinations.

The Same-Different paradigm is even more straightforward than the flanker paradigm, with the added benefit of moving this work closer to visualization displays and paradigms by providing a more direct comparison between shapes. In a Same-Different task, each trial features a set of shapes organized at the center of the screen, and participants simply indicate with a keypress whether all the shapes in the display are the same or whether there are differences. Variance in response times reflects the degree of difficulty participants face in discerning the homogeneity or heterogeneity of the presented objects.

In this same-different study, performance was expected to be fastest for trials in which all shapes were the same; the human visual system rapidly computes summary statistics across the field of vision prior to attentional allocation, and differences in certain channels (curvature, closure, orientation, etc) 'pop out' from uniform distractors. However, based on findings from the previous study, I expected that closed

shapes would be associated with faster RTs and fewer errors than trials with open shapes. Because discrimination between perceptual categories is quicker than discrimination within a category [47], support for that hypothesis would be found if trials with different stimulus elements took longer when the shapes still shared the same open or closed category when compared to stimuli from different feature categories.

Specifically, the hypotheses for this study were as follows:

H1 *Same-shape trials will be faster and less prone to errors than different-shape trials*

H2 *Different-shape trials will be faster and less prone to errors when the shapes differ in boundary closure*

H3 *Same- and different-shape trials with two closed shapes will be faster and less error-prone than those with open shapes*

4.1 Methodology

4.1.1 Participants

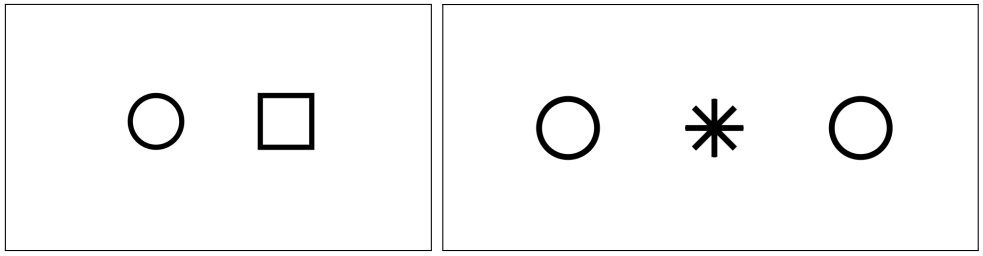
Forty-two student volunteers (thirty-one female, eleven male) were recruited from UNC Charlotte and awarded class credit for participating in approved research studies where relevant. The inclusion criteria required all participants to be over the age of 18, with 20/20 (or corrected to 20/20) vision and no history of visual impairment.

4.1.2 Stimulus Materials

The visual stimuli were presented on an iMac computer with a 17" flat screen LCD monitor. Stimulus presentation and data collection were controlled by SuperLab 4.0.

All Stimuli were created using Javascript and SVG on the same computer to guarantee uniform display properties.

Each trial consisted of fixation and target displays. The background and foreground colors were inverted from the first two experiments to use black targets on white backgrounds, simulating a more 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 two or three 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 two- and three-shape trials.



(a) (b)

Figure 4.1: Displays for the same/different test trials. (a) Two-shape condition with two closed shapes; (b) three-shape condition with open and closed shapes. Participants were shown a fixation display for 500 to 1000 ms, then the stimulus display was presented until a keypress response was made. For a more comprehensive set of stimuli, refer to Appendix B.5, B.6, and B.7

The shapes representing each of the feature categories was expanded from that of the flanker task (Chapter 3) 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 two-shape trials or only three-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 ninety-six trials per block while different feature category and different item same feature category each had forty-eight trials.

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

4.1.3 Procedure

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 four 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. Stimuli remained on the screen rather than being presented briefly as in the flanker task, both to simulate more realistic visualization scenarios and so that participants could take as much time as necessary to respond.

Each participant began with twenty 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. 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.

4.2 Analysis

On average, 2.6% of trials were removed for each participant, and the largest trim proportion was 4%. Data from four participants were removed from the final analysis due to error rates in excess of fifty percent in at least two conditions. For the remaining thirty-eight 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 (two or three shapes). Since there was not a significant block effect when two or three shapes were used ($p = 0.376$) and the number of shapes was not found to interact with the other conditions of interest, the two-and three-shape trials were combined, and the ANOVAs tested for overall differences across feature category and conditions.

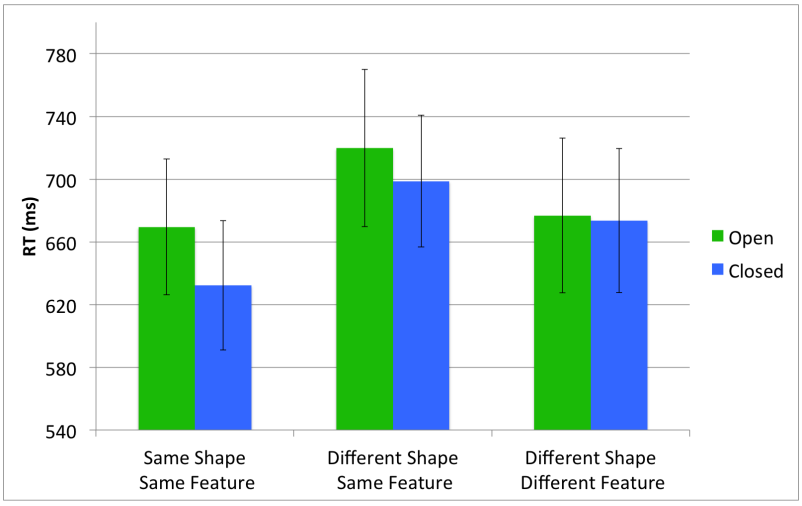


Figure 4.2: The interaction between feature and condition on RTs. Same shape was significantly easier, different-shape/different-feature were close across both features, and different-shape/same-feature trials took the longest. Error bars show 95% confidence intervals.

4.2.1 Reaction Times

The analysis revealed 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 (M = 651 ms), different-item/different-feature trials were the second fastest (M = 675 ms), and different-item/same-feature trials took the longest (M = 709 ms). Follow-up Bonferroni comparisons ($p < .05$) showed that each was significantly different from the other.

Consistent with the findings from the flanker task in the previous chapter, 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 Figure 4.2 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, I examined the trials for each condition separately and compared RTs for each of the individual shapes.

In the same-shape condition, there were significant differences in RTs among the six shapes ($F(5, 185) = 20.836$, $p < .001$, $\eta_p^2 = 0.360$); 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, I again found significant RT differences among the six 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 $\times$ and plus took significantly longer ($M = 760$ ms) than all other shape combinations, and circle/triangle trials were significantly faster ($M = 630$ ms) than the other combinations of closed shapes ($M = 700$ ms for circle/square and $M = 716$ ms for square/triangle). Asterisk/plus-sign trials were also significantly faster ($M = 683$ ms) than square/triangle trials ($M = 716$ ms).

When analyzing the nine different-shape/different-feature conditions a significant main effect of shape pairs was found ($F(8, 296) = 2.836$, $p = 0.012$, $\eta_p^2 = 0.071$), but follow-up Bonferroni tests showed no significant differences among the items.

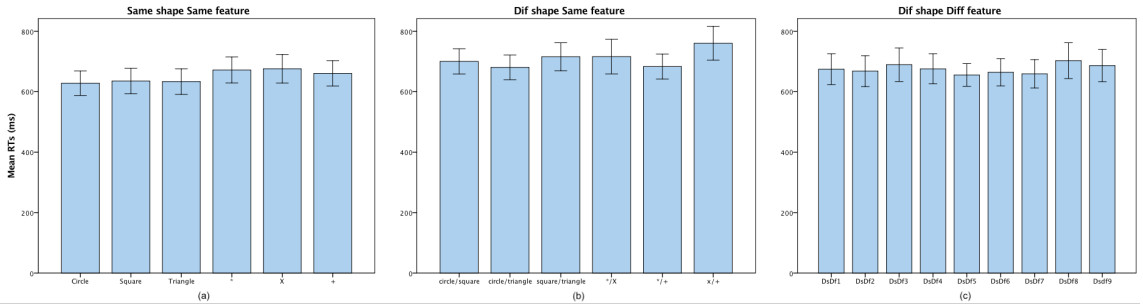


Figure 4.3: Reaction Times (in ms) for each shape combination within each condition. (a) Same shape, same feature - the closed shapes are all significantly faster than the open shapes. (b) Different shape, same feature - the x/plus-sign pairing is significantly slower than all other combinations. (c) Different shape, different feature - no significant differences among pairs. Error bars show 95% confidence intervals.

4.2.2 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%). I 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%).

4.3 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 found in the flanker task in chapter 3 and provide additional evidence for

the existence of open/closed feature categories in visual perception.

Similarly to the flanker study, 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 lends credence to two of the three 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 category. Closed shapes did introduce significantly fewer errors than open shapes, but different-shape/different-feature trials introduced fewer errors than same-shape trials, contrasting the expectation of the first hypothesis. That said, it is not necessarily surprising that participants

were most accurate in their judgments with symbols that differed so significantly.

CHAPTER 5: VISUAL SUMMARY TASKS

The experiments reported in the previous three chapters investigated discrimination within and between open and closed shapes using basic tasks. Those paradigms involved displays with a small number of representative shapes and short exposure times to isolate attentional allocation, and the results provided strong evidence for the categorical distinction between open and closed shapes in early vision. To illustrate the utility of those findings for the visualization community, it was necessary to involve tasks and displays more natural to that context in successive experiments. If the deployment of such encoding strategies could be shown to influence participants' abilities to perform some tasks commonly used in scatterplot displays, then there was justification for expanding this investigation in a number of directions – a larger variety of analysis tasks, additional feature categories, and more comprehensive design recommendations.

A key feature of scatterplot displays is that they allow extraction of generalized, higher-order information through ensemble coding - automatic aggregation of large sets of elements in the visual field. Szafir et al. [115] discuss ensemble coding in the context of data visualization, and provide a categorization of a variety of task types in which this feature of the human visual system can be effectively harnessed. One such example is numerosity estimation, which has been shown to be a preattentive visual property independent of other factors such as texture density or spatial frequency of

elements in the visual field [102, 55, 44].

I chose three visual analysis tasks with relative judgments to test the findings from the flanker and same-different experiments:

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

For each of the tasks, I 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 the previous findings, I had the following hypotheses:

H1 *visualization tasks involving closed symbols will be associated with faster RTs than open symbols*

H2 *when two symbols are used together in tasks requiring discrimination between symbols in a single display, symbols from different feature categories will 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 participants' ability to perform the visual analysis task whereas 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.

5.1 Methodology

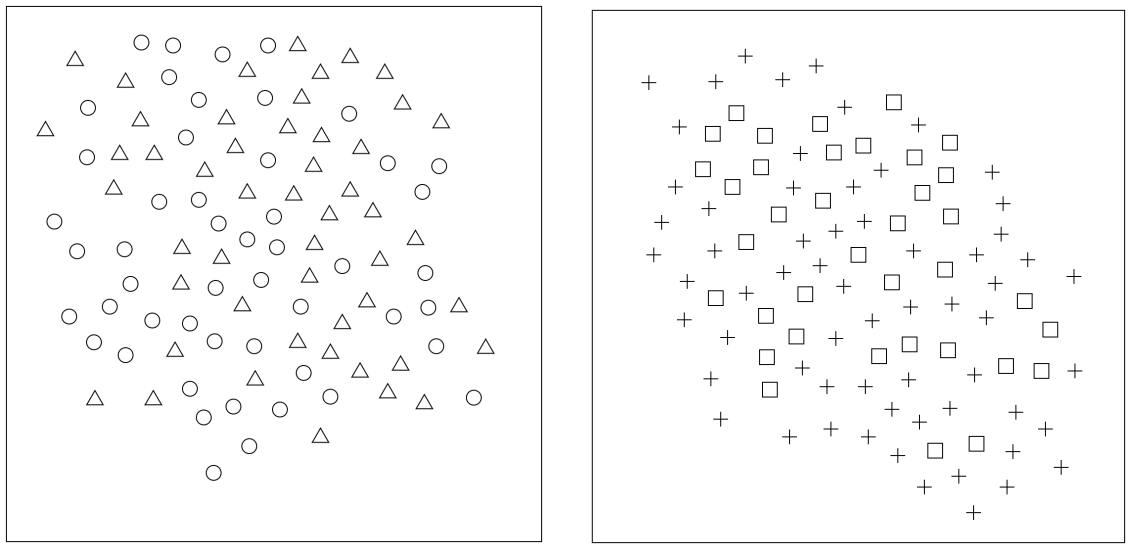
5.1.1 Participants

Twenty-six student volunteers (nineteen were male and seven female) were recruited as participants. Participants were tested individually in 40-minute sessions. 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.

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.

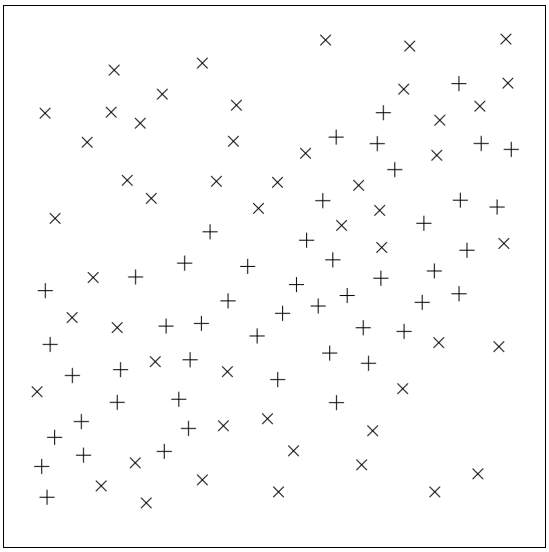
5.1.2 Stimulus Materials

The visual stimuli were presented on an iMac computer with a 17" flat screen LCD monitor. Stimulus presentation and data collection were controlled by SuperLab 4.0. All stimuli were created using Javascript and SVG on the same computer to guarantee uniform display properties.



(a)

(b)



(c)

Figure 5.1: Medium-difficulty single-plot displays for the scatterplot analysis trials. (a) Average Value Task (b) Numerosity Task (c) Linear Relationship Task. Participants were shown a fixation display for 500 to 1000ms, then the stimulus display until a keypress response was made.

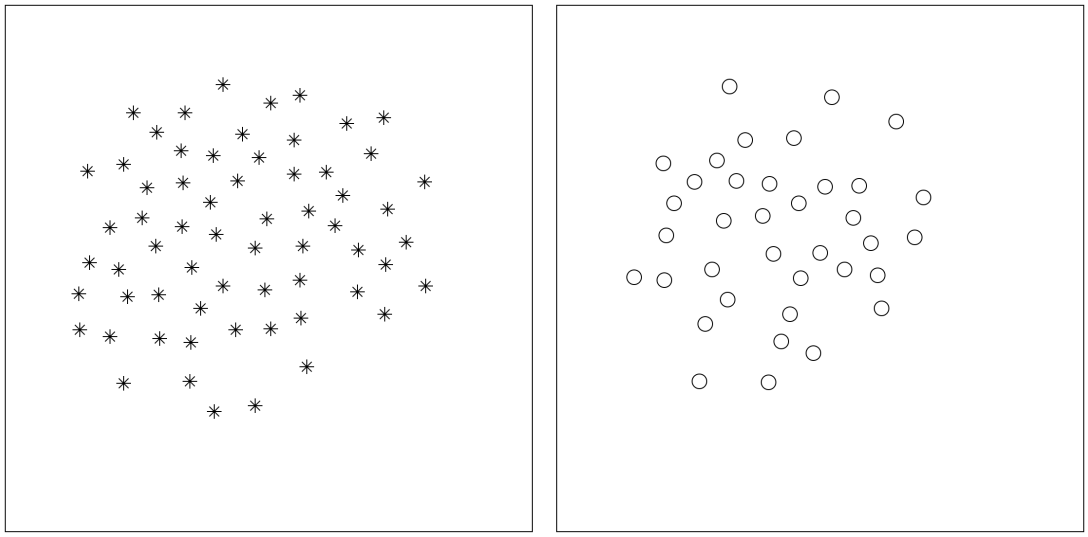


Figure 5.2: Medium-difficulty separate-plot display for the numerosity analysis task. Participants were shown a fixation display for 500 to 1000ms, then the stimulus display until a keypress response was made.

I used the same shape palette as in the Same/Different study (4): circle, square, and triangle for the closed shapes, and asterisk, ×, and plus-sign for the open shapes. Figure 5.1 has examples of the single-plot displays for all the tasks and Figure 5.2 shows the separate-plot stimuli for a numerosity task. All of the displays contained 100 items, 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.

To introduce levels of difficulty, I used the concept of 'delta' for each task, similar to Gleicher et al. [46], 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. Pilot tests were used 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, I adapted the algorithm from Gleicher et al. [46]. First, randomly select the center point of the entire set at a location in the middle third of the display. Then utilize 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, make small vertical adjustments to the resultant sets of points to reach the desired pixel delta for the given difficulty level. The top and bottom shapes were also de-correlated 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, I alternated between shapes until the desired maximum number for the smaller set was reached, then continued drawing the shape from the larger set

until 100 points were drawn overall.

Single-plot target displays for the linear relationship judgment tasks were generated in a fashion similar to the description given by Rensink et al. [97]. I 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, I alternated between the two sets of shapes. For elements in the set of linearly associated shapes, I 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, a pseudo-random number generator was used to produce both x- and y-coordinates. For shapes of either set, small adjustments were made to prevent overlaps and maintain spacing between shapes, and the positions were re-randomized if a satisfactory position could not be achieved with minor adjustments. See Figure 5.1 for examples of each stimulus display.

For the separate-plot target displays for each of the three analysis tasks, the same sequence of steps as the single-plot display generation were taken, but alternate shapes were drawn 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 thirty-six separate-plot trials and seventy-two single-plot trials. Single-plot target displays were split evenly among easy, medium, and hard trials and all six 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 three 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, an equal number of positive and negative correlations were maintained. Within any given sub-block of trials, there was a random arrangement of difficulty levels and shapes.

The presentation order of the three 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.

5.1.3 Procedure

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 twelve 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 twenty-four 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. Keypress mappings for the six shapes were reordered 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.

5.2 Analysis

On average 2% of the trials were trimmed and the data from six participants were removed prior to the analysis due to error rates in excess of 50% in at least two conditions. For the remaining twenty participants, mean correct RTs were computed across the six 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.

5.2.1 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 ms, SD = 1492 ms) than numerosity (M = 662 ms, SD = 205 ms) and linear relationship (M = 632 ms, SD = 149 ms), which did not differ significantly from each other. Unexpectedly, neither a main effect of target feature ($F < 1$, $p = .664$) nor interactions of target feature with any other variables of interest rose to significance: 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 that easy trials (M = 790 ms) were significantly faster than all others, hard trials were the longest (M = 1156 ms) with medium difficulty trials (M = 950 ms) in between, mirroring my expectations and accounts from the literature. Difficulty also interacted with task ($F(4, 76) = 6.09$, $p = .004$, $\eta_p^2 = .243$), as shown in Figure 5.3.

Because of the variability in RTs among the three tasks, I 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 slower than the hard trials. For numerosity ($F(2, 38) = 22.26$, $p < 0.001$, $\eta_p^2 = .540$), the main effect resulted from a significant

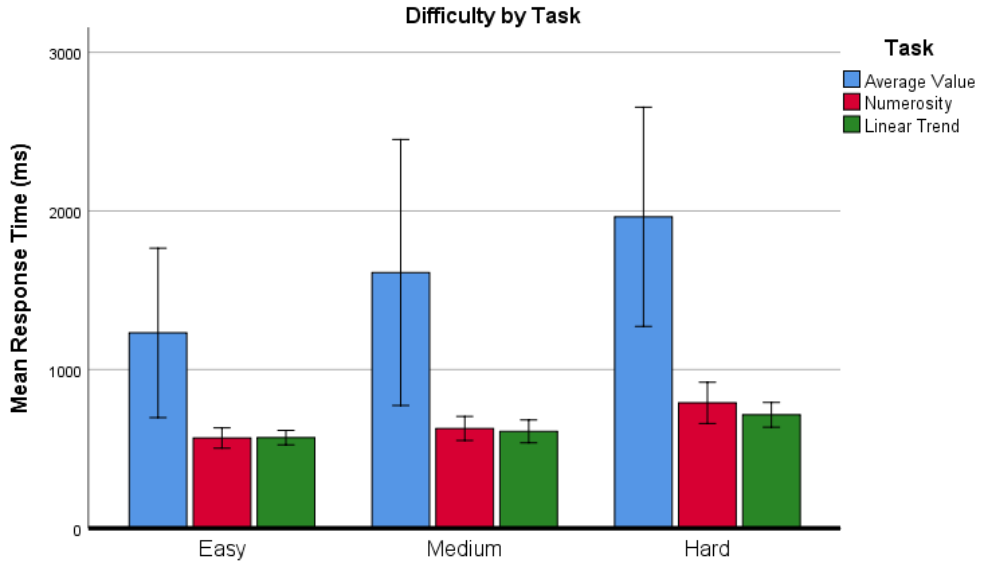


Figure 5.3: Difficulty by task interaction for the side-by-side plots. Average value task required significantly more time and interacted more starkly with difficulty than numerosity or linear tasks, possibly related to the automaticity of the task requirements.

difference among all three levels of difficulty; and for linear relationship ($F(2, 38) = 22.25$, $p < 0.001$, $\eta_p^2 = .539$), 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.

5.2.1.1 Errors

Participant accuracy 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 was 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 were no observable differences in task performance when either open or closed shapes were used as symbols in homogeneous scatterplot displays.

5.2.2 Single-Plot Displays

5.2.2.1 Response Times.

When participants were asked to identify which of the two shapes presented in a single display met the task requirements, a strong, significant influence of task on RTs was found ($F(1.112, 21.326) = 15.673$, $p < 0.001$, $\eta_p^2 = .452$), with follow-up test showing that all three task means differed significantly from each other ($M = 3226$ ms, $SD = 1417$ ms; $M = 2363$ ms, $SD = 404$ ms; $M = 1787$ ms, $SD = 347$ ms for average value, numerosity, and linear relationship respectively). Task was also found to interact with distractor feature ($F(1.513, 28.753) = 6.135$, $p = .01$, $\eta_p^2 = .244$), 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 ($F(2, 38) = 16.046$, $p < 0.001$, $\eta_p^2 = .458$); however, Bonferroni comparisons showed that easy trials ($M = 2261$ ms) differed significantly from medium ($M = 2512$ ms) and hard ($M = 2603$ ms), and the latter two did not differ significantly from each other.

Importantly, distractors sharing features with targets lengthened RTs relative to distractors differing in features, ($F(1, 19) = 52.595$, $p < 0.001$, $\eta_p^2 = .735$), and distractor feature 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 ($F(2, 38) = 3.401$, $p = .044$, $\eta_p^2 = .152$).

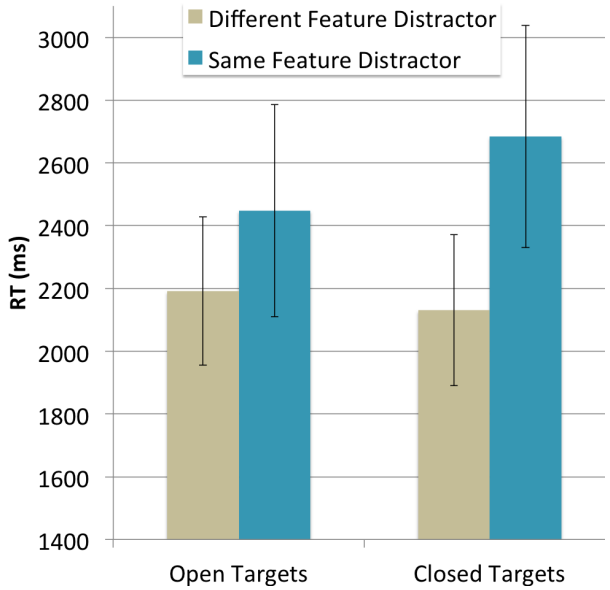


Figure 5.4: Target by distractor feature interaction for the single plot numerosity tasks. Different-featured distractors always decreased RTs, particularly in trials with closed symbols. Same-feature distractors always increased RTs, especially when both symbols were closed. Closed shapes seem more susceptible to influence from distractor shapes overall. Error bars express 95% confidence intervals.

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, no significant main or interaction effects among any of the experimental conditions was found.

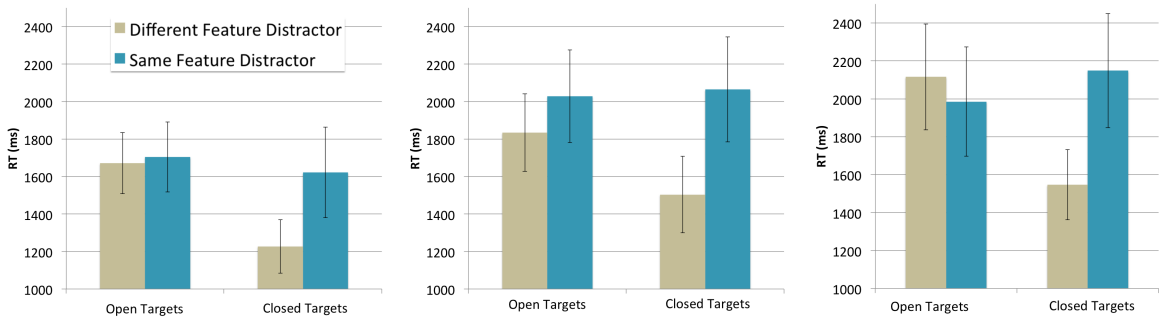
For the numerosity task, however, a number of significant effects emerged. Distractor feature was significant ($F(1, 19) = 48.16$, $p < .001$, $\eta_p^2 = .717$); same-feature distractors took 400 ms longer on average. Difficulty level was also significant ($F(2,$

38) = 8.87, p = .002, η_p^2 = .318); easy trials (M = 2142 ms) differed significantly from both medium (M = 2437 ms) and hard (M = 2511 ms), 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, η_p^2 = .314) indicated that same-feature distractors reliably caused longer reaction times but the effect was modulated by target features (Figure 5.4).

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

Closed target features led to quicker RTs than open targets (F(1, 19) = 20.95, p < .001, η_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, η_p^2 = .571). Additionally, these two variables interacted with each other (F(1, 19) = 26.27, p < .001, η_p^2 = .58) as well as in a three-way interaction with level of difficulty (F(2, 38) = 3.15, p < .054, η_p^2 = .142). Figure 5.5 presents the three-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.*



(a) (b) (c)

Figure 5.5: Three-way interaction of difficulty, target, and distractor features for single plot linear tasks with 95% confidence intervals. (a) easy, (b) medium, and (c) hard conditions. Different-feature distractors were faster than same-feature distractors in all conditions except the hard trials with open targets. In hard trials with relatively low correlation (r = 0.4) among open-featured target symbols, participants were faster on average when the distractors also shared open features; it is not clear why this occurred.

5.2.2.2 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 ($F(1.230, 23.378) = 31.920$, $p < 0.001$, $\eta_p^2 = .627$) 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 ($F(4, 76) = 10.607$, $p < .001$, $\eta_p^2 = .358$) (See Figure 5.6). Distractor feature ($F(1, 19) = 20.083$, $p < .001$, $\eta_p^2 = .514$), its interaction with task

(F(2, 38) = 3.677, p = .035, η_p^2 = .162), and its three-way interaction with task and difficulty (F(4, 76) = 4.002, p = .016, η_p^2 = .174) were all significant effects in the error proportion analysis.

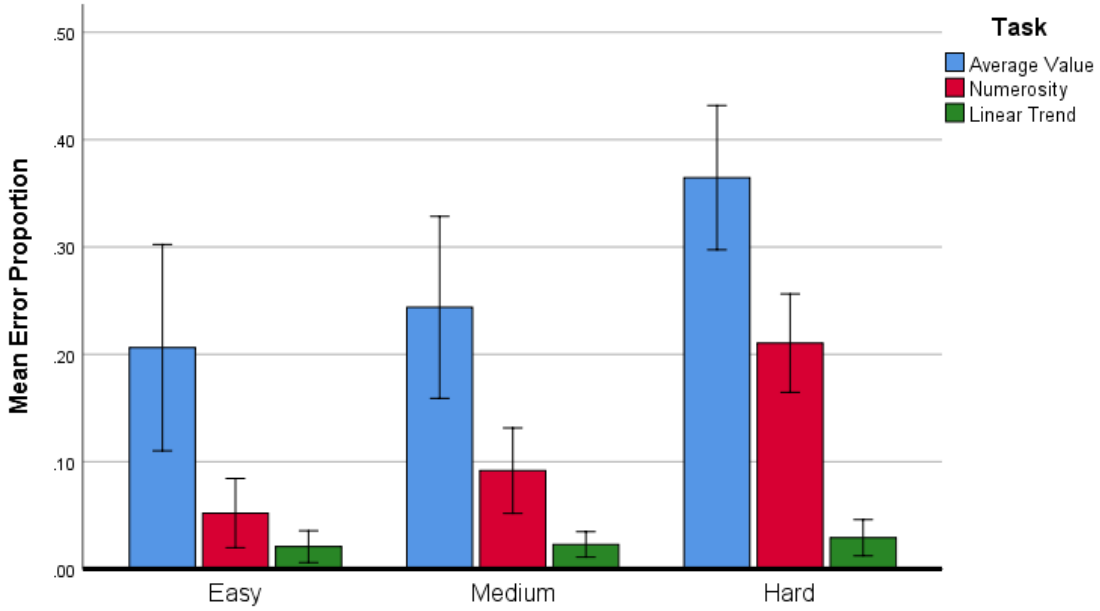


Figure 5.6: Error proportions for task and difficulty interactions in single-plot displays. The linear trend task was far more resilient to difficulty levels, perhaps due to the automaticity of the task.

