Journal of Data Science 0 (0), 1–21
??? 2026
DOI: 10.6339/26-JDS1221
Statistical Data Science

Clusters, Trends, and Choices: Feature Selection in Interactive Statistical Graphics

Dylan Le[1], Rachel Rogers[2], and Emily Robinson[1]

[1]*California Polytechnic State University - San Luis Obispo, United States*
[2]*University of Technology Sydney, Austrailia*

Abstract

This study investigates how user ability to manipulate plot features affects graphical perception, by extending a previous graphical study (Vanderplas and Hofmann, 2017) with an interactive framework. Similar to the original study, statistical lineups included two target patterns (a linear trend and a clustering pattern), as well as eighteen null plots generated from three different mixture proportions of the combined cluster and trend models. Participants were asked to select two plots that they perceived as 'most different', and were able to interact with the graphics by toggling aesthetic features such as cluster coloring, cluster ellipses, linear trendlines, and regression error bands.

We found that toggle workflow varied across participants, revealing a divide between "maximalists," who enabled all features, and "minimalists," who used few or none, with most toggling occurring before the first selection. Starting features aesthetics did not have a significant effect on target choice. A generalized linear mixed model identified mixture proportion as the strongest predictor of target selection, with additional interactions involving the enabled ending features. These findings contribute to understanding how users engage with interactive graphical tools and how such tools support data interpretation in exploratory data analysis.

Keywords *Exploratory Data Analysis; Statistical Lineups; Visual Perception*

1 Introduction

For researchers and their audiences, visual displays are often essential tools for making sense of data. Statistical graphics can transform raw numbers into recognizable patterns and trends, helping to summarize results, reveal statistical relationships, and support scientific arguments. As methods for sharing information have evolved, from static newspaper charts to interactive online dashboards, the potential for audience exploration and direct interaction with the data has expanded. Yet it remains unclear whether this interactivity actually leads to deeper insights.

To investigate this, we recreated and adapted a statistical lineup study that tests how different graphical features affect the detection of patterns in data. Building on the work of VanderPlas and Hofmann (2017), we examined how participants identified clustering and linear trends in simulated datasets when given the ability to toggle specific plot features on and off. In our study, participants toggled plot features on and off, allowing them to actively explore the data rather than passively view static displays as in the original experiment.

*Corresponding author. Email: erobin17@calpoly.edu.

Received August 15, 2025; Accepted February 21, 2026

1.1 Graphical Aesthetics

Designers of statistical graphics must balance clarity and visual appeal to convey their message effectively (Vanderplas et al., 2020). Research in visual perception has identified guiding principles for emphasizing patterns in data, many drawing on Gestalt principles such as grouping by proximity or similarity, enclosing related elements, and perceiving patterns as continuous (Cleveland and McGill, 1987; Lewandowsky and Spence, 1989; Spence, 1990; Glicksohn and Cohen, 2011; Shah and Miyake, 2005; Zeileis et al., 2009).

Aesthetic choices such as adding color, enclosing clusters with ellipses, or overlaying a trendline, can highlight specific trends and shape how viewers interpret a plot. While the decision to include these aesthetics or not are often made by the creator in static graphics, interactive displays can give users control to toggle these features on or off, tailoring the view to their own exploration.

Previous studies have shown that seemingly small choices (e.g., bin width, kernel bandwidth, opacity) and plot type (histogram, density plot, dot plot) can hide or reveal important data characteristics such as spikes, outliers, or gaps (Correll et al., 2019). These findings suggest that no single representation is sufficient and multiple displays may be needed to fully explore a dataset.

1.2 Interactive Graphics

Advances in tools such as Shiny (Chang et al., 2024) and D3.js have made it possible to create interactive graphics in which users can directly modify visual features to highlight different aspects of the data. Here, we define interaction as a user-controlled mechanism that changes what is shown and how it is displayed (Ward et al., 2021). Such mechanisms include filtering, toggling features on and off, zooming, and, in earlier interactive graphical systems, linking and brushing across multiple coordinated views (Swayne and Buja, 2004). Linking and brushing, developed in software like GGobi, allow users to highlight data points in one plot and see them reflected in other connected plots.

