A Design Space of Vision Science Methods for Visualization Research

Madison A. Elliott, Christine Nothelfer, Cindy Xiong, and Danielle Albers Szafir

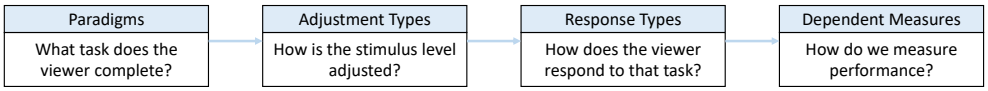

Fig. 1. Overview of design space of experimental methods. We present a four component design space to guide researchers in creating visualization studies grounded in vision science research methods.

Abstract—A growing number of efforts aim to understand what people see when using a visualization. These efforts provide scientific grounding to complement design intuitions, leading to more effective visualization practice. However, published visualization research currently reflects a limited set of available methods for understanding how people process visualized data. Alternative methods from vision science offer a rich suite of tools for understanding visualizations, but no curated collection of these methods exists in either perception or visualization research. We introduce a design space of experimental methods for empirically investigating the perceptual processes involved with viewing data visualizations to ultimately inform visualization design guidelines. This paper provides a shared lexicon for facilitating experimental visualization research. We discuss popular experimental paradigms, adjustment types, response types, and dependent measures used in vision science research, rooting each in visualization examples. We then discuss the advantages and limitations of each technique. Researchers can use this design space to create innovative studies and progress scientific understanding of design choices and evaluations in visualization. We highlight a history of collaborative success between visualization and vision science research and advocate for a deeper relationship between the two fields that can elaborate on and extend the methodological design space for understanding visualization and vision.

Index Terms—Perception, human vision, empirical research, evaluation, HCI

1 Introduction

Visualization researchers are increasingly interested in running hypothesis-driven empirical studies to investigate human visual perception and intelligence for data displays [36, 38, 60, 70]. Efforts such as BeLiV, VISxVISION [48], ETVIS [10], and others continue to promote interdisciplinary science between visualization and experimental psychology. Despite well-established synergy between these fields, there is no shared guide or lexicon for facilitating and designing perceptual visualization experiments. Methodological guides in experimental psychology are either broadly focused [77] or technique-specific [25], limiting their utility for visualization studies. Replete with field-specific jargon, these resources require extensive knowledge from psychology to interpret effectively. The lack of accessible knowledge about experimental methods for understanding visualizations creates barriers for researchers first seeking to engage in experimental research and limits engagement with a broader suite of tools to understand perception and cognition in visualization. There is a desire for rigorous experimentation, yet no common ground or established "basic knowledge" to facilitate or evaluate it.

Our paper provides a set of tools to afford novel, collaborative research between visualization and vision science researchers by establishing a preliminary methodological design space for visualization experiments. Vision science is the study of how vision works and is used. It is most closely associated with psychology, but draws on a range of disciplines, including cognition and neuroscience. While visualization and vision science research have already seen tangible benefits from collaboration in both design and methodology (e.g., [8, 11, 27–29, 62, 69]), increasing both the quality and variety of methods used to understand and evaluate visualization design would more generally benefit visualization research. Adapting methods from vision science can actionably expand our knowledge about user attention, memory, visualization design efficacy, and the nature of visual intelligence used to view data. As a first step towards this goal, we catalogue some of the most useful research methods from vision science and discuss their potential for evaluating user behavior and design for visualizations.

The two main contributions of this paper are: 1) to provide a *design space* of experimental vision science methods that visualization and perception researchers can use to craft rigorous experiments and 2) to provide a shared *lexicon* of experimental techniques to stimulate and engage collaboration between vision and visualization communities. We highlight the potential of using perceptual methods for understanding visualization, discuss prior successes in interdisciplinary research, and identify common methods and opportunities for innovation between the two fields. Through this effort, we aspire to motivate further collaboration and intellectual reciprocity between researchers in both areas of study.

- *Madison A. Elliott is with The University of British Columbia. E-mail: mellio10@psych.ubc.ca.*
- *Christine Nothelfer is with Northwestern University. Email: cnothelfer@gmail.com.*
- *Cindy Xiong is with the University of Massachusetts Amherst E-mail: yaxiong@umass.edu.*
- *Danielle Albers Szafir is with the University of Colorado Boulder. E-mail: danielle.szafir@colorado.edu.*

Manuscript received 30 Apr. 2020; revised 31 July 2020; accepted 14 Aug. 2020. Date of publication 22 Oct. 2020; date of current version 15 Jan. 2021. Digital Object Identifier no. 10.1109/TVCG.2020.3029413

2 Background

In this section, we discuss broad advantages and limitations of using vision science research methods for visualization studies. We highlight past interdisciplinary work and several contributions that each field has made to the other.

2.1 Trade-Offs of Vision Science Research Methods

Visualizations offload cognitive work to the visual system to help people make sense of data, imparting visual structure to data to, for example, surface key patterns like trends or correlations or identify emergent structures. The amount of empirical research in visualization is increasing [38], and evaluations of user performance are among the top three

types of evaluations in published articles. The growing popularity of quantitative experiments for visualization evaluation correlates with a movement to develop empirical foundations for long-held design intuitions. However, meta-analyses of research methods used in visualization studies have called into question the reliability of current studies [17] and even the quality of assumptions from seminal visualization work [36]. Studies have demonstrated that rigorous experimental methods grounded in perceptual theory can expose the limitations of well-accepted conclusions [29, 62, 69].

The important takeaway from these examples is not that visualization evaluations sometimes get things wrong, but instead that science and our subsequent understanding of the world is constantly evolving. The methods used to understand these phenomena dictate the efficacy and reliability of that understanding. Researchers borrowing methods from other disciplines should do so with an awareness of the origin and purpose of those methods in the greater context of what they are used to study. Methods and experimental designs evolve with our knowledge of the world, which is in turn shaped by the progression of theories and evidence. The design of user evaluation and performance studies should not be formulaic and rigid, but rather a creative and thoughtful consideration of the current state of related vision science research that guides careful selection of experimental methods [2]. Vision science methods reflect over a century of work in exploring how people transform real-world scenes and objects into information. This maturity has allowed researchers the time to fail, iterate, improve, converge, and replicate key findings to support well-established theories.

One concern with adapting vision science to visualization is that vision science focuses on basic research, emphasizing understanding the visual system rather than the functions of different designs. However, design and mechanism are not mutually exclusive: understanding how we see data can drive innovative ideas for how to best communicate different properties of data. For example, understanding the features involved in reading quantities from pie charts can drive novel representations for proportion data [37]. Researchers must carefully consider how the experimental designs can capture different intricacies of visualizations, offering opportunities for methodological innovation balancing control and ecological validity through approaches like applied basic research [65].

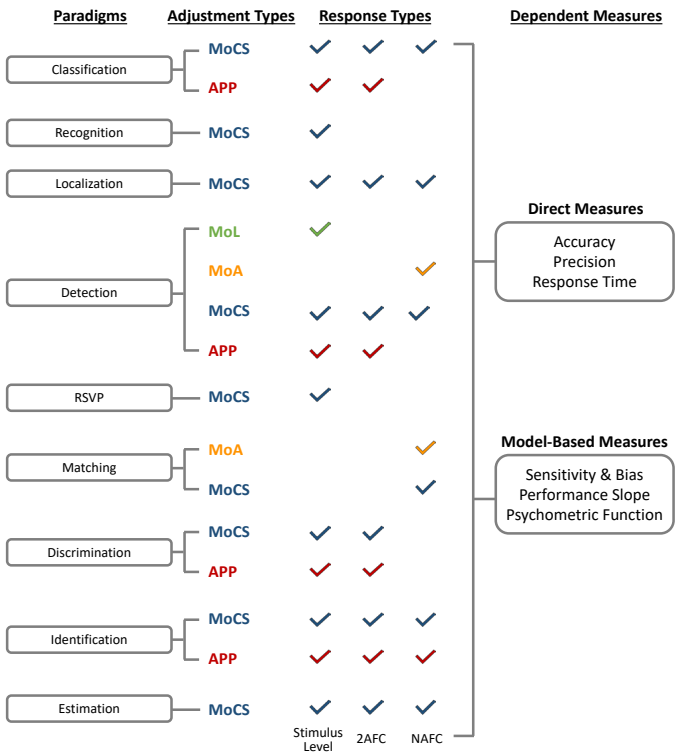

Fig. 2. Summary structure of the design space. Starting from left to right, researchers can select a paradigm and match it with the appropriate adjustment and response types to obtain the desired dependent measures. Not all connections between adjustment types and response types are meaningful. Check marks indicate common combinations of adjustment and response types. Adjustments are abbreviated as follows: Method of Adjustment (MoA), Method of Limits (MoL), Method of Constant Stimuli (MoCS), and Adaptive Psychophysical Procedures (APP).