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

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*, providing further evidence that open and closed shapes are processed differently, particularly in complex displays; more study is needed to investigate this relationship. For linear relationships tasks, the only effect to reach 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.

5.3 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 homogeneous items filled side by side displays. Because visualization tasks facilitate rapid integration of summary statistics of visual information through ensemble coding mechanisms, there may not be as much sensitivity to symbol features and their categorical or topological differences when homogeneous items fill a display (as compared to lower-level perceptual tasks in simpler displays). Feature category differences, however, were more evident in the single-plot displays because participants were required to discriminate among items with different topological features; when dealing with visual clutter, same and different feature categories

had important influences on numerosity and average value judgments.

After reanalyzing the data for each task, it was clear 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 considered with the length of the RTs and the higher error rates, suggests that judgment of average value among sets of marks requires lengthier processing time and more cognitive effort, while numerosity judgment and perception of linear relationship tasks are more automatic. Indeed, the trends in the average value task in this experiment appeared to be moving in the hypothesized directions for the level of difficulty and distractor feature effects, but ultimately did not produce significant main or interaction effects among the experimental conditions.

Numerosity and linear relationship tasks were more interesting: consistent with the stated hypotheses, 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 Figure 5.4).

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 only 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 Figure 5.5).

CHAPTER 6: OVERPLOTTING STUDY

The visual summary experiments described in the previous chapter provided a foundation for moving the more abstract, lower-level studies of the preceding chapters closer to the domain of applied visualization design. The results across the tasks followed expectations to some extent – different combinations of open and closed shapes certainly did exert influence on the speed and accuracy of participants' responses – but also opened the door to additional questions.

One weakness with the methods and results up to this point is the ecological validity of using displays in which symbols did not overlap in the visualization tasks. In particular, the information in the stimulus displays did not reflect an important characteristic of realistic data because all the marks were drawn to positions without any overlap with other symbols. Although the goal is to provide justification and guidance for visualizing realistic displays, I had intentionally maintained a minimum space separating all the symbols in order to control for overlap effects while isolating the open vs closed construct. The general question of symbol discrimination and the application to visualization displays is quite complex, so I tried to move carefully to address the question of overlap in the experiment described in this chapter.

The primary purpose of this experiment is to extend the previous findings with open and closed categories of symbols to address increasingly realistic data characteristics and displays, with a particular focus on overlapping elements. As in the scatterplot

experiments (chapter 5), numerosity and trend judgement tasks were tested with single and side by side displays containing multiple categorical variables encoded with open and closed symbols. This was structured so as to determine whether the open/closed categories had a meaningful influence on RTs and error rates when the data positions induced marks to overlap. The degree of overlap among the symbols was manipulated to create low, medium, and high levels, and I expanded the symbol palette to test more shapes and pairs, but in other respects the experiment was designed to replicate the findings from the previous study.

The following hypotheses were tested:

H1 *Pairs of symbols differing in open and closed features will produce faster RTs and fewer errors than pairs sharing open or closed features*

H2 *Closed-feature target shapes will be more influenced by distractor features*

H3 *Larger proportions of overlap among positions will increase task difficulty*

Mark encodings were varied in multi-class scatterplot displays with specific proportions of overlapping shapes in order to determine (a) if certain types of plotting symbols yield better performance in realistic displays and (b) whether certain combinations of shapes are more susceptible to overdraw effects.

6.1 Methodology

This experiment was conducted to explore how the open and closed feature categories would interact with different levels of overlap and analysis tasks. Two relative judgment tasks were used (numerosity and linear trend), each with low, medium, and

high proportions of overlapping marks in synthetic data distributions. The numerosity task asked participants to determine which of the two sets of shapes contained more elements, and the linear trend task asked participants to determine which set of shapes represented a stronger linear relationship or correlation.

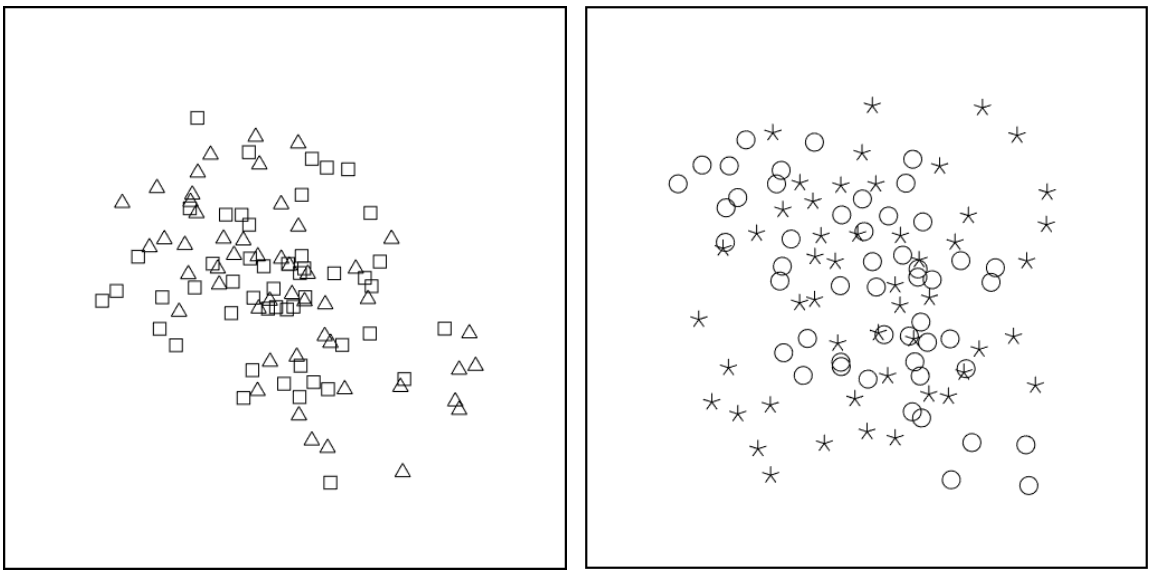
In both tasks, separate-plot displays with side-by-side charts and single-plot displays with variables overlaid on top of each other were used, just as in the previous chapter. Separate-plot displays required participants to make a binary choice between the left and right displays with homogeneous shapes to identify the more numerous or linearly associated set; these were designed primarily to assess participants' task competence and any overall differences between open and closed symbols. Single-plot displays incorporated pairs of shapes to encode the two sets of points, directly assessing the relationship between pairs of open and/or closed symbols.

6.1.1 Participants

Thirty one student volunteers (eighteen female, thirteen male) were recruited from UNC Charlotte and awarded class credit for participation where appropriate. 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.

6.1.2 Stimulus Materials

Stimuli were presented as black on a white background in regions measuring 500 x 500 pixels. Each shape was generated using a bounding circle with a radius of 8 pixels, producing stimuli subtending 0.4043 degrees of visual angle for participants 60cm away from the screen, well within the bounds of symbol size for normal visual



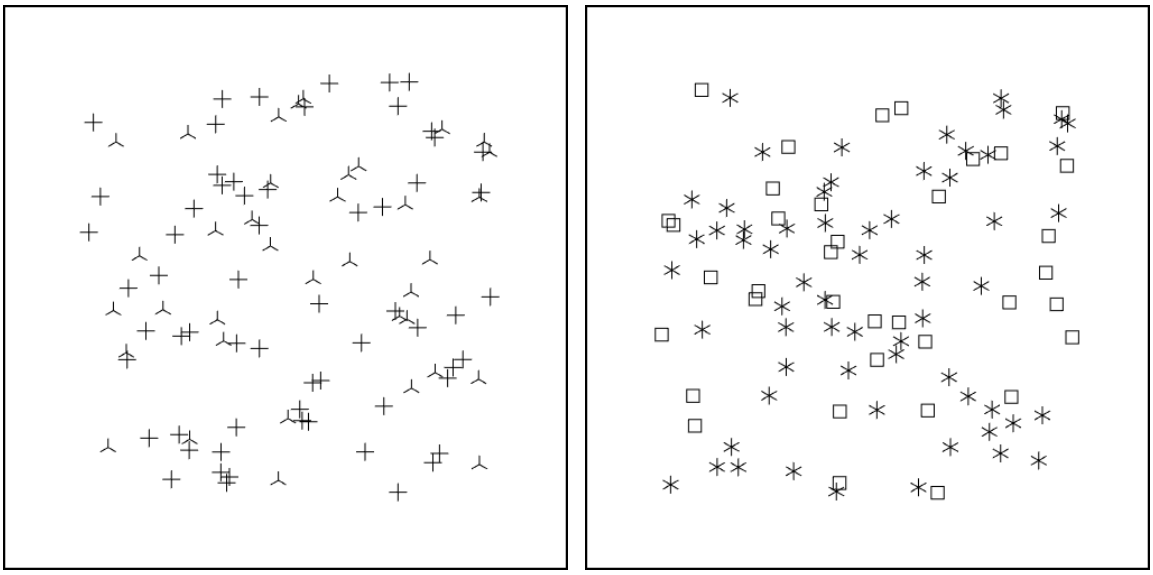
(a) (b)

Figure 6.1: Sample stimulus displays from the Linear Trend task. (a) Triangle and square symbols (i.e. same-feature). (b) Circle and fiveline symbols (i.e. different-feature).

acuity recommended by Li et. al. [77]. Stimulus displays were generated using D3 and presented using Superlab 5.0 on an iMac computer with a 27" flat screen retina display.

Four open and four closed symbols were selected directly from the stimuli used by Li et. al. [77], expanding the palette of symbols used in the previous experiments. Closed symbols included circle, square, triangle, and pentagon, and open symbols were shapes with six, five, four and three radial spokes; see Fig. 6.3.

Previous use of ensemble judgment tasks in the literature [46, 97] and in probing open and closed feature categories [19] involved varying the level of difficulty, with different amounts of symbols for numerosity judgments and higher or lower degrees of correlation for the linear trend judgments. In this experiment, that measure of difficulty was held constant so as not to interfere with varied levels of overlap. All

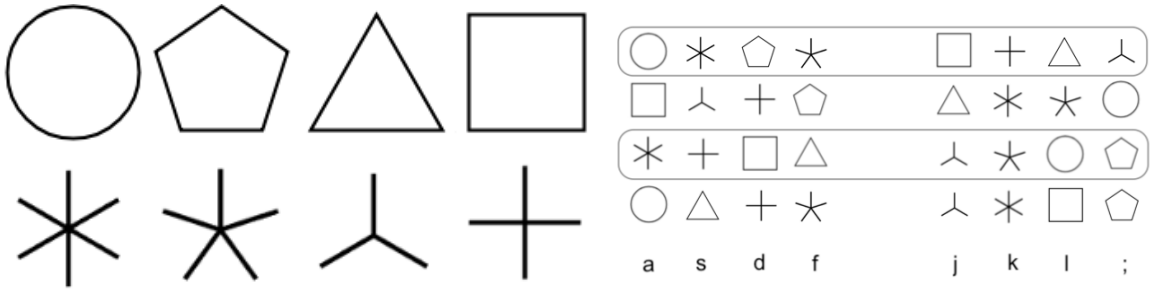


(a) (b)

Figure 6.2: Sample stimulus displays from the Numerosity task. (a) Fourline and threeline (i.e. same-feature). (b) Square and sixline (i.e. different-feature).

linear trend displays contained 100 points split evenly in two, and numerosity task sets were split 37-63. Each linearly correlated set was produced within +/- 0.02 of a +/-0.6 Pearson correlation coefficient. These values correspond to the medium level of difficulty in the previous study, and fall within reasonable ranges used by other researchers in similar contexts.

To carefully vary the overlap among points in the display, I settled on a measure involving the proportion of symbols with a non-zero number of overlaps with other symbols. A pair of symbols were considered to overlap if they were less than two times the bounding radius away from each other, and set low (30%), medium (50%), and high (70%) thresholds for the total proportion of symbols overlapping in the display. A minimum distance between the centers of each point was enforced to avoid cases where symbols were directly on top of each other.



(a) (b)

Figure 6.3: (a) The symbol palette for the current experiment. The top row contains closed symbols and the bottom row contains the open symbols. (b) The four groups mapping keys to symbols. Participants practiced the associations until they were comfortable before performing the experimental trials, and a note remained at the bottom of the screen to serve as a visual reminder of the mappings throughout the study.

For the numerosity task, displays were generated using the following algorithm. Pick two symbols. If the desired number of points in each set is not met, randomly decide to add either an overlapping or non-overlapping point, alternating between symbol 1 and symbol 2, otherwise add whichever type still needs more points. When adding a non-overlapping point, sample new coordinates from a normal distribution until a position is found without any overlaps with existing points. When adding an overlapping point, randomly pick an existing point, generate a random angle around that point, and place a new point along that vector between the minimum distance (2 pixels) and the bounding circle radius (8 pixels). Compare the candidate position to the other points in its vicinity to ensure the minimum distance is not violated. When the desired number of points are drawn into the display, if the overall proportion of overlapping symbols is more than 2.5% away from the desired overlap threshold, start over. See figure 6.2 for two examples; a wider variety of stimuli across conditions is given in the Appendix: B.18, B.20.

For the linear trend task I followed a similar approach to Rensink et. al [96, 97], sampling from a normal distribution and coercing the y coordinates of each point to within +/- .02 of the target +/- .6 correlation for the correlated set of points, randomly deciding whether any given point should overlap or not until the desired overlap threshold was reached. Small perturbations were made to the points to adjust the desired overlap proportions without harming the correlation values, and the whole display was re-randomized if a satisfactory distribution could not be attained. See figure 6.1 for two examples; a wider variety of stimuli across conditions is given in the Appendix: B.17, B.19.

In stimulus displays for both tasks, I re-used the same exact distributional positions with a few different combinations of symbols to produce a number of stimulus displays with similar overlaps and spatial characteristics. The entire set of points was rotated by random increments of 90 degrees within each display region so participants would not recognize them and introduce bias into their responses. For single-plot displays, all points were drawn into the same region of the display, and the overlap was computed across all points. For separate-plot displays, each set of points was drawn into one of two separate regions of the display, and overlap was computed across all points in each region.

Both tasks were organized into separate blocks. Each task held ninety six separate-display trials in one sub-block followed by a second sub-block of 144 single-display trials. Separate-plot blocks contained an even number of each shape and an even split of trials for which left and right was the correct answer. Single-plot blocks contained as close as possible to a fair distribution of pairs selected from the 8-shape palette,

and all blocks contained an even number of low, medium, and high overlaps. Linear trend blocks contained an even number of positive and negative correlations. The display order of the stimuli was randomized in each task block. Presentation order of the two tasks was counterbalanced across participants, and participants were evenly assigned to one of four key mapping groups to guard against biases in key/response mapping.

6.1.3 Procedure

Participants were tested individually in 45-minute sessions. After reading and signing an informed consent document, they were placed 60 cm from the computer screen in a well-lit room. Twenty nine of the thirty one participants completed all four blocks of trials; the other two were not able to finish in the allotted time frame and opted to leave.

Each participant began with a set of general instructions about the key mappings and tasks in the study, and then saw instructions and eight practice trials for the first task's separate-display sub-block to familiarize themselves with the 'left/right' response keys. In the separate-display condition the keypress indicated whether the left/right display contained more symbols, or which one represented a linear trend. After completing the practice trials and the ninety six experimental trials, they saw a second set of instructions and thirty two practice trials for the single-display sub-block to learn the key mappings for the eight shapes, and then performed the 144 experimental trials. On every trial there were two possible responses mapped to one key on the right and one key on the left hand. The keypress indicated which of

the two symbols in the display were more numerous or represented the linear trend. Participants were instructed to take a short break between tasks, and then the same sequence of instructions, practices, and sub-blocks was administered for the second task.

Participants were instructed to respond to all trials as accurately and quickly as possible. If participants answered incorrectly in more than a quarter of the trials in any practice blocks, the block started over. Throughout the experimental session there was a note at the bottom of the screen to remind the participants of the key associations.

Within each experimental trial, participants were first presented with a fixation display for 500, 600, 700, 800, 900, or 1000 milliseconds, then the target display was presented until either a keypress was made or 30 seconds elapsed.

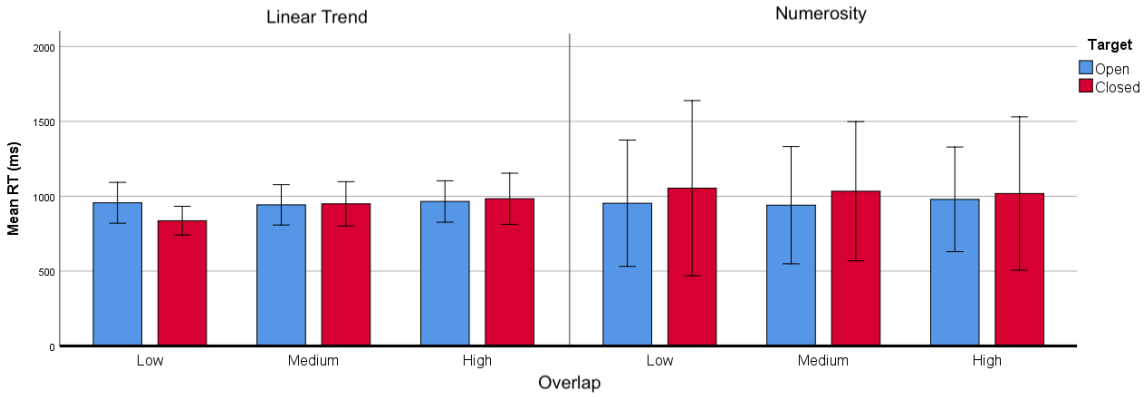


Figure 6.4: Triple interaction of Task, Overlap, and Target feature for separate-plot response times; no main effects or other interactions were significant. Closed targets in the lowest overlap level of the linear trend task were faster than all other conditions, but were indistinguishable from open targets in medium and high overlap cases. All responses were very quick across conditions.

6.2 Analysis

Response Times (RTs) were trimmed if they exceeded 2.5 standard deviations from a participant's mean correct RT in each condition. The mean trimmed correct RTs and proportion of incorrect responses in each condition were analyzed with separate repeated measures analyses of variance (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.

Five participants' data were removed due either due to lack of study completion or error rates in excess of 50% in at least two conditions. An average of 2.3% of trials were removed for each participant, and the largest trim proportion was 4.2%. The ANOVAS for the remaining twenty six participants tested for task (linear, numerosity), target (open, closed), overlap (low, medium, high), and, in the single-plot displays, distractor (same feature, different feature). Follow-up Bonferroni comparisons (at the $p < .05$ level of significance) were also conducted to explore the significant main effect of the overlap factor.

6.2.1 Separate-Plot Displays

6.2.1.1 Response Times

RTs were not found to differ significantly when responding to separate-displays that varied in task (linear trend: M = 939 ms, SD = 62 ms; numerosity: M = 997 ms, SD = 215 ms), target (open: M = 956 ms, SD = 102 ms; closed: M = 979 ms, SD = 128 ms), or overlap conditions (low: M = 950 ms, SD = 124 ms; medium: M

= 967 ms, SD = 110 ms; high: M = 987 ms, SD = 112 ms), nor were any of the two-way interactions significant.

The only interaction to achieve statistical significance was the three-way interaction of task by overlap by target ($F(2, 36) = 4.383, p = .02, \eta^2 = .196$). This complex effect appeared to result from an overlap effect in the linear trend task with closed targets (see Fig. 6.4). Closed targets in the lowest overlap level were much faster than all other conditions, and were indistinguishable from open targets in medium and high overlap cases. Judgments regarding linear correlation between two separate displays were very rapid regardless of features and overlap proportions. In the numerosity task, closed targets took marginally longer than open targets, but no useful interaction with overlap was present, and it is hard to draw conclusions with the amount of overlap among error bars for each target type and at each level of overlap.

6.2.1.2 Error Proportions

Errors were low across all of the experimental conditions, ranging from $0 - 4\%$ of the responses. The analysis did not show any significant main or interaction effects from manipulation of task, target or overlap conditions.

6.2.2 Single-Plot Displays

6.2.2.1 Response Times

Participant response times were not significantly influenced by the effects of task ordering group (p = .535) or the shape group for the key mappings (p = .054) into which they were placed, so I collapsed the data across both between-subjects factors

and analyzed the main and interaction effects that arose.

When two symbols were presented together in a single display and participants were required to discriminate among them, RTs were found to differ in response to all of the expected within-subjects variables. A main effect of task ($F(1, 25) = 18.594, p < .001, \eta^2 = .427$) showed that the linear trend task (M = 4346 ms, SD = 337 ms) took longer on average than the numerosity task (M = 3032 ms, SD = 250 ms).

As hypothesized, the increasing overlap proportions exerted a strong significant effect ($F(1.574, 39.357) = 18.994, p < .001, \eta^2 = .432$), and within-subjects contrasts indicate a significant linear trend in RTs with increasing degree of overlap ($F(1, 25) = 24.787, p < .001, \eta^2 = .498$), with low (M = 3532 ms, SD = 253 ms) and medium (M = 3618 ms, SD = 254 ms) both differing significantly from high (M = 3917, SD = 264) levels of overlap (p < .001, p = .001 respectively).

In addition, RTs were influenced by a significant effect of target feature category ($F(1, 25) = 16.420, p < .001, \eta^2 = .396$). Responses to open targets took longer on average (M = 3811 ms, SD = 244 ms) than closed symbols (M = 3567 ms, SD = 267 ms), indicating that it mattered whether the more numerous or the more correlated set of points were encoded with a closed symbol, regardless of the other symbols at hand.

Among all of the manipulated variables, having distractor symbols from the same or different feature category than the target levied the strongest effect on RTs ($F(1, 25) = 82.890, p < .001, \eta^2 = .768$). Consistent with findings in earlier experiments on open and closed shapes and their relationship in judgment tasks, same-featured distractors significantly lengthened response times (M = 4267 ms, SD = 304 ms) in compari-

son to different-featured distractors (M = 3111 ms, SD = 211 ms); the relationship between same and different featured distractors clearly makes a large difference on target processing.

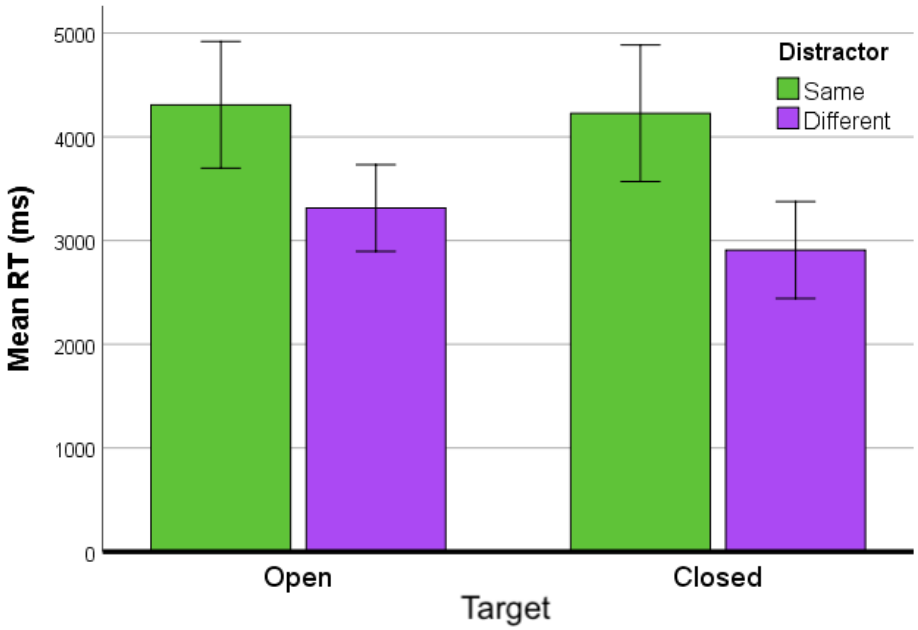


Figure 6.5: The Target by Distractor interaction in the single-plot RT analysis shows that closed targets are more facilitated by different-feature distractors than open targets are. When distractors come from the same feature category, performance is reliably worse regardless of target features. Error bars show 95% confidence intervals.

In addition to their main effects, the relationship between target and distractor features also interacted significantly with a few other manipulated variables.

The Task * Distractor interaction ($F(1, 25) = 13.541, p = .001, \eta^2 = .351$) showed that the linear task was more susceptible to distractor effects than the numerosity task (see figure 6.6), while the Target * Distractor interaction ($F(1, 25) = 6.301, p = .019, \eta^2 = .201$) indicated that closed targets were more susceptible to distractor facilitation effects than open targets (see figure 6.5). This outcome agrees with some of the previous findings, in which closed symbol processing received a more drastic

impact by distractor feature than open symbols.

The distractor * overlap effect ($F(2, 56) = 3.853, p = .028, \eta^2 = .134$) resulted from the fact that different-feature trials showed a steady linear increase in RTs with increasing overlap while the same-feature pairs were more susceptible to distractor effects with high overlap.

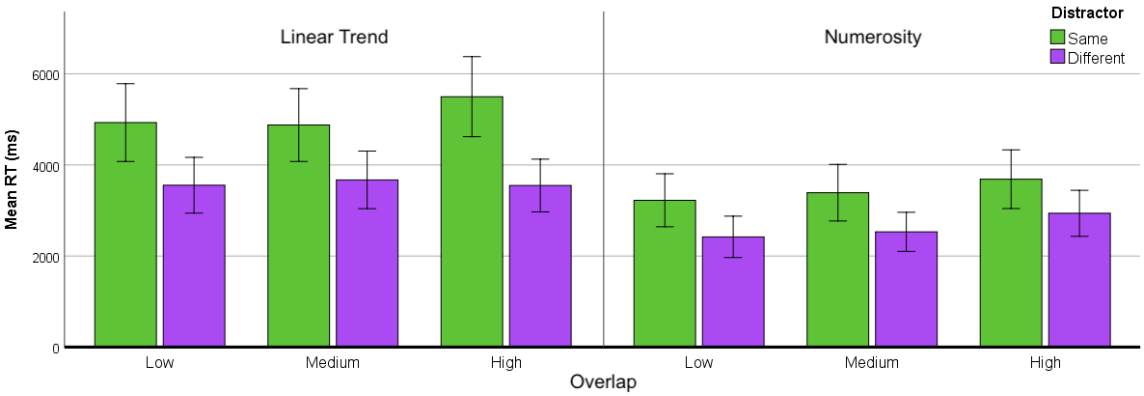


Figure 6.6: Triple interaction of Task, Overlap, and Distractor feature for single-plot response times. Linear Trend tasks took significantly longer than numerosity tasks, and RTs in both tasks were faster when symbols differed in boundary closure. Error bars show 95% confidence intervals.

Finally, there was a triple interaction of Task * Overlap * Distractor ($F(2, 50) = 11.336, p < .001, \eta^2 = .312$); see figure 6.6. To understand this effect and all of the other interactions, a follow-up simple effects analysis was conducted separately on each of the two tasks.

Linear Trend Task

In the analysis on only the Linear Task, strong significant effects of Target ($F(1, 25) = 8.179, p = .008, \eta^2 = .246$), Overlap ($F(1.635, 40.885) = 4.142, p = .030, \eta^2 = .142$), and Distractor feature ($F(1, 25) = 71.228, p < .001, \eta^2 = .740$), and a two-way interaction of Overlap by Distractor ($F(2, 50) = 9.904, p < .001, \eta^2 = .284$) were found.

While the variance of overlap contributed overall, the low (M = 4242 ms, SD = 348 ms), medium (M = 4274 ms, SD = 337 ms), and high (M = 4523 ms, SD = 342 ms) conditions did not differ significantly from each other. Open targets (M = 4487 ms, SD = 329 ms) took significantly longer than closed targets (M = 4206 ms, SD = 351 ms), and pairs with different open or closed features were significantly faster (M = 3591 ms, SD = 290 ms) than same-featured pairs (M = 5101 ms, SD = 398 ms).

As shown in Fig. 6.6 (left side), *different-featured pairs did not vary much across overlaps, and only the high overlap condition significantly lengthened RTs for same-featured pairs.*

However, in contrast to some of the previous findings and the overall interaction effects reported above, Target * Distractor (p = .067) did not rise to statistical significance, nor did the triple Target * Overlap * Distractor interaction (p = .196).

Numerosity Task

Outcomes were similar for the reanalysis of the numerosity task, with significant main effects of Target ($F(1, 25) = 5.033, p = .034, \eta^2 = .168$), overlap ($F(1.352, 33.810) = 24.379, p < .001, \eta^2 = .494$), and distractor ($F(1, 25) = 34.516, p < .001, \eta^2 = .580$). As with the linear trend analysis, same-feature distractors (M = 3435 ms, SD = 294 ms) reliably lengthened RTs in comparison to different-feature distractors (M = 2630 ms, SD = 219 ms), adhering to my expectations, and open targets continued to take significantly longer (M = 3135 ms, SD = 243 ms) than closed targets (M = 2929 ms, SD = 265 ms), albeit both targets were faster overall for numerosity than linear trend. As with the linear trend analysis, target * distractor (p = .092) and target * distractor * overlap (p = .949) did not rise to significance.

Unlike in linear trend trials, each level of overlap (low: M = 2823 ms, SD = 244 ms; medium: M = 2961 ms, SD = 245 ms; high: M = 3313 ms, SD = 270 ms) differed significantly from each other (low/medium: p = .006; medium/high: p = .001; low/high: p < .001), but there was no overlap by distractor interaction (p = .687) in the numerosity task trials.

6.2.2.2 Error Proportions

Just as with the response times, the errors were not significantly influenced by shape group (p = .166), nor did the task ordering group have a significant effect (p = .982) on error proportions. Data were collapsed across these between-subjects factors.

Mean error proportions varied from 3.2% to 23% across the experimental conditions, and many of the same effects and interactions rose to statistical significance as in the RT analyses.

