Old Wine in a New Bottle?
Analysis of Visual Lineups with Signal Detection Theory

Position paper

Sheng Long[*]
Northwestern University

Matthew Kay[†]
Northwestern University

Abstract

This position paper critically examines the graphical inference framework for evaluating visualizations using the lineup task. We present a re-analysis of lineup task data using signal detection theory, applying four Bayesian non-linear models to investigate whether color ramps with more color name variation increase false discoveries. Our study utilizes data from Reda and Szafir's previous work [20], corroborating their findings while providing additional insights into sensitivity and bias differences across colormaps and individuals. We suggest improvements to lineup study designs and explore the connections between graphical inference, signal detection theory, and statistical decision theory. Our work contributes a more perceptually grounded approach for assessing visualization effectiveness and offers a path forward for better aligning graphical inference methods with human cognition. The results have implications for the development and evaluation of visualizations, particularly for exploratory data analysis scenarios. Supplementary materials are available at `https://osf.io/xd5cj/`.

Index Terms: Graphical inference, visual evaluation, signal detection theory.

1 Introduction

How do we evaluate the strength of a "signal" in a visualization? Buja et al. first [2] proposed a framework called "graphical inference" to address this question by "furnish[ing] visual statistical methods with an inferential framework". The lineup task, also known as the "lineup protocol", is the primary method for performing graphical inference. In this task, the *target plot* (i.e., true data plot) is randomly embedded among $m-1$ *decoy plots* (i.e., plots from the "reference distribution"), and the viewer is asked to identify which plot appears the most different. Buja et al. [2] argued that in a lineup task, "[i]f the viewer chooses the plot of the real data, then the [visual] discovery can be assigned a p-value of 0.05" and that the lineup protocol provides "inferential validity". Later extensions of Buja et al.'s ideas [12, 18] used the Binomial distribution and Beta-Binomial distribution respectively to model the distribution corresponding to the "null" hypothesis that the target plot is visually indistinguishable from the decoy plots.

Beyond evaluating the significance of visual discoveries, the graphical inference framework has also been used to assess different visualization designs. By comparing the "power" of lineup tasks across various design alternatives, researchers can determine which design choices lead to better overall performance. This approach has been applied to evaluate various aspects of visualization design, such as the impact of different color schemes [20], coordinate systems [12], and feature hierarchy [24] on the ability to detect significant patterns in data.

We believe that the lineup task, when properly designed, can be considered a type of **signal detection theory** task and should be analyzed as such. We present a re-analysis of lineup task performance using data collected by Reda and Szafir [20]. We built four different Bayesian models to answer whether color ramps with more color name variation lead to more false discoveries. Our results corroborate the original findings from Reda and Szafir but also reveal additional insights into the sensitivity and bias differences across color ramps and individuals. Furthermore, we discuss potential improvements to the experimental design of lineup tasks and explore how graphical inference and signal detection theory connect to statistical decision theory.

2 Background and Related Work

2.1 Graphical inference

Graphical inference, specifically the lineup protocol, has been mainly used for two purposes – to evaluate the validity of graphical findings (i.e., inference), and to evaluate plot designs.

2.1.1 Lineup tasks for inference

Buja et al. first [2] proposed the lineup protocol to "formalize the process of visual discovery" and provide an "inferentially valid" test of whether an observed pattern in a plot is "really there". Majumder et al. [18] proposed a Binomial model for the count of correct observers and Vanderplas et al. [25] extended the Binomial model to a Beta-Binomial model. For more details on the distinction between these models, refer to Appendix A.

Modeled this way, the lineup task has been primarily used as a visual approximation of a hypothesis test to reject "null" assumptions. For example, Loy and Hofmann [16] used the lineup task in hierarchical linear models to determine whether the within-group residual variance is constant across groups. When the observed residual plot stands out, this indicates that the assumption of constant within-group variance may be violated. Kossimer et al. [14] used lineups to test for funnel plot asymmetry. They argued that if the viewer correctly identifies the real funnel plot, then it suggests that the observed data is inconsistent with the null hypothesis of no publication bias.

2.1.2 Lineup tasks for visualization evaluation

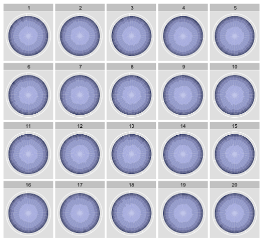

Figure 1: Sample lineup stimuli used in Hofmann et al. [12].

Hofmann et al. proposed to estimate the "power" of a lineup by calculating "the ratio of correct identifications y out of K viewings" [12]. Hofmann et al. [12] used this definition to compare different plot designs and found that Cartesian coordinates result in significantly higher accuracy and shorter response time than polar coordinates for spotting patterns in airport flight data. Vanderplas et al. [24] experimented using lineups with two target plots embedded to test for feature hierarchy.

[*] e-mail: shenglong@u.northwestern.edu

[†] e-mail: mjkay@u.northwestern.edu

They found that aesthetics that emphasized clustering (e.g., color, shape, ellipses) increased the selection of the *cluster* target. In contrast, aesthetics that emphasized continuity (e.g., trend line, error band) increased the selection of the *trend* target. Reda and Szafir [20] studied how the design of color ramps affects people's ability to make inferences from visualized scalar field data. The authors used lineups with $m = 4$ and analyzed error rates for each type of stimuli. They found that participants were "more accurate when viewing colormaps that cross a variety of uniquely nameable colors" [20].

To summarize, when applied to evaluating visualizations, the graphical inference framework primarily relies on two dependent measures – empirical accuracy and response time. We propose an alternative approach: to analyze lineups using well-established techniques from signal detection theory, which can potentially be more informative.

2.2 Signal Detection Theory

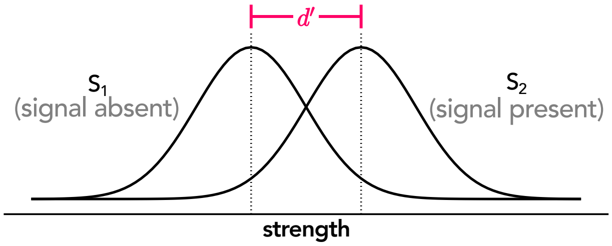

Figure 2: Model of equal variance signal detection theory. Both distributions (i.e., S_1 and S_2) are normal distributions with unit variance. Sensitivity d' is the distance between the mean of the distributions.

