

Project Description

1 Overview

Statistical graphics and models are powerful tools to summarize data and support human decision making; however, empirical research on graphical perception is sparse relative to the number of decisions necessary to make a good chart. When relevant studies are available, they often use incomparable methods and produce conflicting results. Chart design guidelines are often based on opinion, not empirical study, rendering many scientific communications sub-optimal or ineffective. This is alarming: effective science communication is critical for cultivating public trust in the scientific process and ensuring that decision makers accurately interpret supporting information. Addressing these challenges, my long-term career goal is to examine statistical graphics with the goal of *helping people use data more effectively*, and to apply this research to educate and inspire a new generation of scientists while supporting science literacy among the general public.

This CAREER proposal addresses a fundamental research question underpinning this problem: *How do design decisions impact the use, design, and perception of data visualizations?* Three research objectives support this goal:

- **RO1:** Create a framework for comprehensive graphical testing across multiple levels of user engagement.
- **RO2:** Assess the impact of measurement methods on experiments evaluating statistical graphics.
- **RO3:** Empirically validate common chart design guidelines, measuring the impact of design decisions on task performance.

Integrated with these research efforts, the overall **education goal** is to leverage visualization research to motivate statistical learning and improve data-driven decision making in society. Three education objectives (EOs) address this goal:

- **EO1:** Develop and implement experiential learning activities in graphics for undergraduate introductory statistics courses.
- **EO2:** Create graduate course modules for K-12 educators that connect ongoing research to engaging, hands-on classroom activities for teaching statistics, math, and science.
- **EO3:** Improve the penetration of visualization research beyond academia by incorporating summaries of empirical studies in resources used by data scientists, industry analysts, and researchers in STEM disciplines.

Experiential learning activities will connect graphics research to critical concepts within statistics courses at the undergraduate level as well as in K-12 activities provided during graduate coursework for STEM educators. In addition, incorporating research summaries into general visualization resources will not only connect data visualization creators with research; improving these resources will improve teaching materials for statistical computing and will involve undergraduates in research and outreach in graphics and science communication.

The research and educational activities described in this project have the potential to significantly improve how scientists communicate scientific results to each other as well as to the general public, increasing public trust in science and facilitating public decision making based on experimental data and results.

2 Intellectual Merit

This work will expand our understanding of graphical perception and communication by empirically and systematically examining chart design through comprehensive, task-based testing. The proposed studies will be used to generate a framework relating evaluation methods to user engagement with graphics, establish the impact of different experimental design decisions on results, and promote integration of multiple evaluation methods to provide a holistic assessment of visualization effectiveness. Additionally, this project will prioritize inclusion neurodiverse and disabled individuals, ensuring that design guidelines account for accessibility concerns. The results of the systematic examination of different experimental design and testing methods will not only ground design guidelines in empirical results; if successful, the experiments will also help reconcile the results from historical studies with conflicting results. While there are task-based taxonomies for *selection of chart types*, a systematic framework for selecting *testing methods* based on levels of engagement and critical tasks is innovative; we expect that this framework will facilitate well-rounded experiments that examine chart design and use from multiple perspectives, providing nuanced results focused on audience use of graphics. The education activities proposed in this project are closely tied to the research objectives, providing avenues for dissemination of research results as well as inclusion of audiences in graphics research. As a result, education and research activities will combine to support new pedagogical research in experiential learning. This new research will examine the use of statistical graphics as an entry-point to quantitative subjects for individuals who are not traditionally interested in pursuing STEM careers. Previous collaborative research projects have established new and re-imagined old methods for testing statistical graphics; when combined with training and experience in statistics at the intersection of computer science, psychology, and communication, I am well equipped to complete this project supported by collaborations with researchers in cognitive psychology and statistical education.

My long-term career goal is to examine statistical graphics with the goal of *helping people use data more effectively*, and to apply this research to educate and inspire a new generation of scientists while supporting science literacy among the general public.

3 Research Plan

3.1 Overview

Scientific graphics transform quantitative data into image representations that can make use of the human visual system, leveraging our ability to take in and process huge quantities of information with minimal cognitive effort. However, unlike many mathematical data transformations, the transformation to visual space incurs loss both in the rendering of data to image and the transition from image to cognitive representation. That is, when creating data visualizations, we have to be concerned not only with the accuracy of the rendered image, but also with how that image is perceived by the viewer. It is easy to find entire books filled with situations in which the transition from data to image produces results which are misleading [1–3]; identifying scenarios where the transition from image to cognitive representation is suboptimal is more challenging and requires user studies. There have been empirical studies of graphics for at least 100 years [4–6], but the foundational work in graphical perception is [7], which established viewer’s ability to accurately estimate information from simple visual displays. While this work is important, and valuable, it has been synthesized into recommendations and rankings which go far beyond the original experiments [8, 9] with limited empirical verification, though in many cases these extrapolations are based in part on cognitive and perceptual research that is not specific to scientific visualization. It is easy to forget that [7] examined charts with respect to the direct numerical accuracy of quantitative estimates; the results do not necessarily apply if we are interested instead in determining whether differences between quantities can be perceived [10, 11], ordered, remembered [12], or used to reach a reasonable real-world decision [13]. While each of these alternate tasks

has been addressed in user studies of graphics, because the design space of visualization user studies is so large [14–16] and the literature is spread across so many different fields (including psychology, computer science, statistics, design, and communication) with different preferred methods, it is extremely difficult to synthesize the total graphics literature in order to derive empirically driven guidelines for creating graphs that accurately transform the data into an image and also present the data in a form which can be effectively used by the intended audience.

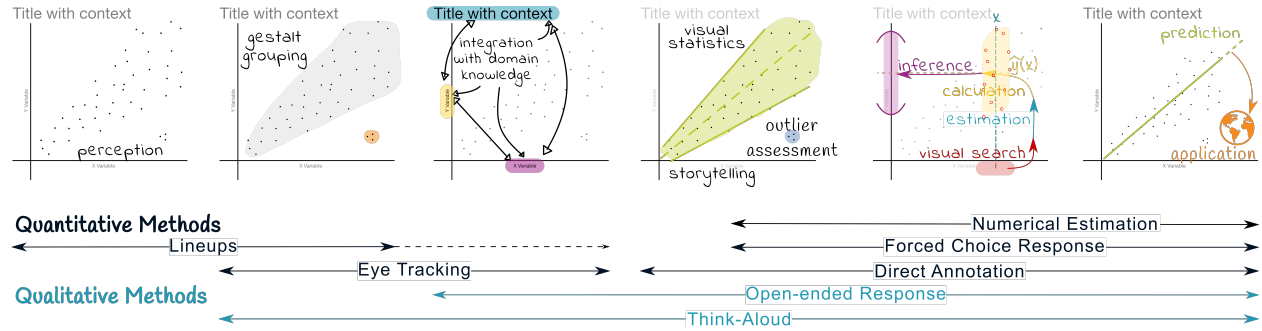


Figure 1: Levels of cognitive engagement with charts, roughly ordered by complexity, time, and effort. Methods which effectively measure (or could be extended to measure) each stage are shown below the charts. Text annotations show examples of the types of operations which involved in each stage.

The research objectives proposed here are designed to lay a foundation for evaluation and testing of scientific visualizations across multiple levels of user engagement. We focus on the integrated cognitive complexity and temporal evolution of user-chart interaction, which is roughly illustrated in Figure 1. Previous hierarchies have focused on the complexity of single graphical tasks [7, 17, 18]; while this is a useful way to determine which chart to use to display data, it does not approach different ways users engage with a single chart: are they perceiving the graphical forms without engaging with the underlying symbolic meaning? Using the chart to understand the underlying natural phenomenon? Doing statistical inference (e.g. visually estimating model parameters from the graph)? Making decisions based on their understanding of the data? Each of these use cases involves different cognitive tasks, and as a result, different graphical testing methods should be used to assess the effectiveness of charts under each type of engagement.