Target feature ($F(1, 25) = 8.114, p = .009, \eta^2 = .243$) and Overlap ($F(2, 50) = 16.772, p < .001, \eta^2 = .402$) were both significant, with higher rates of errors generally mirroring the conditions with longer RTs. Open targets produced significantly more errors (M = .146, SD = .021) than closed targets (M = .112, SD = .014). Trend analysis on the different levels of overlap revealed significance for both linear and quadratic effects, but in this case the dominant effect size was for a quadratic trend in the overall error proportions (linear: $F(1, 25) = 11.368, p = .002, \eta^2 = .313$; quadratic: $F(1, 25) = 24.983, p < .001, \eta^2 = .500$). Low (M = .123, SD = .016) and Medium (M = .103, SD = .014) both differed significantly from High (M = .162, SD

= .022) overlap trials (p = .007, p < .001, respectfully).

Just as with RTs, Distractor ($F(1, 25) = 65.152, p < .001, \eta^2 = .723$) exerted a very strong significant effect on error proportion, with same-feature pairs (M = .173, SD = .020) producing double that of different-feature pairs (M = .086, SD = .014). It also interacted meaningfully in a Target * Distractor interaction ($F(1, 28) = 4.221, p = .049, \eta^2 = .131$) in which open targets produced reliably higher error rates than closed shapes, which produced particularly low error rates when the more numerous or linearly associated symbol was a closed shape paired with an open shape distractor; see figure 6.7.

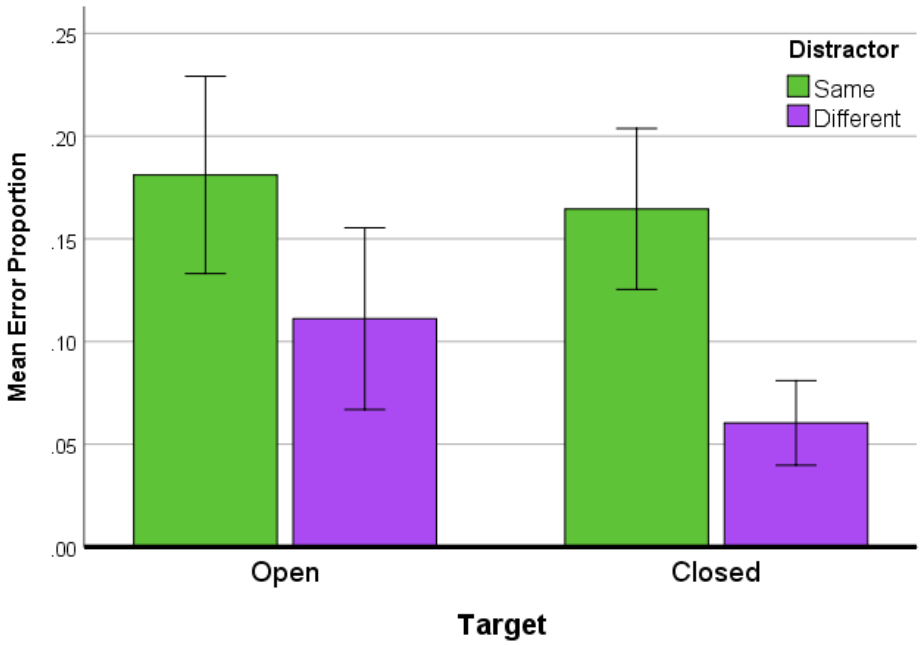


Figure 6.7: The Target by Distractor interaction in the single-plot Error Proportion analysis shows that different-featured distractors caused fewer errors and closed targets induced fewer errors overall, but closed targets with different-feature distractors were by far the most accurate condition. Error bars show 95% confidence intervals.

On the other hand, Task (p = .704) was not a contributing factor to the error proportions overall, but it did interact in the same triple interaction as in the general

RT analysis: Task * Overlap * Distractor ($F(1.560, 39.004) = 3.870, p = .039, \eta^2 = .134$). Same-feature trials clearly caused more errors than different-feature trials (see figure 6.7) and these differences were mediated by task requirements, so, as with the RT analysis, I followed up by reanalyzing the data separately for both tasks to help clarify the relationship of all the factors.

Linear Trend Task

In linear trend tasks, Overlap ($F(2, 50) = 14.050, p < .001, \eta^2 = .360$) and Distractor ($F(1, 25) = 31.487, p < .001, \eta^2 = .557$) both exhibited statistically significant effects on error proportions. Low (M = .119, SD = .024) and Medium overlap (M = .092, SD = .020) both differed significantly from High (M = .166, SD = .030) (p = .010 and p < 0.001, respectively), but Low and Medium did not induce a significant difference in error rates (.202). As hypothesized, the relationship between symbol features played a significant role in determining error rates in the linear trend task. Same-feature distractors (M = .165, SD = .028) had double the error rates of trials differing across the open/closed feature category (M = .087, SD = .020). In addition, Overlap * Distractor produced a significant interaction ($F(2, 50) = .3540, p = .036, \eta^2 = .124$); error proportions increased steadily with overlap level in different-feature trials, but the mean error proportion for low overlap trials with same feature symbols fell evenly between the medium and high levels of overlap; see figure 6.8.

Target (p = .104) no longer bore significance for error rates in linear trend trials, nor did it interact significantly with any other variables of interest.

Numerosity Task

Reanalysis on the numerosity error data echoed the main significant effects of Target

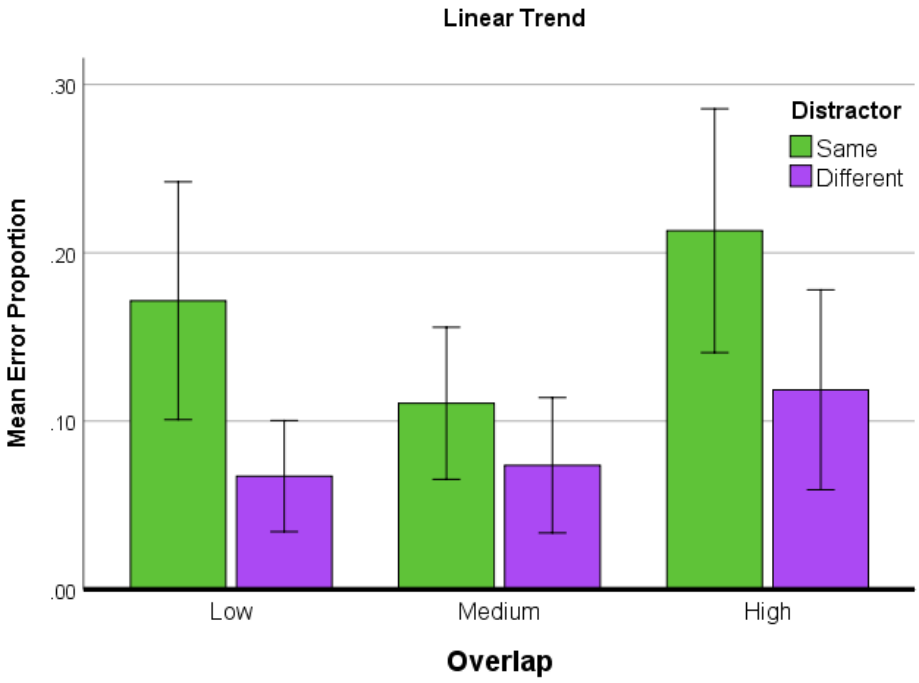


Figure 6.8: In the linear trend task, error rates increased steadily with different-featured symbols and larger overlap proportions, but the low overlap condition saw more errors than the medium overlap condition for same-featured shapes. There is no simple explanation for the performance differences between low and medium overlap trials, as all of the trials were created and analyzed properly. Error bars show 95% confidence intervals.

($F(1, 25) = 8.237, p = .008, \eta^2 = .248$), Overlap ($F(2, 50) = 6.573, p = .003, \eta^2 = .208$), and Distractor ($F(1, 25) = 67.763, p < .001, \eta^2 = .730$). The feature of the more numerous symbol set was influential on error rates, with open targets (M = .152, SD = .018) causing significantly more errors than closed targets (M = .113, SD = .010). As in the Linear trend task, Low (M = .127, SD = .013), Medium (M = .113, SD = .013) and High (M = .158, SD = .018) overlap levels significantly influenced error rates; medium and high levels differed significantly from each other (p = .007) in the numerosity task.

Distractor again proved the strongest determinant of error rates in numerosity tasks, with the continued trend of same-featured symbols (M = .181, SD = .015)

inducing significantly more errors than different-featured trials (M = .085, SD = .013), but distractor did not significantly interact with target (p = .288) in this case.

6.2.3 Shape Pair Analysis

To seek a more granular explanation for the effects and interactions in this experiment, I collapsed the RT and error proportion data across overlap levels for each combination of symbols. I then rank ordered the symbol pairs by their mean RTs across the linear trend and numerosity tasks. The only condition I was unable to consider was pentagon/fiveline; that pair was not presented together to participants in any of the numerosity trials. The resultant ordering found that all different-feature shape pairs produced faster RTs than same-feature pairs, further strengthening the effect of distractor feature reported in the overall RT and error proportion analyses. Due to this clean split, I chose to analyze the same and different-feature pairs separately with an ANOVA that tested for the effect of task by shape-pairs.

6.2.3.1 Response Times

I began with different-feature pairs and found a significant main effect of task ($F(1, 25) = 10.936, p = .003, \eta^2 = .304$) and a significant interaction of task * different-feature pair ($F(5.181, 129.523) = 2.760, p = .020, \eta^2 = .099$), but no significance for different-feature pairs as a main effect (p = .057). In accordance with previous findings, the linear trend task ($M = 3657, SD = 288$) took much longer than the numerosity task ($M = 2738, SD = 220$) in all cases except the circle/sixline pair.

Same-feature pairs also exhibited a significant main effect of task ($F(1, 25) = 24.397, p < .05, \eta^2 = .494$) and task * same-feature pair ($F(6.481, 162.025) = 5.529, p < .05, \eta^2 = .181$). In addition, there were significant differences in RTs among same-feature pairs ($F(5.608, 140.209) = 11.492, p < .05, \eta^2 = .315$). Again, linear trend tasks ($M = 5282, SD = 408$) took much longer than numerosity tasks ($M = 3554, SD = 311$), and many of the same-feature shape pairs were significantly different from each other.

6.2.3.2 Error Proportions

The error proportions for same- and different-feature pairs were moderately high, ranging from 5.6% to 24% with a mean of 12.3%. For different-feature pairs, no main effects or interactions reached significance (task: p = .889; different-feature pairs: p = .276; task*different-feature pairs: p = .812). Task (p = .453) and task * pair interactions (p = .099) were not significant for same-feature pairs either, however I did find a significant main effect of pair ($F(11, 275) = 3.022, p = .001, \eta^2 = .108$). That said, the only significant pairwise difference was between the circle/square and the sixline/fiveline pairs (p = .002).

6.2.4 Combined Measures

One of the assumptions inherent in reaction time studies is that errors will be low so that differences in RTs among conditions reflect the different constructs in the methodology. As mentioned in the analysis on error proportions in section 6.2.3.2, moderately high errors which reflected conditions in the tasks were introduced, weakening the interpretation of the RTs as a varying measure of performance by them-

selves. In cases such as these, it is useful to consider composite measures combining RTs and error rates so that variance in both measures is incorporated in that analysis. The most straightforward approach is to divide the mean trimmed correct RTs by the proportion correct (PC) in each condition, producing Inverse Efficiency Scores (IES) which can be analyzed and interpreted in the same fashion as the RTs and error proportions.

$$PC_{i,j} = 1 - ErrorProportion_{i,j} \tag{6.1}$$

$$IES_{i,j} = \frac{\bar{RT}_{i,j}}{PC_{i,j}} \tag{6.2}$$

(For each participant i and condition j)

In addition, participants were instructed to respond as quickly and accurately as possible to all trials. Speed accuracy tradeoffs (SATs) can arise in experiments such as these, where RTs and error rates can be affected asymmetrically by the cognitive processes under investigation and based on each subject's interpretation and application of the instructions [78, 92, 130]. There was no indication that the results in this study were systematically occluded by SATs; the RTs and error proportions had largely similar effects and pointed in the same direction. Appendix A discusses combined measures designed to account for SATs and presents further analyses on the shape pair data from this study.

6.2.4.1 Inverse Efficiency Scores (IES)

In the rank-ordering of symbol pairs, different-feature symbol pairs all had shorter RTs than pairs with symbols from the same feature category. As task interacted significantly with response times but not error rates in the shape pair analyses, I opted to produce inverse efficiency scores and examine shape pair performance separately for each task. I had run the analyses on error proportions in all previous studies, so I needed to compute PC using eq 6.1 before computing IES. I continued the reanalysis by computing the mean RT and mean PC for each pair of symbols separately for each task, and produced IES for each pair according to eq 6.2.

Tables 6.1 and 6.2 contain the average error rates, response latencies, IES, and their standard deviations for each pair in each of the analysis tasks, rank ordered by mean RT. Figure 6.9 presents the rank order and measure for each of the pairs before and after the IES transformation in both tasks. While many pairs of symbols changed rank, different-feature pairs were faster than same-feature pairs in almost all cases.

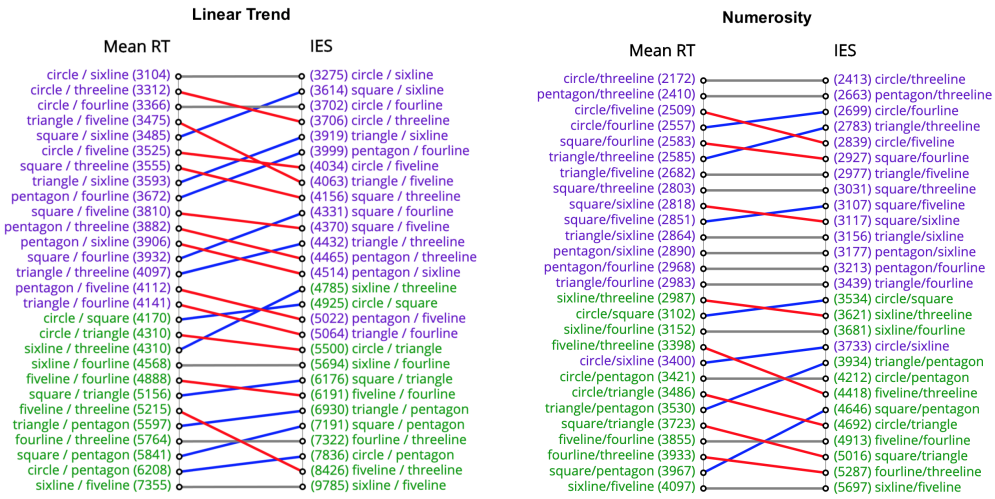
After computing IES for each pair of symbols in each task and examining their respective rank orderings, I analyzed the effect of the pairs on the variance in each task, and explored performance and pairwise significance among the pairs. In the Linear Trend task, shape pairs exerted a significant effect on IES ($F(27, 675) = 12.258, p < .001, \eta^2 = .329$), and many pairs were significantly different from each other. Figure 6.10 compares the magnitudes of each pair's IES, and figure 6.12 displays the significant differences among pairs of symbols. In general, same-feature pairs induced

Table 6.1: Pairwise combinations of symbols rank ordered by mean response times in the **Linear Trend task**. The first 16 purple shape pairs are all different-feature, and all were faster than the successive 12 (17-28) green same-feature pairs.

#	Shape Pair	RT (ms)		ERR		IES (ms)	
		Mean	SD	Mean	SD	Mean	SD
1	circle / sixline	3104	1167	0.048	0.123	3275	222
2	circle / threeline	3312	1293	0.077	0.170	3706	315
3	circle / fourline	3366	2245	0.077	0.161	3702	442
4	triangle / fiveline	3475	1403	0.096	0.188	4063	401
5	square / sixline	3485	2036	0.038	0.092	3614	396
6	circle / fiveline	3525	2323	0.100	0.152	4034	572
7	square / threeline	3555	1690	0.115	0.190	4156	355
8	triangle / sixline	3593	1307	0.085	0.116	3919	253
9	pentagon / fourline	3672	1473	0.077	0.137	3999	289
10	square / fiveline	3810	1668	0.108	0.129	4370	407
11	pentagon / threeline	3882	1829	0.100	0.162	4465	483
12	pentagon / sixline	3906	1894	0.085	0.152	4514	598
13	square / fourline	3932	1902	0.085	0.162	4331	366
14	triangle / threeline	4097	1618	0.062	0.110	4432	344
15	pentagon / fiveline	4112	2178	0.106	0.202	5022	674
16	triangle / fourline	4141	1824	0.125	0.203	5064	655
17	circle / square	4170	1833	0.128	0.178	4925	437
18	circle / triangle	4310	1743	0.147	0.207	5500	653
19	sixline / threeline	4310	2193	0.096	0.171	4785	442
20	sixline / fourline	4568	1591	0.154	0.188	5694	470
21	fiveline / fourline	4888	2036	0.167	0.200	6191	571
22	square / triangle	5156	2404	0.160	0.191	6176	515
23	fiveline / threeline	5215	2092	0.218	0.282	8426	1432
24	triangle / pentagon	5597	3122	0.186	0.185	6930	655
25	fourline / threeline	5764	2791	0.167	0.211	7322	758
26	square / pentagon	5841	2326	0.173	0.185	7191	529
27	circle / pentagon	6208	3213	0.167	0.200	7836	862
28	sixline / fiveline	7355	3459	0.218	0.187	9785	1004

Table 6.2: Pairwise combinations of symbols rank ordered by mean response times in the **Numerosity task**. All but one of the different-feature pairs (purple) was faster than the same-feature pairs (green).

#	Shape Pair	RT (ms)		ERR		IES (ms)	
		Mean	SD	Mean	SD	Mean	SD
1	circle/threeline	2172	694	0.087	0.140	2413	148
2	pentagon/threeline	2410	1117	0.090	0.135	2663	228
3	circle/fiveline	2509	1157	0.128	0.127	2839	209
4	circle/fourline	2557	1436	0.054	0.121	2699	275
5	square/fourline	2583	1342	0.077	0.184	2927	314
6	triangle/threeline	2585	1304	0.051	0.123	2783	274
7	triangle/fiveline	2682	929	0.083	0.127	2977	206
8	square/threeline	2803	1799	0.058	0.107	3031	419
9	square/sixline	2818	969	0.082	0.100	3117	226
10	square/fiveline	2851	1441	0.090	0.135	3107	268
11	triangle/sixline	2864	1778	0.077	0.114	3156	377
12	pentagon/sixline	2890	1231	0.100	0.117	3177	228
13	pentagon/fourline	2968	1351	0.071	0.096	3213	289
14	triangle/fourline	2983	1473	0.115	0.152	3439	327
15	sixline/threeline	2987	1318	0.167	0.163	3621	284
16	circle/square	3102	1418	0.109	0.115	3534	312
17	sixline/fourline	3152	1325	0.109	0.176	3681	317
18	fiveline/threeline	3398	1634	0.231	0.177	4418	386
19	circle/sixline	3400	2770	0.077	0.184	3733	548
20	circle/pentagon	3421	2456	0.179	0.156	4212	571
21	circle/triangle	3486	2392	0.224	0.205	4692	537
22	triangle/pentagon	3530	1579	0.109	0.141	3934	287
23	square/triangle	3723	1650	0.244	0.165	5016	417
24	fiveline/fourline	3855	1589	0.199	0.177	4913	387
25	fourline/threeline	3933	1775	0.186	0.172	5287	805
26	square/pentagon	3967	2368	0.147	0.136	4646	538
27	sixline/fiveline	4097	2106	0.263	0.190	5697	523



(a) (b)

Figure 6.9: Shape pairs rank ordered fastest to slowest (top to bottom) before and after IES transformation in (a) Linear Trend, and (b) Numerosity tasks. Red lines indicate worsening ranks, blue lines indicate increasing ranks, and grey lines indicate no change in rank. Many pairs changed order in both tasks, but most same-feature pairs (green) took longer than different-feature pairs (purple).

significantly higher IES than different-feature pairs. Shape pairs were also a significant determinant of performance in the Numerosity Task ($F(26, 650) = 9.106, p < .001, \eta^2 = .267$), with a number of significant differences among pairs; figure 6.11 shows the rank ordered IES magnitudes and figure 6.13 shows the pairwise significant differences among symbol pairs.

6.2.5 Discussion

6.2.5.1 Separate-plot Displays

Separate-plot displays in the visual summary tasks (chapter 5) found no significant effects of target feature or any interactions with difficulty level, suggesting that the feature categories of the marks were not relevant to the task when positions were encoded in separate same-feature category groups in the display. The response time

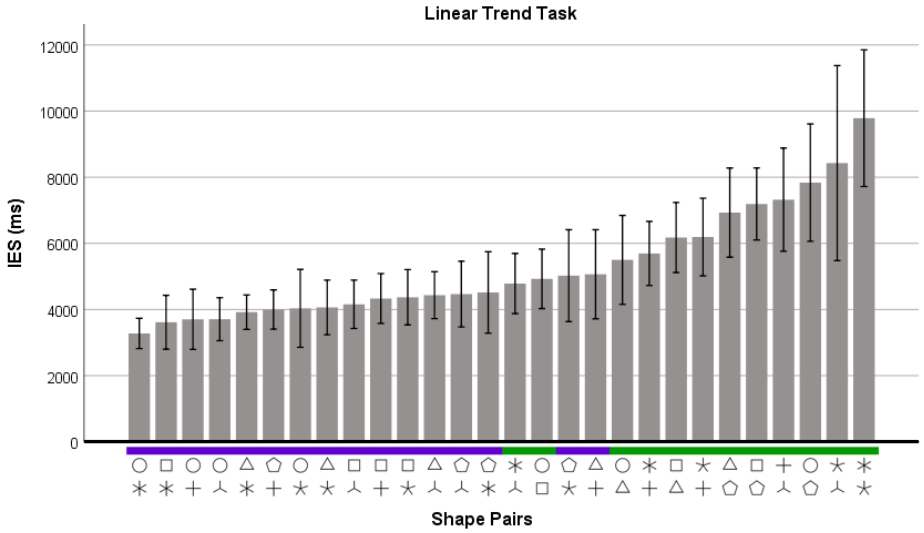


Figure 6.10: Shape pairs ordered by IES in the Linear Trend task. The pairs correspond to the rank orders and values displayed in Table 6.1 and figure 6.9 (a). With two exceptions, same-feature pairs (green) took longer than different-feature pairs (purple). Error bars show 95% confidence intervals.

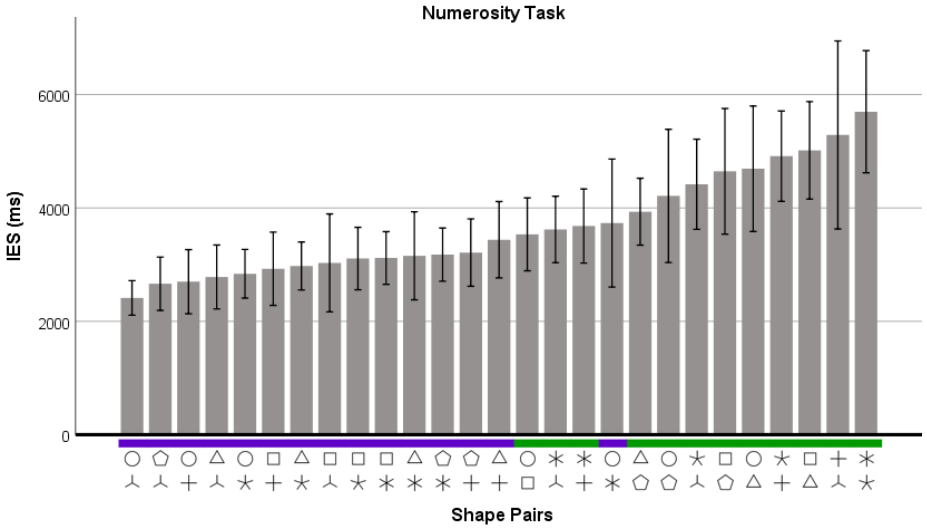


Figure 6.11: Shape pairs ordered by IES in the Numerosity task. The pairs correspond to the rank orders and values displayed in Table 6.2 and figure 6.9 (b). With a single exception, same-feature pairs (green) took longer than different-feature pairs (purple). It is interesting that the circle/sixline pair was the hardest different-feature pair, even harder than three of the same-feature pairs, while the same circle/sixline pair was the easiest symbol pairing in the linear trend task. Error bars show 95% confidence intervals.

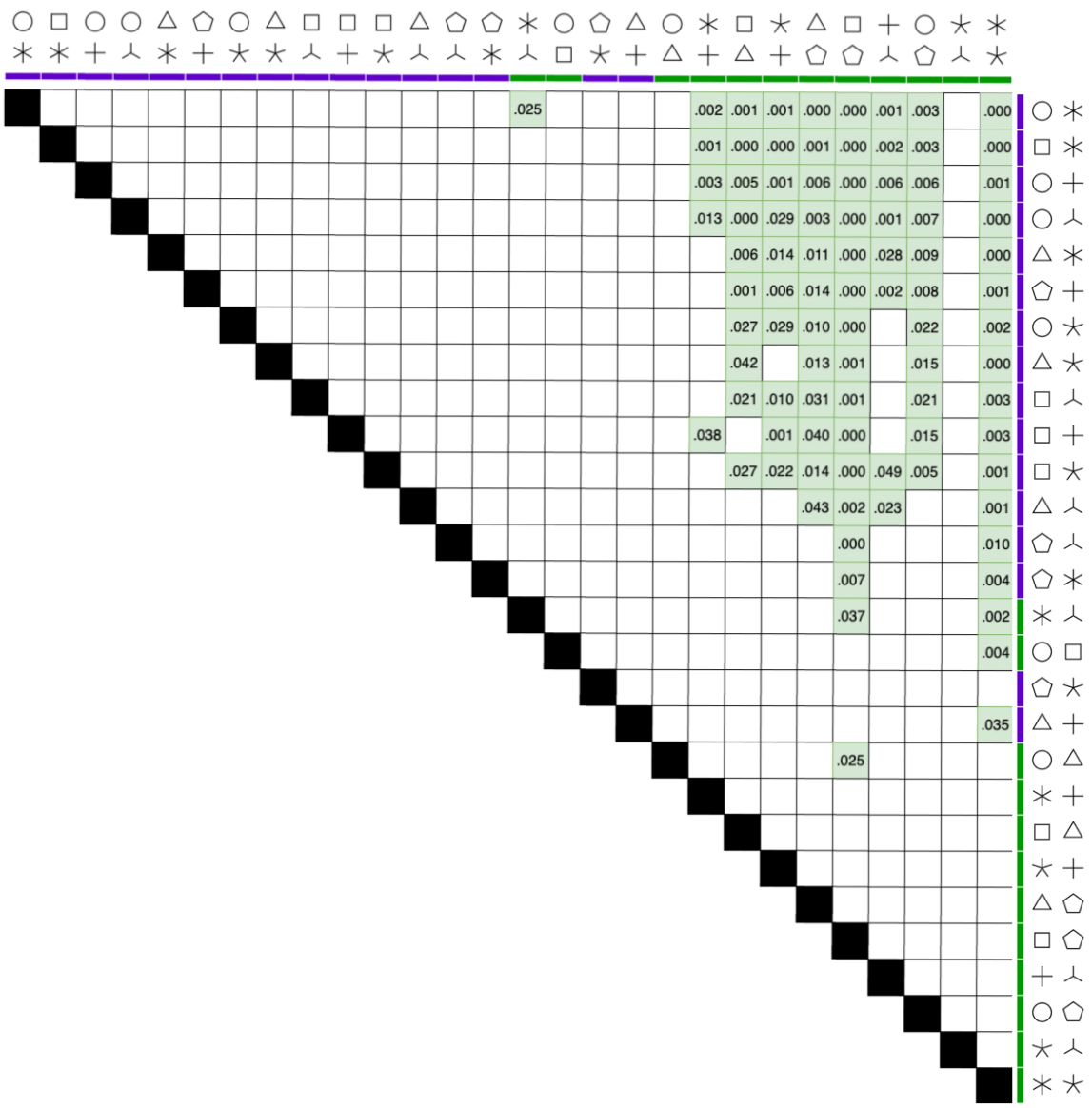


Figure 6.12: Significant differences among IES of shape pairs in the Linear Trend task. Most significance pairwise differences arose where same-feature pairs took longer than different-feature pairs; see figure 6.10 for comparison.