In a review of visualization evaluation techniques, Elliott et al. [8] describe signal detection theory (SDT) as a method for modeling visualization performance as a function of *sensitivity* and *bias*. It uses a probabilistic framework that incorporates inherent uncertainty in decision-making using *noisy* observations [9]. SDT's key assumption is that the strength of sensory and cognitive events is *continuously* variable [17]. In its most basic form, there are two possible states of the world, S_1 and S_2, which correspond to the absence or presence of the signal. Each observation is regarded as a one-dimensional random variable drawn from one of two sensory distributions [9], and errors arise because the sensory distributions overlap. This presents a problem for the observer in choosing a response. The solution is to divide the strength axis into two regions with a *criterion* so that high values lead to "yes" responses and low values lead to "no" responses (Fig. 3) [17].

The roots of SDT can be traced back to Gustave Fechner's work as early as 1860, but the term was initially conceived in a series of technical reports and publications that appeared in 1953 and 1954 [29]. Since then, SDT has inspired conceptual advances in both experimental psychology and cognitive neuroscience [29]. Its applications have expanded beyond these fields, with more recent uses including analysis of criminal lineup data [5].

In the most widely used version of the model (Fig. 2), the two evidence distributions are normal with the same unit variance [19]. The degree of overlap between S_1 and S_2 is termed "**sensitivity**" and defined as

$$d' = \Phi^{-1}(\text{Hit rate}) - \Phi^{-1}(\text{False alarm rate}) \tag{1}$$

where $\Phi^{-1} : [0,1] \to \mathbb{R}$ is the quantile function of the standard normal distribution, otherwise known as the *probit* function. **Hit rate** is defined as $\Pr(\text{"yes"} \mid \text{signal present})$ and **false alarm rate** is defined as $\Pr(\text{ "yes"} \mid \text{signal absent})$.

The location of the criterion is a measure of **response bias**, the tendency towards one response or the other [17]. There are different ways to define response bias (see Appendix B for details); for comparing two evidence distributions, we adopt the definition of **criterion location**[1] c:

$$c = -\frac{1}{2} \cdot \left(\Phi^{-1}(\text{Hit rate}) + \Phi^{-1}(\text{False Alarm Rate})\right) \tag{2}$$

The neutral point is when neither response is favored, which corresponds to $c = 0$ [23]. Negative values of c correspond to *liberal* biases with many "yes" responses, whereas positive values correspond to *conservative* biases with many "no" responses [17].

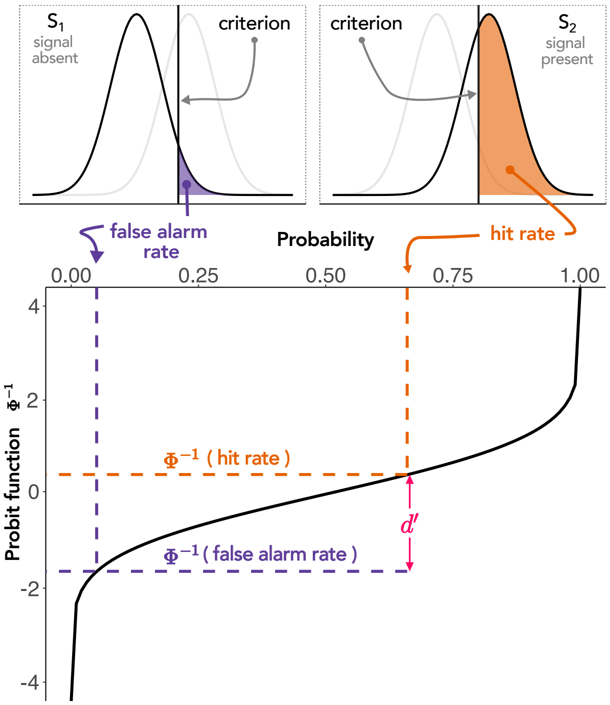

Figure 3: False alarm rate (top left) and hit rate (top right) mapped to the probit function Φ^{-1} (bottom). S_1 and S_2 are unit-variance normal distributions.

3 Analyzing Lineup task using SDT

3.1 Overview

Among the several papers that implement the lineup task, we selected Reda and Szafir [20] given its well-documented nature and readily available data found at `https://osf.io/tck2r/`. We provide a condensed account of the necessary details and encourage the interested reader to check out the original work.

In their paper, Reda and Szafir study how a cognitive metric – *color name variation* (CNV) – impacts people's ability to make model-based judgments [20]. The CNV metric is obtained by extending Heer and Stone's name distance model [11] for a pair of colors to a continuous color ramp. A higher CNV indicates a ramp that combines a variety of distinctly named colors [20]. Reda and Szafir looked at four different color ramp design families: single-hue, multi-hue, divergent, and rainbows.

Reda and Szafir conducted two experiments utilizing the lineup task [2]. Unlike the conventional approach where the lineup task

[1] The criterion location c is defined *relative to the mean of* S_1. Historically, the mean of S_1 is set to be 0, which implies that $c = -\Phi^{-1}(\text{False alarm rate})$. The current definition of c sets the mean of S_1 to be $-\frac{1}{2}d'$. The *exact location* of the mean of S_1 does not matter for computing hit rates and false alarm rates.

is typically formulated with $m = 20$, the authors opted for $m = 4$ (Fig. 4). They justified their choice of a smaller m as a means “to reduce the per-trial response time and allow for a larger number of stimuli per subject” [20].

We are particularly interested in the second experiment. Experiment 2 hypothesized that color ramps with high CNV cause people to detect false differences between visualizations of the same model [20]. In this experiment, half of the trials consisted of four plots from the same target model, and the remaining half of the trials consisted of one plot from the target model and three plots from the decoy model. This design enables us to analyze the collected responses using an SDT approach.

3.2 Data and experiment design

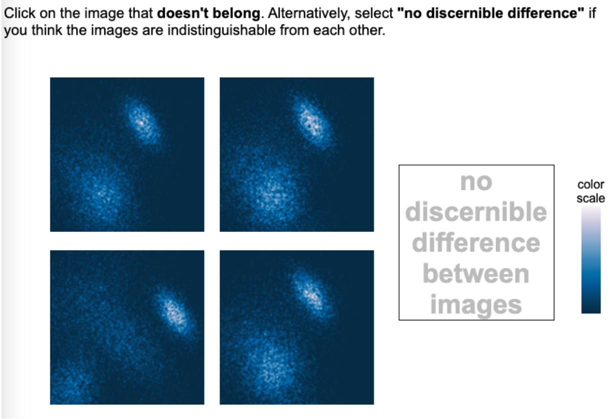
Click on the image that **doesn't belong**. Alternatively, select **"no discernible difference"** if you think the images are indistinguishable from each other.

