Project Description

1 Overview

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.

Rationale and Critical Need: 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 Cleveland & Mcgill [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]. The design space of visualization user studies is incredibly large , and studies may use different numerical measures to address the same basic question. 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, 15] and the literature is spread across so many different fields (including psychology, computer science, statistics, design, and communication) with different standard 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. Such efforts are essential for promoting effective science communication, cultivating public trust in the scientific process, and ensuring that decision makers accurately interpret supporting information.

Goals and Objectives: To that end, the **overall research goal** of this CAREER proposal is to address the fundamental research question underpinning this problem: *How do design decisions impact the use, design, and perception of data visualizations?* Three research objectives (ROs) 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.

We focus our investigation on user engagement represented by the integrated cognitive complexity and temporal evolution of the user-chart interaction, which is roughly illustrated in Figure 1. Previous hierarchies have focused on the complexity of single graphical tasks [7, 16, 17]; 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 parameter values 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 must be used to assess the effectiveness of charts under each type of engagement.

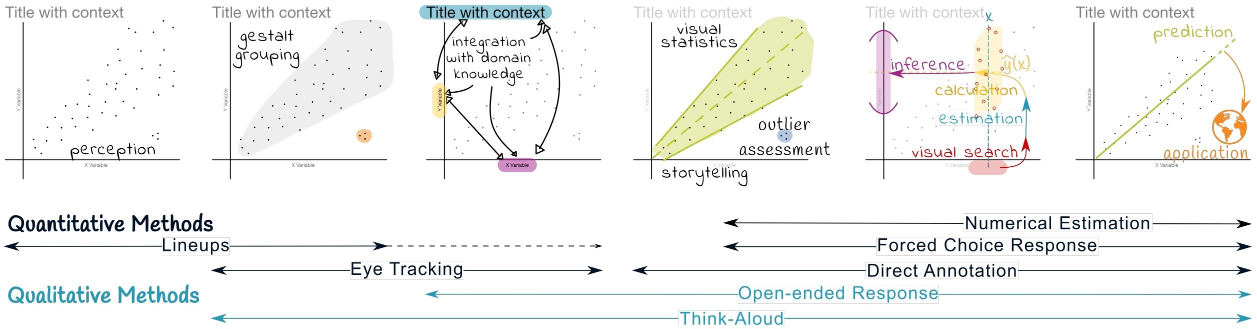

Figure 1: Levels of cognitive engagement with charts, roughly ordered by complexity, time, and effort. Methods which effectively measure (or could be extended to measure) each stage are shown below the charts. Text annotations show examples of the types of operations which involved in each stage.

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.

Expected Impact: Taken together, these objectives are intended to build a user-focused foundation for for measuring and assessing the design and use of data visualizations. The choice to approach testing graphics from the perspective of how the user is 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 scientific communication. In addition, the focus on multiple simultaneous measurement methods within an experiment separates this project from almost all previous graphical research studies, which typically use one evaluation method per experiment. This project will develop methodology for measuring the functional cognition underlying data driven decision making using visual aids. The results from this research will facilitate integration of conflicting historical results, which will contribute to a robust set of empirical evidence that can be leveraged to produce more nuanced, task focused design guidelines for statistical graphics.

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, but improving these resources will also improve teaching materials for statistical computing and will involve undergraduates in research and outreach in graphics and science communication. Ultimately, proposed activities 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.

Relation to the PI's Career Trajectory and Department Goals: This project is a natural evolution of the PI's work to date, repurposing methods and approaches from psychology and psychophysics, integrating them with modern web-based toolkits, and leveraging these new tools to study the perception of statistical graphics. The objectives in this proposal will support the UNL statistics department's new undergraduate programs in statistics and data science, and will contribute to wider university objectives to integrate experiential learning into undergraduate courses and provide undergraduates with research opportunities.

2 Background

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. [18] 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.

Narrow Empirical Support, Broad Guidelines 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 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 [19] has been thoroughly tested [16, 17, 20–22], 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 [23]. 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 [24]. 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 [25].