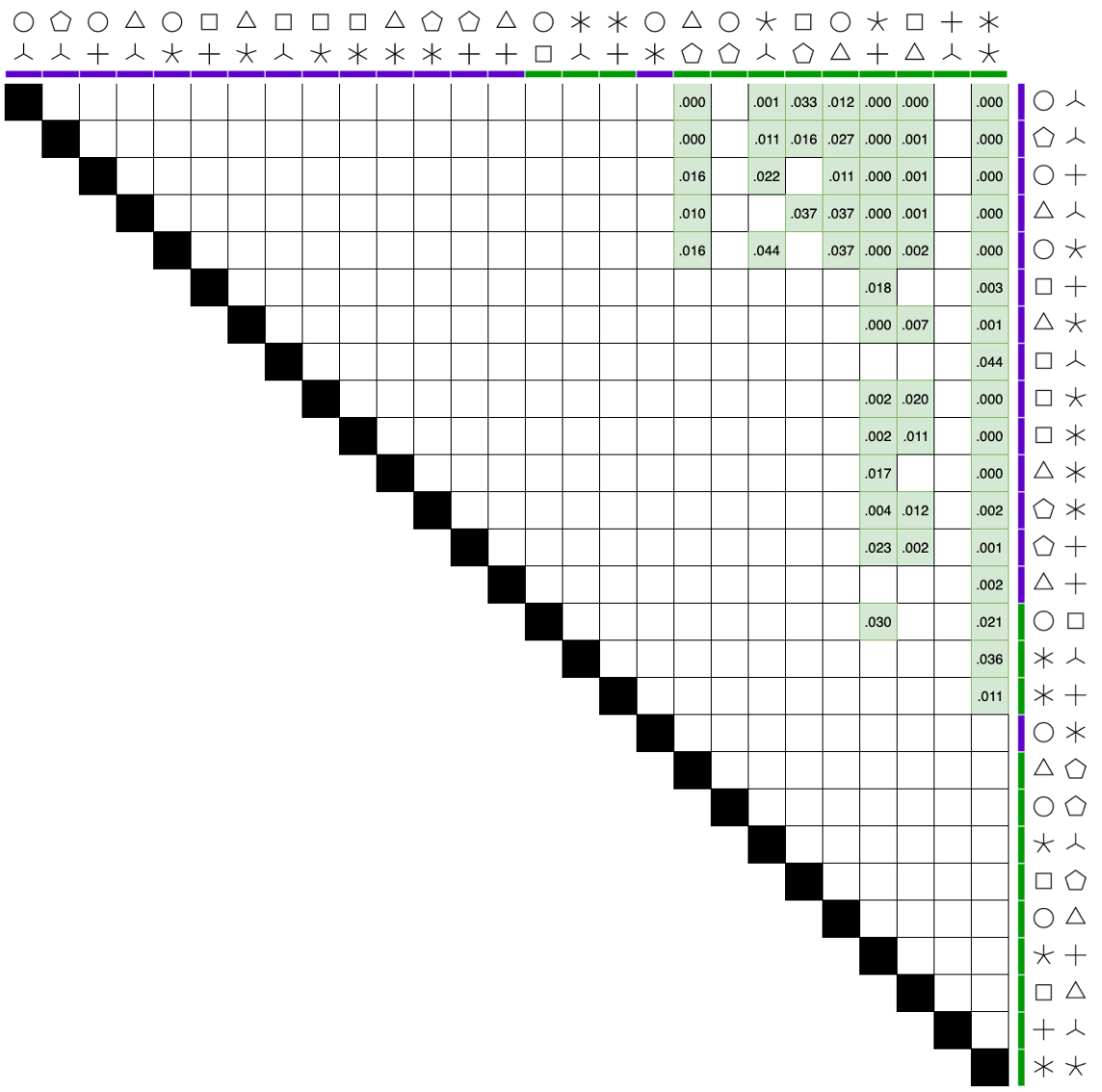


Figure 6.13: Significant differences among IES of shape pairs in the Numerosity task. As with the Linear Trend task, most significance pairwise differences among pairs arose where same-feature pairs took longer than different-feature pairs; see figure 6.11 for comparison.

and error data from this study were consistent with those findings. They show that participants are able to respond quickly and accurately to judgment tasks regardless of the target or overlap conditions when only a single symbol was used to encode marks in both of the displays.

6.2.5.2 Single-plot Displays

There were significant differences between the two tasks in the single-display component of this study, above and beyond the influence of the shape encodings or their feature relationships. Linear trend tasks with multiple shape encodings introduced significantly longer RTs and more errors than judgments on the numerosity task. The previous experiments with these tasks (chapter 5) exhibited the opposite relationship, with linear trend significantly faster than numerosity tasks. The requirements of the tasks and varying characteristics of the displays, particularly the use of realistic, overlapping data distributions, likely influences these findings, but the differences were robust regardless of the shapes and their featural relationships in this experiment.

Across the analyses of RT and error rates, closed targets symbols were processed faster and with fewer errors than open symbols, and displays with two open or two closed (i.e. same-featured) symbols were significantly harder than displays with one of each feature. The more basic perception experiments (chapters 3, 4) found evidence to support a preference for closed symbols, and results from the visual summary tasks (chapter 5) suggested that closed symbols might receive a more drastic impact by distractor feature than open symbols. I found some continued support for both of these findings. The overall analyses of RTs and errors in this study, the analyses of

RTs in both tasks, and the error rates in the numerosity task all support the notion that closed targets are faster and less error-prone than targets encoded with open symbols. Furthermore, the overall RT and error analyses and the analysis of error rates in the linear trend task lend further evidence that closed targets are particularly influenced by the lack of closure of the distractor symbol. Taken together, these results suggest that the closure of the shape encodings segments them more readily and facilitates summary judgments, particularly when their distractors are not as salient, but that the effect does not rely significantly on the features of other items in the display for particular tasks.

I hypothesized that larger levels of overlap among symbols due to overplotting would give subjects a harder time responding to the judgment tasks, and was interested to see whether this would differ depending on the shape features in the displays. Increasing the proportions of overlaps caused RTs to exhibit a linear increase and errors to exhibit a quadratic increase, and overall interactions showed that same-featured displays were more impacted by higher levels of overplotting.

Differences in RTs and error rates due to levels of overplotting were significant in the analysis of both tasks, with particular influence on both measures at the highest level of overlap and for same-feature symbol pairs, especially in the numerosity task. For the overlap by distractor interaction in the linear trend task, I could not find an explanation for the discrepancy between low and medium levels of overlap; all trials appeared randomly, were tagged and scored correctly, and this condition did not exhibit undue levels of trimming, and yet the same-featured trials were faster in the medium level of difficulty. In general, these findings lend further support to

the discriminability between open and closed symbols due to the relative stability heterogeneous encodings exhibited in the face of increased overplotting.

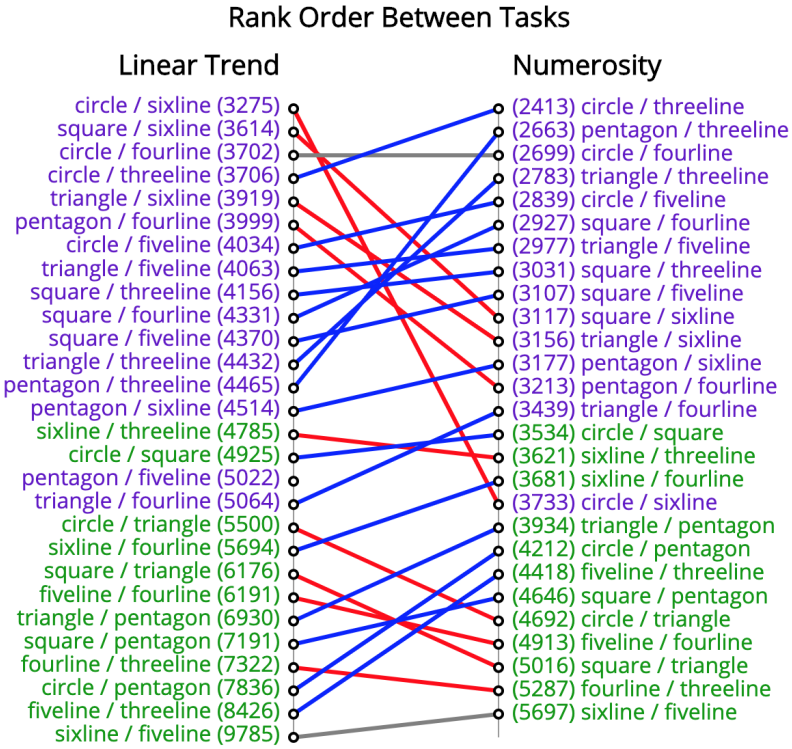


Figure 6.14: Shape pairs rank ordered by IES in both tasks. Different-featured pairs (purple) were almost all easier than same-featured pairs (green), but many of the pairs changed in rank order between the two tasks. Blue lines show a decrease in rank order (i.e. those pairs performed better in the Numerosity task) and red lines show an increase in rank order (i.e. those symbols performed worse in the Numerosity task).

The shape pair analyses provide strong support for the conclusion that it is more difficult to discriminate between pairs from the same open/closed category than from different categories. The rank orders of RT and IES (see figure 6.9) clearly show that different-feature pairs were more performant than almost all of the same pairs. Further, when the size of the symbols are kept constant and discrimination is made between same-featured shape encodings, it is harder to discriminate symbols that

have larger complexity (i.e. larger number of internal angles, more numerous line segments and endings) or more similarity. For example, sixline/fiveline had the worst performance of any pair in both tasks regardless of the measure, and fourline/threeline was also difficult in both tasks; both pairs differ minimally from each other in terms of more basic features, regardless of their topological features. In other cases, the differences in performance of particular shape pairs depended heavily on the task. Figure 6.14 displays the changes in rank order of IES measures between the two tasks. Circle/sixline was the easiest pair in the linear trend task, but was the most difficult different-feature pair in the numerosity task, even taking longer than three of the same-feature pairs. Many other pairs saw stark changes in rank order between the tasks, such as pentagon/threeline, which went from one of the more difficult different-featured pairs in the linear trend task to the second easiest in the numerosity task. While the most robust differences in performance seem to be driven by the categorical difference between featural or topological characteristics of open/closed pairs, the relative complexities of the symbols and the influence of task requirements clearly cause differences as well.

CHAPTER 7: REAL-WORLD DATA STUDY

The basic perception experiments (chapters 3, 4) found that closed symbols elicited better performance than open symbols, and results from the visual summary tasks and overplotting study (chapters 5, 6) found continued support for this and further suggested that closed symbols might receive a more drastic impact by distractor feature than open symbols. One final study was run using real world data to complete the sequence from basic perceptual tasks to application in realistic scatterplot displays, examine the influence of shape and symbol closure, and test previous findings.

More specifically, categorical variables were encoded using specific symbols in a variety of scatterplot displays (a) to test if closed shapes yield better performance in displays with realistic data distributions and (b) to validate the influence of perceptual dissimilarity between open and closed shapes. Support was found for both concepts; the relative distinctiveness of open/closed symbols was more important than differences within each group, and participants responded to the task more quickly and with fewer errors when target symbols were closed.

The following hypotheses were explored:

H1 *Pairs of symbols differing in open and closed features will produce faster RTs and fewer errors than pairs sharing open or closed features*

H2 *Closed-feature target shapes will be faster and more influenced by distractor*

features in comparison to open shapes

7.1 Toxics Release Inventory (TRI)

The Toxic Release Inventory (TRI) was established in 1986 as part of the Emergency Planning and Community Right-to-Know Act (EPCRA), which the United States congress passed in response to two deadly chemical spills in 1984 in Bhopal, India and in 1985 in West Virginia [37]. The TRI program is managed by the Environmental Protection Agency (EPA), which had been instantiated a decade and a half earlier in 1970 to bring climate research, policy-making, and enforcement under the same umbrella in the United States.

Among its specific provisions, the TRI program mandates that facilities report yearly usage above a minimum threshold for a variety of toxic chemicals, and tracks measures of treatment, energy recovery, recycling, and release into the atmosphere, landfills, and waterways for each of those chemicals, among many other variables. The current list of toxic or otherwise harmful chemicals includes over 595 individual chemicals in thirty-three categories, including carcinogens, and persistent bioaccumulative toxic chemicals, which can persist in bodies or the environment over long timespans. Many industrial facilities are required to report to the TRI program every year, from mining and manufacturing to waste management and federal facilities.

Public access to data from the TRI program provides an incentive for improved environmental performance. A variety of tools have been developed to explore the yearly datasets, sift through the complex relationships among the variables collected, and inform citizens, underrepresented groups, and policymakers about chemical usage

in their local areas, states, and country. The social and cultural relevance of this type of information, the volume of data submitted to the program in recent decades, and my previous work using data from TRI made this dataset a reasonable choice to underpin the current study [18, 20].

7.2 Methodology

The primary goal of the present study was to test the influence of open and closed symbols with more realistic displays and analysis contexts. I adapted the stimulus displays to support that goal by increasing the chart areas and decreasing the symbol sizes, and simplified the study design by using a subset of the symbols and only one relative judgment task. Some participants had struggled adapting to the eight symbols in the previous experiment, and the results were muddied somewhat by an effect of symbol ordering between participants, so I wanted to test the hypotheses and find further support for my findings without introducing too much variance in the methodological approach. I elected to use the numerosity task, as it had reliably produced faster responses and fewer errors than the linear trend task in the overplotting study in the previous chapter.

7.2.1 TRI Data Displays

I stepped through the most recent twenty years of TRI datasets and extracted total quantities of recovery, recycling, release, and treatment of each chemical at each unique facility, and stored the data in a nosql database to support flexible query construction for these investigations. These usage categories subsume multiple more specific methods, but they provide a good overview of how a given facility interacted

with chemicals on the TRI list in a given year, and therefore give us enough flexibility for the purposes of this study.

To test the interaction of the shape categories under investigation on real rather than synthetic data distributions while maintaining the use of relative judgment tasks, I generated a number of charts from the TRI data. While the numerosity task selected for the user study did not require any particular knowledge of the TRI data, nor were participants expected to consider chart axes or labels in their judgments, I briefly discussed the source and context of the data to invoke a sense of realism, both for the data distributions and the task itself.

7.2.2 Participants and Stimulus Materials

Sixteen student volunteers (five female, eleven male) were recruited as participants, and tested each individually in 20-minute sessions. As in all the previous studies, all participants were required to be over the age of 18, with 20/20 (or corrected to 20/20) vision and no history of visual impairment.

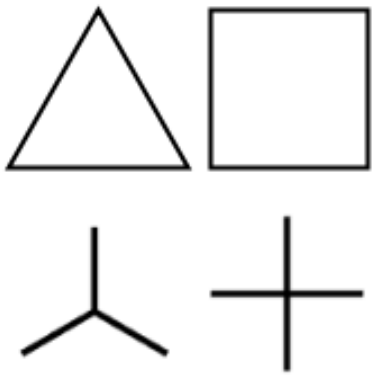
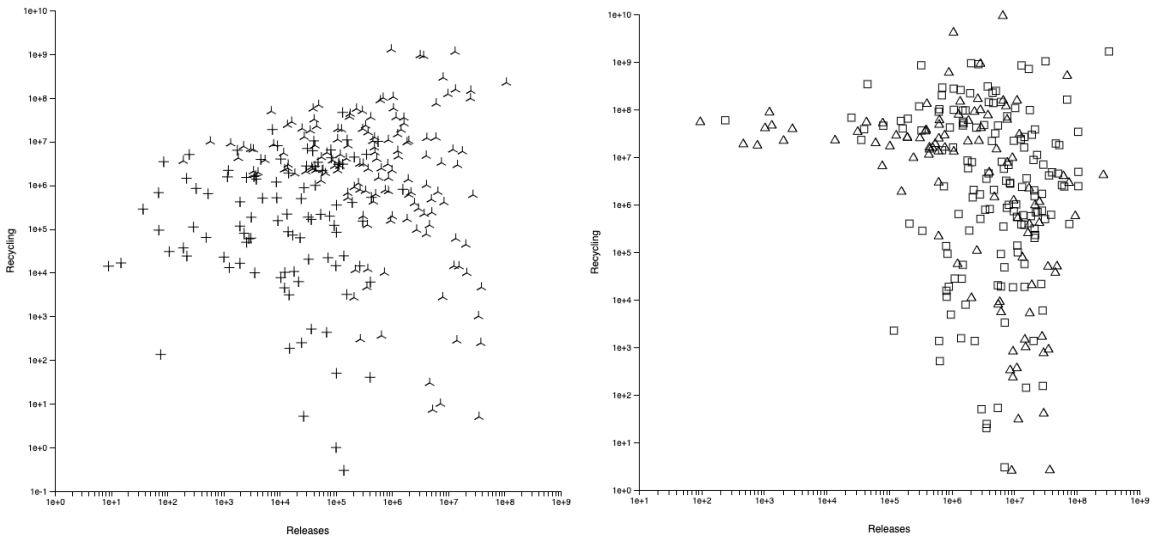


Figure 7.1: The symbol palette for the current experiment. The top row contains closed symbols and the bottom row contains the open symbols. Combinations of square, triangle, fourline, and threeline provided two same-feature pairs and four different-feature pairs.

A palette of four symbols was chosen, with two open and two closed shapes (see Fig. 7.1) that had exhibited similar RT performance when paired together in the numerosity task in the overplotting study in the previous chapter. Combinations of square, triangle, fourline, and threeline were used, providing two same-feature pairs and four different-feature pairs.

Stimuli were presented as black on a white background in regions measuring 700 x 700 pixels. Each shape was generated using a bounding circle with a radius of 6 pixels, producing stimuli subtending 0.3032 degrees of visual angle for participants 60cm away from the screen, within the bounds of symbol size for normal visual acuity [77]. All displays contained 250 points, with a more numerous set of 150 and a less numerous set of 100 symbols for each numerosity judgment. Stimulus displays were generated using D3 and presented using Superlab 5.0 on an iMac computer with a 27" flat screen retina display.

To meet the goal of moving beyond synthetic distributions, data from the TRI program was aggregated as described above and participants were asked to make numerosity judgments in comparisons involving facilities from pairs of US States. Each stimulus display contained a chart with data from two randomly selected states, with the top 150 and 100 facilities respectively contributing the most to two of the four usage metrics in that state, plotted on orthogonal axes using log scales (see Fig 7.2). Each facility was represented using a single mark, and the categorical encoding for each location was selected from the four symbols in the palette. Charts were reused with a few different combinations of symbols to produce a number of stimulus displays with similar overlaps and spatial characteristics.



(a) (b)

Figure 7.2: Sample stimulus displays from the TRI study. Symbols represent states, and each point represents an individual facility from the Toxics Release Inventory dataset [37]. Facilities are plotted based on their aggregate chemical usage: chemicals released into the environment (x-axis) vs quantity recycled (y-axis). (a) The same-feature pair of open symbols; (b) the same-feature pair of closed symbols.

A total of eighty stimulus displays were produced, with an even split of same-feature and different-feature trials, an even number of trials with each shape as the correct answer, and as close as possible to an even split of trials for each unique pair of shapes. The display order of the stimuli was randomized for each participant.

7.2.3 Procedure

After reading and signing an informed consent document, participants were placed 60 cm from the computer screen in a well-lit room. Every participant completed all the trials.

Each participant began with a description of the TRI data and the numerosity task, then saw instructions for the key responses. Shapes were mapped to the 'd, f, j, k' keys, and each hand had one open and one closed symbol. Participants then

completed eight practice trials, receiving feedback after each response to ensure they were comfortable with the tasks and key mappings. After successfully completing the practice trials, they responded to the eighty experimental trials.

Each trial contained two possible responses, and while participants were encouraged to memorize the four key mappings, a visual reminder for the key/shape associations was posted beneath the screen. Participants were instructed to respond to all trials as accurately and quickly as possible.

Within each experimental trial, participants were first presented with a fixation display for 500, 600, 700, 800, 900, or 1000 milliseconds, then the target display was presented until either a keypress was made or 30 seconds elapsed.

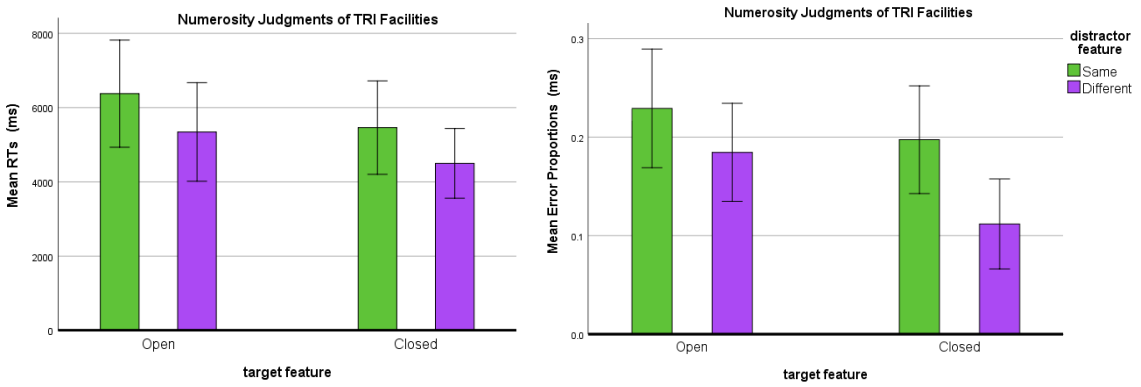
7.3 Analysis

An average of 3.7% of trials were removed for each participant, and the largest trim proportion was 7.5%. The ANOVAS for all sixteen participants tested for effects and interactions of target (open, closed) and distractor feature (same, different).

7.3.1 Target and Distractor Features

7.3.1.1 Response Times

The speed with which participants could respond to the judgment tasks in this experiment was significantly influenced by the feature of the target shape ($F(1, 15) = 9.907, p = .007, \eta^2 = .398$) and the distractor shape ($F(1, 15) = 19.230, p = .001, \eta^2 = .562$). Displays with more numerous open targets ($M = 5860, SD = 631$) took participants significantly longer than those with closed targets ($M = 4980, SD = 497$), and displays in which the symbols shared open or closed features ($M = 5918, SD = 613$)



(a) (b)

Figure 7.3: (a) While the target by distractor interaction was not significant for RTs in this study, main effects of both target feature (open vs closed) and distractor feature (same vs different) exerted significant effects on participants' response latency in this numerosity judgment task. Displays with two open symbols took the longest, and pairing a closed target with an open distractor produced the fastest responses. (b) Error proportions received a similar influence of target and distractor feature. Error bars show 95% confidence intervals.

required more time than displays with different features ($M = 4922, SD = 506$); see figure 7.3. The interaction between target and distractor feature was not significant (p = .875).

7.3.1.2 Errors

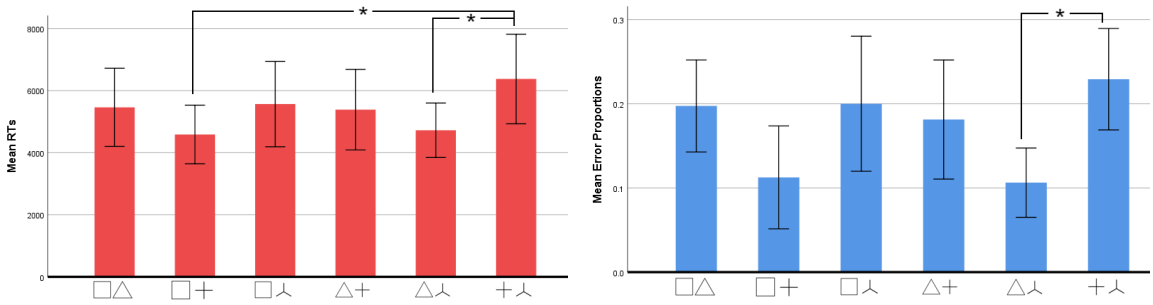
The error proportions across experimental conditions closely followed the effects on RTs, with target feature ($F(1, 15) = 5.819, p = .029, \eta^2 = .280$) and distractor feature ($F(1, 15) = 11.266, p = .004, \eta^2 = .429$) reaching significance and target * distractor feature failing to do so (p = .439); see figure 7.3. In the same pattern as with the response times, open targets ($M = .207, SD = .019$) induced significantly more errors than closed targets ($M = .155, SD = .019$), and subjects made more erroneous judgments with same-feature symbol pairs ($M = .213, SD = .022$) than

with different-feature pairs ($M = .148, SD = .013$).

7.3.2 Shape Pairs

7.3.2.1 Response Times

Each combination of TRI states and facilities was used to generate multiple stimulus displays with the same structure and distribution of data, differing only in the pairs of symbols used as mark encodings. Unsurprisingly, different pairs of symbols led to varying outcomes in RTs in this task ($F(5, 75) = 7.941, p < .05, \eta^2 = .346$). Pairwise differences among symbol pairs appear to be driven primarily by the difficulty of the fourline/threeline pairing and the relative ease of the square/fourline and triangle/threeline pairings, both of which differed significantly from fourline/threeline in follow-up Bonferroni comparisons (see figure 7.4).



(a) (b)

Figure 7.4: A comparison of (a) response times (ms), and (b) error proportions for each pair of symbols used in the TRI study. The fourline/threeline pair was the slowest and most erroneous, adhering to the findings with respect to open symbols and same-feature pairs. Square/fourline and triangle/threeline were the fastest and least error-inducing conditions. Error bars show 95% confidence intervals, and * indicates pairwise significance $< .05$.

7.3.2.2 Error Proportions

Different symbol pairs produced significant differences in error rates in this experiment ($F(5, 75) = 3.138, p = .013, \eta^2 = .173$). As with the RT analysis, fourline/threeline was the hardest, and square/fourline and triangle/threeline were the easiest, although only fourline/threeline and triangle/threeline were significantly different from each other in follow-up Bonferroni comparisons (see figure 7.4).

7.3.3 Inverse Efficiency Scores (IES)

Error rates were moderately high in this study, ranging from 11% to 23% in the target and distractor analysis and 11% to 22% in the shape pair analysis. As discussed in the previous chapter (6.2.4.1), measures combining RTs and error rates can be analyzed if high participant errors reflect the conditions in the experimental task and complicate the interpretation of the main RT measure. In order to have a cleaner analysis of performance in this task, I produced IES for the shape pairs and the target/distractor relationship using equations 6.1 and 6.2, and ran the same analyses as in the previous sections.

7.3.3.1 Target and Distractor Features

IES were significantly influenced by the open or closed features of the target symbol in the displays ($F(1, 15) = 9.111, p = .009, \eta^2 = .378$). Distractor feature ($F(1, 15) = 38.875, p < .05, \eta^2 = .722$) also imposed a strong significant effect on the combined measure. Reflecting the trends seen in the RT analysis in this study and in previous studies, open targets produced higher scores (M = 7486 ms, SD = 801 ms) than closed

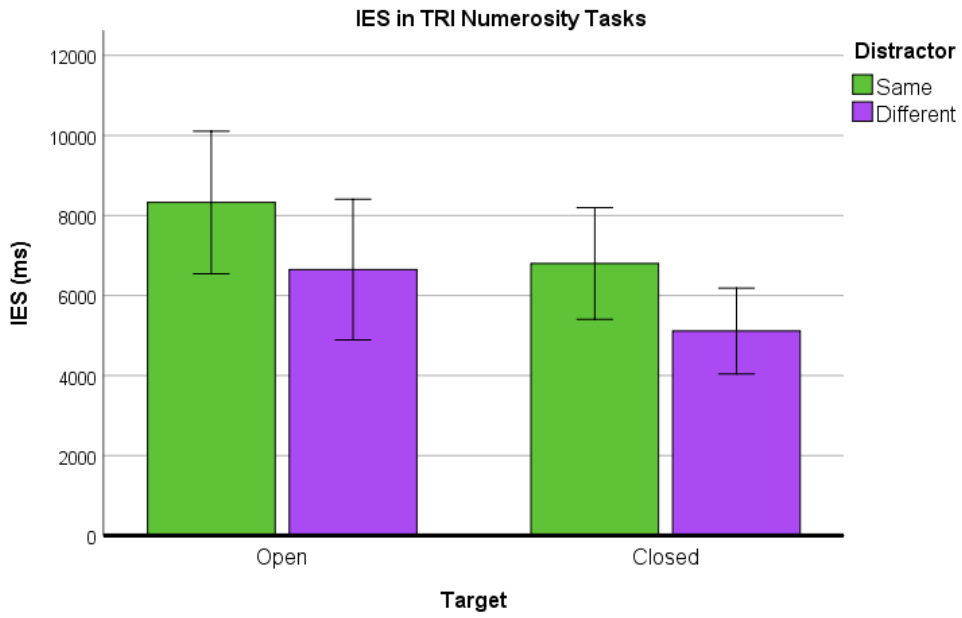
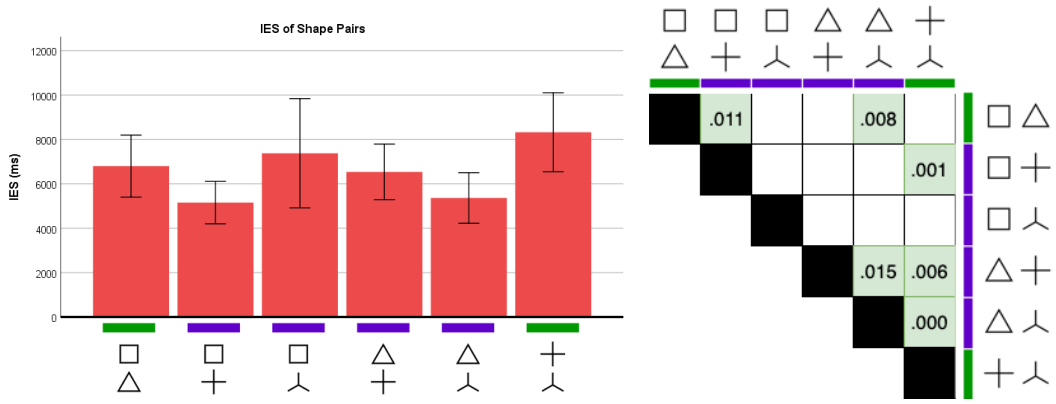


Figure 7.5: While the target by distractor interaction was not significant for IES, it is clear that open targets had worse performance than closed targets, and same-featured distractors produced worse scores than different-featured distractors in this task. Error bars show 95% confidence intervals.

targets (M = 5927 ms, SD = 549 ms), and same-feature pairs (M = 7563 ms, SD = 697 ms) had higher scores than different-feature pairs (M = 5880 ms, SD = 605 ms); see figure 7.5. As in the RT and error proportion analyses, no target by distractor interaction rose to significance (p = .987).

7.3.3.2 Shape Pairs

Reanalyzing IES for pairs of symbols showed a significant influence on participant performance ($F(2.430, 36.454) = 8.588, p < .05, \eta^2 = .364$). Whereas differences from the fourline/threeline pair seemed to drive significant pairwise differences in the RT and error proportion analyses (only two pairs and one pair of symbols were significantly different in these analyses, respectively; see figure 7.4), more pairs of symbols were significantly different from each other in the IES analysis. Figure 7.6



(a) (b)

Figure 7.6: (a) Differences among IES for pairs of symbols in this study. (b) Pairwise significance between pairs of symbols. Error bars show 95% confidence intervals.

displays the differences among pairs of symbols, (a) as magnitudes and (b) with significant differences among pairs.