Figure 4: Screenshot of interface for experiment 2’s training phase. The correct answer is the bottom left plot. This is a signal present (i.e., positive) stimulus.

Reda and Szafir [20] recruited 59 participants from Amazon Mechanical Turk for experiment 2. Each participant was assigned four blocks of trials, with each block corresponding to a unique color ramp. In other words, each participant experienced all four color ramps: *blues* (CNV: 1.2), *cool-warm* (CNV: 2.42), *viridis* (CNV: 2.75), and *RGB-rainbow* (CNV: 4.6). Each color ramp is from a distinct design family. Each block had 24 trials, of which 12 are *positive* trials (consisting of three decoy plots and one target plot), and the remaining 12 are *negative* trials (consisting of four plots from the target model). Data was generated by sampling from a mixture of bivariate normal distributions. The distance between the target model’s dataset and the decoy model’s data set is measured by the Kullback–Leibler (KL) divergence. A distance of 0 corresponds to a negative trial and a non-zero distance corresponds to a positive trial.

In total, each participant saw $24 \times 4 = 96$ trials and 8 engagement trials, resulting in $96 \times 59 = 5664$ lineup judgments for analysis.

Table 1: Stimuli-Response table

	Lineup Stimuli	
Response	1 target + 3 decoys (Positive)	4 targets (Negative)
One of the four plots	True positive **Hit**	False positive **False alarm**
“No discernible difference between images”	False negative Miss	True negative Correct rejection

3.3 Modeling with SDT

3.3.1 Overview

Experiment 2 manipulated two factors:

1. **Requested distance**. This is the requested KL divergence between the data sets of the target and decoy plots. It should be close to the actual distance with minor variations.
2. **Color ramp**. There are a total of 4 different color ramps. It could be modeled as a categorical or a numerical variable (using CNV).

Using SDT, we investigate whether more colorful color ramps increase the false positive rate. We analyze four models. We first calculate point estimates of sensitivity d' and criterion c for a single participant’s responses (Sec. 3.3.2). Then we apply a Bayesian fixed effects model to the same participant’s responses (Sec. 3.3.3). Then we extend the previous model to include responses from all 59 participants and consider random effects (Sec. 3.3.4). The third model uses the requested distance first as a nominal predictor and then as an ordinal predictor with monotonic effects (Sec. 3.3.5). The final model uses rounded distance as an ordinal predictor.

For computing and presenting our results, we used the following R packages: `brms` [4], `ggplot2` [26], `modelr` [27], `tidybayes` [13], and `tidyverse` [28]. We adopted weakly informative priors and ran four chains, each with 5000 iterations for all our models. We inspected the minimal bulk and tail effective sample size (ESS) to ensure reliable estimates. We also examined $\widehat{R}$ values to ensure all of our models converge.

3.3.2 Point estimates for a single participant

In a typical SDT workflow, one starts by aggregating responses across items, participants, or both [22]. We first aggregate across color ramps and calculate the point estimates of the hit rate and false alarm rate. Then, we transform these rates into SDT parameters using the probit function $\Phi^{-1}(\cdot)$ [23].

Table 2: Stimuli-Response table for participant `300`

Positive stimuli?	Picks one of the four plots?	Count
No	No	33
No	Yes	15
Yes	No	12
Yes	Yes	36

To illustrate this concretely, let’s examine the participant with subject id `300` and aggregate their responses across all items. The hit rate is $36/(36+12) = 0.75$ and the false alarm rate is $15/(15+33) = 0.3125$. Using Equations (1) and (2), we can easily calculate **point estimates** of participant `300`’s sensitivity d' and criterion c (averaged over all four color maps): $\widehat{d'} = \Phi^{-1}(0.75) - \Phi^{-1}(0.3125) \approx 1.16$; $\widehat{c} = -1/2(\Phi^{-1}(0.3125) + \Phi^{-1}(0.75)) \approx -0.09$. Likewise, we can also calculate participant `300`’s hit rates and false alarm rates for each color map:

Table 3: Hit rate, false alarm rate, sensitivity d', and criterion c for participant `300`

color ramp	hit rate	false alarm rate	d'	c
blues	8/12	4/12	0.86	0
cool-warm	10/12	5/12	1.18	-0.38
viridis	8/12	3/12	1.11	0.12
RGB rainbow	10/12	3/12	1.64	-0.15

From the point estimates in Table 3, participant `300` is most sensitive to the *rainbow* ramp and least sensitive to the *blues* ramp.

Blues also appear to be least biased, while both *cool-warm* and *rainbow* have liberal criterions and *viridis* has a conservative criterion.

3.3.3 Pooling information for a single participant

As a warm-up, we use a **fixed-effects** Bayesian model to analyze participant `300`'s responses. The full model formula is:

`line 1` $\text{responds yes}_i \sim \text{Bernoulli}(\pi_i)$

`line 2` $\text{probit}(\pi_i) = \alpha_{\text{CMAP}[i]} + \beta_{\text{CMAP}[i]} \cdot \text{signal present}_i \quad (2)$

`line 3` $\alpha_c \sim \mathcal{N}(0,1), \quad c \in \{1,2,3,4\}$

`line 4` $\beta_c \sim \mathcal{N}(0,1), \quad c \in \{1,2,3,4\}$

`line 1` Let i denote the i-th observation in data, and π_i the probability of answering the i-th observation correctly. When participant `300` selects one of the four plots rather than "no significant differences detected", they indicate perceiving a difference. We model this as a "yes" response to whether a signal is present in the stimulus.

`line 2` CMAP[i] denotes the color ramp of the i-th observation, that is, CMAP[i] $\in \{\textit{blues}, \textit{viridis}, \textit{cool-warm}, \textit{RGB rainbow}\}$[2]. *signal present* is the binary predictor variable of whether the stimulus is composed of one target plot + three decoy plots (↔ signal present) or four target plots (↔ signal absent).