We will first identify and evaluate methods for graphical testing across multiple levels of user engagement, comparing methods which examine equivalent stages of graph comprehension and use. @fig-cognition-hierarchy shows some of the methods we intend to assess and compare, along with the rough stages of cognition these methods target. Next, we will establish the impact of different experimental configurations and ways of measuring and recording users’ answers. We expect that this will not only help graphics researchers design and implement new user studies, but we hope to also facilitate comparison of results from past studies, providing context to conflicting conclusions. Finally, we will empirically validate common chart design guidelines, testing whether extrapolated results and aesthetic opinions hold up under critical user studies.

The results from these objectives, taken together, are intended to build a user-focused foundation for measuring and assessing the design and use of data visualizations. The choice to approach testing graphics from the perspective of how the user interacts with and makes decisions based on the visual representation of the data places this research firmly at the intersection of statistics, cognitive science, measurement, and communication. While previous researchers [7, 17, 19] have assessed graphics from the perspective of different estimation or user interaction tasks, the present project is focused on integrating **measurement** methods for different stages of user interaction with graphics. Thus, this project will develop methodology for measur-

ing the functional cognition underlying data driven decision making using visual aids. The results from the proposed research will also allow integration of conflicting historical results, hopefully leading to a robust set of empirical evidence that can be integrated to produce more robust, task-focused design guidelines for statistical graphics.

3.2 Motivation

The first studies experimentally examining the effectiveness of statistical graphics took place approximately 100 years ago; since then, the quantity of charts created, the methods available for creating charts, and the technology available for measuring and evaluating comprehension have evolved in remarkable ways. [20] provides a comprehensive review of studies that experimentally examine the use of statistical graphics as well as the underlying research in cognitive psychology topics such as perception, memory, attention, and executive function which influence our ability to use statistical graphics effectively.

What is remarkable given the ubiquity of statistical graphics in scientific communication is that even after 100 years of empirical graphics research, we still have relatively little empirical evidence to support of some common design guidelines and heuristics; where there are empirical studies, they often conflict or have been over-extrapolated from the design and goal of the original experiments. For example, Tufte’s data-ink ratio [21] has been thoroughly tested [17, 18, 22–24], but results have been decidedly mixed, suggesting that the data-ink ratio is too simplistic; even so, it is still part of the common vernacular and makes its way into many different design guidelines [25]. Another common recommendation is to locate the most important variables along position axes (e.g. x and y in a scatterplot) rather than encoding quantitative information in color; this is because [7] found higher levels of accuracy in these comparisons, but accuracy of numerical estimation is not the only important way people use charts [26], and in fact, it is relatively uncommon for individuals to directly estimate one specific numerical quantity from a chart: for these tasks, a table would be much more appropriate [27].

At a fundamental level, we know that graphics are useful for communicating scientific results and for exploring our data; whether the target audience is ourselves, peers, or the general public, graphics are an invaluable tool. So why do we assess graphics based on things like estimation accuracy or response time [28], and then extrapolate the results to tasks and situations that don’t revolve around estimation accuracy? What is needed instead is a testing framework focused on the user’s level of interaction and purpose for interacting with a chart. [29] divides evaluation scenarios into several user-focused task-based methods for both visualization and data analysis, assessing the utility of several methods for testing these empirically, but stops short of actually performing experiments evaluating the same graphics using multiple different methods. This component of the proposed work is essential because it provides multiple points of experimental control that are not present when aggregating results across experiments: it is possible to keep the same participants, data (or data generating model), and testing conditions across multiple testing methods. In this work, we propose a comprehensive, multi-modal experimental framework for evaluating graphics. This will provide a better alternative to the patchwork testing of individual questions with highly specific methods by empirically assessing how specific charts (or design decisions) function under different tasks and measurement methods.

There are multiple factors that must be considered and evaluated in order to achieve the broader goal of empirically testing design guidelines: the measurement methods and variables used to assess charts are of obvious interest, but other factors are also important. Measurement of numerical information that has passed through the human brain in one form or another can be complicated by the method used to obtain and record the information. Consider the relatively simple case where a participant is asked to estimate the length of a specified bar in a bar chart: the experimenter must determine how this estimate is recorded. Modern web design (assuming our experiment is conducted online or at least that data is entered via a computer interface)

provides multiple options: the user can directly enter a number in a text box or indicate the number on a slider (with or without anchor points); the former requires translation into an explicitly numerical domain, where the latter requires that the participant map the chart onto a spatial domain but does not necessarily require explicit formation of a numerical estimate. Direct entry is subject to rounding effects that increase with participant uncertainty [30, 31]; while these effects can be mitigated [32] through modeling, it might be preferable to make use of numerical inputs that might not trigger rounding, such as slider inputs. Unfortunately, slider inputs are not entirely simple either: they can contain anchor points (or not) that participants may latch on to; the inclusion of these additional annotations may reduce cognitive load, but may provide the opportunity for additional anchoring effects that must be considered and possibly modeled. Most research in this area has examined sliders as inputs for categorical variables[33–37] and suggests that using sliders instead of radio button inputs changes the observed distribution of responses in important ways; while the comparison to radio buttons is not relevant to continuous data, the results of these studies suggest that there is a need to explicitly examine the effects of input methods on participant responses both in the context of visualization evaluation studies and more broadly. This is just one example of the series of decisions experimenters make about the process of elucidating and recording data from participants which do not directly relate to the hypotheses under investigation but that may well impact the results.

Combining the toolbox of methods for testing graphics at different levels of user engagement and the assessment of measurement details that impact research in statistical graphics but are not directly of interest during most graphics experiments, we have a better foundation through which to address the fundamental motivation for this research: **using comprehensive empirical testing to validate common design guidelines**. Many books and papers provide design guidelines along with examples, redesigns, and sometimes, supporting references to empirical studies [19, 21, 29, 38–50]; [51] summarizes the structures and types of guidelines in many of these sources. There have also been empirical assessments of broad themes common to different sets of guidelines: [25] experimentally evaluated two themes (“declutter” and “focus”) using several different assessment methods, finding that focused designs were preferred over decluttered designs, which were preferred over cluttered designs. In addition, we must ensure that our design guidelines are inclusive. We know that expertise as well as disorders such as dyslexia, dyscalculia, and ADHD affect perception, numeracy, and other processes involved in graph comprehension [52–54]. Designers already consider audience and accessibility [55] but have little empirical support assessing graph design choices in these populations.

At a fundamental level, we have a lot to learn about visualization design: the design guidelines that we promote as a discipline are built on fairly limited studies that measure accuracy or response time, instead of examining the multiple different levels at which a user might engage with the chart and the underlying data. We have not sufficiently examined how groups with processing disorders and cognitive differences are affected by our design guidelines; when we consider accessibility, much of the time this is limited to discussions of colorblindness. This project is designed to build a foundation for the next generation of empirical graphical testing by developing a robust set of measurement methods, assessing the impact of different experimental design factors, and leveraging this foundation to examine design guidelines experimentally and inclusively.

3.3 Preliminary Studies

In previous work [56], we have seen that simultaneously collecting quantitative and qualitative data provides the opportunity to gain rich and nuanced insight into how participants respond to graphical tests: a significant proportion of participants in the visual hypothesis test committed a Type III error (the right answer to the wrong question) [57].