2.2 Past Collaborations and Recent Success

Visual perception is widely considered a key element of data visualization. Two common vision science metrics—*accuracy* and *response time*—have been widely used in visualization evaluation studies. While past methodological adaptation offers insight into effective visualization design, several studies fall short of their goals due to methodological flaws (see Kosara [36] and Crisan & Elliott [17] for reviews). For example, early work in graphical perception [13, 39] suffered from mistakes in precision of design and lack of connection to perceptual mechanisms and often can not be replicated [62], making them less useful for creating design guidelines.

Recent interdisciplinary studies between vision science and visualization continue to expand our understanding of how and when visualizations work. For example, color perception and encoding design is largely informed by experimental methods. These studies offer insights into the most effective choices for encoding design [8, 27] and application [64]. While a full survey of topics in visualization experiments is beyond the scope of this paper, studies have investigated basic visual features, like orientation [71], contrast [47], grouping [28], motion [75], and redundant use of color and shape [50]. For example, studying correlation perception in scatterplots using methods from vision science has led to a new understanding of visualization concepts and design [19, 29, 62]. Rensink & Baldridge [62] used psychophysics methods (e.g., see §3.2) to derive just noticeable differences (JNDs) for correlation magnitudes in scatterplots. These JNDs followed Weber's law, meaning that sensitivity to correlation varies predictably and correlation perception is likely a systematic, early (i.e., low-level) visual process and an instance of ensemble coding [61, 76]. This work advances vision science's understanding of ensemble processing, adding correlation to the types of heuristic information that can be processed rapidly and accurately. Later work replicated and extended these methods to understand how viewers perceive correlation in complex displays to inform scatterplot designs [29, 60]. More recent work has leveraged psychophysics models to quantitatively guide visualization design in applications like color encoding design [69] and uncertainty visualization [34].

These studies show the utility of vision methods for visualization and critically demonstrate that visualizations offer useful opportunities for scientific study. The design of visualizations has evolved to work effectively with the human visual system, meaning that they implicitly hold valuable information about how we see and think [60]. Their potential for study is still largely untapped. Excitingly, visualizations will continue to provide novel insight and structural phenomena as we continue to address the need to display larger and more complex types of information effectively.

3 A Design Space of Experimental Methods

Visualization is an inherently interdisciplinary field. Many of its evaluative practices are derived from those used in human computer interaction, psychology, or sociology [17] and leverage qualitative, quantitative, or mixed approaches. Because of this diversity, there has been little consensus or standardization of evaluative practices, and no explicit "handbook" of visualization evaluation procedures exists. Here, we take an important step by proposing a focused design space of methods from vision science, a fundamentally empirical field, in hopes of inspiring new perspectives on visualization design and research.

3.1 Scope

The objective of this paper is to explore quantitative behavioral studies of user performance grounded in methodologically-relevant practices

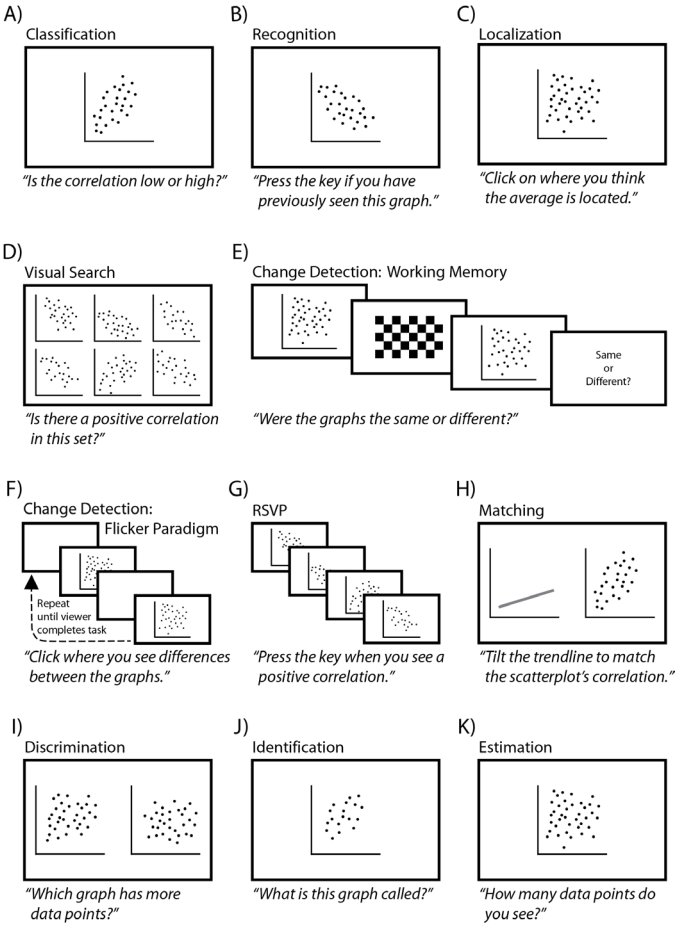

Fig. 3. A running example of scatterplots: eleven ways to study them. Researchers can ask viewers to categorize what they see (A), whether they recognize a stimulus (B), where a target is located (C), whether a target is present or absent (D), if they remember what changed (E) or if they see the change (F), to indicate when they see a target (G), to match a stimulus to another (H), to compare multiple stimuli (I), to identify what they see (J), or to estimate magnitude (K).

from vision science. We focus on quantitative methods because they support better replicability and generalizability and can connect tasks to designs in well-defined and actionable ways. While neural (i.e., brain-based) methods and models are integral for vision science, we do not yet understand how to connect these models to actionable visualization outcomes: visualizations are ultimately created to leverage the visual system and produce optimal viewing *behaviors*, such as facilitating data-driven decisions or gained insights. For these reasons, neural methods are not included here; however, visualizations may offer interesting scenarios for investigating neural activity. Further, physiological measures such as eye-tracking [3] and fNIRs [54] offer objective biometric insights into visualization use; however, these mechanisms require specialized hardware, experimental design, and nuanced interpretations that are beyond the scope of this work. Finally, existing surveys on related topics such as crowdsourcing [4] and statistical analysis [35] provide insight that can augment and influence experimental design, but we focus on broader methodological approaches that can be applied across deployment platforms and can be analyzed using a variety of statistical techniques.

3.2 Structure of the Design Space

Methodological approaches in other fields lie on a spectrum between broad surveys [22, 38, 44, 77] and specific techniques like research through design [83] or rapid ethnography [46]. The design space proposed here balances the complexity of the target problem with the flexibility of a broad survey by organizing relevant experimental techniques [12, 18, 33]. Design spaces identify key variables for a design problem, e.g., creating composite visualizations [33], to provide an actionable structure for systematically reasoning about solutions. While it is tempting to see experimental design as algorithmic and having an "optimal" solution, it is a creative process in practice—different experimental approaches to the same research question may yield different results—making it well-suited to design thinking. The flexibility of our design space allows researchers to adapt vision science methods for various formats of data collection– all response types are agnostic of the manner in which viewers provide their response (e.g., speaking, pressing keys, clicking/moving a mouse, writing/typing), and all paradigms could be used as the core task within a neural or physiological study. Each method is described, connected to the area of vision it was developed to investigate, and then supplemented with relevant concepts in data visualization. To assist in implementing these methods, we also mention relevant analyses and modeling procedures conventionally paired with these methods in order to properly interpret the data and results.

We divide the design space of experimental methods into four categories (see Figure 1 for an overview and Figure 2 for details) corresponding to key design decisions in crafting a visualization experiment:

- **Paradigm:** What task does the viewer have?
- **Adjustment Type:** How is the stimulus level adjusted?
- **Response Type:** How does the viewer respond to that task?
- **Dependent Measures:** How do we measure performance?

The described methods largely come from *psychophysics*. The basic premise of modern psychophysics is testing a viewer's ability to detect a stimulus, identify a stimulus, and/or differentiate one stimulus from another. These methods allow researchers to create descriptive and predictive models of human perception through indirect measurements and probabilistic modeling of responses [21].

4 Paradigms

Paradigms (Figure 3) are the specific visual tasks viewers complete when using a visualization. Below we detail a set of popular vision science paradigms that may be easily extended to visualization research. Some paradigms are defined by only what the viewer is deciding, while more specialized techniques include additional relevant details such as display specifications, timing, and experimental manipulations.

4.1 Classification

In classification tasks, viewers identify a stimulus by categorizing it—usually according to predetermined choices. Viewers may classify an entire display (e.g., is the correlation low or high?) or specific regions or objects (e.g., is the target bar in this bar chart larger or smaller than the rest?) [49].

Advantages: Classification tasks can directly explain how people visually categorize stimuli. Classification is especially useful for both long-term and working memory experiments. A well-designed memory task can be used to disentangle blind guesses, familiar mistakes, and valid/accurate classifications, allowing researchers to quantify what information is retained and confused in memory without having to ask viewers directly (asking directly results in severe bias and noise [63]).

Limitations: Researchers must effectively determine the "ground truth" categories of their stimuli. This sometimes requires a pre-test with a training set or a separate group of viewers rating the test set. Setting categories ahead of time will also bias and constrain viewer responses [40], which must be taken into account during data analysis. For this reason, experience with machine learning and/or Signal Detection Theory [20] is useful in designing and analyzing classification tasks.