`lines 3,4` We expect that intercepts (αs) and slopes (βs) vary with different color ramps, and defined weakly informative priors for each color ramp.

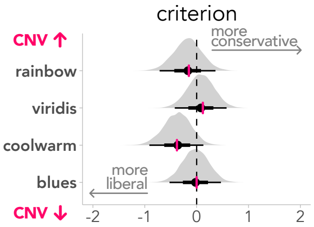

Figure 5: Posterior density, median, 66% and 95% quantile interval of **criterion** for participant `300`. Pink ticks represent the point estimates.

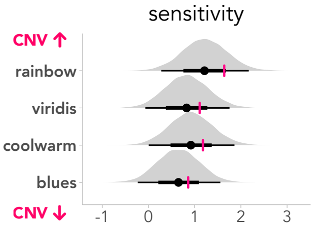

Figure 6: Posterior density, median, 66% and 95% quantile interval of **sensitivity** for participant `300`. Pink ticks represent the point estimates.

The results are comparable to the point estimates in Section 3.3.2, with the added advantage of providing a more precise quantification of uncertainty for SDT parameters. Observe that the results demonstrate model shrinkage (Figs. 5 and 6).

3.3.4 Pooling information with mixed effects

It is reasonable to expect *participant-level* variability in sensitivity. Some participants can differentiate colors better than others, while some may adopt a more conservative criterion, saying "no" to most presented stimuli. When aggregating events across participants, it is implicitly assumed that people have the same sensitivity and criterion. These assumptions are likely too restrictive [22]. To address this limitation, we extend the previous model in Section 3.3.3 by incorporating *participant-level effects* as random effects. The resulting model formula is as follows:

`line 1` $\text{responds yes}_i \sim \text{Bernoulli}(\pi_i)$

`line 2` $\text{probit}(\pi_i) = \alpha_{\text{CMAP}[i],\text{PID}[i]} + \beta_{\text{CMAP}[i],\text{PID}[i]} \cdot \text{signal present}_i$

`line 3` $\alpha_{c,j} = \bar{\alpha}_c + \delta_{\alpha,j}, \quad c \in \{1,\ldots,4\}$

`line 4` $\beta_{c,j} = \bar{\beta}_c + \delta_{\beta,j}, \quad c \in \{1,\ldots,4\}$

`line 5` $\overline{\alpha}_c, \overline{\beta}_c \sim \mathcal{N}(0,1), \quad c \in \{1,\ldots,4\}$

`line 6` $\begin{bmatrix} \delta_{\alpha,j} \\ \delta_{\beta,j} \end{bmatrix} \sim \text{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma\right), \quad j \in \{1,\ldots,59\}$

`line 7` $R \sim \text{LKJcorr}(2)$

`line 2` PID[i] indexes the participant for the i-th observation.

`line 3, 4` $\delta_{\alpha,j}, \delta_{\beta,j}$ denote participant j's random effect for intercept α and slope β respectively.

`line 6` The covariance matrix $\Sigma = \text{diag}(\tau) R \,\text{diag}(\tau)$, where τ is a vector of standard deviations of $\delta_{\alpha,j}$ and $\delta_{\beta,j}$ and R is their correlation matrix. We expect some variance in slopes and intercepts and therefore set $\tau \sim \text{Exponential}(1)$ as priors.

`line 7` LKJcorr(2) implies a weak correlation between participants' slopes and intercepts.

The point estimates were calculated using log-linear correction [10] to account for hit rates equal to 1 and false alarm rates equal to 0. Observe that the point estimates of sensitivity are consistently smaller than the posterior estimates of sensitivity (Fig. 8). This aligns with prior work [21, 22], which has demonstrated that aggregation may lead to an asymptotic underestimation of sensitivity.

The results indicate that an average participant is most sensitive to the *rainbow* color ramp and least sensitive to the single-hue *blues* color ramp (Fig. 8). Figure 7 clearly shows that an average participant is more conservative (i.e., has more tendency to respond "no significant differences detected") to all color ramps except *cool-warm*.

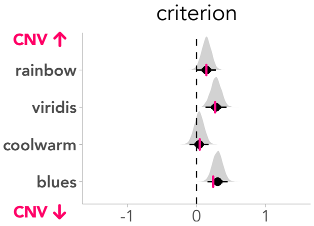

Figure 7: Posterior density, median, 66% and 95% quantile interval of **criterion** for an average participant. Pink ticks represent the point estimates.

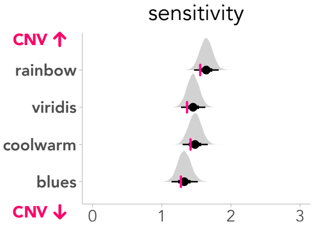

Figure 8: Posterior density, median, 66% and 95% quantile interval of **sensitivity** for an average participant. Pink ticks represent the point estimates.

There is also significant variation in participants' performance. For example, consider participants `310` and `313`. Figure 10 demonstrates that the two participants share similar sensitivity for all color ramps (except for *cool-warm*). However, participant `310` is significantly more conservative than `313` in terms of response criterion (Fig. 9). In other words, `310` tended to respond "no significant differences detected" while `313` tended to respond with one of the four plots, despite the prior instruction that half of the trials are negative trials. This suggests that individual biases can influence the interpretation of color ramps, even though the underlying perceptual sensitivity is similar.

[2] An alternative notation is to subscript by i and j, where i denotes the i-th observation and j indexes which color ramp.

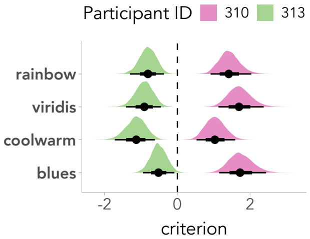

Figure 9: Posterior density, median, 66% and 95% quantile interval of **criterion** for participants 310 and 313.

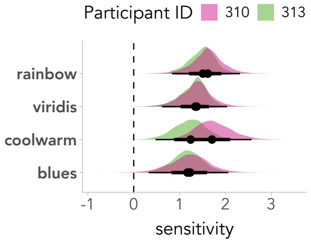

Figure 10: Posterior density, median, 66% and 95% quantile interval of **sensitivity** for participants 310 and 313.