In a more recent series of studies, we expanded this approach, examining the use of log and linear scales to assess exponential time series data across multiple different user tasks: perception, estimation, and prediction. This series of studies, inspired by the COVID pandemic and the lack of empirical research available at that time assessing the effectiveness of log scales, used three different graphical testing methods: statistical lineups, which test whether users can perceive a difference, direct numerical estimation, which assessed whether users could read data off of a chart and use it to perform estimation tasks, and “you-draw-it”, which explored whether users can predict exponential growth. The “you-draw-it” task is a modernized form of hand-drawn regression lines [58] and one example of a direct-annotation method which can be used to provide quantitative information and predictions without requiring participants to convert graphical information to a numerical, real-world domain. We ran this three-part experiment on the same set of participants, and are in the process of publishing the results [59, 60], though initial results from each part of the experiment were published as dissertation chapters [61]. Most empirical visualization studies only use one testing method to assess a design decision, but graphics are *used* for many different purposes; it is important that we test graphics comprehensively, so that empirical guidelines that are appropriate for many different levels of user interaction can be developed.

One challenging part of the estimation task in this series of studies was how to phrase the estimation questions and record participants’ responses. We asked participants to answer five different types of questions requiring estimation of quantities off of an exponentially increasing time series of points. Easy questions required estimation of the conditional value of y given x (or vice versa), two intermediate questions required a calculation on either the additive or multiplicative scale, and a third intermediate question required estimating the time until the population doubled in size. In addition, participants were asked an open-ended question (“describe the data shown in this graph”) before being asked to estimate values. Figure 2 shows the calculations required for one of the intermediate difficulty questions; participant results are shown in Figure 3. In addition to requiring the numerical input value, we provided participants with a scratchpad and basic calculator applet which allowed us to see how some participants solved the problem in greater detail, providing ways to assess logical and estimation errors for a subset of participants who used the scratchpad; methods for integrating the analysis of this additional layer of user data with the direct measures of accuracy are under active development.

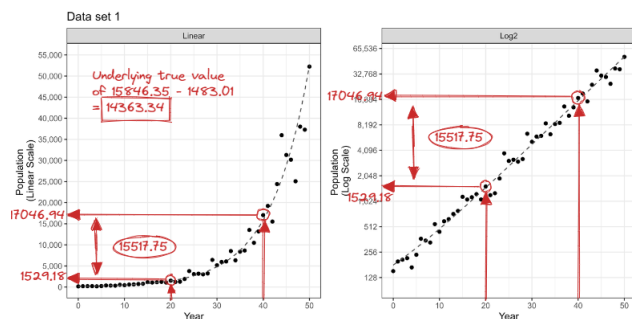


Figure 2: Steps to estimate additive population change.

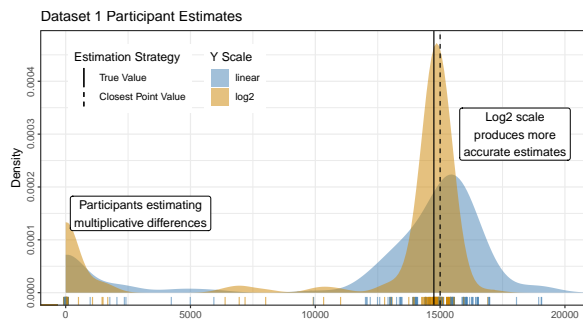


Figure 3: Distribution of participant estimates.

In another study with a different graduate student, we wanted to assess the probability of perceived guilt or innocence of a defendant based on the type of testimony presented by a forensic examiner. Estimation of very small or very large probabilities can be challenging; recording those estimates accurately and consistently may be more so given rounding effects and limitations on input options. To ensure the larger study was properly configured, we conducted a miniature study examining the different ways we could record participant input, using free response, forced binary choice, categorical and numerical input sliders, and numerical inputs for numerator and denominator that calculate a probability of guilt; results in Figure 4 show that the

results are different for different input types, with blank numerical slider inputs biasing results more towards 0.5 and ratio inputs showing evidence of rounding effects. Clearly, the input method used does have some impact on the observed results; there is no reason to expect this to be different for graphical estimation tasks than for a more subjective assessment of guilt probability.

Measures of Probability of Guilt in Jury Studies

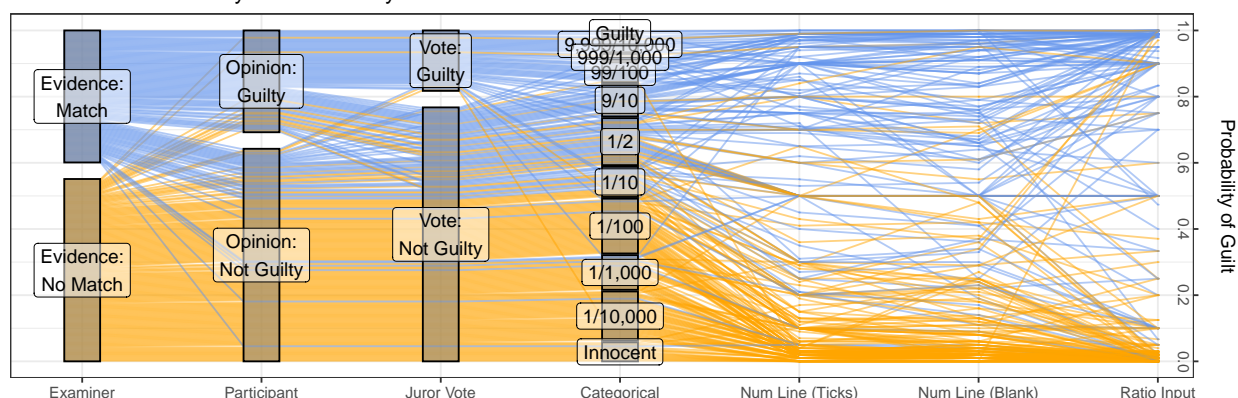


Figure 4: Assessing defendant guilt probability using different input methods.

The final set of recent preliminary data supporting the importance of this research is a study which another graduate student and I conducted to reexamine [7] in light of modern 3D graphical rendering and 3D printing technology that can provide much better 3D charts than the fixed-perspective charts used in the original 1984 study. We wondered whether physical 3D printed bar charts might be less prone to estimation errors than the fixed-angle charts used in [7]. We created charts rendered using modern 2D graphics software, 3D digital renderings [62], and 3D-printed graphics [63]; all three are shown in Figure 5. The study will assess this question using two different populations: one comparable to that used in [7] (department faculty, graduate students, and their housemates), and also in introductory statistics students who will complete the study as part of an experiential learning project discussed in Section 4. Data collection is underway on the latter population as of Summer 2023, but preliminary data is available from the former population.

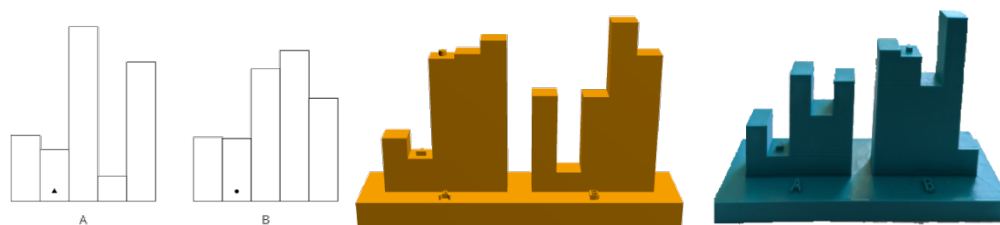


Figure 5: Three renderings of bar charts: 2D (left), 3D digital render (middle), and 3D printed (right).