Interactive graphics can empower users to tailor displays to their needs, discover patterns on their own (Li et al., 2018), and even replicate the decision process of the plot creator, thereby supporting transparency and reproducibility (Weissgerber et al., 2016). In high-dimensional settings, interactivity makes it feasible to explore relationships that would be difficult to show in a single static plot (Liu et al., 2016). These graphics are part of a narrative: design choices inevitably emphasize certain aspects of data while downplaying others.

Interactive interfaces are widely used in exploratory data analysis (EDA) (Komorowski et al., 2016). Commercial tools such as Tableau and PowerBI can even recommend chart types based on the data (Hullman and Gelman, 2021), while open-source technical frameworks like Shiny and D3.js offer greater flexibility for custom design. The recent development of AI-powered tools, such as Shiny Assistant (Chang, 2024), provide guidance for app developers.

Interactive graphics also play a key role in public communication. News organizations such as *The Washington Post*, *The New York Times*, and *Guardian* use them to increase engagement, especially during major events like elections or natural disasters (Li et al., 2018). During the COVID-19 pandemic, for example, Rutter et al. (2021) created an interactive tool allowing readers to vary real-world conditions (e.g., mask-wearing, activity type) and see the corresponding changes in estimated transmission risk.

1.3 Testing Statistical Graphics

One way to study how people interpret data visualizations is to present them with a graphic and ask them to perform a specific task such as selecting the plot that looks most different from others, predicting a future value, or reading a numeric value from a chart. The statistical lineup protocol is a well-established graphical testing framework (Buja et al., 2009). Modeled after police lineups, it places one or more target plots (showing real data) among several decoys generated from a null model. If participants can reliably identify the target plot(s), it suggests the real data differ noticeably from what the null model. In this framework, there is no single data characteristic that participants must focus on. Instead, they might choose a plot to stand out as different because of outliers, a different slope, clustering, or any other visual characteristic that stands out (Loy et al., 2017).

Lineups can also be used to compare the effectiveness of different visual features. In this case, the same underlying data are shown in multiple lineup plots with different design choices, and participants are asked to identify the plot that appears most different. Accuracy of the participant target selection can then be analyzed to assess which features help reveal the target pattern. For example, Hofmann et al. (2012) compared polar versus Cartesian coordinate systems for spotting wind patterns, and other work has used generalized linear mixed models to quantify the impact of design factors on detection rates (VanderPlas and Hofmann, 2017; Reda and Szafir, 2021).

A limitation of the existing lineup approach is that participants are constrained to the display chosen by the lineup creator. Presentation choices can strongly influence the ability to detect differences (VanderPlas and Hofmann, 2017), and no single aesthetic will highlight all aspects of the data (Correll et al., 2019). Interactivity offers a potential remedy by letting users toggle between features such as color and trendlines, focusing on what aesthetics they find most useful. In this study, we extend the lineup framework to an interactive setting, building on VanderPlas and Hofmann (2017) but giving participants direct control over the features displayed on the graph.

1.4 Competing Targets in Lineup Studies

VanderPlas and Hofmann (2017) introduced a variation on the lineup protocol in which two competing targets appear in the same lineup: one showing a linear relationship, the other showing clustered groups of points. The remaining null plots combined elements of both patterns. Plot aesthetics were chosen to emphasize either clustering (color, shape, ellipses) or a linear trend (trendline, error band), following Gestalt principles.

They found a clear masking effect where once participants identified one target (e.g., linear trend), they were less likely to continue searching for the other target (e.g., clustering). Feature effects were also direction-specific where color, shape, and ellipses increased the odds of selecting the cluster target, while a trendline and error band increased the odds of selecting the linear target. They found that these effects were not additive and combining features that favored different patterns did not produce a neutral result.Additionally, adding an extra Gestalt cue tended to increase detection accuracy for its associated pattern.