3.3.5 Going beyond binary – modeling signal strength as ordinal categorical

In the previous models, we treated the signal as a *binary* variable, either present or absent. However, in most psychophysical experiments, researchers are also interested in investigating how the *strength* of the signal impacts sensitivity and criterion. To address this, we extend our analysis by modeling "signal strength" as an **unordered categorical variable** (i.e., nominal). Specifically, we replace the binary variable *signal present* with *requested distance*, a categorical variable with five levels: 0, 0.21, 0.23, 0.24, and 0.26. This changes `line 2` in Section 3.3.4 to

$$\text{probit}(\pi_i) = \alpha_{\text{CMAP}[i],\text{SID}[i]} + \beta_{\text{CMAP}[i],\text{SID}[i]} \cdot \textit{requested distance}_i$$

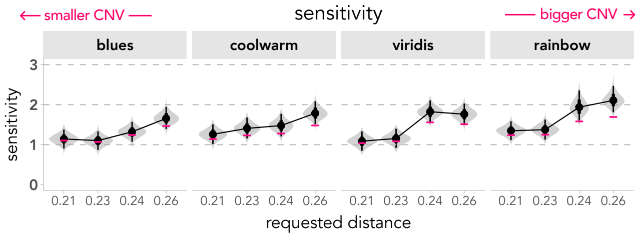

Figure 11: Posterior density, median, 66% and 95% quantile interval of **sensitivity** across different levels of **requested distance** for different color ramps for an average participant. Pink ticks represent the point estimates.

It is clear from Figure 11 that for an average participant, as the requested distance increases, the general trend is an increase in sensitivity across color ramps. However, we also see decreases from 0.21 to 0.23 for *blues* and from 0.24 to 0.26 for *viridis*.

Theoretically, one would expect that sensitivity increases with signal magnitude. Combined with the above results, we next model requested distance as an ordinal predictor **with monotonic effects** [3].

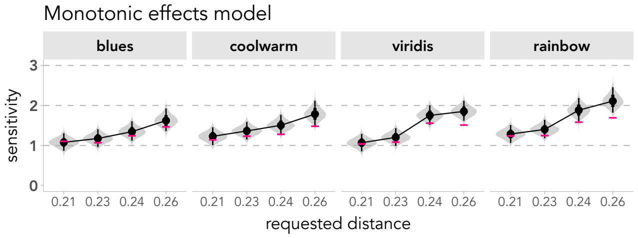

Figure 12: Posterior density, median, 66% and 95% quantile interval of **sensitivity** across different levels of **requested distance** for different color ramps for an average participant. Pink ticks represent the point estimates.

In this new model, the increase in sensitivity across requested distance levels is *more uniform* for the more *perceptually-uniform* color ramps (i.e., *blues* and *cool-warm*) compared to *viridis* and *rainbow* (Fig. 12). For color ramps with higher CNV, the increase in sensitivity between 0.23 and 0.24 seems to be the highest.

One might wonder whether the disproportionate jump in sensitivity from the requested distance of 0.23 to 0.24 (Fig. 12) is due to **aggregation**.

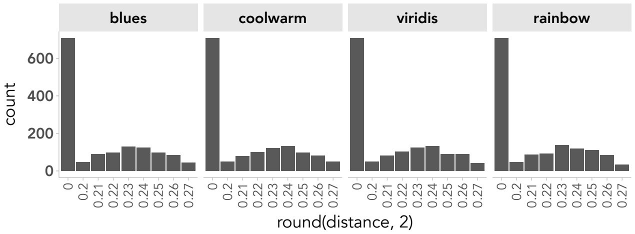

Figure 13: Distribution of **rounded distance**. More trials are administered with rounded distances 0.23 and 0.24 compared to the tail end (0.20 and 0.27).

We extend the prior model by rounding the distance between the target plot and decoy plots and using it as an ordinal predictor with monotonic effects. The new predictor, *rounded distance*, has eight distinct levels ranging from 0.20 to 0.27 with a step size of 0.01.

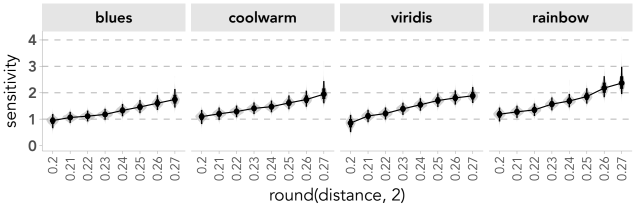

Figure 14: Posterior density, median, 66% and 95% quantile interval of **sensitivity** across different levels of **rounded distance** for different color ramps for an average participant.

The disproportionate jump from 0.23 to 0.24 (Fig. 12) no longer exists in the new model (Fig. 14). The *rainbow* color ramp has more uncertainty compared to the three other color ramps when the rounded distance is 0.27 – this could be due to not enough trials compared to the other color ramps.

Recall that criterion location is defined for a pair of evidence distributions (Eq. (2)). The interpretation of the criterion location is less clear when there are multiple signal distributions. For the sake of an easier interpretation, we plot the posterior estimate of the false alarm rates instead (Fig. 15). Observe that the results demonstrate model shrinkage. Except for *cool-warm*, an average participant has similar false alarm rates for all color ramps.

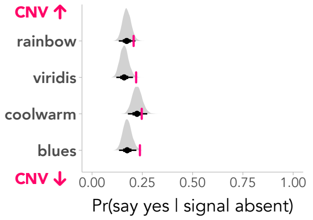

Figure 15: Posterior density, median, 66% and 95% quantile interval of **false alarm rate** across different levels of **requested distance** for different color ramps for an average participant. Pink ticks represent the point estimates.

3.4 Compare to non-SDT models

To analyze the data they collected, Reda and Szafir [20] ran four **logistic regression** models to separately analyze **error rates** for each stimulus type, i.e., true positive ("*sensitivity*") and true negative ("*specificity*") rates. The first two models only examine responses to negative stimuli and the remaining two examine responses to positive stimuli. All models examine a binary response variable of whether the participant correctly identified the stimuli's underlying type (positive vs negative). The two models that are most relevant to our models use *design family*, an unordered categorical predictor variable with four distinctive levels.