4.2 Recognition

This paradigm is used to test retention in short and long term memory. Recognition requires the viewer to indicate whether they saw a stimulus previously. A common recognition task involves presenting a set of images, often one at a time, and then later presenting a second set of

images composed of "old" images (previously seen ones) and "new" images. The viewer then indicates whether each image in the second set is old or new. For example, researchers may show a series of scatterplots with varying numbers of data groups and later show a new set of scatterplot groups, asking the viewer to indicate which scatterplots appear familiar. Another approach presents the viewer with a constant stream of images and ask them to indicate whenever an image repeats. Recognition tasks have already been used to study memorability in visualizations [5, 6] but are potentially useful for other aspects of visualization perception.

Advantages: Recognition tasks are simpler to design and implement (as opposed to classification tasks, for example), and task instructions are easy to explain to viewers. They produce categorical outcomes where the ground truth "correct" response is known by researchers.

Limitations: Recognition tasks cannot be used for continuous dependent measures. They typically use 2AFC or NAFC response types, and therefore follow the same limitations (see §6.2).

4.3 Localization

While classification and recognition are used to understand 'what' people see, localization helps us understand 'where' people see those items. Localization requires the viewer to indicate the location of a stimulus. It can be used to test different aspects of perception by asking where something is (e.g., attention) or where it previously was located (e.g., memory). Viewers can *click* to indicate where an object is or previously appeared. For example, Smart et al. [67] asked viewers to click on the mark in a chart (i.e., a data point in a scatterplot, a heatmap square, or a state in a U.S. map) that they thought encoded a particular value. Viewers can also *press a keyboard key* to specify which region of the screen contains the object of interest. For example, Nothelfer et al. [50] briefly showed viewers screens with many shapes and asked viewers to indicate which quadrant of the screen did not contain a particular shape; viewers responded by pressing one of four keys on a number pad, with each key corresponding to a quarter of the screen. Localization studies align well with experiments testing categorical independent variables.

Advantages: Localization tasks directly measure spatial attention and also indirectly measure display features and structures that guide or capture attention. Understanding where viewers perceive salient items or structure can inform design by predicting viewer behavior.

Limitations: If possible response areas are not explicit (e.g., "click on the box that contained X"), then regions of interest (ROIs) need to be defined to code responses (e.g., any clicks within a 10-pixel radius can be considered a correct localization). Ideally, ROIs are defined *a priori* and spatial overlap is accounted for. Researchers must justify these choices in data cleaning and analysis.

4.4 Detection

Detection requires viewers to indicate whether they perceive the presence of a particular stimulus. For example, researchers might ask a viewer whether they can see data points in a scatterplot in order to determine minimum mark size. Detection can be tested when the stimulus is on the screen (i.e., do you detect the target?) or directly after its presentation (i.e., did you detect the target?). Two common types of detection are visual search and change detection.

4.4.1 Visual Search

Visual search requires viewers to scan a visual scene to detect a target (object of interest) among distractors (irrelevant objects in the scene). For example, viewers could search for the scatterplot with the highest correlation value in a small multiples display. Search targets can be objects that the viewer searches for (e.g., the scatterplot with the highest correlation) or a feature like color (e.g., searching for red dots in a scatterplot with red, blue, and green dots). Visual search has been used to study visual attention for more than 50 years [30, 80].

A basic visual search trial begins with a blank screen showing a small fixation cross. This screen helps control both where the viewer is paying attention (their *spatial attention*) and physically looking by restricting the viewer to the same visual starting point prior to each response screen. Next, the stimulus appears on the screen (the *stimulus display onset*). Viewers typically indicate whether a target is "present" or "absent" as quickly as possible, without sacrificing accuracy. Viewers might then also indicate the location of the target if they reported it as present [49].

One of the key manipulations in a search task is the number of objects present in the visual search scene (calculated as the number of target(s) + distractor(s)), called *set-size*. Set size is the hallmark of a visual search task and is what distinguishes search from localization. Normally, visual search studies vary the number of distractors present over subconditions. For example, if a researcher wanted to know what chart type, connected scatterplot or dual-axis line chart, best facilitates search for positive correlations in a small multiples display, viewers could view small multiples of 5, 10, 15, or 20 connected scatterplots as well as small multiples of 5, 10, 15, or 20 dual-axis line charts. In this case, the study has four set-sizes.

Search tasks measure response time (RT) and accuracy. Performance is understood with search slopes (see "Performance Slopes" in §7.2). The slope of the RT $\times$ set size function describes search efficiency. This function represents the *search rate*—how much *more* time is required with the addition of each distractor to a visual scene. The steeper the slope, the more time is required to search at larger set sizes, which is called *serial* or *inefficient* search. A flat slope close to a value of 0 indicates "pop out," meaning that increasing the set size with distracting information does not affect how quickly people find the target [79].

Visual search studies systematically manipulate set size, and there is normally low error rate across all subconditions [78]. However, some experimental designs induce higher error rates through brief display times or limiting target rates [81]. Logan [41] provides a review of data modeling techniques for visual search data and their implications for models of attention. Visualization researchers could use search tasks to explore many questions, such as which regions of visualizations are salient or important, or how the complexity of a visualization affects what a viewer might attend to.

Advantages: Visual search provides a way to indirectly measure the efficiency of attention. Examining search efficiency could directly inform design guidelines that can help researchers understand how to design more complex displays. Set size manipulations are simple, and the task itself is easy to explain to viewers.

Limitations: Search tasks must be designed carefully, and almost always ask participants to localize a detected target (§4.3) to rule out random guessing. Additionally, while efficient search shows that a target captures attention, to generalize the results to design, researchers must determine what is driving this effect, for example, whether it is the mark's physical properties or its contrast with the background and other data.

4.4.2 Change Detection

Change detection (CD) is used to measure limitations in attention and working memory *capacity*—how much information can be held in mind over a brief interval. CD studies probe whether the viewer was able to notice and remember certain aspects of a stimulus. For example, change detection could help understand what items are attended to in a display and then held in working memory as viewers explore data or view animated transitions. Two common variations are working memory change detection and the flicker paradigm.

Working Memory (WM): A typical working memory change detection study starts with a blank screen showing a fixation cross, followed by a preliminary display screen showing the "before" stimuli. Often, the "before" display is followed by a *mask* (an unrelated image, like a checkerboard) to disrupt iconic memory [14], afterimages (such as the bright spot a viewer sees after appearing in a flash photograph) [7], and rehearsal strategies [57]. After a short delay, an "after" display appears. The "after" display may or may not contain a noticeable change compared to the "before" display. For example, a researcher could show a "before" display with 5 color-coded clusters of data in

a scatterplot. The "after" display could either show the same 5 colors or change one of the cluster colors, and ask whether viewers detect a difference. Depending on what the experiment measures, the nature of the display and its changes will vary. For example, experiments can test item location, feature information (e.g., color or shape), or direction of motion. In some working memory change detection tasks, researchers record response error distance in physical or feature space (see van den Berg et al. [74] for an example). The design possibilities for WM tasks are broad and largely untapped in visualization. Ma et al. [42] surveys what WM can tell us about what people see.

Flicker Paradigm: Tasks in a flicker paradigm continually alternate ("flicker") a display of an original image and a modified image, with a brief blank display in between. The blank display prevents local motion signals from interfering with high-level attentional control [58]. Flicker is used to study both working memory as well as change blindness [51] (see Rensink [59] for a survey). Flicker tasks measure the time it takes a viewer to notice the change and their accuracy in identifying the region of the image that has changed.

Advantages: CD tasks are often easy to design and implement, and viewer task instructions are simple to explain. Some tasks, like flicker paradigms, can be as short as a single trial (*one-shot*). One-shot CD tasks can show inattentional blindness and other illusions or robust failures in perception. WM tasks are critical for quantifying memory capacity, and these studies reduce noise and bias due to viewer habituation by making it hard to anticipate when a change occurs in the display.

Limitations: The biggest limitation in CD tasks lies in modeling the results. There is considerable debate in the working memory research about how to compute WM capacity (known as k) and how to handle response errors in viewer data [42]. This is an active debate among perception researchers and a yet unresolved problem in the field. Analysis and interpretation must be carefully justified by researchers.

4.5 Rapid Serial Visual Presentation (RSVP)

RSVP was designed to explore how viewers comprehend information from a fast series of stimuli [9]. This paradigm presents a set of images, including irrelevant images and at least one target image, in a rapid sequence (see Borkin et al. [5, 6] for an example). The images are shown one at a time at the same screen location. Viewers identify a target or target category (e.g., "name the chart type when you see a positive correlation" in a series of different visualizations or "press the button when you see a positive correlation" in a set of scatterplots).

RSVP experiments manipulate a number of factors, such as timing between stimuli, number of targets present, what type of image precedes a target image, timing between particular irrelevant images and the target image, or timing between target images, known as *lag manipulation*. Lag manipulation is common in RSVP, quantified as the number of irrelevant stimuli which appeared between two target images. For example, if a viewer saw a scatterplot with a positive correlation (the target), then two scatterplots with negative correlations, followed by a positive correlation (the second target), they will have completed a 'Lag 3' trial because the second target appeared 3 images later.