Need for Integrated Testing Methods 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 solely on measures like estimation accuracy or response time [26], and then extrapolate the results to tasks and situations that do not revolve around estimation accuracy or speed? What is needed instead is a testing framework focused on the user's level of interaction and purpose for interacting with a chart. Lam et al. [27] 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 experimental control that is not present when aggregating results across experiments: the same participants, data (or data generating model), and testing conditions can be used 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.

Need to Understand Input Method Choice Impacts There are multiple factors that must be considered and evaluated 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 UI design toolkits provide 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 require explicit formation of a numerical estimate. Direct entry is subject to rounding effects that increase with participant uncertainty [28, 29]; while these effects can be mitigated [30] through modeling, it might be preferable to use a continuous slider input, which might not trigger rounding. 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[31–35] 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 when eliciting and recording data from participants that do not directly relate to the hypotheses under investigation but which may well impact the results.

Validating a toolbox of methods for testing graphics at different levels of user engagement and assessing the impact of measurement decisions will provide a better foundation through which to address the fundamental goal of 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, 27, 36–49]; [50] 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: [23] 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. What is lacking is a series of tests of design guidelines across the different levels of user engagement; each specific experiment referenced above examined one type of user engagement using one measurement method. Another major gap in the existing research is an assessment of how well different guidelines serve different groups of individuals. We know that disorders such as dyslexia, dyscalculia, and ADHD affect perception, numeracy, and other processes involved in graph comprehension [51–53]. Designers already consider audience and accessibility [54] but have little empirical support assessing graph design choices in these populations. It is important that our design guidelines specifically address subpopulations in an inclusive way, so that everyone can benefit from scientific results.

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 Preliminary Studies

In previous work [55], 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 evaluating a visual hypothesis test committed a Type III error: the right answer to the wrong question [56]. This study's results inspired a desire to obtain a more nuanced insight into participants' responses which is reflected throughout more recent work.

Use of Log Scales with Exponential Data 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 [57] 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 [58, 59], though initial results from each part of the experiment were published as dissertation chapters [60]. 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.

Phrasing of Estimation Questions 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.

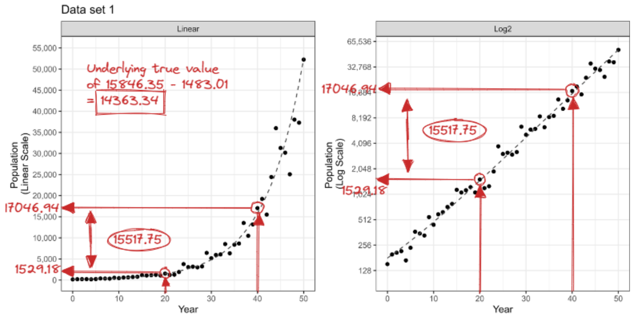

Figure 2: Steps to estimate additive population change.

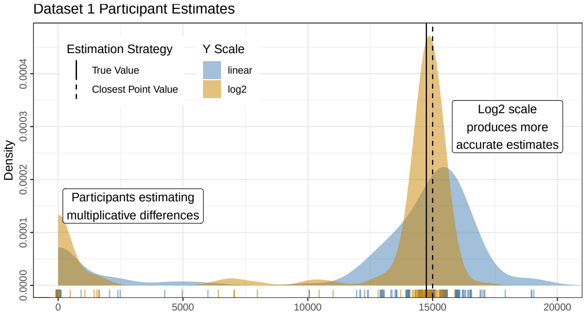

Figure 3: Distribution of participant estimates.