7.3.4 Discussion

Effects of target closure and target-distractor feature relationships adhere to the expectations based on results from all the previous investigations. Relative judgment tasks with shape encodings are much faster and more accurate when symbols differ in boundary closure, and bounded (closed) shapes are easier to respond to than unbounded ones. The preference for closed symbols was fairly stark in this experiment; figure 7.3 shows how displays with two closed symbols took only marginally longer than displays with different-featured symbol categories and more numerous open items, and subjects had the easiest time overall with displays containing more numerous closed targets and open-featured distractors. Reanalysis of targets and distractors using inverse efficiency scores supported the same findings; see figure 7.5.

Results from the shape pair analysis provide mixed support for earlier findings.

Previous experiments found bounded symbols to be faster than open symbols and found a reliable separation between pairs of symbols sharing the open or closed feature category. Figure 7.4 demonstrates that the same-feature open pair was certainly the hardest condition, however the same-feature closed pair actually performed comparably to two of the four different-feature pairs. The two easiest conditions in this study across RTs and error proportions appear to be the square/plus and triangle/threeline pairs, both of which were significantly faster than the hardest condition, and the latter of which introduced significantly fewer errors. It is not clear why the square/threeline and triangle/fourline performed as poorly as the easier same-feature condition while the other two different-feature conditions were so much better.

Reanalysis of the shape pairs using IES revealed more significant pairwise differences. The square/fourline and triangle/threeline were still the easiest conditions, and both differed significantly from both the same-feature pairs with this combined measure. The other two different-feature pairs still did not perform better than same-feature pairs: triangle/fourline was indeed significantly slower than the open-featured fourline/threeline pair but was significantly longer than the different-featured triangle/threeline pair, square/threeline was not significantly different from any of the other pairs. These differences may be due to the displays themselves, as not every data distribution was used for every single pair of symbols due to a need to balance the number of same/different symbol pairs and the number of trials with each symbol as the correct answer, and the overlap was not controlled in the same way as the experiment in the previous chapter.

Overall, the use of the combined measure taking RT and error rates into account

brought the results from this study into closer alignment with findings from previous experiments. The two same-feature pairs produced significantly worse performance than either two or three, respectively, of the four different-feature pairs.

CHAPTER 8: CONCLUSIONS

The world we inhabit is full of natural objects of various shapes and sizes, and we must perceive, recognize, and categorize them in order to effectively navigate our daily lives. Artificial symbols such as letters, numerals, and road signs are also used to communicate more abstract information. Our visual system readily consumes all this stimulus and rapidly derives meaning and context relevant for a variety of tasks. Graphical perception and visual literacy have arisen to exploit these capacities, allowing us to offload cognitive effort onto the automaticity and parallelization of our visual processing systems when engaging with visual displays of complex data and their relationships.

The sequence of experiments described and analyzed in this document encapsulate a systematic approach to exploring a well-bounded subset of human graphical perception: perception of simple, two-dimensional shapes, and particular design decisions related to using shapes as mark encodings in multi-class scatterplot displays. The findings have direct applicability to certain visualization contexts and also provide evidence of interesting relationships and asymmetries in visual perception more generally. The earlier chapters (chapters 2, 3, 4) detail experiments using lower-level paradigms designed to assess perceptual and categorical similarities and differences among shapes. Successive chapters explored the influence of shape categories on higher-level ensemble judgments in visualization contexts (chapter 5), their suscep-

tibility to overplotting effects due to distributional characteristics (chapter 6), and applicability with a real-world dataset (chapter 7). The next sections include a general discussion of the experimental findings across all the studies, connections to other work in complementary areas, and a variety of ideas for future refinement of these ideas and directions for new investigations.

8.1 General Discussion of Experimental Findings

Taken together, the results across all the studies found strong support for the categorical differences between closed (bounded, polygonal) shapes and open (unbounded) shapes composed of line segments, and underscored the importance of these topological features when using simple, two-dimensional symbols as mark encodings in multi-class visualization displays.

There is a preference for processing closed symbols. The Flanker (chapter 3) and Same-Different (chapter 4) experiments both showed evidence that exemplars from the bounded feature category were processed more quickly in basic perceptual tasks. This result was robust in findings from both task paradigms, and the latter study expanded the set of shapes used as stimuli to increase confidence that the results were not tied specifically to the symbols in the Flanker study. Further evidence of a perceptual preference for processing closed symbols was found for the easiest difficulty level in the linear trend task, and to an extent depending on the distractor feature in the other tasks and difficulties, in the Visual Summary study (chapter 5), in both the linear trend and numerosity tasks in the Overplotting study (chapter 6), and in the Real-World Dataset study (chapter 7). Subjects responded more quickly

and with fewer errors when closed symbols were more numerous or more linearly associated, in some cases regardless of the distractor symbol and otherwise (particularly in chapter 5) when the distractors were from the distinct open category. *Overall, these results suggest that closed boundaries of shape encodings serves to segment them more readily, draw heightened attention to them, and facilitate more accurate summary judgments, particularly when distractors differ categorically.*

There is an interesting connection to research by Makovski [86], who provided empirical evidence of an illusory size effect between symbols with and without boundaries. This effect biases perception of symbols toward overestimations of size with open symbols. Perhaps the closedness of a symbol confers some processing advantage related to preattentive scene segmentation or attentional allocation throughout a display, in line with Chen's findings with respect to topological features [26]. On the other hand, maybe the the results obtained in these studies were driven in part by the introduction of this illusory size effect.

Shapes don't matter for separate displays. Results from the side-by-side visualization displays in the Visual Summary and Overplotting studies indicated that there was no difference in processing symbols, regardless of overlap among points or feature category or bounding, when single shapes were used alone to encode a set of points. It was only when different symbols were presented together as visual encodings in the same displays that differences became evident in those tasks. In displays with multiple shape encodings, 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, bolstering the categorical distinction be-

tween these types of symbols. The popularity of side-by-side approaches is evident from chart idioms such as scatterplot matrices, small multiples [126], and scagnostics [132]; these types of displays are great choices for exploring a moderate number of relationships among data, and comparisons can be made fairly rapidly by glancing between elements in the display. The findings here suggest that shape encodings would not meaningfully influence performance in those contexts.

Categorical differences are larger than differences within a category. The shape pair analyses in the Visual Summary Task, Overplotting, and Real-World Data tasks provided strong evidence for the categorical distinction between open and closed symbols. Their results support the notion that it is more difficult to discriminate between symbols within an open or closed feature category than between those categories. Although differences in the relative complexity of symbols caused variations across all pairs, the rank orders of performance highlighted the importance of categorical differences for multi-class displays, as all different-featured pairs were more performant than all same-featured pairs.

Shape pairs are also influenced by symbol complexity. For the individual pairwise differences among symbols across the relative judgment tasks, difficulty was driven by symbol complexity, such as the number of internal angles, and more numerous line segments and endings. The hardest pair in both tasks in the Overplotting study was sixline/fiveline, which paired the most complex symbols from the open feature category together. Other pairs that led to poor performance included fourline/threeline, circle/pentagon, and square/pentagon, each of which share features and have relatively high complexity. This is not surprising, and indeed is predicted through other

work on symbol discrimination in the literature. For example, similarity kernels produced by Demiralp and colleagues [33] projected subjective measures of similarity into rough categories separating circles, triangles of varying orientations, squares and diamonds, and open symbols composed of line segments [33]. After reordering a palette of symbols for maximal perceptual distance, circle and fourline were predicted as the first symbols to choose; data from my Overplotting study indeed showed circle/fourline to be among the easiest pairs. However, it does appear that individual pairwise similarity measures of symbols are not sufficient to predict their visibility in ensemble displays supporting analytical judgment tasks, despite being relatively scalable and robust first passes at the problem. For example, their MDS projections suggest that circle and square are roughly half the perceptual distance apart that circle and triangle are, and yet circle/triangle was harder in both tasks in the Overplotting study. Furthermore, figure 6.14 shows how many pairs of symbols were drastically different in the two tasks in that study. Much more work needs to be done to understand and model the features contributing to the relative perceptual distances among plotting symbols, particularly in conjunction with redundant encoding of visual variables such as size and color, which has been shown to influence segmentation in visualization displays [90].

8.2 General Discussion of Methodological Considerations

In the experiment in the visual summary tasks chapter, shapes were inscribed from a bounding square region, which potentially introduced a size effect since squares were larger than all other symbols. In the overplotting study, shapes were created used a

bounding circle region instead, following the symbol palette from Li et. al. [77]. The types of symbols used as marks have radial symmetry, so the bounding circle region may be more appropriate and grounded in the literature. Despite this modification, bounding circle region may still not be sufficient to control for size effects, let alone to create uniformity of luminance between shapes, especially if line widths are equal. Makovski's [86] findings regarding an illusory size effect for bounded shapes may also have exacerbated the open vs closed distinction in the experiments reported so far.

I originally intended to measure overplotting based on a visibility index approach similar to Urribarri and Castro's work [125], but extending from square filled regions to complex shapes raised more questions than answers. Following the visibility index approach would require discretizing the display and counting the number of 'boxes' (pixels or sub-regions of the display) containing part of more than one mark. It is unlikely that such a granular sum would have the most predictive power for ensemble perception of ensemble displays, not least because my experiments had already yielded evidence that higher-level shape features (i.e. topological characteristics) subsumed differences within a feature category in the basic perception studies described in earlier chapters. Instead, the Overplotting study measured the proportion of symbols that overlapped with other symbols in the display to generate different difficulty levels. Although overlap was constrained to keep pairwise distances within a particular range (i.e. between 25% and 75% of the maximum distance constituting overlap), and each shape had a maximum number of shapes it could overlap with, the measure did not much support nuance in the types of overall distributions or the different amounts of clustering that could occur with real data. Measuring the number of symbols

with overlaps was more appropriate for a rough characterization of clutter in the Overplotting study, but I saved enough data about each stimulus display that I could feasibly rebuild them and test more comprehensive metrics in the future. A significant amount of work still needs to be done to understand and model shape, color, and size channels, and their conjunctive interactions, so there is plenty of space for future refinement in this area alone.

This raises another question I grappled with: how does one define overlap? Do lines need to fully overlap, just touch, or simply have their bounding regions intersect? Full overlap between any pairs of symbols produces new topological features (new line crossings or 'holes'), and there is strong evidence that topological relationships between objects are encoded categorically (and categorical differences are more salient) [82]. Understanding and modeling topological differences en masse in ensemble displays is likely only part of the picture, and I would have been putting the cart before the horse to try and use that as a measure for the purpose of investigating open and closed symbols in cluttered displays. Future work informed by mathematical or perceptual models of symbols, both individually and within scenes, will shed light on this area, not to mention the additional influence of color, task constraints, and different statistical distributions of the underlying data.

It is worth highlighting the fact that differences in task performance changed somewhat in the studies that employed those tasks. Subjects in the Visual Summary tasks (chapter 5) were significantly faster in responding to linear trend tasks than to numerosity tasks, while this relationship was inverted in the Overplotting study (chapter 6), and the tasks in the latter experiment took longer overall. Numerosity tasks took

just under half a second longer on average, while linear trend tasks doubled in time taken in the Overplotting study. Both experiments presented 100 symbols in each display, and the task instructions were effectively the same, so other methodological choices must have contributed to the differences, such as the choice to use more realistic, overlapping data distributions, as well as a few differences in the symbol palette. In this case, distractor symbols were randomly placed throughout the display in the linear trend task in the Visual Summary experiments, while the distractors for the same task in the Overplotting experiment were drawn from a normal distribution. The lack of overlap in the Visual Summary experiments may have meant that the linear target symbol was overrepresented in the center of the display and distractor symbol was more peripheral, potentially biasing subjects' responses to that task, adding to the large differences in performance described earlier. Although I made efforts across all my studies to anticipate and mitigate the types of strategies subjects might employ, and design stimuli to most parsimoniously address the research questions at hand, it is always possible that additional sources of variance crept in.

When comparing the Overplotting study and Real-World Dataset study, a few methodological differences were introduced. The chart region was larger and each symbol was smaller in the latter study in order to test larger data quantities while taking up proportional screen space. As discussed in the section on stimulus materials in chapter 7, the symbol palette was also halved to reduce the cognitive load on the participants, and the displays were not controlled for any measure of overlap. In both of these experiments, I used the exact same sets of points as the foundation for multiple stimulus items with different shape encodings in order to decouple the shape

pair results from the underlying data distributions.

One point of potential interest is the similarity of the structure and orientations of the lines comprising the pairs of symbols in the two easiest conditions in the Real-World Dataset study (square/fourline and triangle/threeline; fig 7.4). The open symbols (fourline, threeline) look rather like the topological shape skeletons [42, 5] of their respective closed symbol pairing (square, triangle), and the lines are either all oriented vertically and horizontally for the square/fourline pair or mostly angled at similar orientations for two of the three arms of the threeline and triangle symbols. I would have guessed that both factors would make the symbols more confusable for each other, and therefore make those two conditions more difficult than other different-feature pairs as a result. Indeed, when considering the ensemble coding of basic image features in early visual areas, large enough differences in orientation of perceived lines and boundaries should pop out 'preattentively,' and yet the more similar different-feature pairs were the best combinations in this study. It remains unclear why these results arose, although they do not meaningfully alter the overall findings with respect to the categorical differences among open and closed shapes.

8.3 Future Work

There are a number of research directions to follow and edge cases to probe beyond the scope of this work. For example, what is it about closed symbols that causes them to be segmented and summarized more rapidly than open symbols? How well can new measures of symbol similarity covary with performance in ensemble displays, and how can this be joined with new taxonomies of tasks in visualization displays?

How can shapes be employed as encodings in other multivariate contexts, perhaps representing ordinal data, and how do redundant encodings with color, size, opacity, and motion build upon or reshape these findings?

Work still needs to be done to understand relative size perception among different types of shapes. The studies in this document relied on bounding square and bounding circular regions, as did the noteworthy experiments in the literature [125, 76]. New methods for generating different shapes in such a way that luminance and size are controlled may be required. One approach is to display pairs of shapes, perhaps in busy displays with other shape or color distractors, and ask participants to adjust sliders controlling dominant shape features (stroke width of lines, bounding region, etc.) until they look identical at different levels of gaussian filtering. Validating results from such an experiment could use a same-different task, compute just-noticeable-differences, and construct psychometric functions to model the influence of stroke width and bounding region on perceptibility. Such an approach could hone in on uniformity among symbols, and could provide an opportunity to refine subjective or modeled similarity metrics for existing or arbitrary symbols, and extend work on adjusting sizes to minimize overplotting effects.

It would be valuable to have a meta-analysis combining the basic analytics tasks and their sub-tasks from taxonomies, either specifically for scatterplots [104] or more generally [108, 21, 17, 107, 103], with experimental studies that have employed them to see if any trend arises in RTs, PCs, or combined measures, or whether gaps exist that could be explored. Work by Saket and colleagues [103] is particularly relevant, as they analyzed the types of basic visual analysis tasks tasks best supported by

a variety of standard visualization approaches. The relative judgments used in my studies are only a small sampling of the types of tasks that are useful to consider, and asymmetries are likely to arise in a variety of contexts I did not consider directly.

Another avenue for future work would be to investigate whether certain shapes or types of symbols are more susceptible to over- or underestimation in busy visual displays. The numerosity task gave some data on how quickly and accurately subjects could determine the larger set of symbols, but the relative judgment and the fixed delta and set sizes did not provide the opportunity to explore this question.

One useful direction will be to continue examining the parsimony of the open/closed construct. It is possible that the results we obtained are not tied to specific shapes or combinations of primitive features, but rather luminance and size effects for the stimuli used in our tasks. Smart and Szafir [110] found asymmetric influence of shape on size and color perception which differed based on symbol features, Li et. al. found size effects to dominate shape differences in symbol discrimination [76, 77], Bergen and Adelson [8] suggested that relative texture-element size has high explanatory power over low-level features like line endings and crossings, and Chen [26] has argued a compelling case for more intermediate perceptual organization (instead of focusing on lowest-level features). We attempted to control for these types of effects when designing our stimulus materials by constraining the bounding region of each shape and testing the selections in short pilot studies, but the base-level similarity among symbols and the mapping of these features into visualization contexts is by no means a solved problem.

A discussion of symbol and shape differences considering topological differences

may also yield fruit, perhaps in tandem with investigations of other models of shape or symbol and their relative similarity. Lovett and Franconeri [82] studied changes in categorical relationships among shapes (overlap, above, etc) and found that these differences were more noticeable than equal metric distance changes that did not alter categorical relationships. This may have implications for the outcomes we obtained from our Overplotting and Real-World Dataset studies, where the influence of overlapping points in analysis contexts was explored. Topological characteristics of shapes and topological relationships among those shapes likely influence the relative discriminability of regions in cluttered displays, biasing perception of numerosity or clustering, or drawing saccadic eye movements under brief viewing conditions. Future studies could focus specifically on whether variations in emergent topological relationships covary significantly with performance in visualization tasks. Additionally, computational vision models predicting fixations and saccadic eye movements based on low- and intermediate-level features in data displays should be explored, either in order to refine those models with task-relevant information or to help explain human performance in those contexts [94].

Exploration of more general computational approaches to understanding human shape perception could also inform future work on more basic shapes, and the influence of those shapes in visual analysis contexts. The current state-of-the-art models of biological shape perception involve medial axis skeletons, and it is noteworthy that the open and closed symbols considered in the experiments described in this document look like basic shapes and their own topological skeletons [42, 40]. Future work is required to understand whether mathematical models of symbols can correlate with

or predict their similarity, and work in vision science can help illuminate the processes underlying their perception.

In visualization design, the shape channel is most commonly used to represent categorical variables; symbols are not drawn from a continuous shape space, nor do we tend to perceive different shapes along such a continuum. Aside from some consideration by Chung et. al. [27], who found that star-like shapes with varying numbers of points were perceived as more orderable than polygonal shapes, very little work has been done to explore shapes as encodings for other types of data. Future studies could consider combinations of categorical and ordinal data, using closed and open symbols respectively, and test participant performance in multivariate displays.

Shapes can also play a role in more complex visual environments than two-dimensional charts. Encoding the relative positions and trajectories of multiple dynamic objects, such as drones, planes, or people, either in two- or three-dimensional space, is likely to differ in efficacy depending on the features and differences among the symbols used. The results from my studies support the notion that closed symbols are more salient, even in larger quantities, so it would be interesting to test whether fill, boundary closure, and other topological characteristics follow similar patterns in visualization of motion.

REFERENCES

[1] G. Aisch. Mastering multi-hued color scales with chroma. js. *Online. Accessed*, 23, 2013.

[2] R. Amar, J. Eagan, and J. Stasko. Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pages 111–117. IEEE, 2005.

[3] D. Ariely. Seeing sets: Representation by statistical properties. *Psychological science*, 12(2):157–162, 2001.

[4] F. Attneave and M. D. Arnoult. The quantitative study of shape and pattern perception. *Psychological bulletin*, 53(6):452, 1956.

[5] V. Ayzenberg and S. F. Lourenco. Skeletal descriptions of shape provide unique and privileged perceptual information for object recognition. *bioRxiv*, page 518795, 2019.

[6] B. Balas. Seeing number using texture: How summary statistics account for reductions in perceived numerosity in the visual periphery. *Attention, Perception, & Psychophysics*, 78(8):2313–2319, 2016.

[7] S. Bateman, R. L. Mandryk, C. Gutwin, A. Genest, D. McDine, and C. Brooks. Useful junk?: the effects of visual embellishment on comprehension and memorability of charts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2573–2582. ACM, 2010.

[8] J. R. Bergen and E. H. Adelson. Early vision and texture perception. *Nature*, 333(6171):363–364, 1988.

[9] J. Bertin. Semiology of graphics: diagrams, networks, maps. 1983.

[10] E. Bertini and G. Santucci. Quality metrics for 2d scatterplot graphics: automatically reducing visual clutter. In *International Symposium on Smart Graphics*, pages 77–89. Springer, 2004.

[11] E. Bertini and G. Santucci. Give chance a chance: modeling density to enhance scatter plot quality through random data sampling. *Information Visualization*, 5(2):95–110, 2006.

[12] E. Bertini, A. Tatu, and D. Keim. Quality metrics in high-dimensional data visualization: An overview and systematization. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2203–2212, 2011.

[13] R. Borgo, J. Kehrer, D. H. Chung, E. Maguire, R. S. Laramee, H. Hauser, M. Ward, and M. Chen. Glyph-based visualization: Foundations, design guidelines, techniques and applications. In *Eurographics (STARs)*, pages 39–63, 2013.

[14] D. Borland and R. M. T. Ii. Rainbow color map (still) considered harmful. *IEEE computer graphics and applications*, 27(2):14–17, 2007.

[15] M. Bostock, V. Ogievetsky, and J. Heer. D^3 data-driven documents. *IEEE transactions on visualization and computer graphics*, 17(12):2301–2309, 2011.

[16] R. Brath. Metrics for effective information visualization. In *Information Visualization, 1997. Proceedings., IEEE Symposium on*, pages 108–111. IEEE, 1997.

[17] M. Brehmer and T. Munzner. A multi-level typology of abstract visualization tasks. *IEEE transactions on visualization and computer graphics*, 19(12):2376–2385, 2013.

[18] D. Burlinson, K. Koehrn, K. Subramanian, and A. Lu. Are environmental regulations working? a visual analytic approach to answering their impact on toxic emissions. In *Proceedings of the Workshop on Visualisation in Environmental Sciences*, pages 17–21. Eurographics Association, 2016.

[19] 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, 2018.

[20] D. Burlinson, K. Subramanian, and A. Lu. Tri-direct: Interactive visual analysis of tri data. *Electronic Imaging*, 2016(1):1–8, 2016.

[21] S. M. Casner. Task-analytic approach to the automated design of graphic presentations. *ACM Transactions on Graphics (ToG)*, 10(2):111–151, 1991.

[22] L.-Y. Chang, Y.-C. Chen, and C. A. Perfetti. Graphcom: A multidimensional measure of graphic complexity applied to 131 written languages. *Behavior research methods*, 50(1):427–449, 2018.

[23] H. Chen, S. Engle, A. Joshi, E. D. Ragan, B. F. Yuksel, and L. Harrison. Using animation to alleviate overdraw in multiclass scatterplot matrices.

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

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

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

[27] D. H. Chung, D. Archambault, R. Borgo, D. J. Edwards, R. S. Laramee, and M. Chen. How ordered is it? on the perceptual orderability of visual channels. In *Computer Graphics Forum*, volume 35, pages 131–140. Wiley Online Library, 2016.

[28] W. S. Cleveland, P. Diaconis, and R. McGill. Variables on scatterplots look more highly correlated when the scales are increased. *Science*, 216(4550):1138–1141, 1982.

[29] 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.

[30] J. Cottam, A. Lumsdaine, and P. Wang. Overplotting: Unified solutions under abstract rendering. In *Big Data, 2013 IEEE International Conference on*, pages 9–16. IEEE, 2013.

[31] Q. Cui, M. Ward, E. Rundensteiner, and J. Yang. Measuring data abstraction quality in multiresolution visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):709–716, 2006.

[32] L. S. Davis. Understanding shape: Angles and sides. *IEEE Trans. Computers*, 26(3):236–242, 1977.

[33] 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.

[34] G. Dutilh, D. van Ravenzwaaij, S. Nieuwenhuis, H. L. van der Maas, B. U. Forstmann, and E.-J. Wagenmakers. How to measure post-error slowing: a confound and a simple solution. *Journal of Mathematical Psychology*, 56(3):208–216, 2012.

[35] J. Elder and S. Zucker. The effect of contour closure on the rapid discrimination of two-dimensional shapes. *Vision Research*, 33(7):981–991, 1993.

[36] G. Ellis and A. Dix. A taxonomy of clutter reduction for information visualisation. *IEEE transactions on visualization and computer graphics*, 13(6):1216–1223, 2007.

[37] EPA. Toxics release inventory. `https://www.epa.gov/toxics-release-inventory-tri-program/`.

[38] 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.

[39] 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.

[40] J. Feldman and M. Singh. Bayesian estimation of the shape skeleton. *Proceedings of the National Academy of Sciences*, 103(47):18014–18019, 2006.

[41] S. Few and P. Edge. Solutions to the problem of over-plotting in graphs. *Visual Business Intelligence Newsletter*, 2008.

[42] C. Firestone and B. J. Scholl. 'please tap the shape, anywhere you like': Shape skeletons in human vision revealed by an exceedingly simple measure. *Psychological Science*, 25(2):377–386, 2014.

[43] 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.

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

[45] P. Garrigan. The effect of contour closure on shape recognition. *Perception*, 41(2):221–235, 2012.

[46] 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.

[47] R. L. Goldstone. Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, 123(2):178, 1994.

[48] N. Gueugneau, T. Pozzo, C. Darlot, and C. Papaxanthis. Daily modulation of the speed–accuracy trade-off. *Neuroscience*, 356:142–150, 2017.

[49] J. Haberman and D. Whitney. Seeing the mean: ensemble coding for sets of faces. *Journal of Experimental Psychology: Human Perception and Performance*, 35(3):718, 2009.

[50] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Advances in neural information processing systems*, pages 545–552, 2007.

[51] A. Harrison and R. Etienne-Cummings. An entropy based ideal observer model for visual saliency. In *2012 46th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2012.

[52] A. Harrison, M. A. Livingston, D. Brock, J. Decker, D. Perzanowski, C. Van Dolson, J. Mathews, A. Lulushi, and A. Raglin. The analysis and prediction of eye gaze when viewing statistical graphs. In *International Conference on Augmented Cognition*, pages 148–165. Springer, 2017.

[53] 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.

[54] M. Harrower and C. A. Brewer. Colorbrewer. org: an online tool for selecting colour schemes for maps. *The Cartographic Journal*, 40(1):27–37, 2003.

[55] C. G. Healey, K. S. Booth, and J. T. Enns. High-speed visual estimation using preattentive processing. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 3(2):107–135, 1996.

[56] 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*, pages 203–212. ACM, 2010.

[57] A. Inselberg. The plane with parallel coordinates. *The visual computer*, 1(2):69–91, 1985.

[58] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research*, 40(10-12):1489–1506, 2000.

[59] B. Johnson and B. Shneiderman. *Tree-maps: A space-filling approach to the visualization of hierarchical information structures*. IEEE, 1991.

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

[61] B. Julesz and J. Bergen. Textons, the fundamental elements in preattentive vision and perception of textures, readings in computer vision: issues, problems, principles, and paradigms, 1987.

[62] 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.

[63] D. A. Keim, M. C. Hao, U. Dayal, H. Janetzko, and P. Bak. Generalized scatter plots. *Information Visualization*, 9(4):301–311, 2010.

[64] Y. Kim and J. Heer. Assessing effects of task and data distribution on the effectiveness of visual encodings. In *Computer Graphics Forum*, volume 37, pages 157–167. Wiley Online Library, 2018.

[65] K. Koffka. *Principles of Gestalt psychology*. Routledge, 2013.

[66] J. Krause, A. Dasgupta, J.-D. Fekete, and E. Bertini. Seekaview: An intelligent dimensionality reduction strategy for navigating high-dimensional data spaces. In *2016 IEEE 6th Symposium on Large Data Analysis and Visualization (LDAV)*, pages 11–19. IEEE, 2016.

[67] H. Lam, M. Tory, and T. Munzner. Bridging from goals to tasks with design study analysis reports. *IEEE transactions on visualization and computer graphics*, 24(1):435–445, 2017.

[68] D. Lamy, L. Alon, T. Carmel, and N. Shalev. The role of conscious perception in attentional capture and object-file updating. *Psychological science*, 26(1):48–57, 2015.

[69] D. Lamy and H. E. Egeth. Attentional capture in singleton-detection and feature-search modes. *Journal of Experimental Psychology: Human Perception and Performance*, 29(5):1003, 2003.

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

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

[72] P. A. Legg, E. Maguire, S. Walton, and M. Chen. Quasi-hamming distances: An overarching concept for measuring glyph similarity. 2015.

[73] P. A. Legg, E. Maguire, S. Walton, and M. Chen. Glyph visualization: A fail-safe design scheme based on quasi-hamming distances. *IEEE computer graphics and applications*, 37(2):31–41, 2017.

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

[75] J. Li, J.-B. Martens, and J. J. Van Wijk. Judging correlation from scatterplots and parallel coordinate plots. *Information Visualization*, 9(1):13–30, 2010.

[76] 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*, pages 2553–2562. ACM, 2010.

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

[78] H. R. Liesefeld, X. Fu, and H. D. Zimmer. Fast and careless or careful and slow? apparent holistic processing in mental rotation is explained by speed-accuracy trade-offs. *Journal of experimental psychology: learning, memory, and cognition*, 41(4):1140, 2015.

[79] H. R. Liesefeld and M. Janczyk. Combining speed and accuracy to control for speed-accuracy trade-offs. *Behavior research methods*, 51(1):40–60, 2019.

[80] A. Light and P. J. Bartlein. The end of the rainbow? color schemes for improved data graphics. *Eos, Transactions American Geophysical Union*, 85(40):385–391, 2004.

[81] G. Loffler. Perception of contours and shapes: Low and intermediate stage mechanisms. *Vision research*, 48(20):2106–2127, 2008.

[82] A. Lovett and S. L. Franconeri. Topological relations between objects are categorically coded. *Psychological science*, 28(10):1408–1418, 2017.

[83] M. R. Luo, G. Cui, and B. Rigg. The development of the cie 2000 colour-difference formula: Ciede2000. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur*, 26(5):340–350, 2001.

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

[85] J. Mackinlay, P. Hanrahan, and C. Stolte. Show me: Automatic presentation for visual analysis. *IEEE transactions on visualization and computer graphics*, 13(6):1137–1144, 2007.