Advantages: RSVP tasks are especially useful for examining the time course of attention as well as modeling temporal shifts in attention and their impact on working memory. Well-designed RSVP tasks control eye movements and spatial attention, and would therefore be especially useful for investigating animation or displaying changes over time in data, including live, dynamic displays [73].

Limitations: RSVP tasks are limited by an inability to control gaze fixations due to the nature of the display. As we move our eyes, we miss intermediate visualizations (*saccadic blindness* [56]). One way to mitigate this is by using masking in the stream of images (§4.4.2).

4.6 Matching

In matching paradigms, the viewer adjusts one stimulus until it matches another. Viewers can match sub-features of a stimulus (e.g., adjust the luminance of one population in a two-class scatterplot until it matches the other) or match the entire stimuli (e.g., adjust the height of the leftmost bar until it matches the value of the middle bar in a bar chart). Given its versatility, the matching paradigm can be used to study a wide variety of perception and attention topics.

This paradigm is well-suited to understanding how well viewers aggregate data across visualization types. For example, experiments might ask viewers to adjust the angle of a trend line until it matches the correlation of the scatterplot [16] or to adjust the bar in a bar chart on the left until it matches the mean value of a swarm plot on the right. In Nothelfer & Franconeri [49], viewers adjusted the height of a bar to indicate the average delta in a dual bar chart.

Advantages: Matching paradigms are optimal for comparing data across visualization types and could be used to evaluate the utility of different design idioms. Matching also indirectly probes whether or not a viewer's mental representation is consistent across designs.

Limitations: Matching trials often require unlimited viewing time, so total experiment time could be long. Adjustment methods must be considered carefully (see §5.2 for details).

4.7 Discrimination

In the discrimination paradigm, viewers make comparative judgements about the magnitude of (typically) side-by-side stimuli, such as asking viewers to indicate which of two scatterplots contains more data points. This can be measured at multiple levels of data point numerosity.

Discriminations can be performed across separate stimuli (e.g., two scatterplots on a screen) or within the same stimulus (e.g., two data groups in the same scatterplot). Nothelfer & Franconeri [49] showed viewers dual bar charts, and viewers judged whether there were more increasing or decreasing bar pairs in each display. Rensink & Baldridge [62] asked viewers which of two scatterplots contained a higher correlation—a method extended by Harrison et al. [29] to rank other visualizations of correlation, including parallel coordinates, donut charts, and stacked area charts. Gleicher et al. [26] asked viewers to indicate which of two data groups in a scatterplot had the higher mean.

Advantages: Discrimination tasks are highly flexible and lend well to adaptive psychometric procedures (see §5.4). They are the preferred paradigm for evaluating perceptual precision (see §7.1) and can be used with complex stimuli such as dashboards.

Limitations: Potential limitations of discrimination tasks are largely contingent on the Adjustment Type (§5) used in their implementation. Researchers should be aware that using discrimination to measure accuracy is unnecessarily time-consuming and may be inefficient for subjective measures like preference.

4.8 Identification

Identification paradigms require viewers to respond with the identity of the stimulus using open-ended responses. In identification paradigms, experiments typically do not provide a recognition "template" or training set. Instead, identification tasks are used to study how viewers name stimuli. Viewers can be asked to identify an entire stimulus (e.g., what would you call this chart type?) or to identify a specific feature (e.g., name the color of the less correlated marks in this display).

Advantages: Identification paradigms offer less biased insight into perceived categories than classification tasks. They are also useful when predetermined categories are unavailable. For example, understanding how viewers segment color bins may provide insight on how to design better multihue palettes [55]. Identification tasks can work as a *pre-task* for classification tasks to generate the categories that are later fed into a classification study. Whereas classification task provide a mental template (categories), identification tasks require viewers to access their individual long term memory store to identify stimuli.

Limitations: Because viewers are not provided with a mental template or a set of categories, identification trials may take longer than classification trials. Additionally, in some cases researchers should be prepared to account for a wider variety of response categories since they are not constrained ahead of time. This has implications for coding data before analysis that must be carefully considered [20].

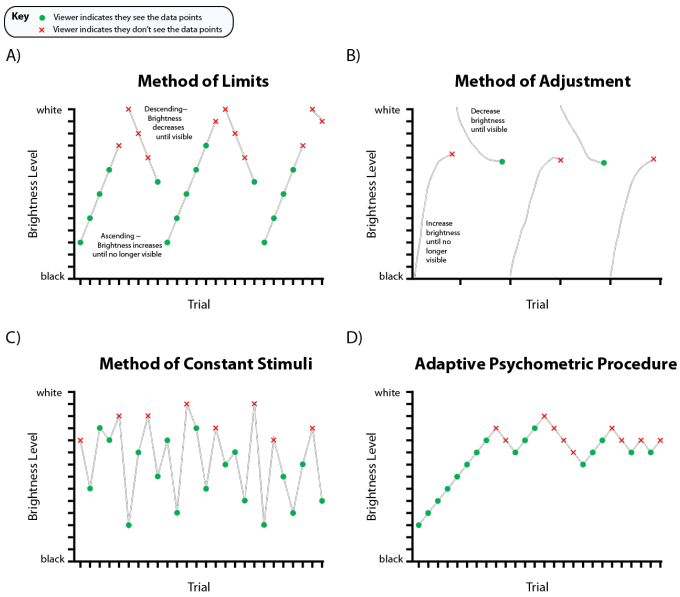

Fig. 4. Four adjustment types. Researchers can adjust the brightness of scattered dots on a white background until participants report that they are no longer visible (A), have participants adjust the brightness level of the scattered dots until they are just visible (B), present participants with random brightness levels and ask them to report whether the dots are visible or not (C), or find a visibility threshold by adjusting brightness until the viewer can reliably detect it 75% of the time (D).

4.9 Estimation

Estimation paradigms require viewers to directly estimate some value of a continuous feature in a display. Magnitude production is the most common type of estimation task, where viewers are required to estimate the magnitude of a stimulus with a numeric response. Estimation tasks are different from classification tasks, which ask viewers to categorize stimuli. For example, if a researcher wished to know how viewers perceive correlation strength in a scatterplot, viewers could *classify* the correlation as ”low” or ”steep” or *estimate* it at ”0.2” or ”0.8”.

Advantages: Estimation tasks measure accuracy, have intuitive instructions, and are amenable to various Adjustment Types. To obtain a full psychometric function, the level of the stimulus can be systematically manipulated to understand how close the viewer's response is to its true value at different magnitudes. This function can be used to evaluate, generalize, and predict future estimation performance.

Limitations: Estimation paradigms do not capture precision (see §4.7) and should not be used to obtain objective magnitudes from viewers. This is because perceptual estimates of most feature properties are systematically biased (e.g., we underestimate mid levels of correlation magnitude [61]). This bias is often modeled as an instance of Steven's Law, Ekman's Law, or Fechner's Law [72].

5 Adjustment Types

Psychophysical adjustment types (Figure 4) define the overall structure of perceptual experiments by determining the manner in which the stimulus level will be adjusted and responded to. Here, we discuss the three main types—Method of Limits, Method of Adjustment, and Method of Constant Stimuli—and adaptive psychophysical procedures which aid their use.

5.1 Method of Limits [MoL]

The goal of the Method of Limits is detection: the researcher wishes to identify the level at which people see a target property in an image by steadily changing that property until the viewer sees (or no longer sees) the target property. For example, to detect the upper bound for colors of scatterplot points, an experiment may start with a scatterplot of white dots on a white background and slowly decrease the lightness of the marks until viewers can perceive them. The result gives researchers the highest detectable lightness level that can be used to draw scatterplots. This can be done using either *ascending* or *descending* methods, and it is common to use both in a single experiment. Ascending MoL tasks start at a low level of magnitude (often zero) and increase the level of the stimulus over time, requiring viewers to indicate when they can perceive it. Descending MoL tasks start at a high level of the stimulus and decrease its level, requiring viewers to indicate when they can no longer perceive it. In the scatterplot example above, an ascending MoL design would start with black dots and show viewers increasingly lighter dots until the marks were no longer visible on a white background. A descending design would start with white marks and decrease lightness until viewers report that the marks become visible.

Advantages: MoL tends to be easy to implement and easy for viewers to understand: studies need to provide viewers with a single value to observe, such as the visibility of marks. These features collectively mean that MoL studies are often fast, affording more trials in a short time. Precisely manipulating a single feature also allows experiments to collect precise measurements about specific phenomena (e.g., color perception) in context using a single stimuli.

Limitations: While MoL provides precise per-trial measures, viewers can quickly habituate to trials and often begin predicting when the stimulus will become perceivable or imperceivable. These predictions lead to premature responses, called anticipation errors, that are not precise or accurate representations of perception. Habituated viewers may also become less sensitive to the stimulus overall. Techniques like staircase procedures (§5.4) help address these limitations in practice.