Manipulating Input Type In another study, we assessed the probability of perceived guilt or innocence of a defendant based on the type of testimony presented by a forensic examiner. To establish a baseline, we began with a calibration 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 has an impact on the observed results.

```r
source("code/rachel-response.R")
library(ggpcp)
library(ggplot2)
results2 |>
  filter(!is.na(logGuiltCalc), is.finite(logGuiltCalc)) |>
  pcp_select(conclusion_nice, opinion_guilt, guilty, fixed_like, prob_vis, prob_hide, guilt_
  pcp_scale() |>
  pcp_arrange() |>
  ggplot(aes_pcp(color = conclusion)) +
  theme_bw() +
  geom_pcp(alpha = 0.5) + geom_pcp_boxes(color = "black", linewidth = 0.5) + geom_pcp_labels
  scale_color_manual("Evidence", guide = 'none', values = c("NoMatch" = "orange", "Match" = 
  scale_x_discrete(labels = c("Examiner", "Participant", "Juror Vote", "Categorical", "Num L
  scale_y_continuous("Probability of Guilt", position = "right", breaks = seq(0, 1, length.o
  theme(axis.text.y.right = element_text(angle = -90, hjust = 0.5, vjust = -7), axis.title.x
  ggtitle("Measures of Probability of Guilt in Jury Studies")
```

```
Scale for x is already present.
Adding another scale for x, which will replace the existing scale.
```

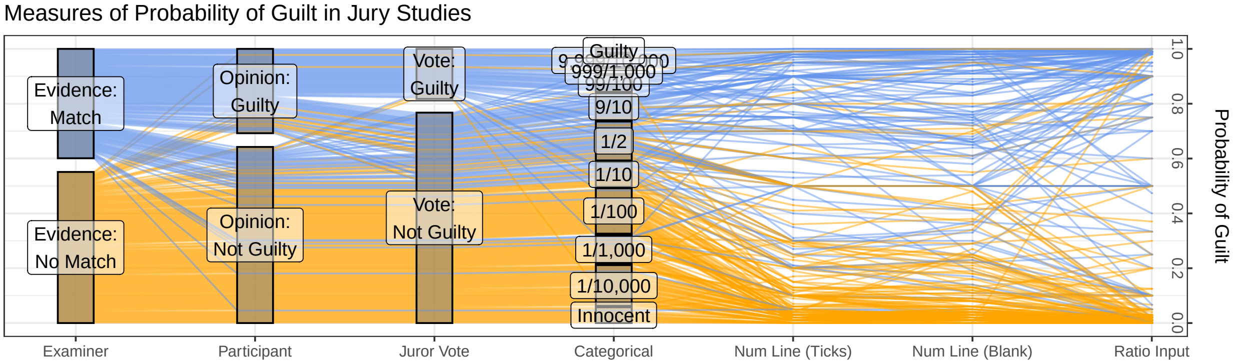

Figure 4: Assessing defendant guilt probability using different input methods.

Examining 2D vs. 3D Design Guidelines The final set of recent preliminary data supporting the importance of this research is a study reexamining [7] in light of modern 3D graphical rendering and 3D printing technology. 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 shown in Figure 5 rendered using modern 2D graphics

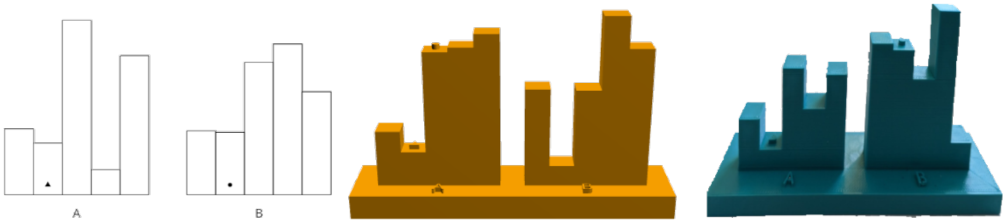

Figure 5: Three renderings of bar charts: 2D (left), 3D digital render (middle), and 3D printed (right).

software, 3D digital renderings [61], and 3D-printed graphics [62]. 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 [63, 64], 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. This is another reason that it is important to revisit previous graphical studies in light of new technology, graphical software, and testing platforms.

While this study's results are primarily relevant to the third research objective of this 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. [17] 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.