Our initial investigation found few differences between 2D, 3D digital, and 3D printed charts in comparison accuracy; one explanation for this is simply the limited power of the small sample size in our pilot study, but an alternate explanation is that the 3D virtual rendering environment is not really comparable to the fixed-angle 3D plots used in the original study. To assess this possibility, we are expanding the study to assess an additional 3D fixed-angle projection condition that does not allow for realistic manipulation, which is more similar to the 3D renderings used in the original study. Misapplied depth perception has been implicated in other graphical mis-perceptions [64, 65], and it is entirely possible that 3D charts that are interactive can be accurately perceived while artificial 3D fixed-angle projections into 2D space lead to inaccurate

perceptions. If this is the case, the guidelines to avoid 3D graphics may be entirely misguided given the interactive rendering environments available today.

While this study’s results are primarily relevant to the third aim of this research proposal, the supporting literature, which contains conflicting results, also contains conflicting estimation procedures, in addition to differing on input data and underlying population of interest. [18] had participants “position the cursor [along a number line] so that the horizontal line is divided in proportion to the apparent sizes of the elements”, which is essentially estimating the proportion $A/(A + B)$, while [7] asked participants to “judge what percentage the smaller was of the larger”, which is A/B , using a numerical estimate (rather than a slider). This difference in input methodology and estimation quantity may explain the conflicting findings between the two papers; more importantly, this is a factor that must be investigated both to assist with the interpretation of past studies and to inform the design of future studies.

3.4 Methods

This project is designed to lay a foundation for robust experimental evaluation of statistical graphics: we will examine graphical evaluations that can assess common ways charts are used in practice, simultaneously developing and validating methods focused on practical evaluation and providing empirical support for nuanced, user-focused design guidelines. In support of this goal, we will first compare the insights from testing methods which address different levels of user engagement, developing toolkits for implementing empirical studies of graphics and assessing which methods can be combined to produce a holistic assessment of how a chart is used to support decision-making. These experiments are described in Section 3.4.1.

As many different smaller factors, such as measurement and recording methods, can have an outsized influence on the results of graphical testing experiments, we will also conduct a thorough comparison of the effects of these decisions by revisiting previous studies and manipulating the measurement methods. If successful, this will provide contextual information which we can use to reconcile conflicting results from historical studies with slightly different methods. Even if this portion of the project does not produce ideal results, we will still gain greater insight into the ideal design of inputs for user testing, which will facilitate better study design in the future. Experiments relating to this second research objective are described in Section 3.4.2.

Finally, we will leverage the foundation of multi-modal user testing and better understanding of inputs to empirically assess common design guidelines at multiple levels of user engagement with statistical graphics. This process will not only include assessment of graphics using undergraduate populations or internet surveys, but will also include specific assessment of the accessibility of graphics for those who require disability accommodations due to neurodivergence, visual impairment, or learning disabilities. The results of these studies will directly tie into outreach activities that will inform data visualization practitioners about best practices based on empirical results. The experiments which will contribute to the third research objective are described in [?@sec-methods-guidelines](#).

While these goals are related, they are not dependent: there are already sufficient methods in the literature for testing graphics to allow us to complete a task-focused evaluation of design guidelines, and while a critical examination of methods for numerical input in graphics studies will be useful, it is not essential in order to empirically assess design guidelines with a focus on accessibility and neurodiversity in addition to normative visual cognition.

3.4.1 Multimodal Task-Based Testing Framework The first research objective for this project is to create a framework for comprehensive graphical testing across multiple levels of user engagement. Our overall hypothesis is that by combining multiple methods of user testing within the same experiment, we can gather information which spans multiple levels of user engagement with acceptably small impact on participant cognitive load.

To this end, we will conduct a series of experiments which incorporate multiple testing methods into empirical assessments of statistical graphics. Many methods commonly employed for testing graphics can be combined in the same experiment; think-aloud protocol are commonly combined with eye-tracking and other assessment methods to provide qualitative information about the user experience in combination with more quantitative assessments [66–68]. There are two fundamental limits to the combination of multiple methods: method incompatibility and what a single participant can reasonably be asked to do in a single experiment. For instance, statistical lineups involve multiple sub-plots, of which only one is composed of real data; this is incompatible with direct numerical estimation, because the framework for statistical inference under a randomization test necessarily removes the focus from the “real” data. We can also expect that asking participants to complete too many tasks with a single plot will result in poorer results than optimizing the methods to maximize information gain while minimizing participant effort. However, because this type of multi-method research is relatively rare (other than collection of open-ended opinions after quantitative data is recorded), we do not know where this limit is. If this aim is successful, we will be able to recommend one or more sets of measures which will produce a holistic picture of how users interact with graphics and use them to complete different tasks.

As the primary goal of these experiments is to assess the methodology, here we focus on describing the set of experiments and methodological comparisons; we will use data and graphical design comparisons from past experiments in the field. In the first set of experiments, we will focus on combinations of statistical lineups and other measurement methods and engagement levels. In general, we will conduct in-person experiments with a target participant sample size of 30-50 undergraduate psychology students; this sample size will likely adjust as the PI gains experience with eye-tracking studies and the sample size needed for statistical power using the appropriate analysis methods for eye-tracking data. Online experiments will typically be conducted using a platform such as Prolific, with a target sample size of 300 participants; we have successfully conducted previous multiple-method studies with sufficient power at this sample size. Explicit power calculations are reasonable for studies that can be evaluated with a statistical model or hypothesis test, however, with multiple testing and evaluation methods, some of which are qualitative, direct power calculations are not as useful. Instead, we rely on past experience and budget for as many participants as we can reasonably afford.

In **Experiment A1**, we will examine whether there is value in adding contextual information (axis scales, labels, titles) to lineups. Lineup studies typically do not include contextual information that would require participants to evaluate the plots using domain knowledge; instead, lineups in most studies lack axis labels and even titles [56, 69–71]; participants are encouraged to pick the plot which is the most different (which does not require understanding any data context). Experiment 1 will assess whether individuals use contextual information when deciding which plot is the most different by manipulating axis scales (which are usually controlled) as part of the experiment; we will also analyze user explanations to see if information in the plot and axis titles are referenced.

In experiments A2-3, we will establish the use of lineups with eye tracking and direct annotation (A3 only). While lineups have been used with eye tracking before [72], the measures used were not sophisticated, and eye tracking technology has improved considerably in the past decade XXX is this true? XXX. During **Experiment A2**, we will examine the effect of making lineup decisions under cognitive load, mimicking conditions where people use graphics in daily life with distractions. If successful, Experiment A2 will allow us to assess the process of decision-making and specifically identify which data features attract the most attention XXX need to talk to Michael about this XXX. **Experiment A3** will expand upon experiment A2, asking participants to directly annotate interesting features in a lineup using JavaScript-based web tools. There is the possibility that this additional task will add too much cognitive load, as well as that the additional motion required for the annotation will disrupt the eye tracking results; both of these outcomes provide useful

information. If successful, Experiment A3 will demonstrate whether there is added value from using both eye tracking (which requires in person testing) and direct annotation (which can be completed online) together. **Experiment A4** will validate the use of lineups with direct annotation, establishing if there is added value in including direct interaction with lineups in a more typical setting for visual inference studies (online). If successful, this will provide an easy way for visual inference researchers to gain additional value and insight about participants’ decisions; however, if this is not successful, we will have a better understanding of the cognitive demands of lineup evaluation. This experiment also has an additional benefit: visual inference has been suggested as one solution to the problem of overfitting during exploratory data analysis [73–75]; direct annotation could be easily integrated into analysis software in combination with automated lineup generation to provide a physical way analysts can record observations and examine those observations via hypothesis testing.