5.2 Method of Adjustment [MoA]

The Method of Adjustment operates on the same principles as MoL, but instead of the researcher manipulating a target feature, viewers directly adjust properties of a visualization until they reach a perceptual criteria such as “present/detectable,” “absent/indetectable,” or “equal to X.” This task type is repeated over many trials, and the difference between the correct stimulus level and the viewer response is typically recorded and averaged over all trials as a measure of perceptual sensitivity. In the MoL example, viewers would adjust the lightness of scatterplot points until they are just visible. This trial type could be repeated for marks with different starting colors to determine an absolute threshold. Averaging errors at each level of lightness tested would indicate perceptual sensitivity at each level of lightness across different color categories.

Advantages: The method of adjustment allows for a broader sampling of space of possible responses, datasets, and designs since researchers can vary the distance between the adjustable stimuli (e.g., the color of a mark) and the defined objective (e.g., matching to different colors or backgrounds) over many trials. The potential for many interleaved conditions and trial types means less risk of habituation and possible increases in statistical power. Viewers can also perform either absolute threshold detection (e.g., adjusting to a fixed value) or relative threshold detection (e.g., adjusting to match a target stimulus). This method works well for target tasks measured along continuous levels and can provide highly sensitive results using a relatively small number of stimuli from a precise sampling of errors across viewers, resulting in concrete, numerical guidelines such as the grid-line alpha values in [1].

Limitations: Experiments using MoA require complex design and implementation. Researchers must choose the levels of target variable to test, as well as the starting distances between adjustable properties of the display and target stimuli both from-above and from-below. MoA studies are also sensitive to how people interact with visualizations during the experiment. For example, some studies leverage keyboard inputs (e.g., using the arrow keys to increase or decrease a value) or sliders; however, the design of these inputs may affect how precisely viewers adjust a visualization [43]. Finally, viewers may develop a motor pattern of adjustment (i.e., “muscle memory”) that biases their responses over time, sometimes using arbitrary heuristics like, “10 presses increasing the lightness should be enough.” To prevent these

kinds of predictions and habits, researchers can jitter the adjustment values non-linearly so that one increase or decrease increment is not the same value as the next. Clear task instructions, practice trials, and comprehension checks can help ensure viewer task understanding.

5.3 Method of Constant Stimuli [MoCS]

The Method of Constant Stimuli is among the most common methods in modern psychophysics experiments. Like MoL, MoCS is optimized for detection paradigms (§4.4) and is also commonly used for classification, recognition, or identification. MoCS presents viewers with random levels of a target property, presented randomly across trials, and asks them to draw inferences about that property. Following the previous scatterplot example, researchers can present scatterplots with marks at different lightness levels and ask viewers to indicate when the dots are present or absent. This method is often used so viewers can assess different properties of the data or display, such as determining which feature (e.g., color or shape) influences average estimation in a scatterplot [26].

Advantages: MoCS enables a diverse range of response types: viewers can be asked to detect an absolute threshold (e.g., "present"/"absent"), much like the method of limits, but they can also be asked to identify relative thresholds based on exemplars (e.g., "greater than x") or even perform stimuli classification (e.g., "red/blue/green"). Responses can be recorded as binary responses (yes/no), along a continuous (e.g., a magnitude from 0-100), or on a categorical scale (e.g., a color category). MoCS also allows the researcher full control over how the stimuli are sampled (e.g., how wide of a difficulty range is used) to afford creating a full psychometric function or testing large response space. These experiments generally provide greater response precision and objectivity due to less viewer habituation: randomizing the order of stimulus levels and interleaving trials with different properties of interest can prevent trial-to-trial response predictions.

Limitations: Experiments using this method can be complicated to implement. Because the target property is sampled rather than estimated by the viewer, it requires a greater number of trials per viewer than other methods. Researchers must also decide how to sample the space of possible datasets and visualization designs. This can be modeled through psychometric functions with Ideal Observer Analysis [24, 66]. Researchers need to decide and justify how broadly and evenly they sample across variable levels in MoCS experiments.

5.4 Adaptive Psychophysical Procedures [APP]

Because perception is neither perfect nor absolutely precise, viewers may never detect a given magnitude of a stimulus 100% of the time [34]. APPs help researchers find absolute, intensive thresholds in perception by adapting the stimulus level sampling procedure used in the above methods based on viewer responses. Researchers using APPs can adjust the visualizations presented to viewers based on their current performance relative to a threshold (e.g., 75% correct detection) to capture and represent the perceptual processes being measured [15].

The most common APP is staircasing, where experiments increase or decrease the discriminability of presented stimuli depending on the viewer response in the current trial. For example, Rensink & Baldridge [62] use staircasing to find correlation JNDs in scatterplots. They asked viewers to indicate which of two scatterplots has a higher correlation. If viewers respond correctly, the next pair of plots would have closer correlation values; if they respond incorrectly, the next pair would have a larger difference in their correlations. In staircasing, the adjustment continues until some steady-state criteria is met (e.g., 50% accuracy over n trials). Several algorithmic variants of staircasing and similar techniques are reviewed in detail by Otto & Weinzierl [52].

Advantages: APPs improve measurement quality. They allow researchers to collect precise performance estimates using fewer trials sampled optimally from the possible levels of the target variable.

Limitations: APPs can be difficult to implement, and researchers must justify their choice in algorithm as well as their choice in steady-state criteria. They also result in varied experiment duration that is viewer performance dependent.

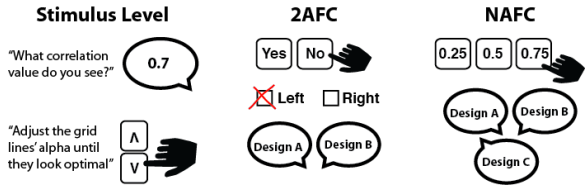

Fig. 5. Three ways to elicit responses from viewers during an experiment: stimulus level reporting (left), two-alternative forced-choice (2AFC; center) response, and multiple-alternative forced-choice (NAFC; right).

6 Response Types

There are three popular ways to elicit responses from viewers during an experiment: stimulus level reporting, two-alternative forced-choice (2AFC) response, multiple-alternative forced-choice (NAFC) (Figure 5). These response types can be used in a variety of paradigms and can output different dependent measures.

6.1 Stimulus Level

Researchers can elicit direct reports, such as perceived values, from viewers. For example, in Xiong et al. [82], viewers reported the average vertical positions of lines and bars in a chart by drawing a line indicating perceived mean value on the screen. In [1], viewers adjusted the alpha value of gridlines until they considered it to be optimal. Viewers can report stimulus level verbally, by typing in a specific value, or visually by recreating the stimulus level.

Advantages: Stimulus reports enable researchers to measure the specific amount of deviation or bias of a percept from ground truth, called error. Because viewers directly report a value instead of selecting from several alternatives, researchers can quantify and model the specific amount viewer reporting deviates from ground truth.

Limitations: Reporting stimulus level can introduce biases like motor inertia (see §5.2) and whole-number bias [31]. For instance, if asked to verbally report scatterplot correlation values, whole-number bias or proportion judgment bias may cause viewers to exclusively report correlation in coarse increments such as 0.25, 0.5, and 0.75.

6.2 2AFC

Two-alternative forced-choice tasks give viewers two options after perceiving certain stimuli. Common choices for the two alternatives are comparison and categorization. In 2AFCs designed for comparison, viewers typically identify which of two alternatives measures better on a certain metric. For instance, A/B tests commonly use designs where viewers see two designs and determine which one they prefer. Another type of 2AFC is the "-er" task, where viewers decide which of the two alternatives is [e.g., dark]-"er" than the other. Cleveland & McGill used this approach in their canonical study [13], where viewers had to determine which of two values were smaller.

The choices could be presented verbally or visually. For a verbal task, viewers might be given a series of dashboard interfaces to determine whether each shows an increasing trend in sales. The viewer would indicate yes/no upon seeing each dashboard. For a visual manipulation, viewers might be presented with two configurations of a stimulus to choose from. For example, the researcher could present a viewer with two designs of a dashboard and ask them to select the one showing a greater increasing trend.

The stimuli and choices in a 2AFC task can be presented over space or over time. In a spatial presentation, viewers see both alternatives at once to make their decision. In a temporal presentation, the stimuli are shown in the same location on the screen but over a certain time interval. For example, to test which interface (A or B) shows a larger trend, a spatial 2AFC design would show both interfaces at the same time, one on the left, and one on the right, while a temporal 2AFC design would show one interface for a certain duration on screen, take it away, then show another interface for a certain duration.

Advantages: 2AFC experiments afford straightforward data analysis. The binary response input allows researchers to classify correct hits, correct rejections, misses, and false alarms. Signal Detection Theory can be used to infer sensitivity and bias [20], which can in turn help us describe which trends, patterns or visual characteristics are apparent or preferred for a viewer. Another critical advantage of 2AFC tasks is that the researcher can control the rate of criterion they present. With only two choices, 2AFC tasks also motivate viewers to scrutinize the presented stimuli to capture subtle differences.