[86] T. Makovski. The open-object illusion: size perception is greatly influenced by object boundaries. *Attention, Perception, & Psychophysics*, 79(5):1282–1289, 2017.

[87] J. Matejka, F. Anderson, and G. Fitzmaurice. Dynamic opacity optimization for scatter plots. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 2707–2710. ACM, 2015.

[88] T. Munzner. *Visualization Analysis and Design*. CRC Press, 2014.

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

[90] 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, 2017.

[91] L. T. Nowell. *Graphical encoding for information visualization: using icon color, shape, and size to convey nominal and quantitative data*. PhD thesis, Virginia Tech, 1997.

[92] D. Paoletti, M. D. Weaver, C. Braun, and W. van Zoest. Trading off stimulus salience for identity: A cueing approach to disentangle visual selection strategies. *Vision research*, 113:116–124, 2015.

[93] D. G. Pelli, C. W. Burns, B. Farell, and D. C. Moore-Page. Feature detection and letter identification. *Vision research*, 46(28):4646–4674, 2006.

[94] R. J. Peters and L. Itti. Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007.

[95] R. A. Rensink. On the prospects for a science of visualization. In *Handbook of human centric visualization*, pages 147–175. Springer, 2014.

[96] R. A. Rensink. The nature of correlation perception in scatterplots. *Psychonomic bulletin & review*, 24(3):776–797, 2017.

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

[98] H. Reuss, A. Kiesel, and W. Kunde. Adjustments of response speed and accuracy to unconscious cues. *Cognition*, 134:57–62, 2015.

[99] B. E. Rogowitz and L. A. Treinish. Data visualization: the end of the rainbow. *IEEE spectrum*, 35(12):52–59, 1998.

[100] R. Rosenholtz, J. Huang, A. Raj, B. J. Balas, and L. Ilie. A summary statistic representation in peripheral vision explains visual search. *Journal of vision*, 12(4):14–14, 2012.

[101] R. Rosenholtz, Y. Li, and L. Nakano. Measuring visual clutter. *Journal of vision*, 7(2):17–17, 2007.

[102] J. Ross and D. C. Burr. Vision senses number directly. *Journal of Vision*, 10(2):10–10, 2010.

[103] B. Saket, A. Endert, and C. Demiralp. Task-based effectiveness of basic visualizations. *IEEE transactions on visualization and computer graphics*, 2018.

[104] A. Sarikaya and M. Gleicher. Scatterplots: Tasks, data, and designs. *IEEE transactions on visualization and computer graphics*, 24(1):402–412, 2018.

[105] A. Satyanarayan, D. Moritz, K. Wongsuphasawat, and J. Heer. Vega-lite: A grammar of interactive graphics. *IEEE transactions on visualization and computer graphics*, 23(1):341–350, 2017.

[106] J. Schneidewind, M. Sips, and D. A. Keim. Pixnostics: Towards measuring the value of visualization. In *Visual Analytics Science And Technology, 2006 IEEE Symposium On*, pages 199–206. IEEE, 2006.

[107] H.-J. Schulz, T. Nocke, M. Heitzler, and H. Schumann. A design space of visualization tasks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2366–2375, 2013.

[108] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *The craft of information visualization*, pages 364–371. Elsevier, 2003.

[109] D. Skau and R. Kosara. Arcs, angles, or areas: Individual data encodings in pie and donut charts. In *Computer Graphics Forum*, volume 35, pages 121–130. Wiley Online Library, 2016.

[110] 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*, page 669. ACM, 2019.

[111] S. S. Stevens. On the psychophysical law. *Psychological review*, 64(3):153, 1957.

[112] M. Stone, D. A. Szafir, and V. Setlur. An engineering model for color difference as a function of size. In *Color and Imaging Conference*, volume 2014, pages 253–258. Society for Imaging Science and Technology, 2014.

[113] H. Strasburger, I. Rentschler, and M. Jüttner. Peripheral vision and pattern recognition: A review. *Journal of vision*, 11(5):13–13, 2011.

[114] D. A. Szafir. Modeling color difference for visualization design. *IEEE transactions on visualization and computer graphics*, 24(1):392–401, 2018.

[115] 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.

[116] A. Tatu, P. Bak, E. Bertini, D. Keim, and J. Schneidewind. Visual quality metrics and human perception: an initial study on 2d projections of large multidimensional data. In *Proceedings of the International Conference on Advanced Visual Interfaces*, pages 49–56. ACM, 2010.

[117] A. Torsello and E. R. Hancock. A skeletal measure of 2d shape similarity. *Comput. Vis. Image Underst.*, 95(1):1–29, July 2004.

[118] A. Treisman and S. Gormican. Feature analysis in early vision: Evidence from search asymmetries. *Psychological review*, 95(1):15, 1988.

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

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

[121] J. Tsotsos, I. Kotseruba, and C. Wloka. A focus on selection for fixation. *Journal of Eye Movement Research*, 9(5):1–34, 2016.

[122] E. R. Tufte. *Envisioning information.* Graphics Press, 1990.

[123] E. R. Tufte. *The visual display of quantitative information*, volume 2. Graphics press Cheshire, CT, 2001.

[124] S. Uddenberg, G. Newman, and B. Scholl. Perceptual averaging of scientific data: Implications of ensemble representations for the perception of patterns in graphs. *Journal of Vision*, 16(12):1081–1081, 2016.

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

[126] S. van den Elzen and J. J. van Wijk. Small multiples, large singles: A new approach for visual data exploration. In *Computer Graphics Forum*, volume 32, pages 191–200. Wiley Online Library, 2013.

[127] 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.

[128] A. B. Watson. Perimetric complexity of binary digital images: Notes on calculation and relation to visual complexity. 2011.

[129] D. Whitney, J. Haberman, and T. Sweeny. From textures to crowds: multiple levels of summary statistical perception. *The new visual neurosciences*, pages 695–710, 2014.

[130] W. A. Wickelgren. Speed-accuracy tradeoff and information processing dynamics. *Acta psychologica*, 41(1):67–85, 1977.

[131] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[132] L. Wilkinson, A. Anand, and R. Grossman. Graph-theoretic scagnostics. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pages 157–164. IEEE, 2005.

[133] C. Wloka, I. Kotseruba, and J. K. Tsotsos. Saccade sequence prediction: Beyond static saliency maps. *arXiv preprint arXiv:1711.10959*, 2017.

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

[135] K. Wongsuphasawat, D. Moritz, A. Anand, J. Mackinlay, B. Howe, and J. Heer. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *IEEE transactions on visualization and computer graphics*, 22(1):649–658, 2016.

APPENDIX: OVERPLOTTING BIS ANALYSIS

A Speed Accuracy Tradeoffs

Speed accuracy tradeoffs (SATs) can arise when response times and error rates are measured as dependent variables in a study. RTs and error rates can be affected asymmetrically by the cognitive processes under investigation and based on each subject's interpretation and application of the instructions [78, 92, 130], and can vary across trials or within specific conditions, wittingly or otherwise [48, 98], or due to phenomenon such as post-error slowing [34]. SATs can confound or mask the effects being studied, potentially leading to conflicting or spurious conclusions.

One approach for mitigating the effect of SATs in behavioral psychology experiments is to combine response times and proportion of correct (PC) responses into a new measure for analysis. Perhaps the most common approach is to divide the mean trimmed correct RTs by the proportion correct in each condition, producing Inverse Efficiency Scores (IES), as shown in the section on combined measures 6.2.4 in chapter 6 and the section on IES 7.3.3 in chapter 7. Liesefeld and Janczyk have more recently proposed the Balanced Integration Score (BIS) as another method for combining RTs and PCs data, and have shown that it is less susceptible to the types of speed accuracy tradeoffs subjects can employ [78, 79].

Although there were no specific indications of SATs in the overplotting study, I chose to run further analyses using BIS to explore the data and findings. This decision was due in part to the RT and error proportion analyses, which found different interactions with task and same- or different-feature pairs. Those differences are likely

to be simply reflections of individual performances, but the procedure is nevertheless instructive.

B Balanced Integration Scores (BIS)

I computed BIS for each subject and condition from the original shape pair data using eq A.1 and explored the main effects and interactions, as with the IES values, using repeated measures ANOVAs. First, I took a birds-eye view and examined all shape pairs and tasks together, and found significant main effects of task ($F(1, 25) = 13.287, p = .001, \eta^2 = .347$) and shapePair ($F(26, 650) = 22.810, p < .05, \eta^2 = .477$), and a significant task * shapePair interaction ($F(26, 650) = 2.528, p < .05, \eta^2 = .092$).

$$BIS_{i,j} = Z_{PC_{i,j}} - Z_{\bar{RT}_{i,j}}$$
$$\text{with } Z_{x_{i,j}} = \frac{x_{i,j} - \bar{x}}{S_x} \tag{A.1}$$

(For each participant i and condition j)

Because BIS is a standardized score computed over all observed mean RTs and PCs, its range of values reflects relative performance compared to the average performance. The previous examination of the effect of the two tasks in the Overplotting study showed linear trend producing longer RTs or higher error proportions, and this analysis comports with that effect by showing how the linear trend task produced below-average performance (M = -.284, SD = .167) and the numerosity task produced above-average performance (M = .286, SD = .101) compared to the overall average (see fig A.2).

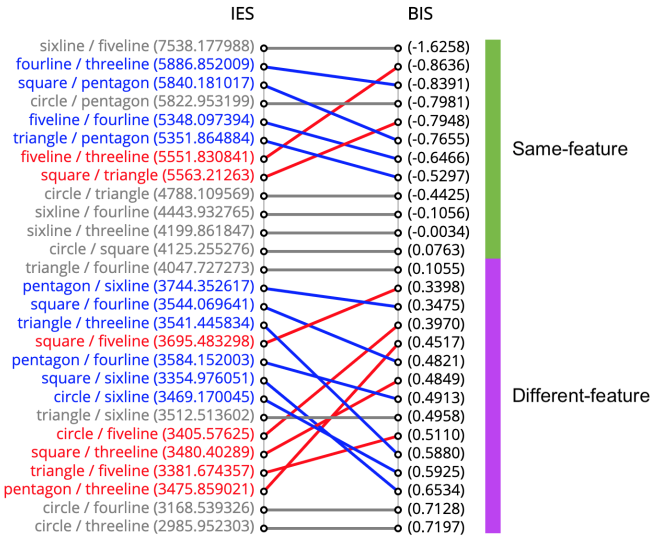


Figure A.1: Shape pairs rank ordered for BIS transformations. Same-feature pairs all took longer than different-feature pairs. Lines are colored blue or red if they decreased or increased in rank order, respectively. BIS are measured as standardized values (see eq A.1).

Many of the symbol pairs differed significantly from each other in the overall analysis. Ranking the shape pairs based on their BIS value produced a number of changes in order, but none of the rank changes crossed the boundary between same- and different-feature pairs; see figure A.1. I reanalyzed the BIS values separately for both types feature pairs.

The analysis found a significant main effect of task ($F(1, 25) = 7.582, p = .011, \eta^2 = .233$) for different-featured pairs, but no significant main effect of differentPair (p = .067) or task * differentPair interaction (p = .084). Keeping in mind that the different-featured pairs were all easier than same-featured pairs, the linear trend task (M = .268, SD = .157) was closer to the mean performance than the numerosity task (M = .715, SD = .099) for this subset. Beyond the influence of task, the differences among the different-featured pairs did not rise to significance, nor did they interact

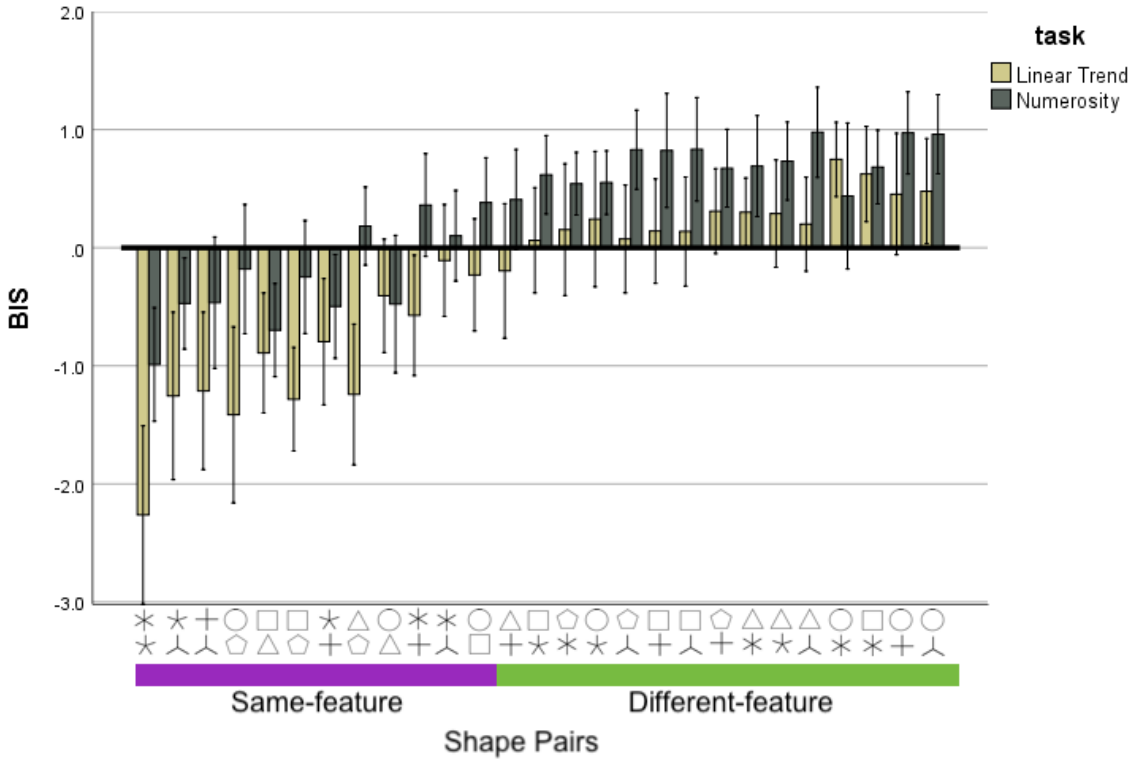


Figure A.2: BIS Task * ShapePair interaction across all symbol pairs. Each symbol pair is ranked according to the mean BIS across both tasks, with poorer scores to the left and better scores to the right. Keeping in mind that BIS reflect relative performance compared to the average across all conditions, it can clearly be seen that the linear trend task induced below-average performance and the numerosity task induced above-average performance across the majority of symbol pairs. In addition, it is noteworthy that the 12 left-most pairs are all same-feature pairs, and the 15 right-most pairs are all different-feature pairs.

meaningfully with task.

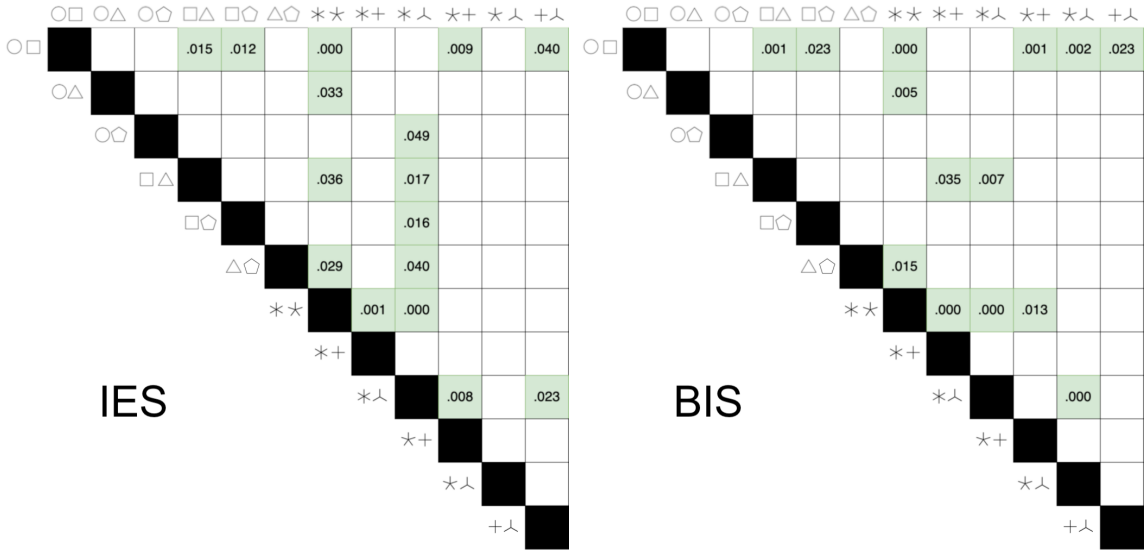


Figure A.3: Significant differences among same-feature shape pairs in the Overplotting study (chapter 6) compared between (a) IES and (b) BIS. P values $< .05$ from pairwise Bonferroni comparisons are shown.

For the same-feature pairs, both main effects of task ($F(1, 25) = 16.865, p < .05, \eta^2 = .403$) and samePair ($F(7.142, 178.541) = 9.22, p < .05, \eta^2 = .269$) reached significance, as did their interaction ($F(11, 275) = 3.057, p = .001, \eta^2 = .109$). The linear trend task (M = -.973, SD = .193) was much further from the average performance than the numerosity task (M = -.250, SD = .120) for the same-featured pairs; circle/triangle was the lone exception. Figure A.2 shows the task * shapePair interaction, with the 12 left-most (i.e. worst performing) pairs including all 12 same-feature pairs. Within the same-feature pairs, many pairs differed significantly from each other; figure A.3 shows the significant pairwise differences.

APPENDIX: SAMPLE STIMULUS MATERIALS

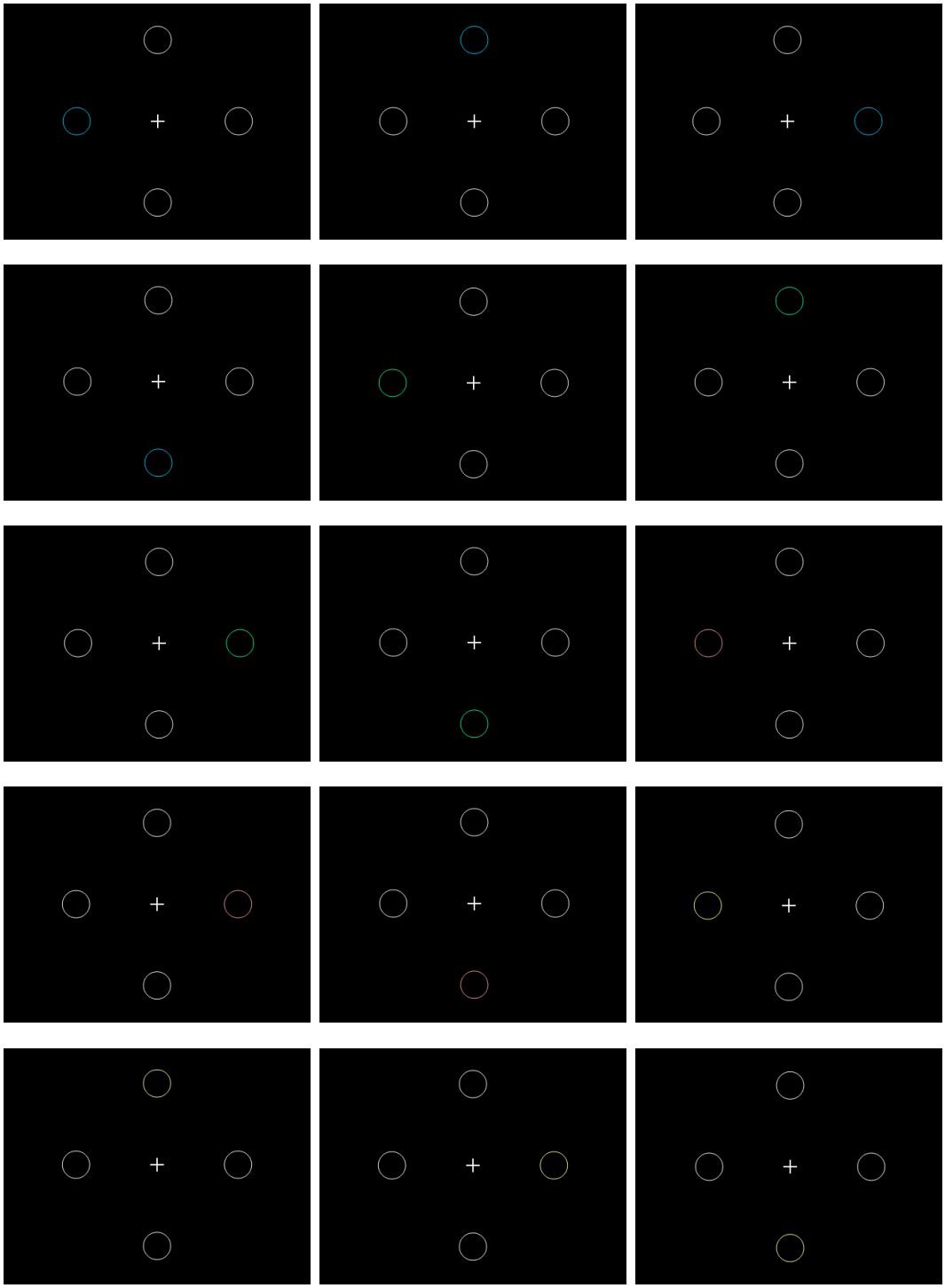


Figure B.1: Stimulus Materials: color cue displays for the liminal perception study (Chapter 2). While not an exhaustive array of stimulus displays, figures are sampled from all experimental conditions. Figures are trimmed to show the important feature differences among conditions.

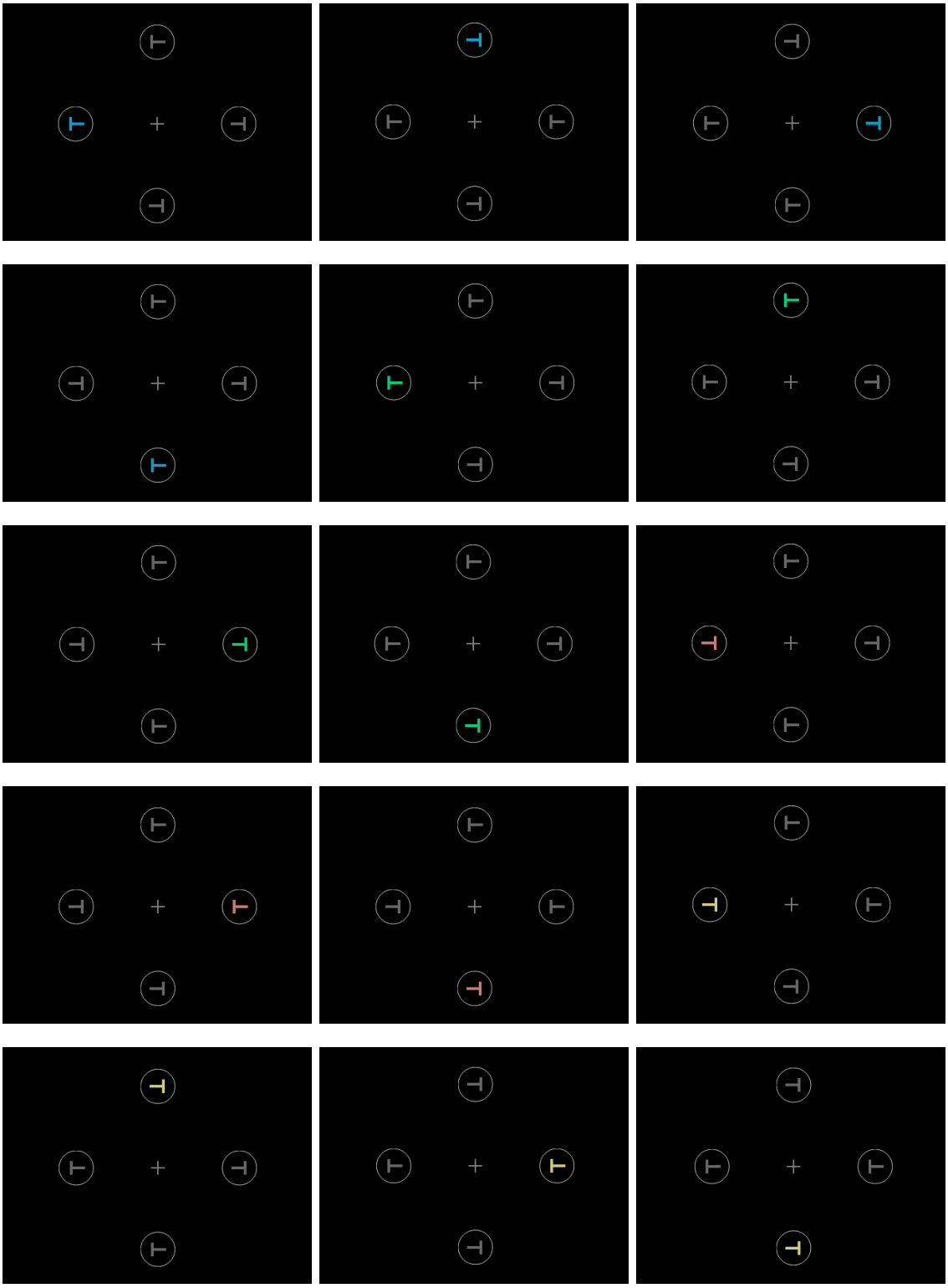


Figure B.2: Stimulus Materials: target displays for the liminal perception study (Chapter 2). Figures are sampled from all experimental conditions. Figures are trimmed to show the important feature differences among conditions.

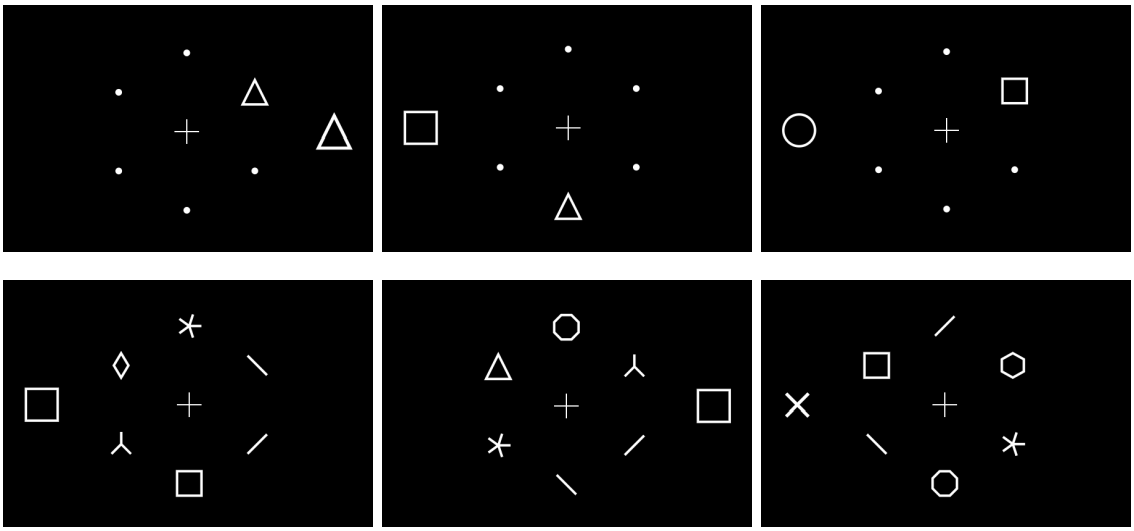


Figure B.3: Stimulus Materials: stimulus displays from the Square/Triangle block of the Flanker study (Chapter 3). Figures are sampled from each experimental condition: load [low (top), high (bottom)], flanker compatibility [compatible (left), incompatible (middle), neutral (right)]. Figures are trimmed to show the important feature differences among conditions.

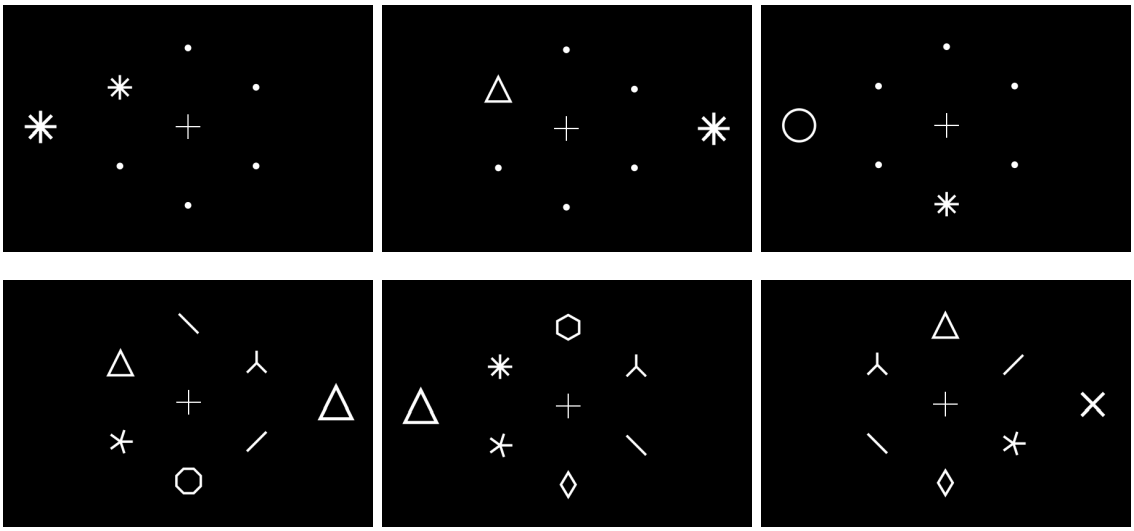


Figure B.4: Stimulus Materials: stimulus displays from the Asterisk/Triangle block of the Flanker study (Chapter 3). Figures are sampled from each experimental condition: load [low (top), high (bottom)], flanker compatibility [compatible (left), incompatible (middle), neutral (right)]. Figures are trimmed to show the important feature differences among conditions.