As visual inference with lineups is qualitatively different than experiments examining a single plot, experiments A5 through A9 will focus on methods for examining single plots, which allows us to test graphics in ways that directly mimic how they are used for decision support. **Experiment A5** will combine eye tracking, numerical estimation, and direct annotation: users will answer a set of questions requiring estimation of data from the chart, but the direct annotation component of the task will be varied across three levels (no annotation, annotation without numerical feedback, annotation with numerical feedback). This will allow us to assess the flow of attention during the estimation process as well as the effect that direct annotation has on the participants. Providing numerical feedback from the direct annotation will allow us to assess how much of the participant’s estimation accuracy is due to the transformation from spatial to numeric information. In previous numerical estimation studies, we found that providing a “scratchpad” and calculator produced a rich source of data that provided insight into participants’ estimation strategies [61, Ch 4]; participant annotation is a more natural method to record the same information. A small subset of participants in Experiment A5 may be asked to also think aloud as they complete the task; this will provide some preliminary information allowing us to compare eye tracking, direct annotation, and think-aloud protocols, with any interesting results explored in more depth in a follow up study. XXX May want to include a picture from estimation task showing response variability XXX

Think-aloud protocols, which ask participants to talk through the process of making a decision, have been proposed as an alternative to eye tracking for usability studies [66]; this is intriguing for experimental evaluation of graphics because think-aloud tasks can be performed through a modern web browser with minimal experimenter labor using APIs to automate transcription [76] and response coding [77]. In **Experiment A6** we will examine the overlap between direct annotation and think-aloud protocols using an online platform; automatic transcription APIs will be validated using manual transcription performed by undergraduate research assistants. If successful, this will validate think-aloud and direct annotation for use when testing chart usability online; the implementation in Shiny[78, 79] will be published in an R [80] or python [81] package to facilitate use by others in the graphics research community.

One of the major focuses of this project is exploring how people use charts to support real-life decision-making; as a result, it is important to include forced-choice questions in our battery of tests available for examining graphics. **Experiment A7** will investigate real-world decision making by examining numerical estimation, forced-choice questions, and open-ended responses, with the potential to include or substitute the use of direct annotation or think-aloud methods based on the results of previous experiments. Participants will be directly asked to make a decision based on data and real-world consequences, such as “is X product safe for consumer use” or “do levels of Y meet the threshold for regulatory action” based on a scenario and sample data. Participants will be asked to estimate a relevant numerical quantity that should inform the decision making process and then use open-ended responses to explain their reasoning on the forced-choice decision task. Follow up experiments may be used to explore the effect of uncertainty [16, 82] and other

important factors on this decision-making process, but the primary goal of this experiment is to compare results for the different measures used to assess this real-world decision-making process.

Graphics also support inferential processes in the visual domain (distinct from visual inference using lineups). **Experiments A8 and A9** will examine the process of using graphics to support visual statistical inference calculations using eye-tracking (Ex A8, in person) and direct annotation (Ex A9, online). In the case that the direct annotation protocols used to support Experiment A6 do not work, additional methods for assessing inferential processes in online usability testing may be explored in Experiment A9. Experiments A8 and A9 will also include forced-choice real-world decisions that should be supported by the inference participants are asked to complete. If successful, these experiments will demonstrate the relative benefits of using eye tracking and direct annotation to assess inferential processes supported with visualizations.

Finally, we have distinguished between visual inference and statistical inference, but the primary difference between them is that in visual inference, the null model is embedded in the lineup generation process, where in statistical inference, the null model is embedded in the scenario description and the cognitive load of inferring the graphical consequences is placed on the participant. **Experiment A10** directly compares results from these two tasks through a head-to-head comparison of lineups and graphical inference, where both tasks are observed through direct annotation or think-aloud protocols. Participants will be provided with multiple scenarios and will be given either a visual inference (lineup) or a statistical inference (single graph) task that requires evaluation of the same hypothesis. We will examine not only the power of each method in a statistical sense, but also the richness of the additional information provided through annotation or think-aloud protocols.



Figure 6: Methods used in each proposed experiment along with targeted level of engagement from Figure 1.

Figure 6 provides a high-level pictographic summary of the different methods which will be used in each proposed experiment contributing to this aim. Taken together, the results of the proposed experiments will validate these methods for observing and evaluating how users leverage graphs for decision support by examining each stage of engagement while accounting for different tasks. While we have provided limited details about data sets and types of graphs tested using these methods, we will leverage past studies extensively to construct scenarios that balance the desire to assess real-world processes with the need for experimental control.

Specific outcomes from each experiment have been briefly outlined above, however, the whole of these proposed experiments is greater than the sum of the individual experimental outcomes. If successful, the methodological developments of these 10 experiments as well as any follow-up experiments will result in an R or python package which implement Shiny modules for including direct annotation capabilities in graphical testing (x-axis estimates, y-axis estimates, drawn regression lines, and interval estimation), recording these annotations and think-aloud audio inputs as data, and transcribing audio inputs to text. Additional functions for analysis of eye-tracking data may also be included depending on the functionality currently available in the eye-tracking lab software stack. Development of this software will provide graduate and undergraduate statistics students with the opportunity to learn open-source software development practices and to contribute back to the community.

In addition to making these methods available for other experimenters through open-source software, we will be able to compare the types and quality of information gained using each method through statistical

and qualitative analyses. Each experiment will also include questions designed to assess the cognitive load of participants through user reflection questions, in order to establish any limits on concurrent measurement methodology imposed by working memory and attention resource constraints. Taken together, these experiments, even if individual experiments are not successful, will establish the limits of using multiple graphical evaluation methods in parallel when assessing charts experimentally.

As this aim is specifically designed to examine the limits of the use of multiple testing methods simultaneously in graphics studies, most experiments are set up so that success and failure are both informative. However, there are a few components of this plan which contain potential obstacles.

First, I have not previously used eye tracking methods to explore how we use graphics. While I am approaching this research space from a background primarily in Statistics, the project sits at the intersection of Human-Computer Interaction, Statistics, and Cognitive Psychology; as a result, I have enlisted Dr. Michael Dodd at UNL, an expert in eye tracking, attention, and cognition, to mentor me as I become familiar with eye tracking equipment and methodology. This collaboration will also allow me to access the undergraduate psychology participant pool, which will ensure that I can recruit students for the eye tracking studies, as these cannot be conducted over the internet.

Another obstacle which may impact the results is that the direct annotation software framework does not yet exist for many of the types of annotations required for the described experiments. Currently, direct annotations can be used to draw trend and smooth lines and extrapolate beyond provided data points [60], but additional functionality will have to be implemented in order to allow participants to highlight individual data points or regions of the plot, select positions along the x or y axis, and indicate regions for inferential purposes. This functionality exists in other interactive software [83], which should ensure that we can borrow from that implementation to create a similar interactive toolkit in Shiny. Some of the desired features are in the process of being added to the `youdrawitR` package under development through Google Summer of Code 2023; Emily Robinson and I are mentoring an undergraduate data scientist and introducing him to open-source software development. I expect that additional functionality can be added during this proposal's review cycle, but if not, the schedule allows for time to implement the necessary features before they are needed.

Finally, while there are packages for audio recording using JavaScript, I am not aware of any dedicated Shiny implementation, so we will need to write code to interface between an appropriate JavaScript library and Shiny. I have experience connecting similar JavaScript libraries to Shiny (including the JavaScript code used to implement the `youdrawitR` package under development), so this is not expected to be a significant obstacle, but in previous studies, there have been issues with browser permission conflicts causing Shiny to crash. We typically address potential issues like this during the pilot study before an experiment is officially deployed, but we will need to take special care that both the participant recruitment and the Shiny application are set up properly to ensure that we can successfully record this information.