Limitations: 2AFC tasks may be subject to response bias. When two alternatives are presented to the viewer, they could interact with each other via anchoring effects. Viewers might become more sensitive to the first alternative they see, causing their judgment criteria to change by the time they view the second alternative. Another limitation of the 2AFC task is that it requires multiple viewers or replications of trials to counterbalance stimulus presentation order (to control for order effects). In a preference task, if the two alternatives are equally preferred, the researchers need to aggregate the results of multiple trials and then compare the number of preferences for each alternative to see if they are different. In other words, while it is easy to assess percentage correct in a 2AFC task, obtaining the exact error is difficult without formally modeling the data.

6.3 NAFC

NAFC (or N-alternative forced-choice) scales up a 2AFC task. Instead of presenting the viewer with two alternatives, the researcher shows multiple (N) alternatives. For example, in a correlation classification task, given a ground-truth correlation of 0.5, in a 2AFC task, the researcher might ask the viewer whether the correlation is 0.5 or not, while in a NAFC task, the researcher could ask the viewer to choose the correct correlation from a set of N values: 0.25, 0.5, and 0.75 (N=3).

Advantages: NAFC tasks increase how efficiently experiments can detect random guessing. For example, if four alternatives are presented, random chance drops to 25%. An NAFC task also measures the degree of bias more precisely. For example, the researcher could provide the viewer with five options depicting the difference between A and B: A much greater than B, A slightly greater than B, A equals B, A slightly smaller than B, and A much smaller than B.

Limitations: Although NAFC provides more fine-grained information to measure bias, it is still limited by the size of N. Limitations on human visual attention suggests that viewers should be provided no more than six alternatives [32]. With six alternatives, it becomes difficult to quantify the specific amount that viewer perception deviates from the ground truth. As with 2AFC tasks, while assessing the percentage correct is straightforward, modeling the exact amount of error in response is complicated. Further, the options provided to the viewers could interact with each other or the presented stimulus in memory to cause memory decay, biasing the accuracy of the final response.

7 Dependent Measures

Dependent measures provide metrics for assessing parameters of a visualization, as shown in Figure 6. They help researchers concretely quantify how people process visualized data. Experiments should use dependent measures that allow designers to make informed and generalizable decisions about visualizations across a breadth of relevant designs. Once computed, researchers can use a plethora of statistical methods to interpret the resulting outcomes (c.f., Kay et al. [35]).

7.1 Direct Dependent Measures

We traditionally think of dependent measures as a single number directly measuring how well people process visual information, such as how quickly or how accurately they found a statistic in their data. While visualization largely relies on time and accuracy, efforts such as BELIV have encouraged an expanded library of techniques and measures for assessing visualizations. In vision science, time and accuracy are likewise dominant (though often are only part of the total measure, §7.2). We refer to measures whose distributions capture performance as direct dependent measures. Common direct measures include:

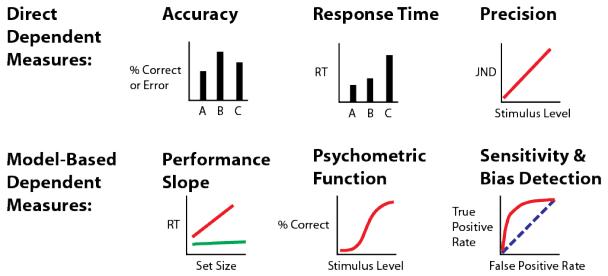

Fig. 6. Researchers can analyze dependent measures based on direct measures (top) and model-based measures (bottom). They can analyze viewers' percentage correct or degree of error (top left) to understand how close viewers' judgments are to the true value, how quickly viewers can complete a task (top center), variability of viewers' judgments (top right), how robust is a visualization for a given task (bottom left), how decision behavior changes as a function of changes in a stimulus (bottom center), and functions of sensitivity and bias in a ROC curve (bottom right).

Accuracy: ***How close to the true value are people's judgements?*** Visualization experiments conventionally measure accuracy in two ways: percentage correct (how often do I get the answer right?) and error (how close is my estimate to the true value?). Percentage correct provides a more coarse estimate of visualization effectiveness as a binary correct/incorrect; however, it enables faster responses and greater control over parameters like difficulty. Error offers more precise methods for gauging people's abilities to infer statistics from data; however, it offers little control over parameters such as task difficulty due to potential confounds and typically requires more time per trial.

Accuracy provides an intuitive metric for assessing visualizations, but the simplicity of mean accuracy may hide more sophisticated relationships between visualization design and perception. For instance, while accuracy can tell us whether a value is over- or under-estimated, it cannot tell us how precisely that value is perceived compared to others. Most model-based dependent measures, such as psychometric functions or sensitivity and bias, use accuracy to form more nuanced insights into performance.

Response Time: ***How quickly can people complete a task?*** Visualizations typically aim to communicate information both quickly and accurately. Response time (RT) characterizes the time it takes to complete a task with one stimulus. Studies typically use response time either on its own (for simpler tasks) or in conjunction with accuracy (for more challenging tasks [53]) to understand how readily people can infer information from a visualization. While ill-suited for paradigms requiring rapid presentation, RT has often been used in visualization and vision science studies to measure how long it takes people to process visual information with lower response times typically implying more efficient visual processing.

RT provides an intuitive measure well-aligned with traditional visualization goals. However, it requires careful control and, at the time scales of many visualization tasks, may be subject to significant individual differences. As with accuracy, raw RT can inform visualization design; however, it better serves to ground models that allow designers to use experimental outcomes to tailor visualizations to data and tasks.

Precision: ***How variable are people's judgements?*** Precision is typically quantified as a threshold of performance. Threshold measures correspond to the bounds within which we can reliably observe a particular property of a visualization, such as the level of brightness at which dots in a scatterplot are visible. Studies typically measure thresholds through either adjustment tasks (§5.2) where thresholds are derived from the range of values provided by viewers [62] or classification tasks (§4.1) where thresholds are determined by adjusting parameters of the data or visualization until a desired effect is observed [21]. One of the most common threshold measures is a just noticeable difference (JND). JNDs are the threshold at which people can detect a difference with a given reliability (typically 50% or 75%).

Thresholds provide probabilistic bounds on the resolution of what people can see in a visualization, allowing designers to reason about how much information is communicated (e.g., the ability to discern the height of a bar or the level of correlation). However, threshold tasks require a significant number of trials due to the amount of expected noise in viewers' responses and the parameter adjustment required to precisely estimate thresholds.

7.2 Model-Based Dependent Measures

We can use factors like time and accuracy to model viewer behavior. We build these models by aggregating accuracy or RT according to systematically-varied attributes of the visualization or data. We can then use statistical comparisons to draw conclusions from the models. These models allow us to make more precise claims about how the visual system processes information but typically require more data and a more complex experimental design. Model-based measures include:

Performance Slope: ***How robust is the visualization for a given task?*** One method of using accuracy or RT to provide more nuanced insight into a visualization is by modeling how these measures change as a function of the difficulty of a task. A common example of this in vision science is in visual search (§4.4.1): how quickly and accurately can I locate a target as the number of targets increases? Experiments measure search slope by systematically varying the number of data points (or another aspect of task difficulty) and tracking accuracy and/or RT at each level. A linear regression would fit a line to the resulting pattern, and the resulting slopes correspond to how sensitive performance is to changes in difficulty. Lower slopes corresponding to more robust designs and a slope of 0 indicates that people can do the task robust to the chosen difficulty level (e.g., finding a point that "pops-out" [79]).

Slopes give us quantitative insight into the relationship between the data and the design by measuring how performance changes with different data characteristics. However, they also assume that performance changes linearly with difficulty and measures changes in RT and accuracy, emphasizing robustness over overall performance. For example, a bar chart may offer lower RT slopes than a dot plot for finding the largest value but only because the bar chart aggregates away the largest value, making it impossible to find.

Psychometric Function: ***How does performance change as a function of design?*** Psychometric functions model change in decision behavior (e.g., the number of correct or incorrect decisions) as a function of continuous changes in a stimulus (e.g., the luminance range of a colormap) using a logistic function [21]. The magnitude of the function at a given point estimates performance for that design setting whereas the spread of the function correlates with noise imparted during the task and the inflection point corresponds to key decision making thresholds. For example, Kale et al. [34] use psychometric functions to compare JNDs and noise in trend estimation in uncertainty visualization.

Psychometric functions offer a way to model decision making behavior as a function of data and design. They capture values at which people predictably make a decision (e.g., when we can reliably estimate which of two marks is larger [67]?) and the noisy space between where behavior is less well-defined (e.g., how frequently a mark will be estimated as larger?). While psychometric functions offer a powerful tool for modeling perceived values and decision making behaviors, they also require careful experimental control to correctly map the relationship between relevant aspects of a visualization and require tasks that can reliably be modeled as a binary decision (e.g., 2AFC).