4 Research Plan

This project will lay a foundation for robust experimental evaluation of statistical graphics by examining simultaneously developing and validating methods focused on practical evaluation of chart use. In addition, we will empirically evaluate design guidelines to identify for nuanced, user-focused guidelines that have empirical support and address accessibility. The results of these studies will directly tie into outreach activities that will inform data visualization practitioners about best practices based on empirical results.

While related, each of the below objectives can be completed independently from each other. The project is designed to allow results from the first two objectives to enrich our approach to the third, but there are already sufficient methods in the literature for testing graphics to allow us to complete a task-focused evaluation of design guidelines. In addition, while a critical examination of methods for numerical input in graphics studies will be useful, it is not essential to empirically assess design guidelines.

4.0.1 Research Objective 1: Create a framework for comprehensive testing across multiple levels of user engagement 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.

Method Combination Rationale 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 protocols can be combined with eye-tracking and other assessment methods to provide qualitative information about the user experience in combination with more quantitative assessments [65–67]. 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.

Participant Considerations and Sample Size As the primary goal of these experiments is to assess the measurement 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 or generate new data when necessary to push the limits of the measurement methodology. In general, we will conduct in-person experiments with a target participant sample size of 50 undergraduate 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 less useful. Compounding this problem, we do not have any idea about relevant effect sizes, and these would be expected to change with chart type and the particulars of each experiment. Instead, we rely on past experience and have planned for as many participants as possible while limiting experiment costs and data collection time; this approach has worked successfully for the preliminary studies.

Eye Tracking Metrics In experiments where eye trackers are utilized, we will assess dwell time (time spent on each fixation area), both in total and on the first run, first fixation time for each fixation area, run count, and order of interest areas, both across and between participants. As the PI gains experience with eye tracking, additional metrics may be incorporated where these metrics can reasonably be expected to provide valid insight based on the experimental design.

Augment Lineups The first collection of experiments will assess expansions and augmentations of the statistical lineup protocol. **Part A1** will examine whether there is value in adding contextual information (axis scales, labels, titles) to lineups, expanding lineups beyond basic perception and grouping. 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 [55, 68–70]; participants are encouraged to pick the plot which is the most different (which does not require understanding any data context). We will manipulate axis scales (which are usually controlled) to determine whether viewers use this information; we will also analyze user explanations to see if information in the plot and axis titles are referenced. While lineups have been used with eye tracking before [71], the technology used samples at a low rate and does not allow for collection of measures such as fixation length, time to return, and other quantities of interest. During **Part A2**, we will examine the lineup perception and decision-making process using eye tracking, with and without cognitive load, mimicking conditions where people use graphics in daily life with distractions. **Part A3** will expand upon experiment A2, asking participants to directly annotate interesting features in a lineup using JavaScript-based web tools while in an eye-tracker. 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. **Part 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).

Augmenting Single-Plot Studies As visual inference with lineups is qualitatively different than experiments examining a single plot, the second collection of experiments (A5-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. **Part 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; this is a more natural source of information than the "scratchpad" used in [60, Ch 4]. 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 [65]; this is intriguing for experimental evaluation of graphics because think-aloud tasks can be performed through a web browser with minimal experimenter labor using APIs to automate transcription [72] and response coding [73]. In **Part 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.

One major focus of this CAREER program is exploring how people use charts to support real-life decision making. **Part A7** adds forced-choice questions to the battery of methods for examining graphics, 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 [74, 75] and other important factors on this decision-making process, but the primary goal of this experiment is to compare empirical 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). **Part A8** (in-person, eye-tracking) and **Part A9** (online, direct annotation) will examine the process of using graphics to support visual statistical inference calculations; these experiments will also include forced-choice real-world decisions supported by these inference calculations.

Visual vs. Statistical Inference 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. The cognitive demands of the two inferential procedures are different: in visual inference, the participant must infer the null model; in statistical inference, the participant must work out the graphical and real-world interpretation of the model. The final planned experiment, **Part 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.