1.5 Objectives

We aim to replicate and extend the findings of VanderPlas and Hofmann (2017) by using statistical lineups with an interactive graphics framework. In our study, participants toggled visual aesthetics (color, ellipses, trendline, and error band) while evaluating the lineups to identify the target plots, simulating the exploratory data analysis workflow.

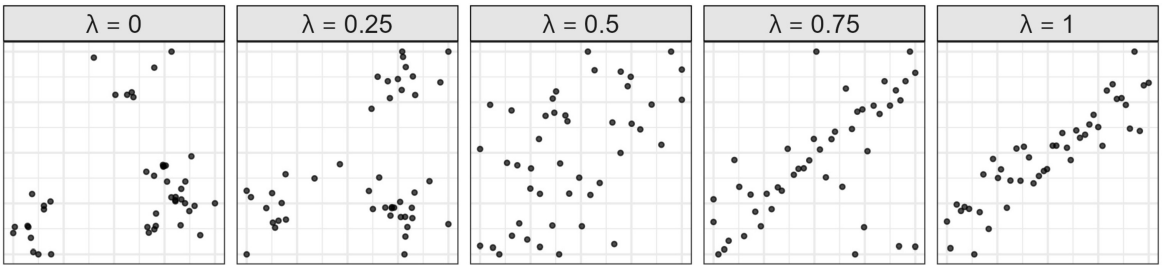


Figure 1: Example datasets generated with different values of λ. A $\lambda = 0$ produces the cluster target, $\lambda = 1$ produces the linear target, and intermediate values generate mixture-model null plots with varying amounts of clustering and linear trend. Lower λ values appear more clustered, while higher values appear more linear.

We investigate how often and in what ways participants engage with these interactive features. We also examine whether the features shown at the start of each lineup influence which targets are selected, and whether the features in place at the end of each trial are associated with detection accuracy. Finally, we analyze participants' interaction patterns to understand different data exploration styles.

In this paper, we first outline the data simulation, study platform, and participant recruitment. We then report results for target selection accuracy and toggle interaction workflows, followed by a discussion of findings, limitations, and future directions.