Sensitivity & Bias Detection: ***How well does what we see match our data and bias our decisions?*** Signal detection theory provides a method for modeling performance as a function of sensitivity (how does performance change as the data or design changes?) and bias (do viewers have a tendency towards a certain response?). Signal detection begins by identifying true and false positive and negative responses at different levels of a target independent variable. Once these responses are computed, the resulting patterns are modeled to construct curves showing how sensitivity changes over the corresponding variables, typically with one curve per level of categorical independent variable. Bias corresponds to the curve intercepts, and sensitivity corresponds to the parameters of the curve. Signal detection is explained in more detail in other sources [20, 23, 24, 45, 68].

As with psychometric functions, sensitivity and bias detection allow us to measure how well a visualization performs under different conditions and statistically disentangles meaningful aspects of performance from noise. It also allows researchers to measure potential bias in visualization interpretation and can apply to tasks beyond conventional decision making and detection as well as those that may not have a linear or logistic pattern. Sensitivity and bias detection enable researchers to use more traditional algorithmic analyses, such as ROC curves [68], to analyze their results. However, like other model-based measures, using these measures requires experiments that are carefully structured to systematically manipulate relevant variables, and comparing sensitivity parameters is a level of abstraction removed from more direct metrics like accuracy or slope.

8 Discussion & Future Work

In this paper, we discuss the value of using perceptual methods in rigorous visualization experiments. Our design space can be applied to facilitate novel research by systematically examining viewer behavior and to produce replicable and generalizable design guidelines. This paper is non-exhaustive and prioritizes breadth and structure of methods over depth. This is not a complete handbook on how to study perceptual mechanisms; however, we do anticipate our design space being highly useful for experimenters, reviewers, and readers in critical planning of new studies and evaluation of past work.

We cover task design topics extensively but exclude fundamentals of behavioral research such as experimental control, the basics of hypothesis testing, implementation (e.g., experiment software), and materials (e.g., hardware). Additionally, essential modeling techniques such as Signal Detection Theory, Ideal Observer Analysis, and the statistical methods used to analyze experimental data (e.g., Bayesian inference) are beyond the scope of this work. We strongly encourage researchers to consider these topics as part of any experimental design.

Visualization research can sample this design space to construct experiments that measure key components of visualization design and help bridge findings from vision science. We have included a supplementary guide showcasing how an experiment could be designed following this design space at https://visxvision.com/using-the-design-space/. Note that while experiments can use these methods in common configurations (Fig. 2), we hope that the structure provided by these four phases will also yield novel and innovative studies. Shared methodologies and associated vocabularies can help researchers understand what makes visualization comprehension unique from other visual experiences. For example, these methods may explore critical thinking affordances in visualizations or biases from visual illusions. Understanding visualization as a function of design and of the visual mechanisms used to process those designs may lead to broadly generalizable guidelines and more effective visualization practices. By providing a library of common techniques and relevant terminology, we hope to bridge lexical divides between visualization and vision science and better facilitate these innovations.

9 Conclusion

We provide a design space of vision science methods for visualization research, along with a shared lexicon for facilitating deeper conversation between researchers in both fields. Our goal was to synthesize and organize the most popular experimental tools to provide a foundation for evaluating and conducting future research. We hope that the visualization community sees the value in diversifying experimentation, and embracing new approaches to advance our knowledge about both human behavior and guidelines for design.

Acknowledgments

We thank Dr. Ronald Rensink for his feedback and suggestions and Dr. Tamara Munzner, Dr. Juergen Bernard, Zipeng Liu, Michael Oppermann, Francis Nguyen, and our reviewers for their thoughtful comments. This work was supported by NSF #1764092, #1764089, #1657599, and the UBC 4 Year Ph.D. Fellowship.

REFERENCES

[1] L. Bartram and M. C. Stone. Whisper, don't scream: Grids and transparency. *IEEE Transactions on Visualization and Computer Graphics*, 17(10):1444–1458, 2010.

[2] W. I. B. Beveridge. *The Art of Scientific Investigation*. Blackburn Press, Caldwell, NJ, 2004.

[3] T. Blascheck, K. Kurzhals, M. Raschke, M. Burch, D. Weiskopf, and T. Ertl. Visualization of eye tracking data: A taxonomy and survey. In *Computer Graphics Forum*, vol. 36, pp. 260–284. Wiley Online Library, 2017.

[4] R. Borgo, L. Micallef, B. Bach, F. McGee, and B. Lee. Information visualization evaluation using crowdsourcing. In *Computer Graphics Forum*, vol. 37, pp. 573–595. Wiley Online Library, 2018.

[5] M. A. Borkin, Z. Bylinskii, N. W. Kim, C. M. Bainbridge, C. S. Yeh, D. Borkin, H. Pfister, and A. Oliva. Beyond Memorability: Visualization Recognition and Recall. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):519–528, Jan. 2016. doi: 10.1109/TVCG.2015.2467732

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

[7] C. Bradley and J. Pearson. The sensory components of high-capacity iconic memory and visual working memory. *Frontiers in psychology*, 3:355, 2012.

[8] C. A. Brewer, G. W. Hatchard, and M. A. Harrower. Colorbrewer in print: a catalog of color schemes for maps. *Cartography and geographic information science*, 30(1):5–32, 2003.

[9] D. E. Broadbent and M. H. Broadbent. From detection to identification: Response to multiple targets in rapid serial visual presentation. *Perception & psychophysics*, 42(2):105–113, 1987.

[10] M. Burch, L. Chuang, B. Fisher, A. Schmidt, and D. Weiskopf. *Eye tracking and visualization: Foundations, Techniques, and applications. etvis 2015*. Springer, 2017.

[11] Z. Bylinskii, M. A. Borkin, N. W. Kim, H. Pfister, and A. Oliva. Eye fixation metrics for large scale evaluation and comparison of information visualizations. In *Workshop on Eye Tracking and Visualization*, pp. 235–255. Springer, Cham, 2015.

[12] S. K. Card and J. Mackinlay. The structure of the information visualization design space. In *Proceedings of VIZ'97: Visualization Conference, Information Visualization Symposium and Parallel Rendering Symposium*, pp. 92–99. IEEE, 1997.

[13] 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. doi: 10.2307/2288400

[14] M. Coltheart. Iconic memory and visible persistence. *Perception & psychophysics*, 27(3):183–228, 1980.

[15] T. N. Cornsweet. The staircrase-method in psychophysics. *The American Journal of Psychology*, 75:485–491, Sept. 1962.

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

[17] A. Crisan and M. Elliott. How to Evaluate an Evaluation Study? Comparing and Contrasting Practices in Vis with Those of Other Disciplines : Position Paper. In *2018 IEEE Evaluation and Beyond - Methodological Approaches for Visualization (BELIV)*, pp. 28–36. IEEE, Berlin, Germany, Oct. 2018. doi: 10.1109/BELIV.2018.8634420

[18] N. Diakopoulos, F. Kivran-Swaine, and M. Naaman. Playable data: characterizing the design space of game-y infographics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1717–1726, 2011.

[19] M. A. Elliott. Interference in the perception of correlation in two population scatterplots. Master's thesis, University of British Columbia, 2016.

[20] D. S. Emmerich. *Signal Detection Theory and Psychophysics* . David M. Green , John A. Swets. *The Quarterly Review of Biology*, 42(4):578–578, Dec. 1967. doi: 10.1086/405615

[21] B. Farell and D. Pelli. Psychophysical Methods, or how to Measure a Threshold, and why. *Vision Research: A Practical Guide to Laboratory Methods*, Jan. 1999. doi: 10.1093/acprof:oso/9780198523192.003.0005

[22] A. Følstad, E. Law, and K. Hornbæk. Analysis in practical usability evaluation: a survey study. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2127–2136, 2012.

[23] W. S. Geisler. Sequential ideal-observer analysis of visual discriminations. *Psychological Review*, 96(2):267–314, 1989. doi: 10.1037/0033-295X.96.2.267

[24] W. S. Geisler. Ideal observer analysis. *The visual neurosciences*, 10(7):12–12, 2003.

[25] G. A. Gescheider. *Psychophysics: the fundamentals*. Psychology Press, 2013.

[26] 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.

[27] C. C. Gramazio, D. H. Laidlaw, and K. B. Schloss. Colorgorical: Creating discriminable and preferable color palettes for information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):521–530, Jan. 2017. doi: 10.1109/TVCG.2016.2598918

[28] S. Haroz and D. Whitney. How capacity limits of attention influence information visualization effectiveness. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2402–2410, 2012.

[29] 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, Dec. 2014. doi: 10.1109/TVCG.2014.2346979

[30] C. G. Healey and J. T. Enns. Attention and Visual Memory in Visualization and Computer Graphics. *IEEE Transactions on Visualization and Computer Graphics*, 18(7):1170–1188, July 2012. doi: 10.1109/TVCG.2011.127

[31] J. Hollands and B. P. Dyre. Bias in proportion judgments: the cyclical power model. *Psychological review*, 107(3):500, 2000.

[32] S. S. Iyengar and M. R. Lepper. When choice is demotivating: Can one desire too much of a good thing? *Journal of personality and social psychology*, 79(6):995, 2000.