3.4.2 Experiment Configuration Effects The second research aim of this project is to thoroughly examine the impact of experimental design factors, such as question phrasing and user input, on the results of graphical testing experiments. The history of experiments evaluating the effectiveness of graphics is filled with studies going back and forth arguing about e.g. the relative utility of pie charts and stacked bar charts [4, 5, 7, 18]. It is very easy to read these studies and conclude that all of these experimental evaluations are useless and easily manipulated (even unintentionally) by the setting of the experiment; another possibility is that the differences in these studies are due to the underlying test populations. More recently, confined to the niche of experimental evaluation of uncertainty visualization, [16] diagrammed 384 different paths taken by 82 papers through different goals, measures, input types, analysis methods, and other design decisions; while such explorations are valuable, this aim does not just seek to record the different languages of the

world; instead, our goal is to assemble a Rosetta stone by which we can compare and interpret historical studies as well as guiding the design and implementation of future visualization evaluation work. In this aim, we will examine the impact of different ways of obtaining numerical estimates from participants as well as the impact of different methods for prompting participants to provide a specific estimate.

The experiments laid out under this aim do not cover the full space of input options or ways to phrase estimation questions, however, we aim to provide details for the initial studies, with the expectation that additional follow up studies will be necessary in order to better understand the reasons for observed effects. This portion of the research is tightly integrated with the education plan; in order to conduct a thorough review of the different testing practices in the literature, I will involve undergraduate students both during the summer and during the course of the academic year. During the academic year, undergraduate researchers will examine studies that may influence design guidelines, differences in research methodology across studies, the ultimate conclusions, and track how those conclusions were interpreted when referenced in later studies, as part of [?@sec-educ-visgallery](#). These records will inspire the summer studies which will be completed in years 3-5; I have selected two critical needs for examination during years 1 and 2.

In the first two years, we will begin with a comprehensive assessment of different methods for recording numerical input. These projects are designed to be accessible to undergraduates interested in STEM or advanced high school students (hereafter, ‘new researchers’) and of limited complexity so that they can be reasonably completed over 8-10 weeks of summer. Due to the large space of different design decisions in graphics experiments, I am confident that undergraduate researchers during the academic year will uncover additional important experimental design factors to assess during years 3-5.

Experiment B1 will examine input methods for continuous estimates. Continuous numerical estimates are more complicated than one might expect: in some problems, there is a defined $[A, B]$ input range that is relevant, while in other problems, estimates may be located along the entire real line (though typically, there is a range within that where the experimenter expects most values to fall). We sometimes want participants to generate high-precision estimates, but on other occasions we need them to estimate quantities over several orders of magnitude (e.g. multiplicative or “by a factor of” estimation) where precision on a log scale is more important than on a linear scale. Providing participants with appropriate cues that indicate which characteristics apply to the problem at hand is important, but we do not want to waste valuable participant cognitive resources on understanding and manipulating the user interface. This first series of experiments will be conducted by 1-2 undergraduate summer researchers and will examine the factorial combination of input range, desired type of precision, and use of different input technologies. We will start with the assessment of segmented scales, unsegmented scales, numerical range inputs that can assess uncertainty, numerical estimates (e.g. typing in 98.7 to a text input field), and direct annotations on charts that we would expect to lessen cognitive load. The new researchers will design scenarios across the different conditions described above (defined range, whole real line, order of magnitude precision, linear precision) and develop simulated data appropriate for testing the different input measures. We will execute the designed experiment and participants will clean, process, visualize, and model the data in order to assess the relative benefits of each input method.

Experiment B2 will examine the impact of question phrasing for comparative judgments and fractional estimation, such as those in [7, 18]. Students will complete an assessment of the different studies which have examined this question and will assemble a list of commonly used methods for assessing comparative judgments. We will then work together to design and execute an appropriately controlled experiment that allows us to compare the accuracy of these methods on both a raw accuracy scale and using appropriate psychophysical models (e.g. Stephens’ law as used in [18] compared to the corrected log2 midmeans method in [7]).

At least one potential topic for subsequent years' explorations may include the effect of other cognitive load manipulations (distractions, working memory tasks, etc.) to simulate graphical estimation and decision-making under the more chaotic conditions where we typically use graphics in daily life. Experiment A2 will briefly explore this within the context of lineup studies in an eye tracker, but cognitive loading manipulations are not common in most empirical studies of real-world chart usability and comprehension.

While I have not fully described the implementation details of the two studies which are detailed here, this is primarily because I hope to involve undergraduate and advanced high school students in this research; providing them with the agency to design the studies (within reason) and determine the course of the investigation. This “scaffolded” approach [84] focuses the research process on inquiry and the quest for understanding, and students learn the skills necessary to complete the tasks because they are interested, rather than as a prerequisite to doing an interesting project.

In addition, I anticipate studies B3, B4, and B5 will be designed as outgrowth of the literature review conducted as part of [?@sec-educ-visgallery](#). These summer projects will be inspired by undergraduate research, and designed and implemented by new researchers.

While the studies outlined here are not likely to dramatically change the direction of graphical perception and user testing research over the next decade, when combined with the methodological research in the first aim and the design guideline based research discussed in the next section, this forms a critical part of the foundational research necessary to be able to critically assess historical studies while accounting for their methodological differences.

4 Education Plan

4.1 Overview

4.2 Design and methods

4.3 Evaluation

4.4 Integration of Research and Education

5 Timeline

6 Broader Impacts

7 Results from Prior NSF Support

References

1. Cairo A (2016) The truthful art: Data, charts, and maps for communication.
2. Cairo A (2019) How charts lie: Getting smarter about visual information.
3. Huff D (1954) How to lie with statistics.
4. Croxton FE, Stryker RE (1927) Bar Charts Versus Circle Diagrams. *Journal of the American Statistical Association*, 22(160):473–482. <https://doi.org/10.2307/2276829>
5. Eells WC (1926) The Relative Merits of Circles and Bars for Representing Component Parts. *Journal of the American Statistical Association*, 21(154):119–132. <https://doi.org/10.2307/2277140>
6. Wickham H (2013) Graphical criticism: Some historical notes. *Journal of Computational and Graphical Statistics*, 22(1):38–44. <https://doi.org/10.1080/10618600.2012.761140>
7. Cleveland WS, McGill R (1984) Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):531–554. <https://doi.org/10.1080/01621459.1984.10478080>
8. Mackinlay J (1986) Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics*, 5(2):110–141. <https://doi.org/10.1145/22949.22950>
9. Franconeri SL, Padilla LM, Shah P, Zacks JM, Hullman J (2021) The science of visual data communication: What works. *Psychological Science in the Public Interest*, 22(3):110–161. <https://doi.org/10.1177/15291006211051956>
10. Lu M, Lanir J, Wang C, Yao Y, Zhang W, Deussen O, Huang H (2022) Modeling Just Noticeable Differences in Charts. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):718–726. <https://doi.org/10.1109/TVCG.2021.3114874>
11. Rensink RA, Baldridge G (2010) The Perception of Correlation in Scatterplots. *Computer Graphics Forum*, 29(3):1203–1210. <https://doi.org/10.1111/j.1467-8659.2009.01694.x>
12. Borkin MA, Bylinskii Z, Kim NW, Bainbridge CM, Yeh CS, Borkin D, Pfister H, Oliva A (2016) Beyond memorability: Visualization recognition and recall. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):519–528. <https://doi.org/10.1109/TVCG.2015.2467732>
13. Keller C, Siegrist M (2009) Effect of Risk Communication Formats on Risk Perception Depending on Numeracy. *Medical Decision Making*, 29(4):483–490. <https://doi.org/10.1177/0272989x09333122>
14. Abdul-Rahman A, Chen M, Laidlaw DH (2020) A survey of variables used in empirical studies for visualization. *Foundations of data visualization*, :161–179. https://doi.org/10.1007/978-3-030-34444-3_7
15. Bolte F, Bruckner S (2020) Measures in visualization space. *Foundations of data visualization*, :39–59. https://doi.org/10.1007/978-3-030-34444-3_3
16. Hullman J, Qiao X, Correll M, Kale A, Kay M (2019) In Pursuit of Error: A Survey of Uncertainty Visualization Evaluation. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):903–913. <https://doi.org/10.1109/TVCG.2018.2864889>
17. Carswell CM (1992) Choosing specifiers: An evaluation of the basic tasks model of graphical perception. *Human factors*, 34(5):535–554. <https://doi.org/10.1177/001872089203400503>
18. Spence I (1990) Visual psychophysics of simple graphical elements. *Journal of Experimental Psychology: Human Perception and Performance*, 16:683–692. <https://doi.org/10.1037/0096-1523.16.4.683>
19. Shneiderman B (1996) The eyes have it: A task by data type taxonomy for information visualizations. *Proceedings 1996 IEEE Symposium on Visual Languages*, :336–343. <https://doi.org/10.1109/VL.1996.545307>
20. Vanderplas S, Cook D, Hofmann H (2020) Testing statistical charts: What makes a good graph? *Annual Review of Statistics and Its Application*, 7(1):61–88. <https://doi.org/10.1146/annurev-statistics-031219-041252>
21. Tufte E (2001) The visual display of quantitative information.