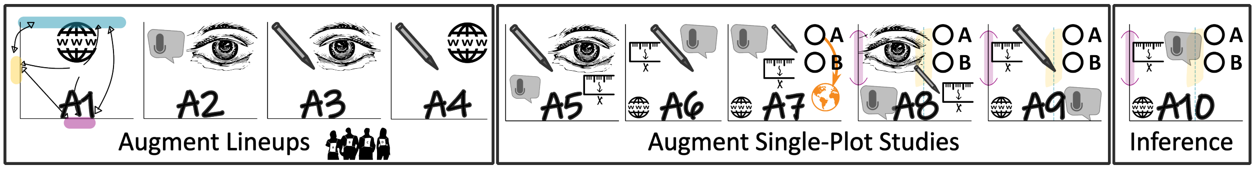

Figure 6: Methods used in each proposed experiment along with targeted level of engagement from Figure 1.

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.

Expected Results and Significance: If successful, the lineup augmentation experiments will validate lineups for assessing contextual information, leverage eye-tracking to examine attention and feature comparison during lineup evaluation, and explore use of direct annotation with lineups. Taken together, these methods will provide rich ways to gain additional information from lineup studies conducted in person or online. The single-plot augmentation experiments will allow us to assess the flow of attention during the estimation process, examine the information provided by direct annotation and think-aloud protocols, and explore the different components that contribute to estimation accuracy. We will also be able to examine the inferential process and compare the relative benefits of lineups, eye tracking, and direct annotation methods within this context. If these experiments are not all successful, however, we will still gain a better understanding of the cognitive demands of lineup and single plot evaluation, and we will be able to determine what combination of measurement methods best covers the range of user tasks. Taken together, these experiments will establish the limits of using multiple graphical evaluation methods in parallel when assessing charts experimentally.

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 and use of Shiny apps with eye trackers 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.

Potential Pitfalls and Alternative Strategies: 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. 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 (see letter). 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 that 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 [59], but additional functionality will have to be implemented 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 [76], 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. 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. In addition, if direct annotation is not effective during Part A6, additional methods for assessing inferential processes in online usability testing may be explored in Part A9.

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.

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. 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

17. 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

18. 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

19. Tufte E (2001) The visual display of quantitative information.

20. 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*,

21. Gillan DJ, Richman EH (1994) Minimalism and the Syntax of Graphs. *Human Factors*, 36(4):619–644. https://doi.org/10.1177/001872089403600405

22. 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

23. Ajani K, Lee E, Xiong C, Knaflic 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

24. 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

25. 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

26. 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

27. 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

28. 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

29. 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

30. 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

31. 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

32. 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

33. 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

34. 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

35. 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

36. (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

37. 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

38. Kosslyn SM (2006) Graph design for the eye and mind.

39. 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

40. 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

41. 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

42. Carr DA (1999) Guidelines for Designing Information Visualization Applications. *Proceedings of the 1999 Ericsson Conference on Usability Engineering*,

43. Munzner T (2014) Visualization Analysis and Design. https://doi.org/10.1201/b17511

44. 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

45. 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

46. 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

47. Steele J, Iliinsky N (2010) Beautiful visualization: Looking at data through the eyes of experts.

48. Yau N (2013) Data points: Visualization that means something.

49. Wong DM (2010) The wall street journal guide to information graphics: The dos and don'ts of presenting data, facts, and figures.

50. 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

51. 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/10.1080/875656

41.2018.1481068

52. 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.

53. 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

54. 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

55. 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

56. 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

57. 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

58. 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

59. 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

60. Robinson EA (2022) Human Perception of Exponentially Increasing Data Displayed on a Log Scale Evaluated Through Experimental Graphics Tasks.

61. Murdoch D, Adler D (2023) Rgl: 3D visualization using OpenGL.

62. Marius Kintel (2023) OpenSCAD documentation. OpenSCAD. https://openscad.org/documentation.html

63. 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

64. 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

65. 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

66. 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

67. 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

68. 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

69. 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

70. 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

71. 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

72. 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

73. 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

74. 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

75. 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

76. Sievert C (2020) Interactive web-based data visualization with r, plotly, and shiny. https://plotly-r.com