2 Methods

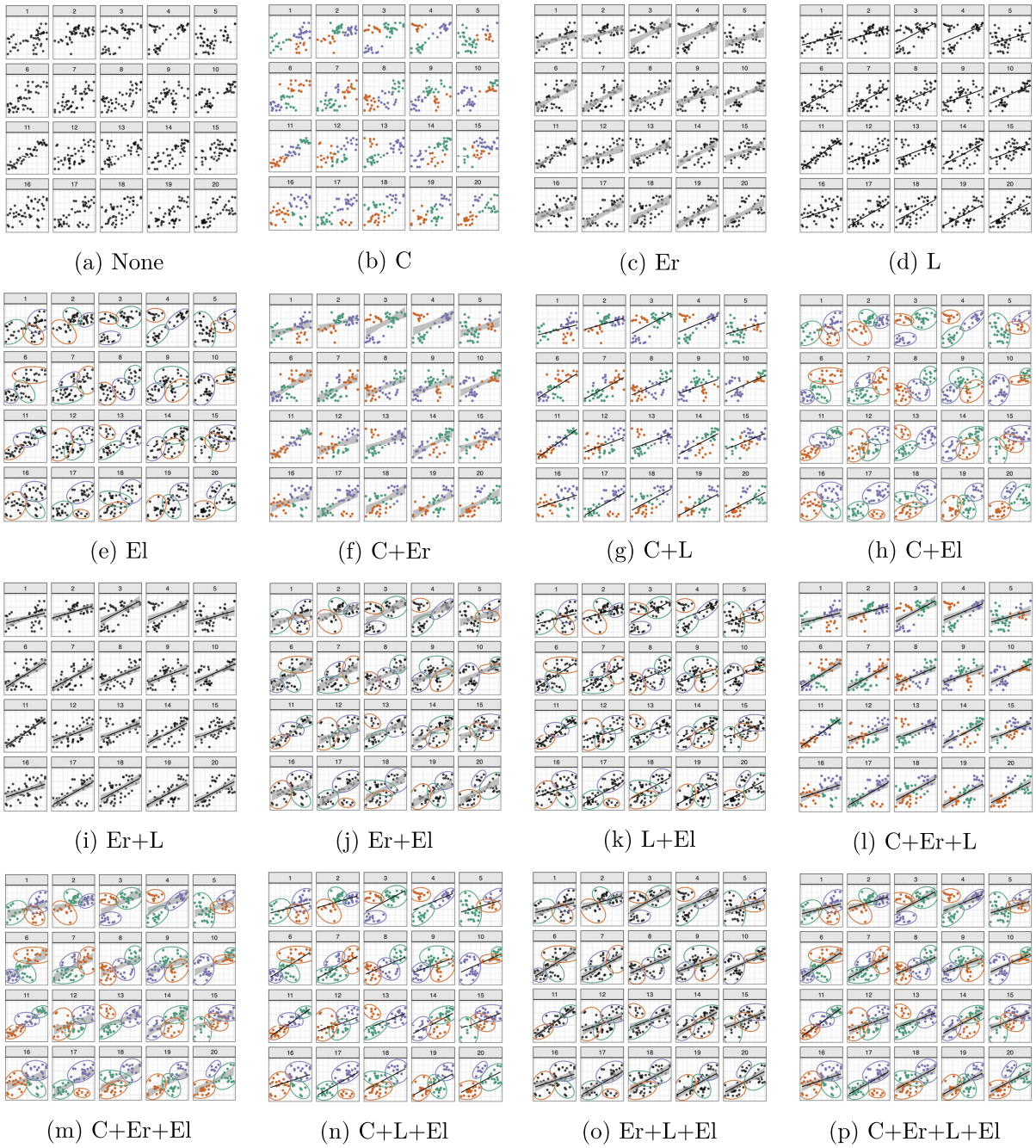
We used R v4.5.0 (R Core Team, 2024) to simulate data, generate lineup plots, build the study application, and conduct statistical analyses. The following subsections describe the simulation process, study platform, experimental design, and analysis methods.

2.1 Data Simulation and Lineup Generation

We simulated data following the approach in VanderPlas and Hofmann (2017), using three models: a linear trend, a clustered pattern, and a mixture of the two (See Appendix A). Each lineup contained 20 plots with one linear target, one cluster target, and 18 mixture-model null plots with $N = 45$ points per plot. Mixture plots were generated by sampling a proportion, λ, of points from each model, where lower values of λ made the cluster pattern more prominent and higher values made the linear pattern more prominent. We used $\lambda \in \{0.25, 0.5, 0.75\}$ Figure 1. Note that a $\lambda = 0$ corresponded to a cluster target and a $\lambda = 1$ to a linear target.

Following pilot testing, we set the number of clusters $K = 3$ and standard deviations $\sigma_C = 0.25$ and $\sigma_T = 0.25$ to make the target plots detectable without being trivial or impossible to identify. For each lineup, the positions of the two targets were randomly assigned, and null plots were randomly placed in the remaining positions. We created 12 unique lineups with four replicates for each λ for use in the study.

Each lineup was plotted so that participants could toggle four graphical features on or off (1) point color by k-means cluster (Color), (2) ellipses around clusters (Ellipse), (3) a best-fit trendline (Line), and (4) a regression error band (Error). Figure 2 shows all possible feature combinations.



(a) None (b) C (c) Er (d) L

(e) El (f) C+Er (g) C+L (h) C+El

(i) Er+L (j) Er+El (k) L+El (l) C+Er+L

(m) C+Er+El (n) C+L+El (o) Er+L+El (p) C+Er+L+El

Figure 2: All aesthetic feature combinations available for users to toggle on and off in the interactive lineups. Features: point color by k-means cluster (C), ellipses (El), trendline (L), and regression error band (Er). Shown from no features to all four enabled.