22. Kelly JD (1988) The Data-Ink Ratio and Accuracy of Information Derived from Newspaper Graphs: An Experimental Test of the Theory. *Visual Communication Division of the Association for Education in Journalism and Mass Communication*,
23. Gillan DJ, Richman EH (1994) Minimalism and the Syntax of Graphs. *Human Factors*, 36(4):619–644. <https://doi.org/10.1177/001872089403600405>
24. Gillan D, Sorensen D (2009) Minimalism and the Syntax of Graphs: II. Effects of Graph Backgrounds on Visual Search. *Human Factors and Ergonomics Society Annual Meeting Proceedings*, 53:1096–1100. <https://doi.org/10.1518/107118109X12524443344998>
25. Ajani K, Lee E, Xiong C, Knafllic CN, Kemper W, Franconeri S (2022) Declutter and Focus: Empirically Evaluating Design Guidelines for Effective Data Communication. *IEEE Transactions on Visualization and Computer Graphics*, 28(10):3351–3364. <https://doi.org/10.1109/TVCG.2021.3068337>
26. Bertini E, Correll M, Franconeri S (2020) Why Shouldn't All Charts Be Scatter Plots? Beyond Precision-Driven Visualizations. *2020 IEEE Visualization Conference (VIS)*, :206–210. <https://doi.org/10.1109/VIS47514.2020.00048>
27. Gelman A, Wainer H, Briggs WM, Friendly M, Kwan E, Wills G (2011) Why Tables Are Really Much Better Than Graphs [with Comments and Rejoinder]. *Journal of Computational and Graphical Statistics*, 20(1):3–40. <https://doi.org/10.1198/jcgs.2011.09166>
28. Hullman J, Adar E, Shah P (2011) Benefitting InfoVis with Visual Difficulties. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2213–2222. <https://doi.org/10.1109/TVCG.2011.175>
29. Lam H, Bertini E, Isenberg P, Plaisant C, Carpendale S (2012) Empirical Studies in Information Visualization: Seven Scenarios. *IEEE Transactions on Visualization and Computer Graphics*, 18(9):1520–1536. <https://doi.org/10.1109/TVCG.2011.279>
30. Ruud PA, Schunk D, Winter JK (2014) Uncertainty causes rounding: An experimental study. *Experimental Economics*, 17(3):391–413. <https://doi.org/10.1007/s10683-013-9374-8>
31. Honda H, Kagawa R, Shirasuna M (2022) On the round number bias and wisdom of crowds in different response formats for numerical estimation. *Scientific Reports*, 12(1):8167. <https://doi.org/10.1038/s41598-022-11900-7>
32. Wang B, Wertelecki W (2013) Density estimation for data with rounding errors. *Computational Statistics & Data Analysis*, 65:4–12. <https://doi.org/10.1016/j.csda.2012.02.016>
33. Thomas M, Kyung EJ (2019) Slider Scale or Text Box: How Response Format Shapes Responses. *Journal of Consumer Research*, 45(6):1274–1293. <https://doi.org/10.1093/jcr/ucy057>
34. Liu M, Conrad FG (2019) Where Should I Start? On Default Values for Slider Questions in Web Surveys. *Social Science Computer Review*, 37(2):248–269. <https://doi.org/10.1177/0894439318755336>
35. Funke F (2016) A Web Experiment Showing Negative Effects of Slider Scales Compared to Visual Analogue Scales and Radio Button Scales. *Social Science Computer Review*, 34(2):244–254. <https://doi.org/10.1177/0894439315575477>
36. DeCastellarnau A (2018) A classification of response scale characteristics that affect data quality: A literature review. *Quality & Quantity*, 52(4):1523–1559. <https://doi.org/10.1007/s11135-017-0533-4>
37. Couper MP, Tourangeau R, Conrad FG, Singer E (2006) Evaluating the Effectiveness of Visual Analog Scales: A Web Experiment. *Social Science Computer Review*, 24(2):227–245. <https://doi.org/10.1177/0894439305281503>
38. (1915) Joint committee on standards for graphic presentation. *Quarterly Publications of the American Statistical Association*, 14(112):790–797. <https://doi.org/10.1080/15225445.1915.10503668>
39. Brewer CA (1994) Guidelines for use of the perceptual dimensions of color for mapping and visualization. *Color hard copy and graphic arts III*, 2171:54–63. <https://doi.org/10.1117/12.175328>
40. Kosslyn SM (2006) Graph design for the eye and mind.
41. Kelleher C, Wagener T (2011) Ten guidelines for effective data visualization in scientific publications. *Environmental Modelling & Software*, 26(6):822–827. <https://doi.org/10.1016/j.envsoft.2010.12.006>