Reda and Szafir [20] found that:

- "... *cool-warm* ramp exhibited significantly lower [true negative rate]"
- Pairwise comparison with Tukey's adjustment show *cool-warm* to be worse than *viridis*.
- "both *RGB rainbow* and *cool-warm* led to higher true positive rate than *viridis* and *blues*"

To calculate SDT parameters, our models examine the hit rate (i.e., true positive rate) and false alarm rate (i.e., false positive rate = 1 − true negative rate). We found that:

- An average participant is most sensitive to *rainbow* and least sensitive to *blues*. We also find the *cool-warm* ramp to be the least biased, while the remaining three color ramps are all slightly biased towards responding "no significant differences detected" (Sec. 3.3.4).
- As KL divergence increases between the target plot and the decoy plots, an average participant becomes more sensitive for all examined color ramps (Sec. 3.3.5).
- Variations exist both within an individual (e.g., different criterion for each color ramp) and between individuals (e.g., different sensitivity to color ramp)

Returning to the original question that Reda and Szafir posed, *do rainbow color ramps lead to more false discoveries, compared to other color ramps?* While an average participant is most sensitive to the *rainbow* color ramp (Fig. 8), they still have a conservative criterion (i.e., participants tended to respond "no difference detected") (Fig. 7). From our results, we agree with Reda and Szafir in that rainbow color ramps do not lead to more false discoveries, despite their higher sensitivity compared to other color ramps.

3.5 Discussion and limitations for modeling

We used both the probit function and the logit function as link functions in our models. The probit and logit functions generally yield similar results, with the main difference being in how the parameters are scaled [6]. While these are the most common choices, other link functions are also available, e.g., complementary log-log link [6]. It's important to note that **each link function embodies a different assumption about the underlying sensory distributions**. The probit link assumes a normal distribution, the logit link assumes a logistic distribution, and the log-log link assumes an extreme value distribution.

Our models assume equal variance between the two sensory distributions. While this is a common assumption, it may not always be appropriate [21]. Works such as Lages [15] extended the equal-variance hierarchical SDT model to an unequal-variance model by constraining the sensitivity and criterion distributions at the population level. Future work could explore this extension, as well as the application of multi-dimensional SDT, as our current model assumes a one-dimensional scale for the perception of an "odd" plot. Additionally, we were unable to implement the Receiver Operating Characteristic (ROC) curve analysis due to the lack of confidence rating data in Reda and Szafir's work [20].

An alternative way to analyze data collected via the typical lineup task is to analyze it as an m-alternative forced choice (mAFC) task, where $m-1$ stimuli are drawn from the signal-absent distribution and another stimulus is drawn from the signal-present distribution. Bias is typically ignored when analyzing mAFC data due to the intractability it introduces into modeling, yet choosing to not model bias *does not* imply its non-existence [10]. A potential approach to incorporate bias into analyzing mAFC data was introduced by Decarlo [7]. In a typical lineup with $m = 20$ plots, we will need nineteen criteria to carve up the decision space to incorporate response bias into the model. Let $b_i, i \in [19]$ denote the bias towards the response associated with plot i and let $\Psi_i, i \in [20]$ denote the realized perception of plot i. The decision rule is to respond

- plot 1 if $\Psi_1 + b_1 > \max\{\Psi_2 + b_2, \Psi_3 + b_3, ..., \Psi_{20}\}$
- plot 2 if $\Psi_2 + b_2 > \max\{\Psi_1 + b_1, \Psi_3 + b_3, ..., \Psi_{20}\}$
- ...
- plot 20 if $\Psi_{20} > \max\{\Psi_1 + b_1, \Psi_2 + b_2, ..., \Psi_{19} + b_{19}\}$

To fit all the above parameters would likely require more data and have potential issues with overfitting and multicollinearity.

4 General Discussion

4.1 Suggestions for better lineup experiment design

The lineup task is an innovative approach for evaluating visualizations, but it is essentially a reinvention of an existing task that falls within SDT. When designed and analyzed correctly, lineup tasks can provide valuable insights. However, the traditional lineup experimental design and analysis are lacking because they expose participants only to the signal-present distribution – that is, participants are only shown trials that include one target plot and $m-1$ decoy plots. Measuring sensitivity requires exposing participants to signal-present and signal-absent distributions, as sensitivity depends on *both* hit rate and false alarm rate [10].

To enable SDT analysis and improve lineup experiment design, we propose integrating elements from Reda and Szafir's experiment design:

1. Include a response option for "no significant differences detected".
2. Introduce stimuli where all plots are drawn from the same decoy or target model.
3. Employ payoffs to incentivize participants to adopt better decision-making criteria.
4. Continue to collect participant confidence ratings for their selections.

These modifications will expose participants to both signal-present and signal-absent distributions, allowing for SDT analysis, ROC curve analysis, and potentially improving lineup task interpretability.

4.2 Beyond lineups

Beyond analyzing the lineup task, SDT can also be applied to analyze data collected through various tasks [8]. Elliott et al. [8] introduced a design space categorization that divides dependent measures for assessing visualization into two categories: *direct dependent measures* and *model-based dependent measures*. Direct dependent measures, such as **accuracy** (measured by percentage correct or error), **response time**, and **precision**, appear frequently in existing visualization experiments. Model-based dependent measures, such as **performance slope**, **psychometric function**, and **senstivity & bias detection** (modeled via SDT), appear less frequently but enable more precise claims about the visual system. However, there is a trade-off – model-based dependent measures typically require more data and more complex experimental designs [8]. Ultimately, it is up to the designer to choose the most appropriate measures for their specific goals, balancing the costs and benefits of each approach.

4.3 Connections to statistical decision theory

Graphical inference and SDT are both inspired by statistical decision theory. However, despite a brief mention of the Rorschach protocol by Buja et al. [2], existing models of graphical inference predominantly focus on rejecting some null hypotheses.

We believe that SDT provides a better model for the ideas in Buja et al. [2]. SDT has been influenced by statistical decision theory since its conception [29], and there is a clear analogy between an observer in a detection experiment and a statistician deciding between data sources. The observer must determine whether the simulation arose from the signal-present distribution or the signal-absent distribution. Similarly, a statistician must decide whether the data is best explained with an alternative hypothesis (i.e., a real difference in the world) or a null hypothesis (i.e., a difference arising from sampling variability).

Compared to the graphical inference framework, SDT is more explicit about its assumptions about human perception. Every theory of the mind embraces some sets of assumptions, either explicitly or implicitly [29]. There is no free lunch. Employing SDT enables us to separate the world of stimuli and their perturbations from that of the decision process [10]. By separating the factors that influence the effectiveness of plot designs into sensitivity and bias, we can gain deeper, more nuanced insights (e.g., non-uniform increase in sensitivity across different color ramps in Section 3.3.5) and make more informed decisions.