Figure 3: Example lineup trial in the Shiny study application, showing a statistical lineup, toggle controls, and task instructions.

2.2 Study Platform

The study was implemented in a Shiny application (Chang et al., 2024). Participants first completed two practice lineup trials, followed by six study trials. One practice trial contained two targets with the same signal (both clusters or both trends), and the other contained competing signals (one cluster, one linear trend). For training purposes, we reduced the variability in the simulated data by lowering the standard deviations to $\sigma_C = 0.15$ and $\sigma_T = 0.2$, and generated the null plots from a mixture model with $\lambda = 0.5$. These adjustments made the targets more visually salient, helping participants learn the toggle controls, the meaning of "most different," and become familiar with the requirement to select two plots. If a participant chose non-target plots during a practice trial, a notification displayed the correct answers (See Appendix B).

In each study trial, participants viewed a 20-panel statistical lineup with toggle controls to enable or disable four graphical features: point color by k-means cluster (C), ellipses around clusters (El), a fitted trendline (L), and a regression error band (Er). They could change features as much or as little as they wished and were required to select at least two panels before proceeding. Similar to VanderPlas and Hofmann (2017), we did not set a maximum value for the number of plots participants could select. We recorded each toggle action, the timing of actions, and the order and location of plot selections. Figure 3 shows a single study trial shown to participants during the experiment.

After completing the study lineups, participants answered open-ended questions about their selection reasoning and toggle strategies, as well as demographic questions, such as age group, gender, education level, and if they work in a STEM field.

2.3 Study Design

When a participant first saw each lineup, a pre-selected set of aesthetic features was turned on. This allowed us to test whether initial features influenced target choice or whether participants made their own adjustments. We used eleven starting-feature combinations: no features, each of the four features alone, and all two-way combinations of features (C+El, C+Er, C+L, El+Er, El+L, Er+L), based on a 2^4 factorial structure considering up to two-way interactions.

Each participant viewed two lineups for each $\lambda \in \{0.25, 0.5, 0.75\}$, for a total of six study trials. Starting-feature assignments followed a balanced incomplete block design created in JMP v18.0 (SAS Institute Inc., 2023), with $t = 11$ treatments, $b = 11$ blocks of size 6, and each treatment appearing in 6 blocks. Each participant was randomly assigned one block, and trial order was randomized. For the study, we intended for 550 participants to complete the study, such that each block was replicated 50 times. However, due to 155 participants reloading the page or not completing the study, we did not achieve perfect balance.

2.4 Statistical Analysis

We focused our analysis on three questions: (i) how accurately participants identified targets (overall and by target type), (ii) how starting and ending feature sets related to target choice, and (iii) how people used the interactive toggles over time.

To compare cluster vs. trend detection, we fit generalized linear mixed models (GLMMs) with a binomial response and logit link function using `lme4` (Bates et al., 2015). For each lineup evaluation, we coded the response as $Y = 1$ if the participant's first two selections included the trend target only, and $Y = 0$ if they included the cluster target only. For the models, we included all evaluations where only the linear trend or only the cluster trend were selected. This analysis aims to replicate the GLMM cluster vs. trend analysis conducted in VanderPlas and Hofmann (2017).