42. Craft B, Cairns P (2005) Beyond guidelines: What can we learn from the visual information seeking mantra? *Ninth International Conference on Information Visualisation (IV'05)*, :110–118. <https://doi.org/10.1109/IV.2005.28>
43. Carr DA (1999) Guidelines for Designing Information Visualization Applications. *Proceedings of the 1999 Ericsson Conference on Usability Engineering*,
44. Munzner T (2014) Visualization Analysis and Design. <https://doi.org/10.1201/b17511>
45. Munzner T (2009) A nested model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics*, 15, 6:921–928. <https://doi.org/10.1109/TVCG.2009.111>
46. Brehmer M, Munzner T (2013) A multi-level typology of abstract visualization tasks. *IEEE transactions on visualization and computer graphics*, 19(12):2376–2385. <https://doi.org/10.1109/TVCG.2013.124>
47. Card SK, Mackinlay J (1997) The structure of the information visualization design space. *Proceedings of VIZ '97: Visualization Conference, Information Visualization Symposium and Parallel Rendering Symposium*, :92–99,. <https://doi.org/10.1109/INFVIS.1997.636792>
48. Steele J, Iliinsky N (2010) Beautiful visualization: Looking at data through the eyes of experts.
49. Yau N (2013) Data points: Visualization that means something.
50. Wong DM (2010) The wall street journal guide to information graphics: The dos and don'ts of presenting data, facts, and figures.
51. Kandogan E, Lee H (2016) A Grounded Theory Study on the Language of Data Visualization Principles and Guidelines. *Electronic Imaging*, 28:1–9. <https://doi.org/10.2352/ISSN.2470-1173.2016.16.HVEI-132>
52. Cheng D, Xiao Q, Chen Q, Cui J, Zhou X (2018) Dyslexia and dyscalculia are characterized by common visual perception deficits. *Developmental Neuropsychology*, 43:497–507. <https://doi.org/10.1080/87565641.2018.1481068>
53. Chity N, Harvey J, Quadri S, Pete S, Stein S (2012) Thinking Differently. Assistive Technology as a Complement to the Learning Style of Post-Secondary Students with ADHD: Recommendations for Design Opportunities.
54. Hokken MJ, Krabbendam E, van der Zee YJ, Kooiker MJG (2023) Visual selective attention and visual search performance in children with CVI, ADHD, and Dyslexia: A scoping review. *Child Neuropsychology*, 29(3):357–390. <https://doi.org/10.1080/09297049.2022.2057940>
55. Bako HK, Liu X, Battle L, Liu Z (2023) Understanding how Designers Find and Use Data Visualization Examples. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):1048–1058. <https://doi.org/10.1109/TVCG.2022.3209490>
56. VanderPlas S, Hofmann H (2017) Clusters Beat Trend!? Testing Feature Hierarchy in Statistical Graphics. *Journal of Computational and Graphical Statistics*, 26(2):231–242. <https://doi.org/10.1080/10618600.2016.1209116>
57. Kimball AW (1957) Errors of the Third Kind in Statistical Consulting. *Journal of the American Statistical Association*, 52(278):133–142. <https://doi.org/10.1080/01621459.1957.10501374>
58. Mosteller F, Siegel AF, Trapido E, Youtz C (1981) Eye Fitting Straight Lines. *The American Statistician*, 35(3):150–152. <https://doi.org/10.1080/00031305.1981.10479335>
59. Robinson EA, Howard R, VanderPlas S (2022) Eye Fitting Straight Lines in the Modern Era. *Journal of Computational and Graphical Statistics*, 0(0):1–8. <https://doi.org/10.1080/10618600.2022.2140668>
60. Robinson EA, Howard R, VanderPlas S (2023) “You Draw It”: Implementation of Visually Fitted Trends with R2d3. *Journal of Data Science*, 21(2):281–294. <https://doi.org/10.6339/22-JDS1083>
61. Robinson EA (2022) Human Perception of Exponentially Increasing Data Displayed on a Log Scale Evaluated Through Experimental Graphics Tasks.
62. Murdoch D, Adler D (2023) Rgl: 3D visualization using OpenGL.
63. Marius Kintel (2023) OpenSCAD documentation. OpenSCAD. <https://openscad.org/documentation.html>

64. VanderPlas S, Hofmann H (2015) Signs of the sine illusion—why we need to care. *Journal of Computational and Graphical Statistics*, 24(4):1170–1190. <https://doi.org/10.1080/10618600.2014.951547>
65. Hofmann H, Vendettuoli M (2013) Common angle plots as perception-true visualizations of categorical associations. *IEEE Transactions on Visualization & Computer Graphics*, (12):2297–2305. <https://doi.org/10.1109/TVCG.2013.140>
66. Guan Z, Lee S, Cuddihy E, Ramey J (2006) The validity of the stimulated retrospective think-aloud method as measured by eye tracking. *Proceedings of the SIGCHI conference on Human Factors in computing systems - CHI '06*, :1253. <https://doi.org/10.1145/1124772.1124961>
67. Kulhavy, Pridemore, Stock (1992) Cartographic Experience and Thinking Aloud about Thematic Maps. *Cartographica*, 29(1):1–9. <https://doi.org/10.3138/H61J-VX35-J6WW-8111>
68. Ratwani RM, Trafton JG, Boehm-Davis DA (2008) Thinking graphically: Connecting vision and cognition during graph comprehension. *Journal of Experimental Psychology: Applied*, 14(1):36. <https://doi.org/10.1037/1076-898X.14.1.36>
69. Hofmann H, Follett L, Majumder M, Cook D (2012) Graphical tests for power comparison of competing designs. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2441–2448. <https://doi.org/10.1109/TVCG.2012.230>
70. Loy A, Follett L, Hofmann H (2016) Variations of Q-Q Plots: The Power of Our Eyes! *The American Statistician*, 70(2):202–214. <https://doi.org/10.1080/00031305.2015.1077728>
71. Majumder M, Hofmann H, Cook D (2013) Validation of visual statistical inference, applied to linear models. *Journal of the American Statistical Association*, 108(503):942–956. <https://doi.org/10.1080/01621459.2013.808157>
72. Yifan Zhao, Cook D, Hofmann H, Majumder M, Chowdhury NR (2013) Mind Reading: Using an Eye-Tracker to See How People are Looking at Lineups. *International Journal of Intelligent Technologies & Applied Statistics*, 6(4):393–413. <https://doi.org/10.6148/IJITAS.2013.0604.05>
73. Hullman J, Gelman A (2021) Designing for Interactive Exploratory Data Analysis Requires Theories of Graphical Inference. *Harvard Data Science Review*, <https://doi.org/10.1162/99608f92.3ab8a587>
74. Cook D, Reid N, Tanaka E (2021) The Foundation is Available for Thinking about Data Visualization Inferentially. *Harvard Data Science Review*, <https://doi.org/10.1162/99608f92.8453435d>
75. VanderPlas S (2021) Designing Graphics Requires Useful Experimental Testing Frameworks and Graphics Derived From Empirical Results. *Harvard Data Science Review*, 3(3)<https://doi.org/10.1162/99608f92.7d099fd0>
76. Hannun A, Case C, Casper J, Catanzaro B, Diamos G, Elsen E, Prenger R, Satheesh S, Sengupta S, Coates A, Ng AY (2014) Deep Speech: Scaling up end-to-end speech recognition. <https://doi.org/10.48550/arXiv.1412.5567>
77. da Silva Franco RY, Santos do Amor Divino Lima R, Monte Paixão R do, Resque dos Santos CG, Serique Meiguins B (2019) UXmood—A Sentiment Analysis and Information Visualization Tool to Support the Evaluation of Usability and User Experience. *Information*, 10(12):366. <https://doi.org/10.3390/info10120366>
78. Chang W, Cheng J, Allaire J, Sievert C, Schloerke B, Xie Y, Allen J, McPherson J, Dipert A, Borges B (2022) Shiny: Web application framework for r. <https://CRAN.R-project.org/package=shiny>
79. Posit PBC (2023) Shiny for Python. *Shiny for Python*,
80. R Core Team (2023) R: A language and environment for statistical computing. <https://www.R-project.org/>
81. Van Rossum G, Drake FL (2009) Python 3 reference manual.
82. Hofman JM, Goldstein DG, Hullman J (2020) How Visualizing Inferential Uncertainty Can Mislead Readers About Treatment Effects in Scientific Results. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, :1–12. <https://doi.org/10.1145/3313831.3376454>
83. Sievert C (2020) Interactive web-based data visualization with r, plotly, and shiny. <https://plotly-r.com>

84. Perrella A, Dam H, Martin L, MacLachlan JC, Fenton N (2020) Between Culture and Curricula: Exploring Student and Faculty Experiences of Undergraduate Research and Inquiry. *Teaching & Learning Inquiry*, 8(2):90–113. <https://doi.org/10.20343/teachlearningqu.8.2.7>