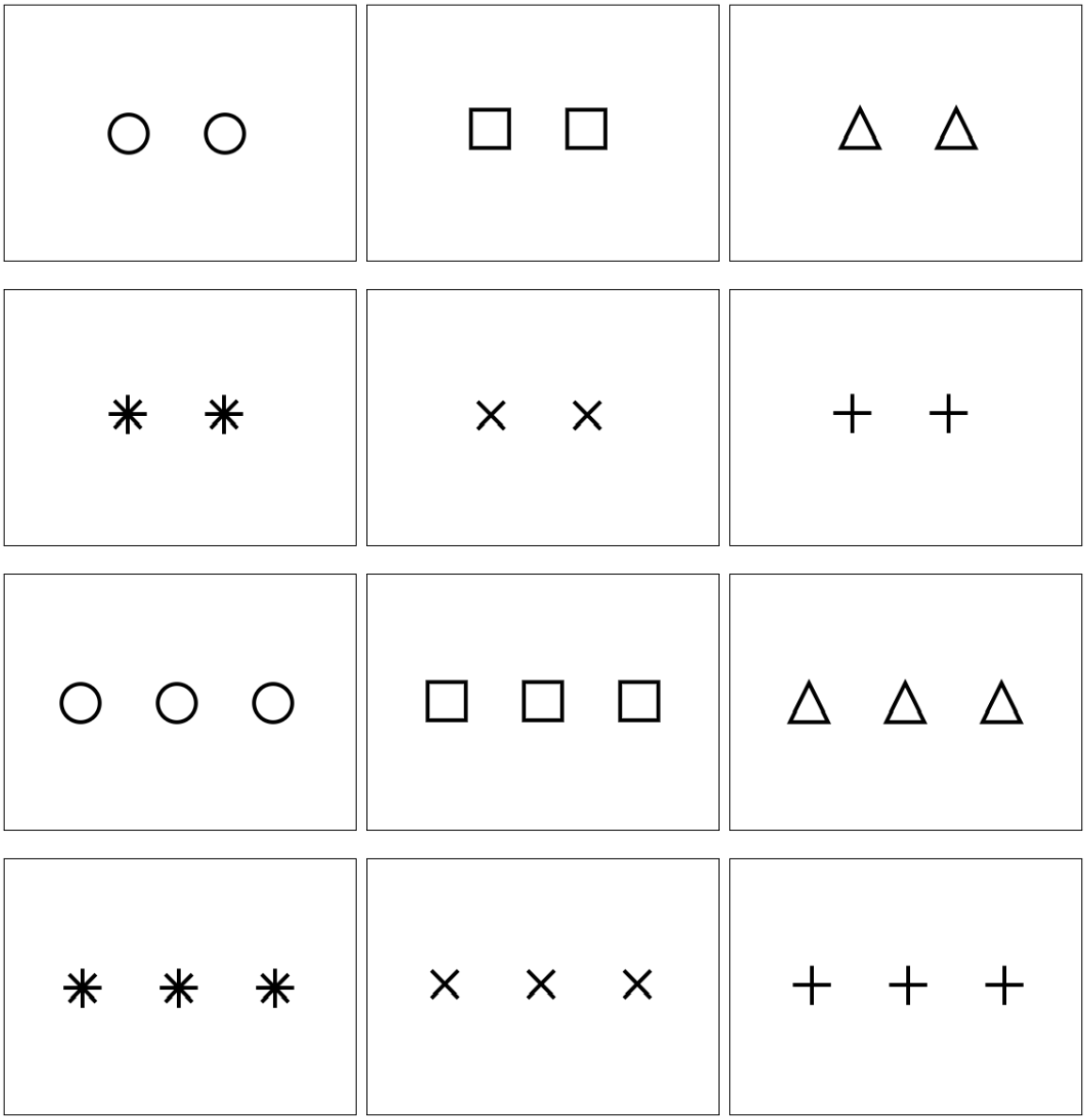


Figure B.5: Stimulus Materials: All target displays for the same-shape conditions in the same-different study (Chapter 4). Figures are trimmed to show the important feature differences.

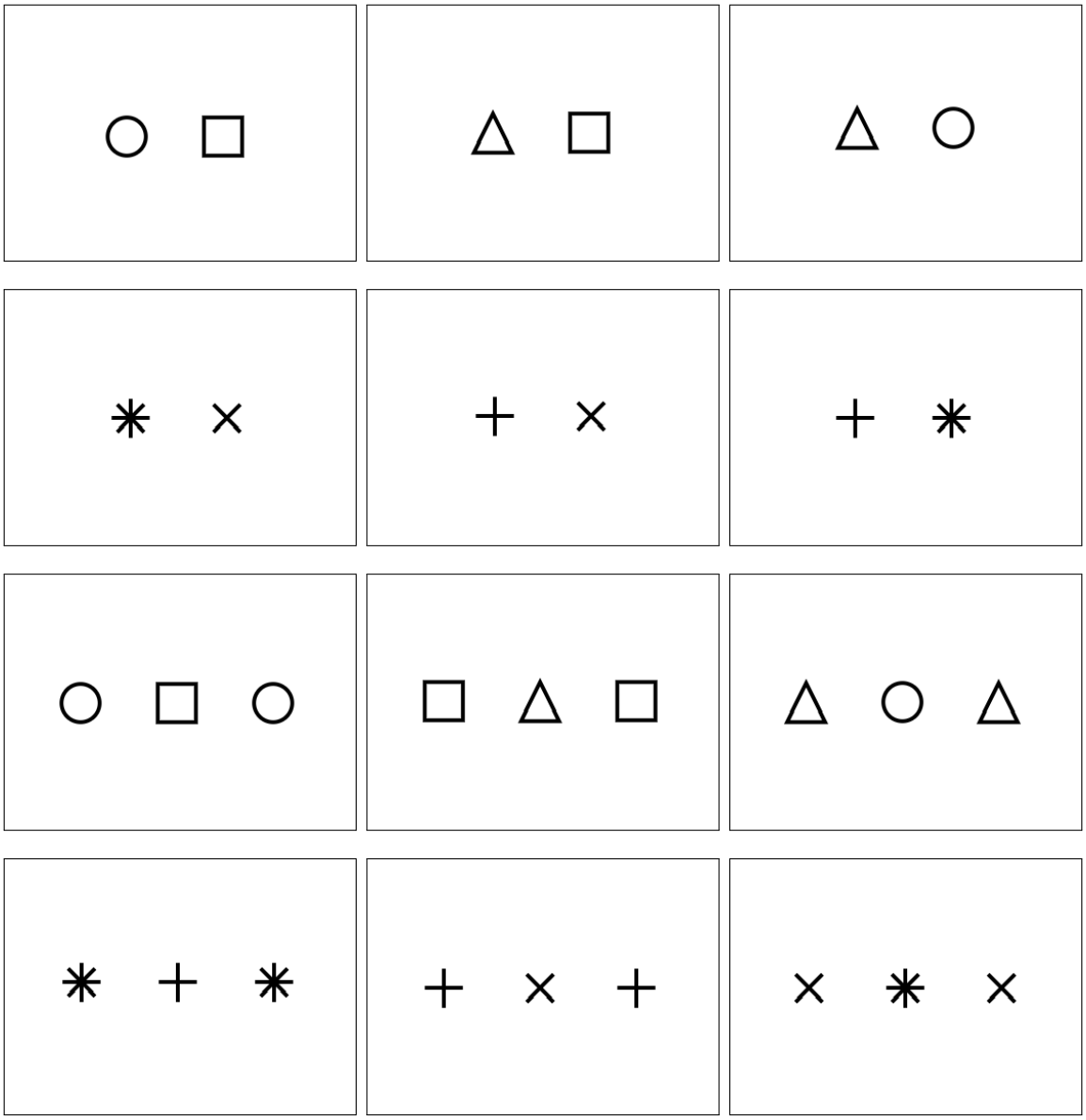


Figure B.6: Stimulus Materials: A subset of target displays for the different-shape, same-feature condition in the same-different study (Chapter 4). Figures are trimmed to show the important feature differences.

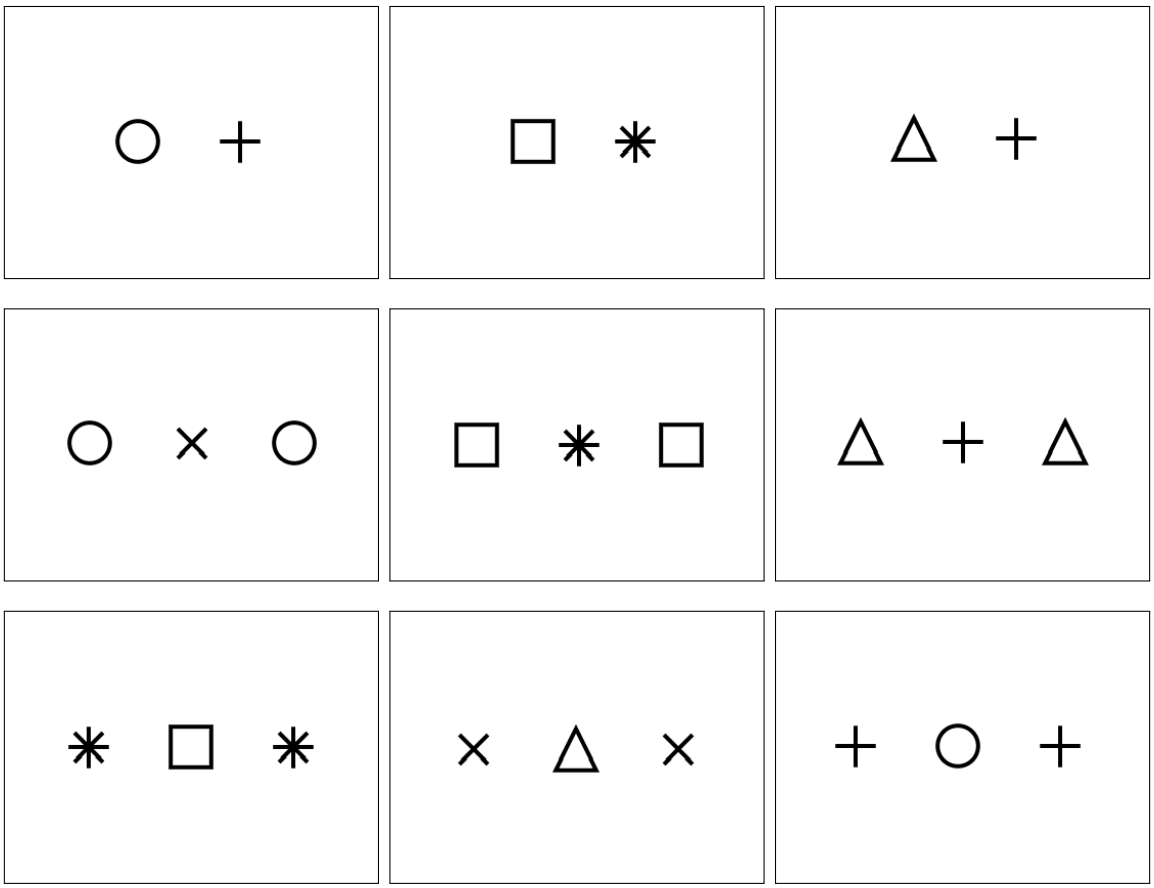


Figure B.7: Stimulus Materials: A subset of target displays for the different-shape, different-feature condition in the same-different study (Chapter 4). Figures are trimmed to show the important feature differences.

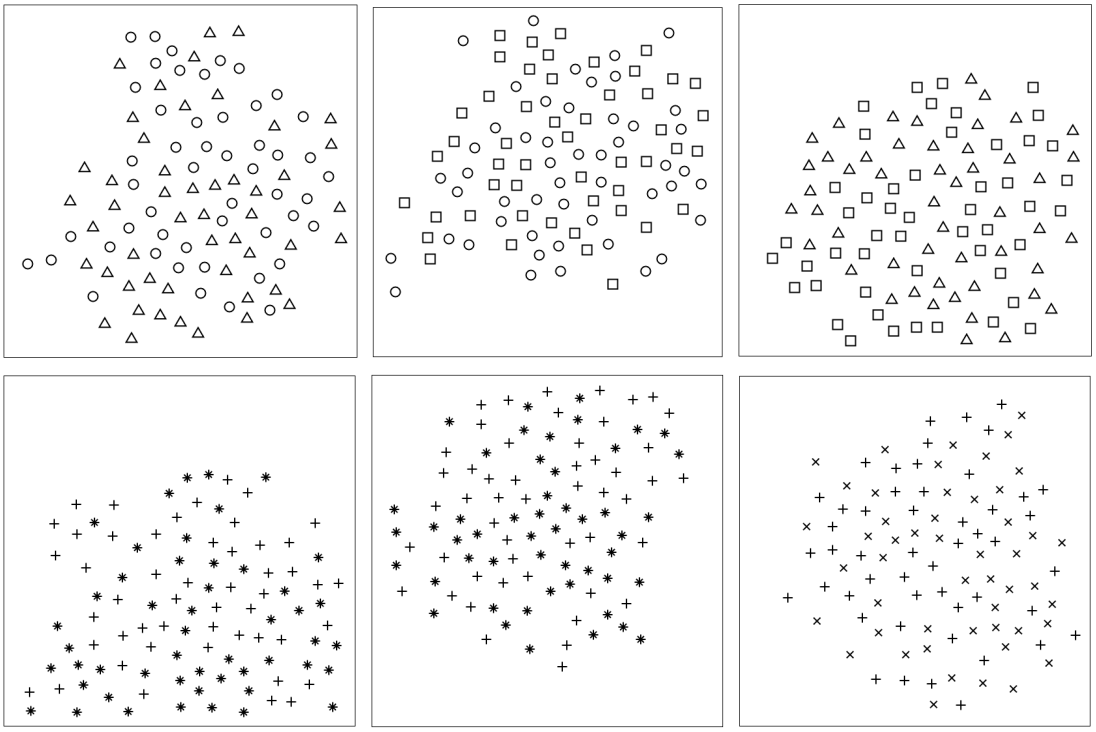


Figure B.8: Stimulus Materials: A subset of single-plot displays for same-feature shapes in the average value judgment task from the scatterplot study (Chapter 5). Columns left to right show easy, medium, and hard conditions.

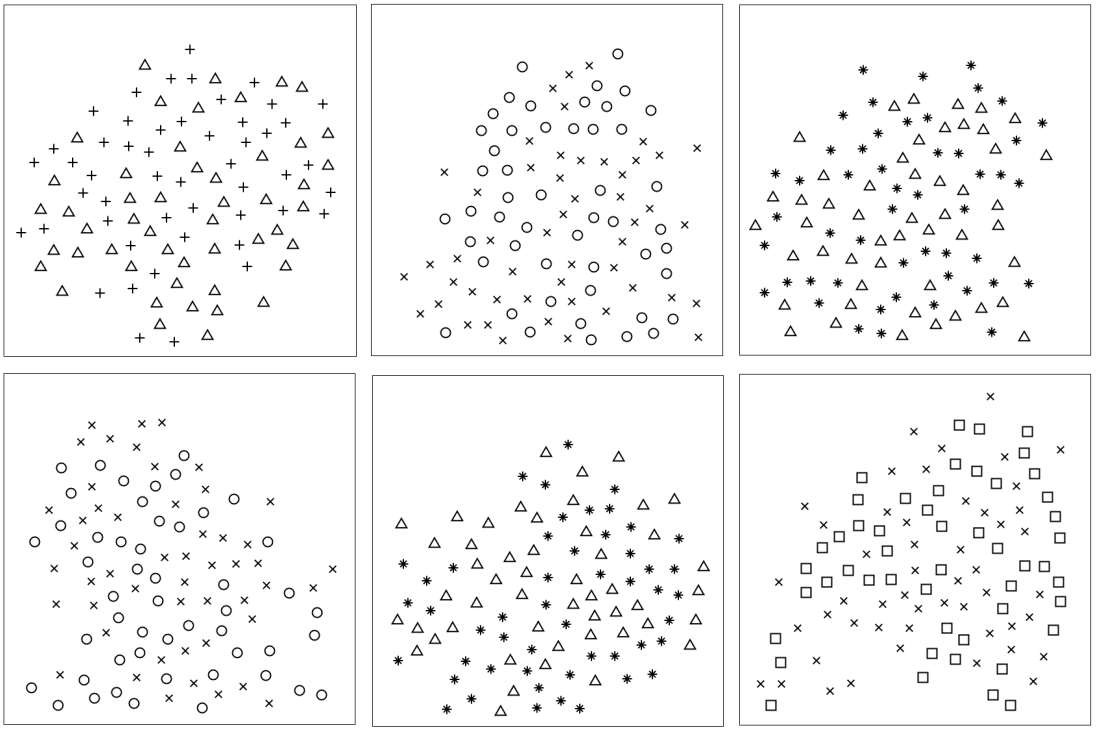


Figure B.9: Stimulus Materials: A subset of single-plot displays for different-feature shapes in the average value judgment task from the scatterplot study (Chapter 5). Columns left to right show easy, medium, hard conditions.

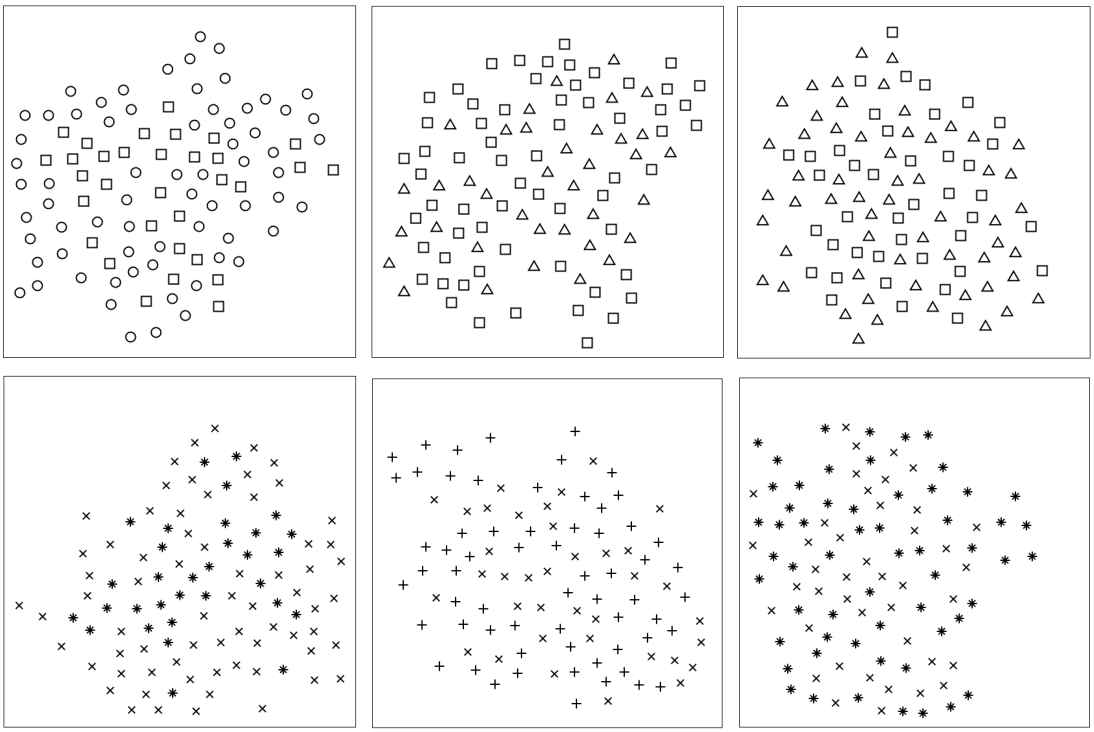


Figure B.10: Stimulus Materials: A subset of single-plot displays for same-feature shapes in the numerosity judgment task from the scatterplot study (Chapter 5). Columns left to right show easy, medium, and hard conditions.

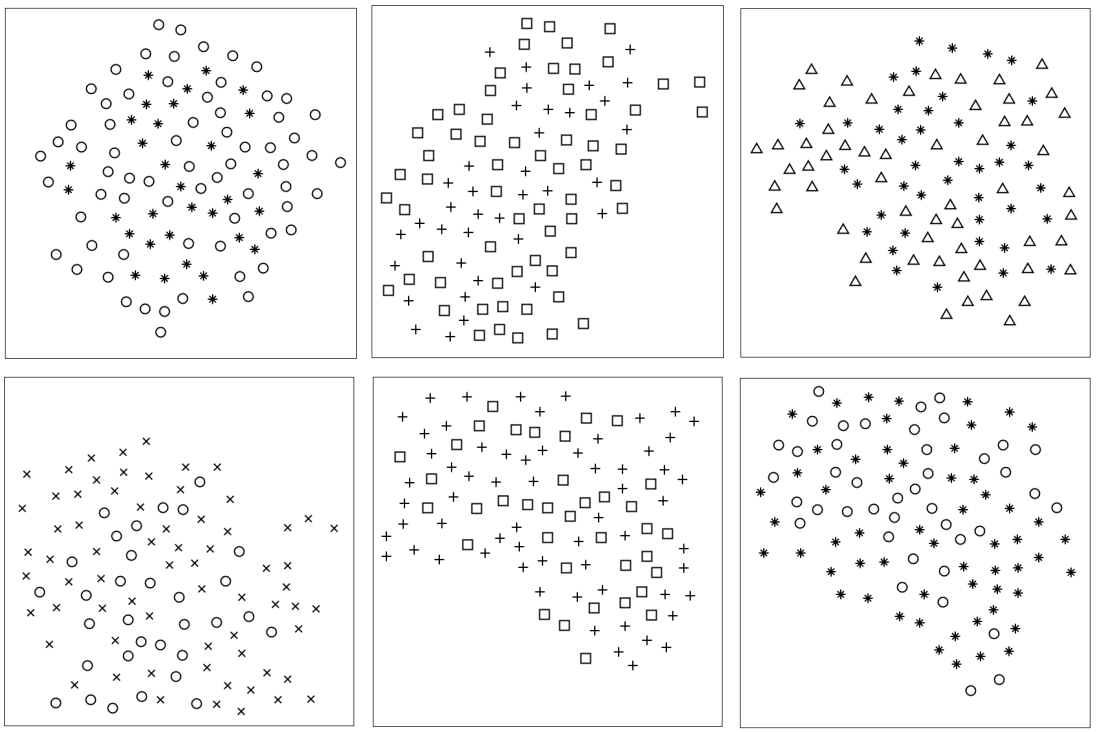


Figure B.11: Stimulus Materials: A subset of single-plot displays for different-feature shapes in the numerosity judgment task from the scatterplot study (Chapter 5). Columns left to right show easy, medium, hard conditions.

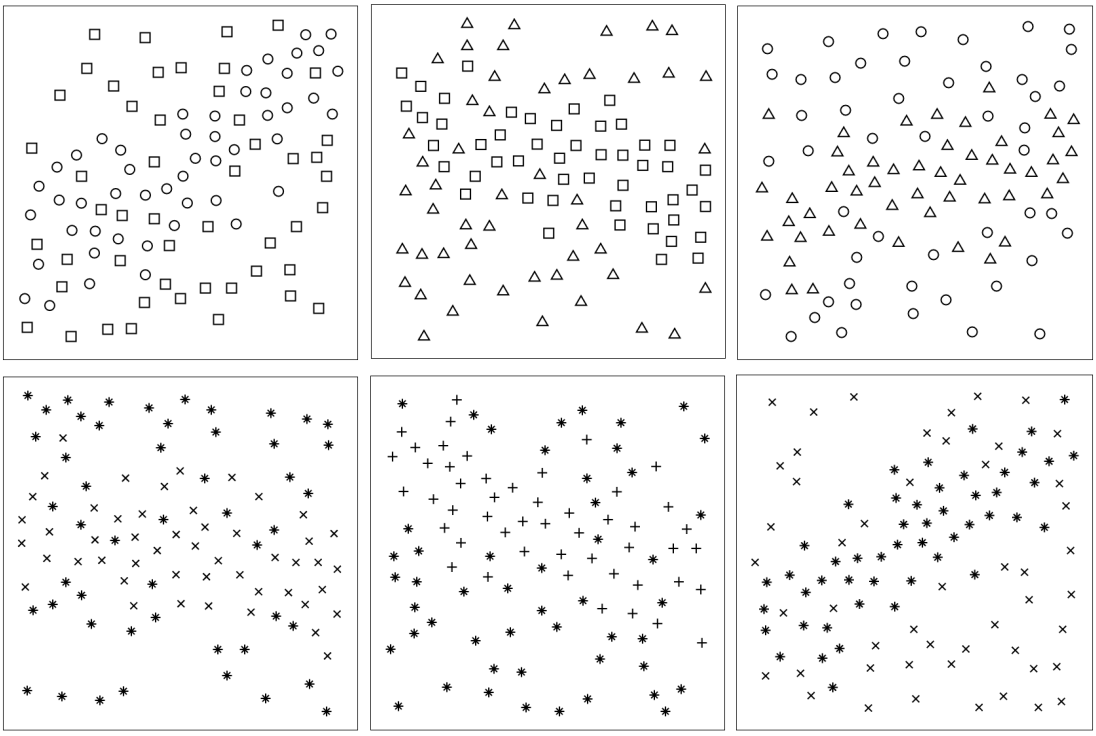


Figure B.12: Stimulus Materials: A subset of single-plot displays for same-feature shapes in the linear trend judgment task from the scatterplot study (Chapter 5). Columns left to right show easy, medium, and hard conditions.

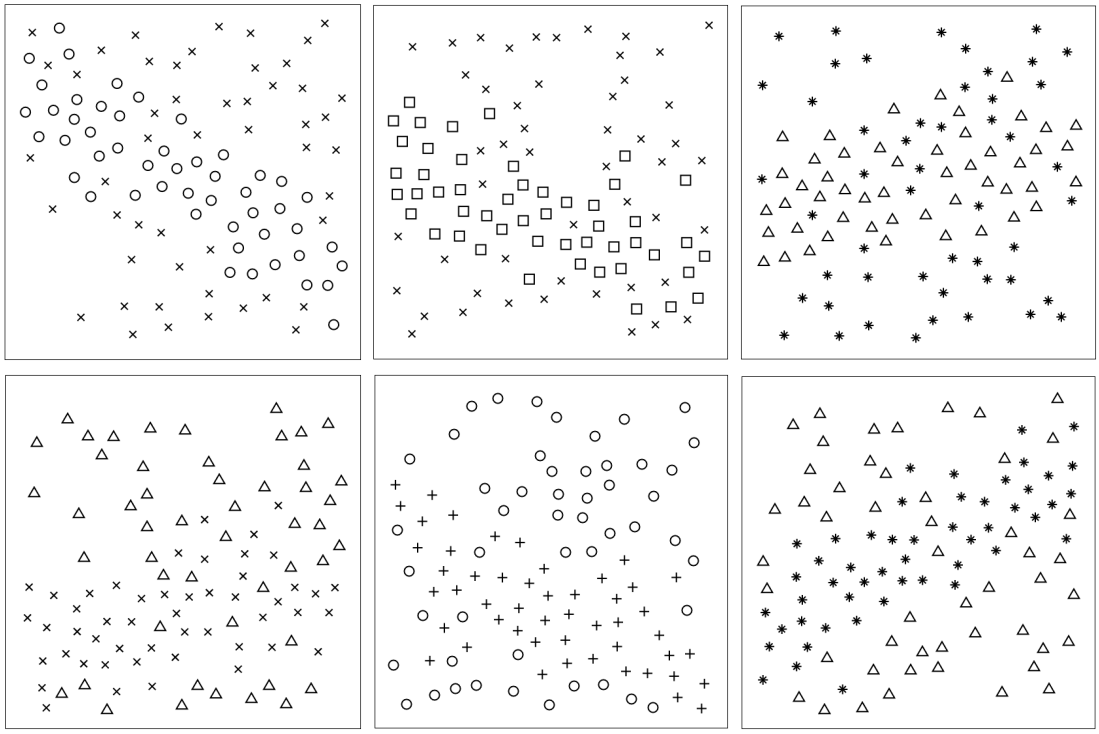


Figure B.13: Stimulus Materials: A subset of single-plot displays for different-feature shapes in the linear trend judgment task from the scatterplot study (Chapter 5). Columns left to right show easy, medium, hard conditions.