We conducted two GLMMs investigating the starting features (features preset in the study at the start of the trial) and the ending features (features enabled by the user before submitting their selected panels). For each model, the fixed effects included the four aesthetics (color, ellipses, trendline, error band), all two-way interactions among these features, the mixture proportion λ, and interactions of λ with each feature main effect and two-way feature interaction (i.e., up to three-way interactions involving λ). We included random effects for participant, λ within lineup ID, and the interaction between the participant and lineup ID to account for any over-dispersion. We specified the model to follow the experimental design of the study and binomial response type; further, we evaluated model fit and tested for over-dispersion using the `DHARMa` library in R (Hartig, 2024). We reported significant effects at an $\alpha = 0.05$ level and display pairwise comparisons using compact letter displays after using Tukey's adjustment to control for multiplicity. In both analyses, 1855 lineup observations were included with 21 main and interaction effects for a total of 32 estimated fixed effect parameters and three estimated variances due to random effects. The specifications for the GLMM models is given by:

$$\begin{aligned}\text{logit } P(Y_{ijklmno}) = \eta &+ \delta_i + \gamma_j + \tau_k + \theta_l + \lambda_m \\ &+ (\delta\gamma)_{ij} + (\delta\tau)_{ik} + (\delta\theta)_{il} + (\gamma\tau)_{jk} + (\gamma\theta)_{jl} + (\tau\theta)_{kl} \\ &+ (\delta\lambda)_{im} + (\gamma\lambda)_{jm} + (\tau\lambda)_{km} + (\theta\lambda)_{lm} \\ &+ (\delta\gamma\lambda)_{ijm} + (\delta\tau\lambda)_{ikm} + (\delta\theta\lambda)_{ilm}\end{aligned}$$

$$+ (\gamma\tau\lambda)_{jkm} + (\gamma\theta\lambda)_{jlm} + (\tau\theta\lambda)_{klm}$$
$$+ s_n + d(\lambda)_{mo} + (sd)_{no}.$$

Where:

- η is the baseline average probability of selecting a trend target
- δ_i, γ_j, τ_k, θ_l, and λ_m is the main effects of the visual features: color ($i = 1, 2$), trendline ($j = 1, 2$), error band ($k = 1, 2$), ellipses ($l = 1, 2$), and lambda condition ($m = 1, 2, 3$), respectively.
- $(\delta\gamma)_{ij}$, $(\delta\tau)_{ik}$, $(\delta\theta)_{il}$ $(\gamma\tau)_{jk}$, $(\gamma\theta)_{jl}$, $(\tau\theta)_{kl}$ are the two-way interactions between graphical features
- $(\delta\lambda)_{im}$, $(\gamma\lambda)_{jm}$, $(\tau\lambda)_{km}$, $(\theta\lambda)_{lm}$ are the two-way interactions between λ and the graphical feature main effects
- $(\delta\gamma\lambda)_{ijm}$, $(\delta\tau\lambda)_{ikm}$, $(\delta\theta\lambda)_{ilm}$, $(\gamma\tau\lambda)_{jkm}$, $(\gamma\theta\lambda)_{jlm}$, $(\tau\theta\lambda)_{klm}$ are the three-way interactions between all two-way feature combinations and λ
- $s_n \sim \mathrm{N}\left(0, \sigma^2_{participant}\right)$ is the random effect for participant characteristics
- $d(\lambda)_{mo} \sim \mathrm{N}\left(0, \sigma^2_{lineup}\right)$ is the random interaction for lineup characteristics and λ
- $(sd)_{no} \sim \mathrm{N}\left(0, \sigma^2\right)$ is the random interaction between the nth lineup and o^{th} participant.

To examine participants' toggle interaction workflows, we used `ggplot2` (Wickham, 2016) to visualize the timing of toggle actions within each lineup evaluation, separating actions made before the first selection, between the first and second selections, and after the second selection. Additionally, we investigate favored feature combinations and individual case studies on participant workflows.

3 Results

3.1 Data Collection & Participant Characteristics

We recruited participants via Prolific (prolific.com), in April 2025. A total of 603 individuals began the study, and 564 unique participants completed it, resulting in 3,866 study trials. Occasional server connection errors caused 83 participants to reload the page, resulting in incomplete submissions or duplicate completions. In 33 cases, participants intentionally or accidentally restarted the study after finishing, and completed more than one full study. To maintain consistency, we retained only the first complete study for each participant.