5 Conclusion

In this position paper, we re-examined the graphical inference framework and its primary method, the lineup task, for evaluating visualizations. By re-analyzing data from Reda and Szafir's [20] study using Bayesian signal detection theory (SDT), we demonstrated that the SDT approach can provide a more nuanced understanding of the factors influencing the effectiveness of visualizations. Furthermore, we suggest integrating elements of Reda and Szafir's experiment design into typical lineup task designs to enable SDT analysis and potentially improve lineup task interpretability. Our approach demonstrates the general utility and benefits of applying SDT for evaluating and comparing visualization designs using visual lineups.

A Distinction between Buja et al. [2] and Majumder et al. [18]

Buja et al.'s [2] lineup protocol is about evaluating and quantifying the significance of graphical findings. It involves embedding the true data plot (i.e., *target plot*) among $m = 20$ *decoy plots* generated from "simple generic null hypotheses" [2]. The act of selecting the target plot as the most different is considered as *rejecting a null hypothesis*, and the "[visual] discovery can be assigned a p-value of 0.05 (=1/20)" [2].

In a follow-up work, Majumder et al. [18] proposed a model where a lineup is shown to K independent observers (as opposed to the single observer in Buja et al.). Each observer is assumed to have the same probability π of correctly identifying the target plot. The number of observers Y who correctly identify the target plot follows a **Binomial** distribution, as Y is the sum of K independent **Bernoulli** random variables. In more formal terms, $Y \sim \text{Binomial}(K, \pi)$. Given Y's distribution, the probability that Y exceeds a particular value y, i.e., the *complementary cumulative distribution function* (CCDF) of Y is

$$\begin{aligned}\Pr(Y \geq y) &= \sum_{i=y}^{K} \Pr(Y = i) \\ &= \sum_{i=y}^{K} \binom{K}{i} \pi^i (1-\pi)^{K-i} \qquad (3)\end{aligned}$$

Under the assumption that the target plot is indistinguishable from the decoys (`ASMP. 1`), the probability of a participant correctly identifying the target plot by chance is $\pi = 1/m = 1/20$. The minimum number of correct responses $\widetilde{y}$ needed to confidently reject `ASMP. 1`, that is, $\widetilde{y}$ such that $\Pr(Y \geq \widetilde{y} \mid$ `ASMP. 1` $) =$ small significance level 0.01, can be solved for via the inverse Binomial CCDF.

In Majumder et al. [18]'s framework, the **null hypothesis** H_0 is `ASMP. 1`, which implies that the **null distribution** follows a $\text{Binomial}(K, 1/m)$ distribution. The "p-value of a lineup of size m evaluated by K observers"[18] is the Binomial CCDF in Equation (3) with $\pi = 1/m$, and the **test statistic** is y, the observed number of participants who correctly identify the target plot. The **significance level** α, once defined, can help us solve for $\widetilde{y}$, where $\widetilde{y}$ is defined such that $\Pr(Y \geq \widetilde{y} \mid H_0) \leq \alpha$.

Therefore, inferential validity achieved via Majumder et al.'s [18] Binomial model deviates from the original ideas in Buja et al. that "plots take on the role of test statistics, and human cognition the role of statistical tests" [2]. The distribution of the perception of plots and the distribution of counts of correct participants under `ASMP. 1` are two *different* distributions.

B Additional definitions of criterion

There are different ways of measuring bias. We have already introduced criterion location (Eq. (2)). We can also define **relative criterion location** c' [10]:

$$c' = \frac{c}{d'}$$

Another definition of criterion is the **likelihood ratio** β [10]:

$$\beta = \frac{f(x|S_2)}{f(x|S_1)}$$

C Interpreting SDT model parameters

Consider a simplified version of `line 2` in Section 3.3.4:

$$\text{probit}(\pi_i) = \alpha_i + \beta_i \cdot \textit{signal present}_i$$

When no signal is present, then false alarm rate $\pi_i = \Phi(\alpha_i)$. When there is signal present, then hit rate $\pi_i = \Phi(\alpha_i + \beta_i)$. This implies

$$\begin{aligned}\alpha_i &= \Phi^{-1}(\text{False alarm rate}_i) \\ \alpha_i + \beta_i &= \Phi^{-1}(\text{Hit rate}_i) \\ \implies \beta_i &= \Phi^{-1}(\text{Hit rate}_i) - \Phi^{-1}(\text{False alarm rate}_i) \equiv d'\end{aligned}$$

D brms formulae

D.1 Section 3.3.3

The corresponding brms formula is

```
responds_yes ~ 0 + colormap + colormap:signal_present,
family = bernoulli("probit"),
data = filter(df, subjectid == 300)
```

D.2 Section 3.3.4

The corresponding brms formula is

```
responds_yes ~ 0 + colormap + colormap:signal_present +
            (0 + colormap + colormap:signal_present | subjectid),
family = bernoulli("probit"),
data = df
```

D.3 Section 3.3.5

The corresponding brms formula is

```
responds_yes ~ 0 + colormap + colormap:req_dist +
            (0 + colormap + colormap:req_dist | subjectid),
family = bernoulli("probit"),
data = df
```

The modified brms formula for **monotonic effects** is

```
responds_yes ~ 0 + colormap + colormap:mo(req_dist) +
               (0 + colormap + colormap:mo(req_dist) | subjectid),
family = bernoulli("logit"),
data = df
```

Hereon we use the logit function for easier model fitting. The results were then transformed using an approximation of the logit to the probit [1] to calculate SDT parameters.
The modified brms formula for **rounded distance** is

```
responds_yes ~ 0 + colormap + colormap:mo(round_dist) +
               (0 + colormap + colormap:mo(round_dist) | subjectid),
family = bernoulli("logit"),
data = df
```

Supplemental Materials

All supplemental materials are available on OSF at https://osf.io/xd5cj/. They include 1) all model fits, 2) graphs used in this paper, and 3) analysis files used to generate and plot model fits.

Figure credits

Figure 1 is taken from Mahbubul Majumder's website, which is in the public domain. Figure 4 is taken from Reda and Szafir's [20] experiment interface, posted at https://osf.io/tck2r/.