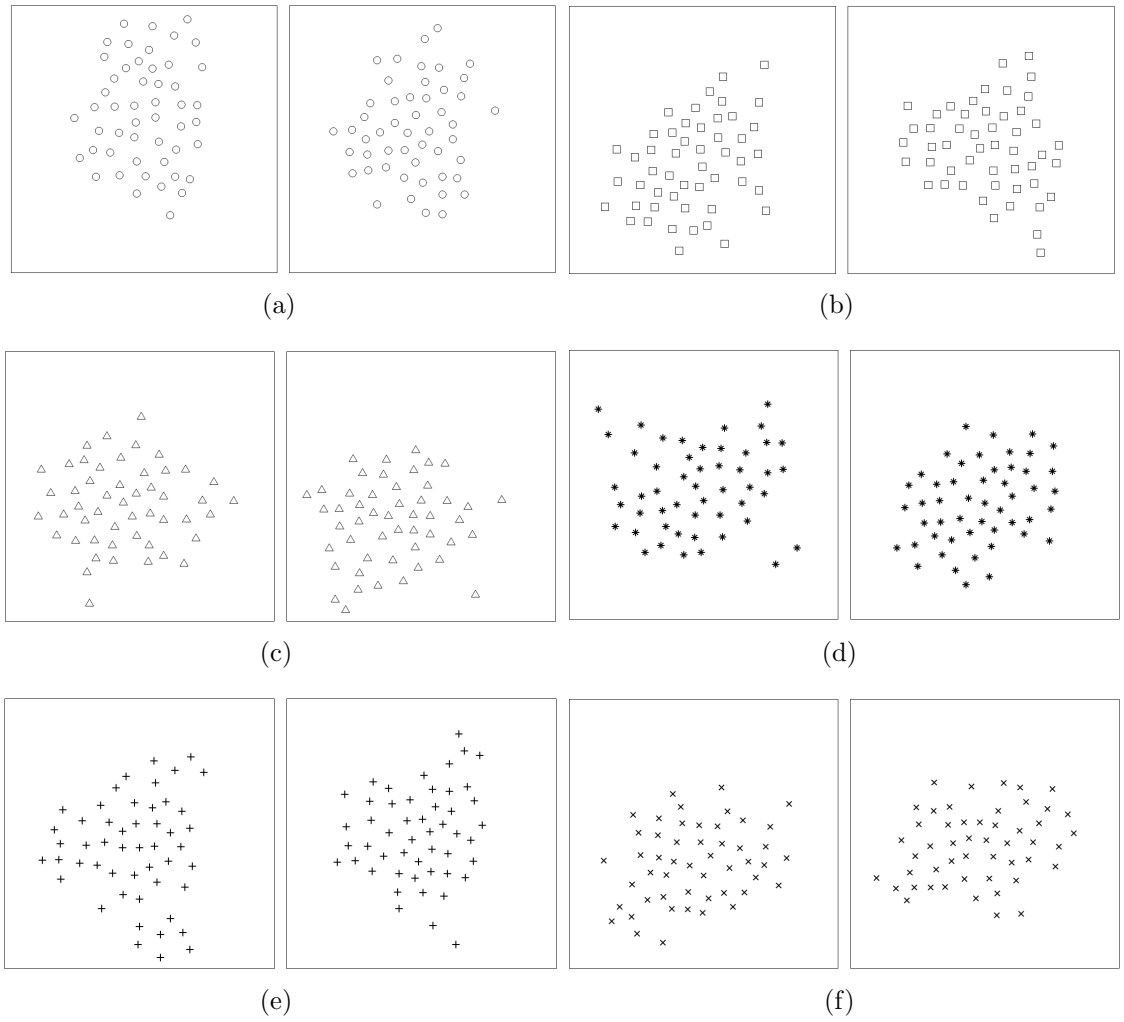


Figure B.14: Stimulus Materials: A subset of medium-difficulty separate-plot displays from the average value task from the scatterplot study (Chapter 5).

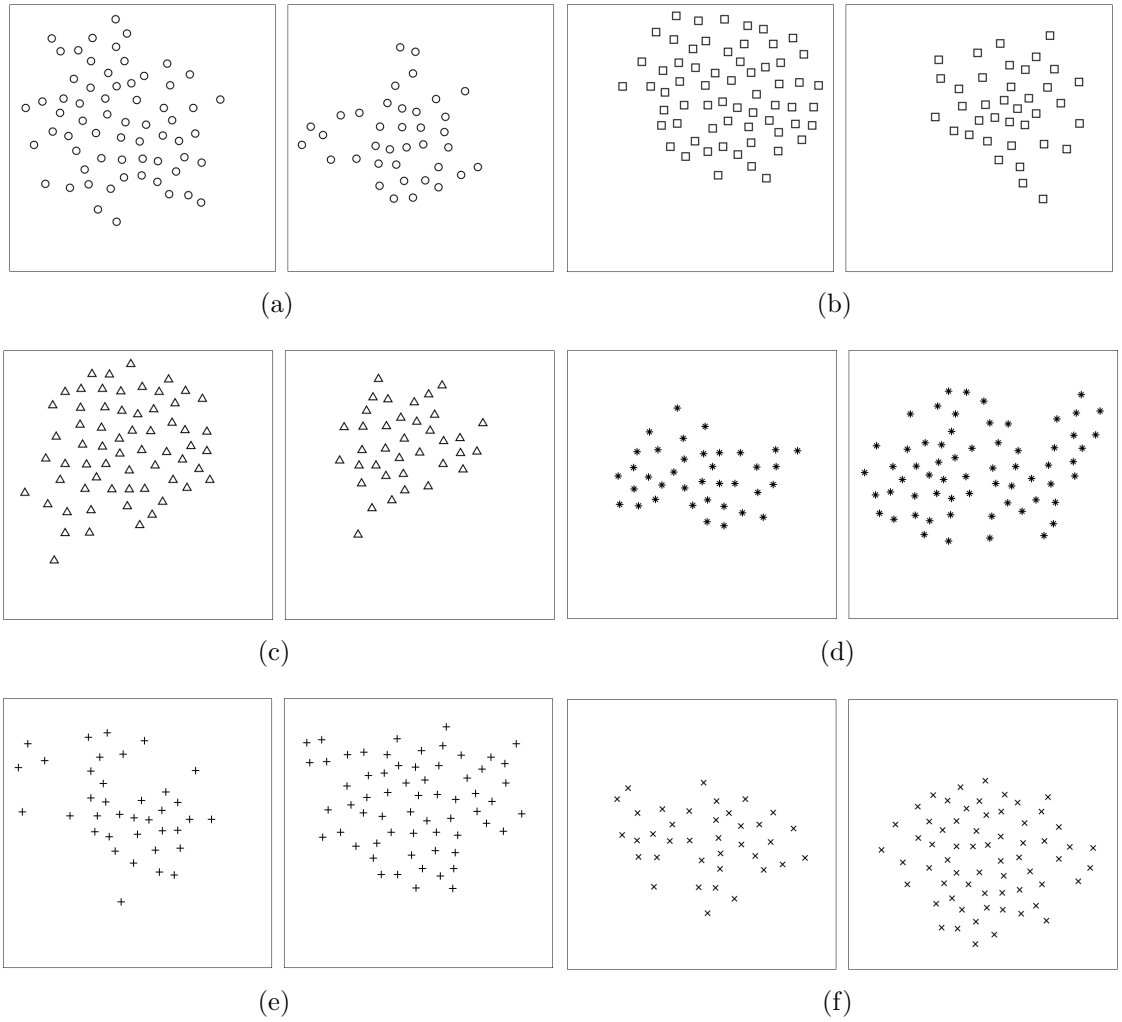


Figure B.15: Stimulus Materials: A subset of medium-difficulty separate-plot displays from the numerosity task from the scatterplot study (Chapter 5).

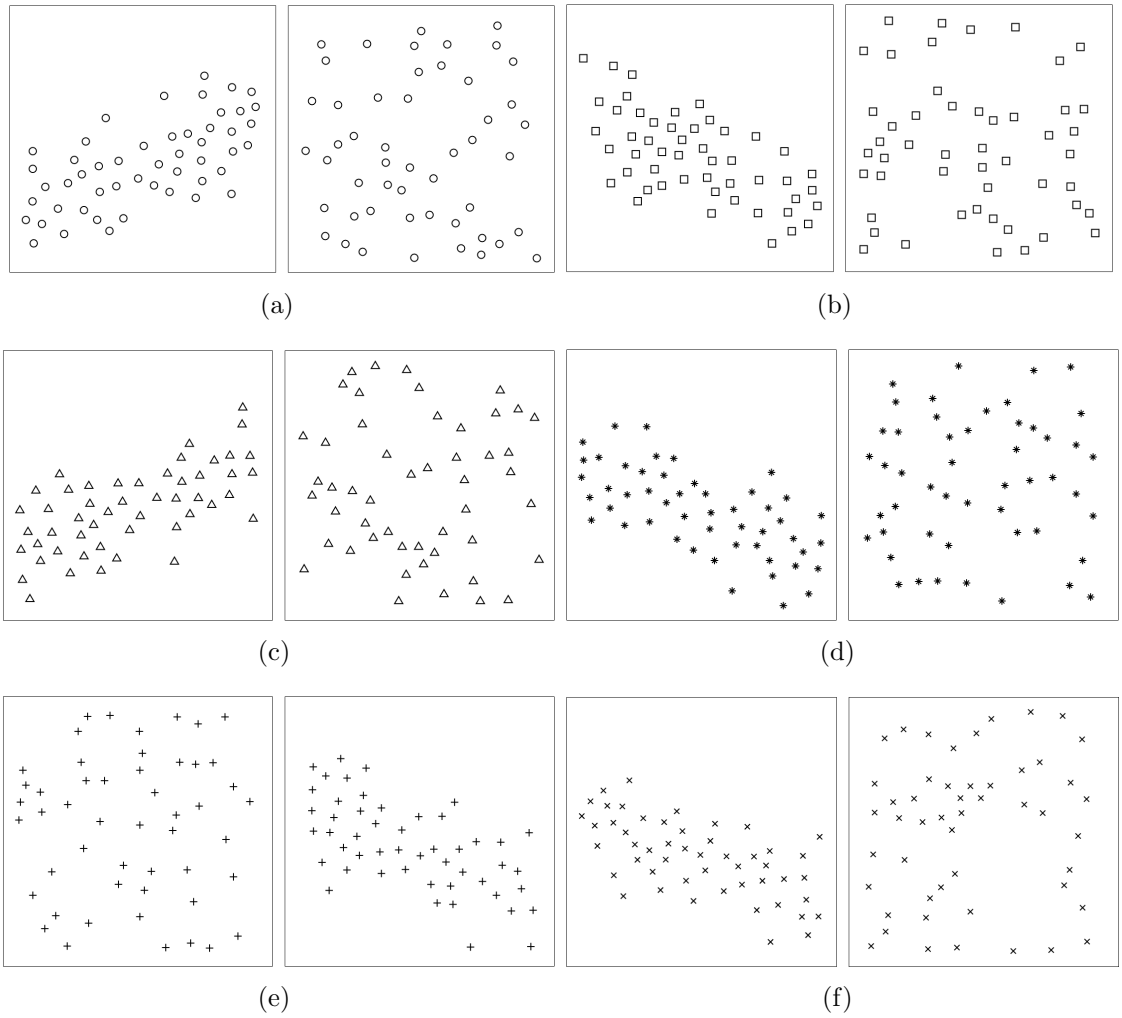


Figure B.16: Stimulus Materials: A subset of medium-difficulty separate-plot displays from the linear trend judgment task from the scatterplot study (Chapter 5).

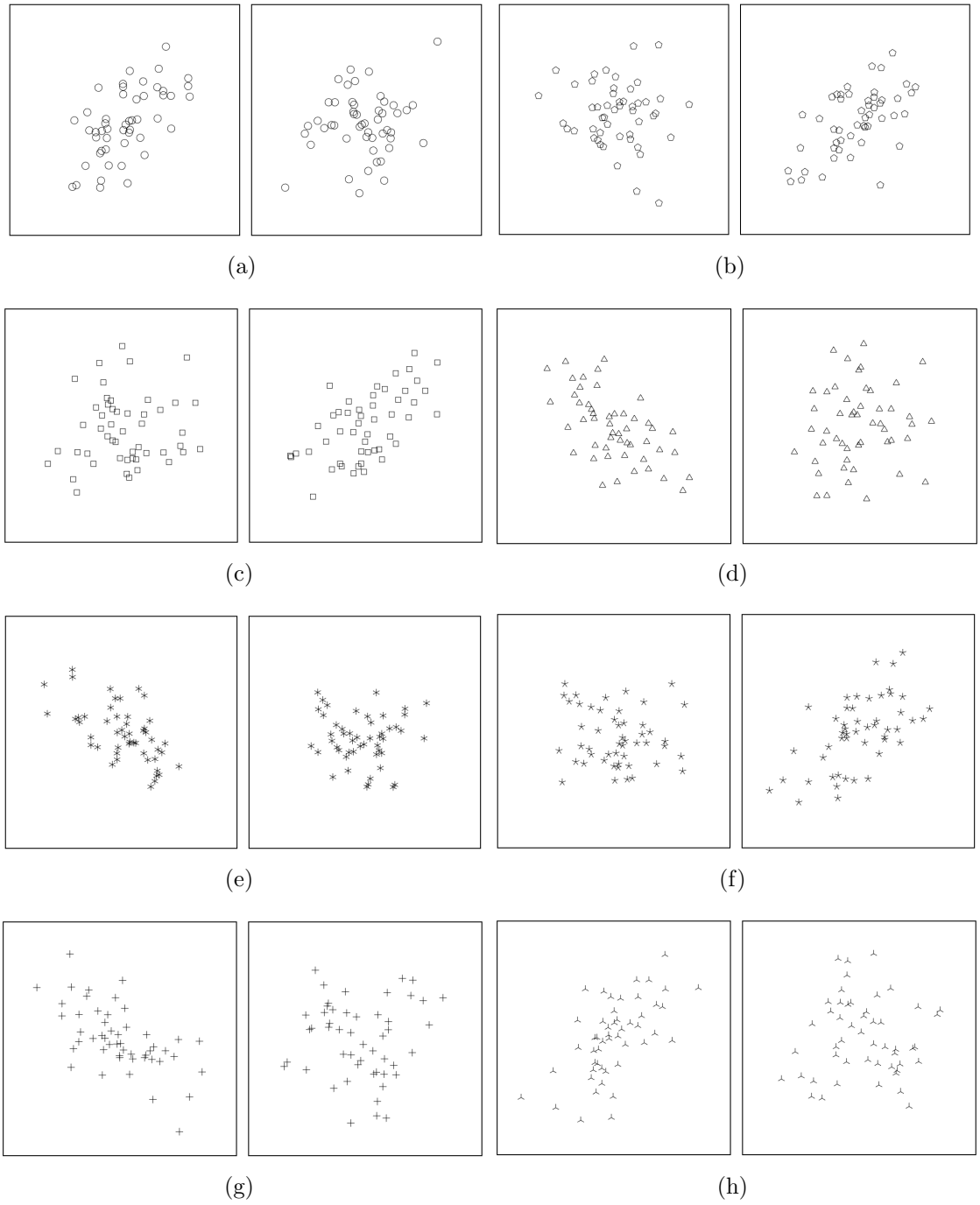


Figure B.17: Stimulus Materials: A subset of separate-plot displays from the linear trend judgment task from the overplotting study (Chapter 6). Stimulus displays contained pairs of adjacent charts; a-d contain closed symbols, and e-h contain open symbols.

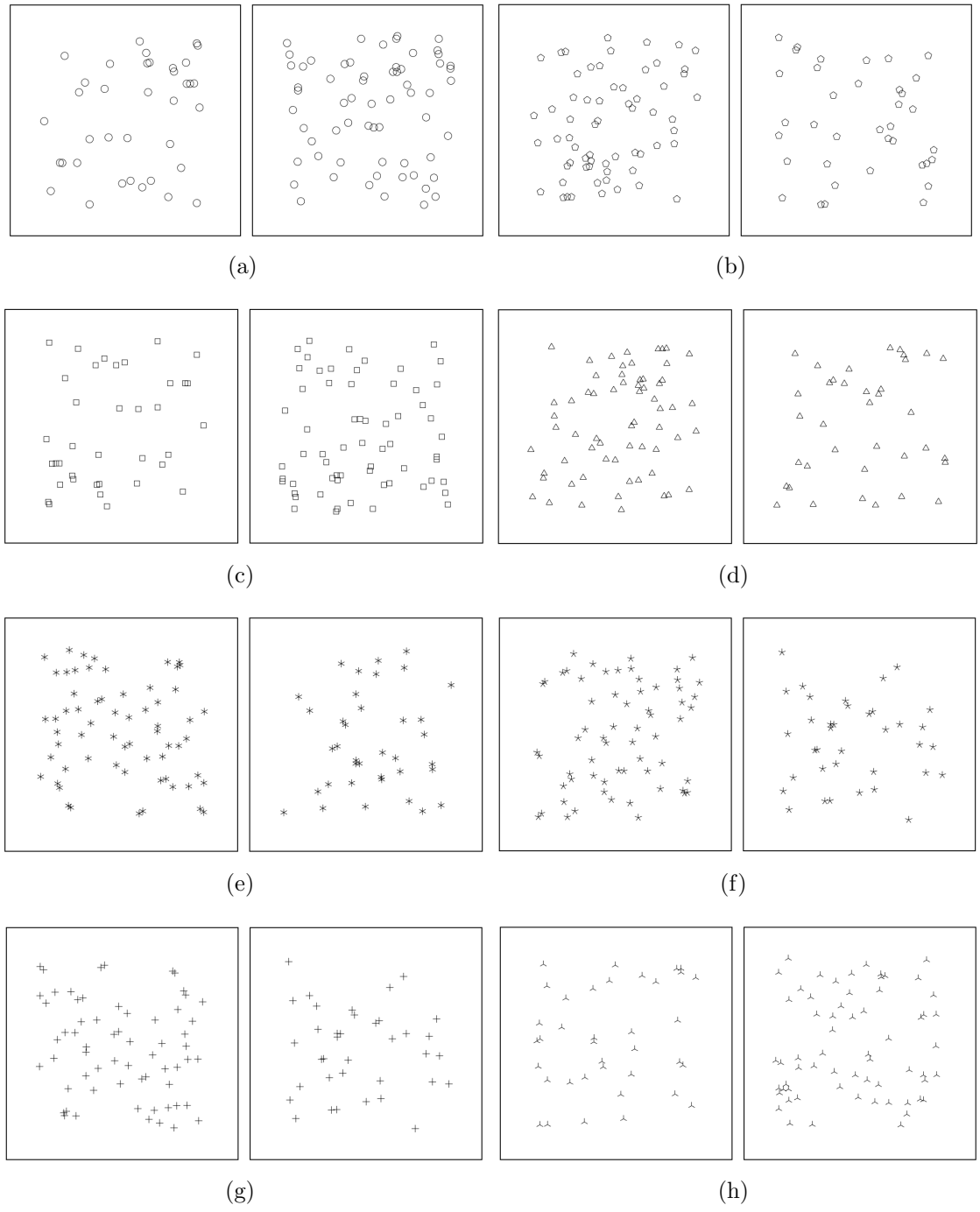


Figure B.18: Stimulus Materials: A subset of separate-plot displays from the numerosity judgment task from the overplotting study (Chapter 6). Stimulus displays contained pairs of adjacent charts; a-d contain closed symbols, and e-h contain open symbols.

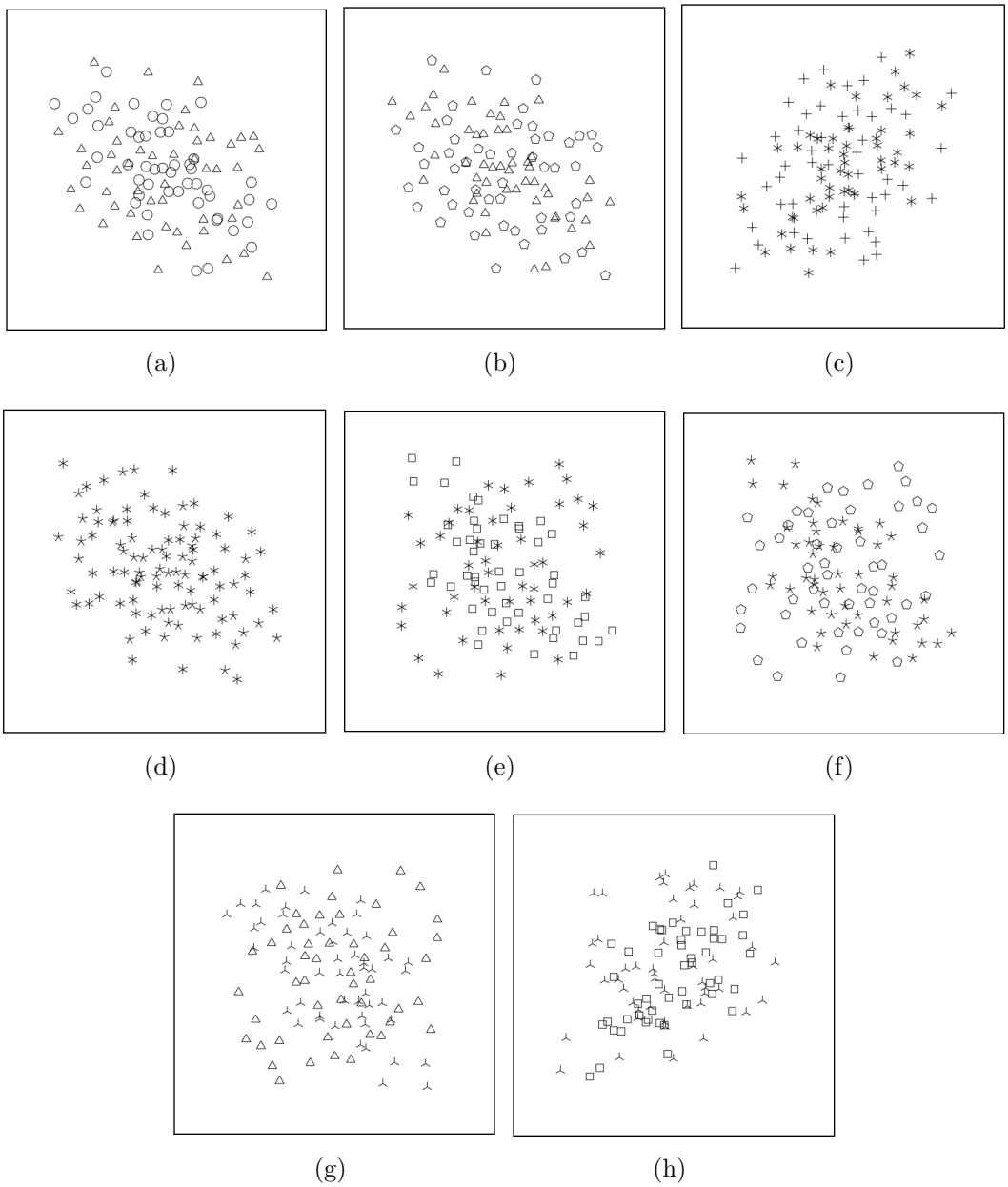


Figure B.19: Stimulus Materials: A subset of single-plot displays from the linear trend judgment task from the overplotting study (Chapter 6). a-d are same-feature pairs, e-h are different-feature pairs.

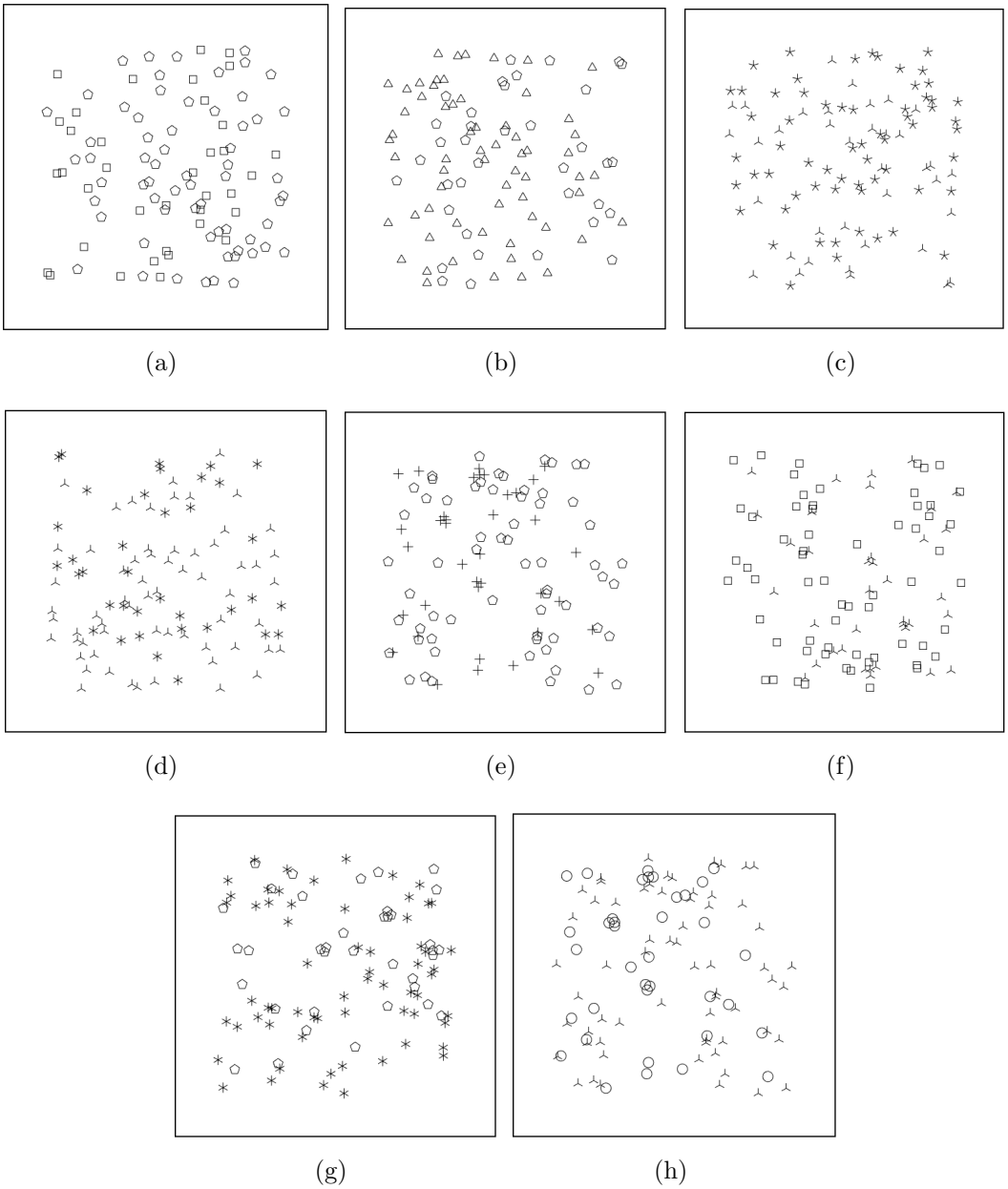


Figure B.20: Stimulus Materials: A subset of single-plot displays from the numerosity judgment task from the overplotting study (Chapter 6). a-d are same-feature pairs, e-h are different-feature pairs.

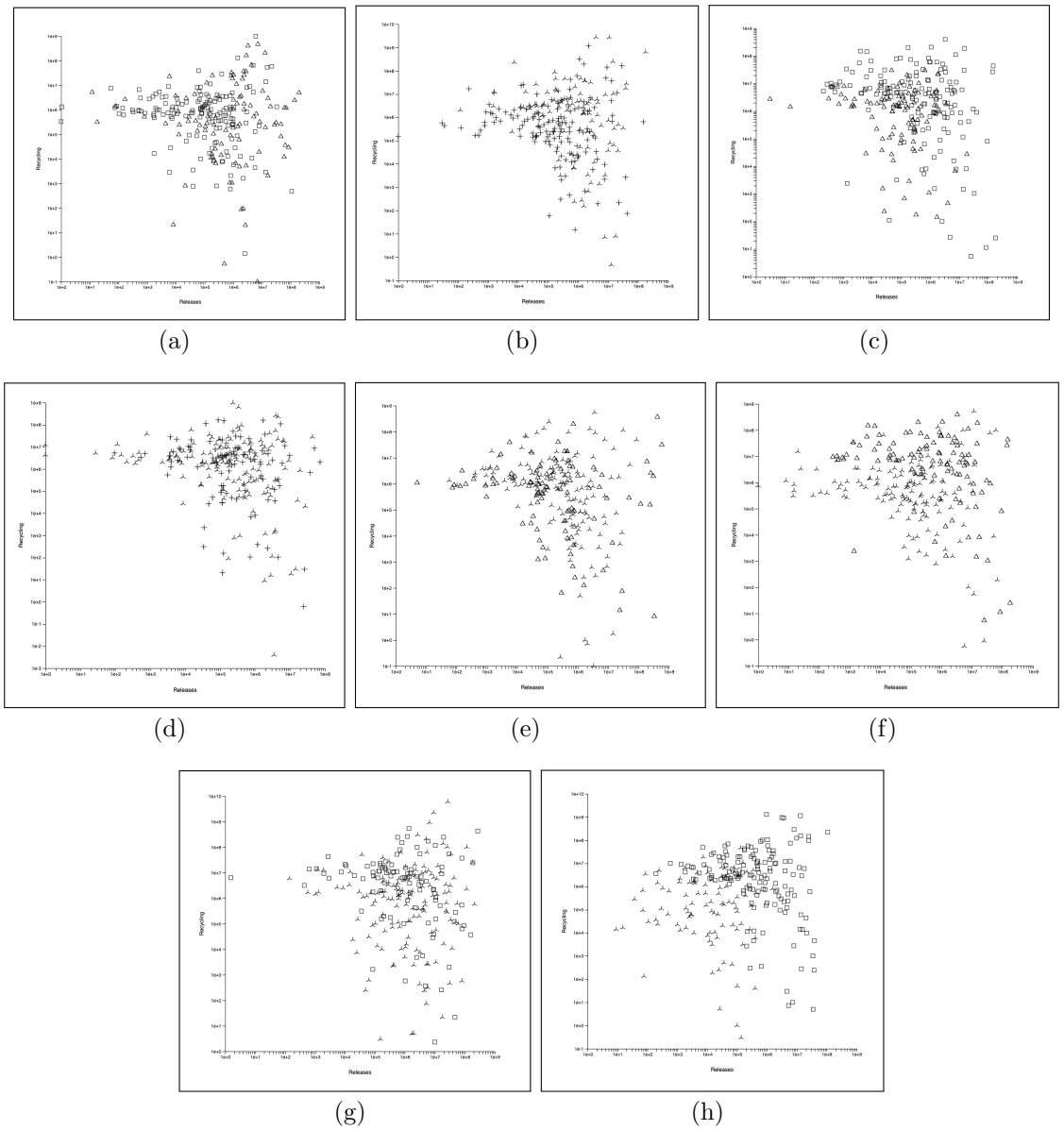


Figure B.21: Stimulus Materials: A subset of single-plot displays from the numerosity judgment task from the real-world dataset study (Chapter 7). a-d are same-feature pairs, e-h are different-feature pairs.