[33] W. Javed and N. Elmqvist. Exploring the design space of composite visualization. In *2012 ieee pacific visualization symposium*, pp. 1–8. IEEE, 2012.

[34] A. Kale, F. Nguyen, M. Kay, and J. Hullman. Hypothetical outcome plots help untrained observers judge trends in ambiguous data. *IEEE transactions on visualization and computer graphics*, 25(1):892–902, 2018.

[35] M. Kay, G. L. Nelson, and E. B. Hekler. Researcher-centered design of statistics: Why bayesian statistics better fit the culture and incentives of hci. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 4521–4532, 2016.

[36] R. Kosara. An Empire Built On Sand: Reexamining What We Think We Know About Visualization. In *Proceedings of the Beyond Time and Errors on Novel Evaluation Methods for Visualization - BELIV '16*, pp. 162–168. ACM Press, Baltimore, MD, USA, 2016. doi: 10.1145/2993901.2993909

[37] R. Kosara. Circular part-to-whole charts using the area visual cue. In *21st Eurographics Conference on Visualization, EuroVis 2019-Short Papers*, 2019.

[38] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. Empirical Studies in Information Visualization: Seven Scenarios. *IEEE Transactions on Visualization and Computer Graphics*, 18(9):1520–1536, Sept. 2012. doi: 10.1109/TVCG.2011.279

[39] S. Lewandowsky and I. Spence. Discriminating Strata in Scatterplots. *Journal of the American Statistical Association*, 84(407):682–688, Sept. 1989. doi: 10.1080/01621459.1989.10478821

[40] K. R. Livingston, J. K. Andrews, and S. Harnad. Categorical perception effects induced by category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24(3):732, 1998.

[41] G. D. Logan. Cumulative progress in formal theories of attention. *Annual Review of Psychology*, 55:207–234, 2004. doi: 10.1146/annurev.psych.55.090902.141415

[42] W. J. Ma, M. Husain, and P. M. Bays. Changing concepts of working memory. *Nature Neuroscience*, 17(3):347–356, Mar. 2014. doi: 10.1038/nn.3655

[43] J. Matejka, M. Glueck, T. Grossman, and G. Fitzmaurice. The effect of visual appearance on the performance of continuous sliders and visual analogue scales. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 5421–5432, 2016.

[44] J. E. McGrath. Methodology matters: Doing research in the behavioral and social sciences. In *Readings in Human–Computer Interaction*, pp. 152–169. Elsevier, 1995.

[45] D. McNicol. *A Primer of Signal Detection Theory*. Psychology Press, Jan. 2005. doi: 10.4324/9781410611949

[46] D. R. Millen. Rapid ethnography: time deepening strategies for hci field

research. In *Proceedings of the 3rd conference on Designing interactive systems: processes, practices, methods, and techniques*, pp. 280–286, 2000.

[47] S. Mittelstädt, A. Stoffel, and D. A. Keim. Methods for compensating contrast effects in information visualization. In *Computer Graphics Forum*, vol. 33, pp. 231–240. Wiley Online Library, 2014.

[48] C. Nothelfer, Z. Bylinskii, M. Elliott, C. Xiong, and D. A. Szafir. Vision science meets visualization. *IEEE VIS Workshop*, 2017. https://visxvision.com/.

[49] C. Nothelfer and S. Franconeri. Measures of the benefit of direct encoding of data deltas for data pair relation perception. *IEEE transactions on visualization and computer graphics*, 26(1):311–320, 2019.

[50] 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.

[51] J. K. O'regan, R. A. Rensink, and J. J. Clark. Change-blindness as a result of 'mudsplashes'. *Nature*, 398(6722):34, 1999.

[52] S. Otto and S. Weinzierl. Comparative simulations of adaptive psychometric procedures. p. 4, 2009.

[53] J. Palmer. Attentional limits on the perception and memory of visual information. *Journal of Experimental Psychology. Human Perception and Performance*, 16(2):332–350, May 1990.

[54] E. M. M. Peck, B. F. Yuksel, A. Ottley, R. J. Jacob, and R. Chang. Using fnirs brain sensing to evaluate information visualization interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 473–482, 2013.

[55] P. S. Quinan, L. Padilla, S. H. Creem-Regehr, and M. Meyer. Examining implicit discretization in spectral schemes. In *Computer Graphics Forum*, vol. 38, pp. 363–374. Wiley Online Library, 2019.

[56] J. E. Raymond, K. L. Shapiro, and K. M. Arnell. Temporary suppression of visual processing in an RSVP task: An attentional blink? *Journal of Experimental Psychology: Human Perception and Performance*, 18(3):849–860, 1992. doi: 10.1037/0096-1523.18.3.849

[57] J. S. Reitman. Without surreptitious rehearsal, information in short-term memory decay. 1974.

[58] R. A. Rensink. Change blindness: Implications for the nature of visual attention. In *Vision and attention*, pp. 169–188. Springer, 2001.

[59] R. A. Rensink. Change detection. *Annual review of psychology*, 53(1):245–277, 2002.

[60] R. A. Rensink. Visualization and Human Vision: A Tale of Two Systems. In *2014 Second IEEE Working Conference on Software Visualization*, pp. xv–xv. IEEE, Victoria, BC, Canada, Sept. 2014. doi: 10.1109/VISSOFT.2014.36

[61] R. A. Rensink. The nature of correlation perception in scatterplots. *Psychonomic Bulletin & Review*, 24(3):776–797, June 2017. doi: 10.3758/s13423-016-1174-7

[62] R. A. Rensink and G. Baldridge. The Perception of Correlation in Scatterplots. *Computer Graphics Forum*, 29(3):1203–1210, Aug. 2010. doi: 10.1111/j.1467-8659.2009.01694.x

[63] J. N. Rouder, R. D. Morey, C. C. Morey, and N. Cowan. How to measure working memory capacity in the change detection paradigm. *Psychonomic bulletin & review*, 18(2):324–330, 2011.

[64] K. B. Schloss, C. C. Gramazio, A. T. Silverman, M. L. Parker, and A. S. Wang. Mapping Color to Meaning in Colormap Data Visualizations. 25(1):810–819, 2019.

[65] B. Shneiderman. *The new ABCs of research: Achieving breakthrough collaborations*. Oxford University Press, 2016.

[66] C. R. Sims, R. A. Jacobs, and D. C. Knill. An ideal observer analysis of visual working memory. *Psychological Review*, 119(4):807–830, 2012. doi: 10.1037/a0029856

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

[68] H. Stanislaw and N. Todorov. Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, 31(1):137–149, Mar. 1999. doi: 10.3758/BF03207704

[69] D. A. Szafir. Modeling Color Difference for Visualization Design. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):392–401, Jan. 2018. doi: 10.1109/TVCG.2017.2744359

[70] 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, Mar. 2016. doi: 10.1167/16.5.11

[71] J. Talbot, J. Gerth, and P. Hanrahan. An empirical model of slope ratio comparisons. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2613–2620, 2012.

[72] R. Teghtsoonian. On the exponents in stevens' law and the constant in ekman's law. 1971.

[73] B. Tversky, J. B. Morrison, and M. Betrancourt. Animation: can it facilitate? *International journal of human-computer studies*, 57(4):247–262, 2002.

[74] R. van den Berg, H. Shin, W.-C. Chou, R. George, and W. J. Ma. Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences of the United States of America*, 109(22):8780–8785, May 2012. doi: 10.1073/pnas.1117465109

[75] R. Veras and C. Collins. Saliency deficit and motion outlier detection in animated scatterplots. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2019.

[76] D. Whitney and A. Yamanashi Leib. Ensemble Perception. *Annual Review of Psychology*, 69(1):105–129, 2018. doi: 10.1146/annurev-psych-010416-044232

[77] J. T. Wixted and S. L. Thompson-Schill. *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience, Language and Thought*, vol. 3. John Wiley & Sons, 2018.

[78] J. Wolfe and T. S. Horowitz. Visual search. *Scholarpedia*, 3(7):3325, 2008.

[79] J. M. Wolfe. Guided Search 2.0 A revised model of visual search. *Psychonomic Bulletin & Review*, 1(2):202–238, June 1994. doi: 10.3758/BF03200774

[80] J. M. Wolfe and T. S. Horowitz. Five factors that guide attention in visual search. *Nature Human Behaviour*, 1(3):0058, Mar. 2017. doi: 10.1038/s41562-017-0058

[81] J. M. Wolfe, T. S. Horowitz, M. J. Van Wert, N. M. Kenner, S. S. Place, and N. Kibbi. Low target prevalence is a stubborn source of errors in visual search tasks. *Journal of experimental psychology: General*, 136(4):623, 2007.

[82] C. Xiong, C. R. Ceja, C. J. Ludwig, and S. Franconeri. Biased Average Position Estimates in Line and Bar Graphs: Underestimation, Overestimation, and Perceptual Pull. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):301–310, 2019.

[83] J. Zimmerman, J. Forlizzi, and S. Evenson. Research through design as a method for interaction design research in hci. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 493–502, 2007.