Acknowledgments

We thank Khairi Reda and Danielle Albers Szafir for sharing their data, without which this paper would not have been possible. We also thank Jessica Hullman for discussions related to graphical inference at the early stages of this project.

References

[1] T. Amemiya. Qualitative response models: A survey. *Journal of economic literature*, 19(4):1483–1536, 1981. 8

[2] A. Buja, D. Cook, H. Hofmann, M. Lawrence, E.-K. Lee, D. F. Swayne, and H. Wickham. Statistical inference for exploratory data analysis and model diagnostics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4361–4383, 2009. doi: 10.1098/rsta.2009.0120 1, 2, 6, 7

[3] P.-C. Bürkner and E. Charpentier. Modelling monotonic effects of ordinal predictors in bayesian regression models. *British Journal of Mathematical and Statistical Psychology*, 73(3):420–451, 2020. doi: 10.1111/bmsp.12195 5

[4] P.-C. Bürkner. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28, 2017. doi: 10.18637/jss.v080.i01 3

[5] A. L. Cohen, J. J. Starns, and C. M. Rotello. sdtlu: An r package for the signal detection analysis of eyewitness lineup data. *Behavior Research Methods*, 53:278–300, 2021. doi: 10.3758/s13428-020-01402-7 2

[6] L. T. DeCarlo. Signal detection theory and generalized linear models. *Psychological methods*, 3(2):186, 1998. doi: 10.1037/1082-989X.3.2.186 6

[7] L. T. DeCarlo. On a signal detection approach to m-alternative forced choice with bias, with maximum likelihood and bayesian approaches to estimation. *Journal of Mathematical Psychology*, 56(3):196–207, 2012. doi: 10.1016/j.jmp.2012.02.004 6

[8] M. A. Elliott, C. Nothelfer, C. Xiong, and D. A. Szafir. A design space of vision science methods for visualization research. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1117–1127, 2020. doi: 10.1109/TVCG.2020.3029413 2, 6

[9] T. Griffith, S.-A. Baker, and N. F. Lepora. The statistics of optimal decision making: Exploring the relationship between signal detection theory and sequential analysis. *Journal of Mathematical Psychology*, 103:102544, 2021. doi: 10.1016/j.jmp.2021.102544 2

[10] M. J. Hautus, N. A. Macmillan, and C. D. Creelman. *Detection theory: A user's guide*. Routledge, 2021. doi: 10.4324/9781003203636 4, 6, 7

[11] J. Heer and M. Stone. Color naming models for color selection, image editing and palette design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1007–1016, 2012. doi: 10.1145/2207676.2208547 2

[12] H. Hofmann, L. Follett, M. Majumder, and D. Cook. Graphical tests for power comparison of competing designs. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2441–2448, 2012. doi: 10.1109/TVCG.2012.230 1

[13] M. Kay. *tidybayes: Tidy Data and Geoms for Bayesian Models*. R package version 3.0.5. doi: 10.5281/zenodo.1308151 3

[14] M. Kossmeier, U. S. Tran, and M. Voracek. Visual inference for the funnel plot in meta-analysis. *Zeitschrift für Psychologie*, 2019. doi: 10.1027/2151-2604/a000358 1

[15] M. Lages. A hierarchical signal detection model with unequal variance for binary responses. *Psychonomic Bulletin & Review*, pp. 1–24, 2024. doi: 10.3758/s13423-024-02504-5 6

[16] A. Loy and H. Hofmann. Diagnostic tools for hierarchical linear models. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5(1):48–61, 2013. doi: doi.org/10.1002/wics.1238 1

[17] N. A. Macmillan. Signal detection theory. In *Stevens' Handbook of Experimental Psychology: Methodology in Experimental Psychology, Vol. 4, 3rd Ed.*, pp. 43–90. John Wiley & Sons, Inc., 2002. doi: 10.1002/0471214426.pas0402 2

[18] M. Majumder, H. Hofmann, and D. Cook. Validation of visual statistical inference, applied to linear models. *Journal of the American Statistical Association*, 108(503):942–956, 2013. doi: 10.1080/01621459.2013.808157 1, 7

[19] B. Paulewicz and A. Blaut. The bhsdtr package: a general-purpose method of bayesian inference for signal detection theory models. *Behavior Research Methods*, 52(5):2122–2141, 2020. doi: 10.3758/s13428-020-01370-y 2

[20] K. Reda and D. A. Szafir. Rainbows revisited: Modeling effective colormap design for graphical inference. *IEEE transactions on visualization and computer graphics*, 27(2):1032–1042, 2020. doi: 10.1109/TVCG.2020.3030439 1, 2, 3, 5, 6, 7, 8

[21] J. N. Rouder and J. Lu. An introduction to bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic bulletin & review*, 12(4):573–604, 2005. doi: 10.3758/BF03196750 4, 6

[22] J. N. Rouder, J. Lu, D. Sun, P. Speckman, R. Morey, and M. Naveh-Benjamin. Signal detection models with random participant and item effects. *Psychometrika*, 72(4):621–642, 2007. doi: 10.1007/s11336-005-1350- 3, 4

[23] H. Stanislaw and N. Todorov. Calculation of signal detection theory measures. *Behavior research methods, instruments, & computers*, 31(1):137–149, 1999. doi: 10.3758/BF03207704 2, 3

[24] S. VanderPlas and H. Hofmann. Clusters beat trend!? testing feature hierarchy in statistical graphics. *Journal of Computational and Graphical Statistics*, 26(2):231–242, 2017. doi: 10.1080/10618600.2016.1209116 1

[25] S. VanderPlas, C. Röttger, D. Cook, and H. Hofmann. Statistical significance calculations for scenarios in visual inference. *Stat*, 10(1):e337, 2021. doi: 10.1002/sta4.337 1

[26] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. doi: 10.1007/978-3-319-24277-4 3

[27] H. Wickham. *modelr: Modelling Functions that Work with the Pipe*, 2023. R package version 0.1.11, https://github.com/tidyverse/modelr. 3

[28] H. Wickham, M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. L. Pedersen, E. Miller, S. M. Bache, K. Müller, J. Ooms, D. Robinson, D. P. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, and H. Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686 3

[29] J. T. Wixted. The forgotten history of signal detection theory. *Journal of experimental psychology: learning, memory, and cognition*, 46(2):201, 2020. doi: 10.1037/xlm0000732 2, 7