Each participant completed six study trials, and each of the 12 unique lineups were evaluated between 260 and 307 times. The three λ mixture proportions were each evaluated 1,128 times, and each starting-feature combination appeared in 290–325 evaluations. The median completion time for the entire study was 7.9 minutes (IQR = 6.3 minutes). Participants spent the most time on the first practice trial, with evaluation times shortening slightly over the next few trials and then remaining fairly consistent, reflecting growing familiarity with the task while avoiding fatigue effects observed in pilot testing Figure 4.

Of the 564 participants, 64.93% were 19–35 years old; 56% identified as female, 44% as male, and two participants selected "Variant/Nonconforming." Approximately half reported working in a STEM field (46.6% STEM, 46.7% No STEM, and 6.6% Preferred not to answer).

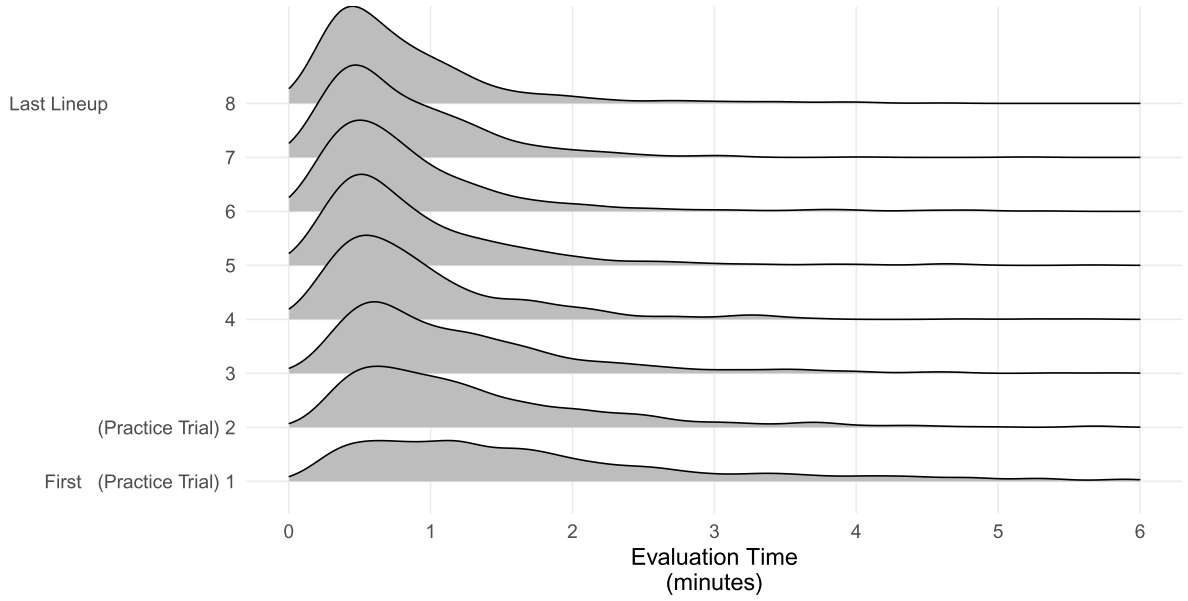


Figure 4: Amount of time (in minutes) that participants spent on each lineup, including the two practice trials.

3.2 Target Detection Accuracy

Participants were required to select at least two panels in each lineup but could select more if they wished. 2,885 evaluations involved exactly two selections, aligning with the number of targets and the training sessions. In 12.3% of evaluations, participants selected more than two panels, sometimes as many as ten. To define target detection consistently, we considered an evaluation "accurate" if one of the first two selections was a target panel.

Across the six study trials, participants identified at least one target in 4.04 trials on average, corresponding to an overall accuracy of 67.3%. This suggests that the task was challenging but achievable. Selecting both targets in the same lineup was relatively rare as 55.9% of participants never did so. The overall rate of "both correct" selections (where participants correctly identified both the cluster and trend target plots) is 11.8%, which is noticeably higher than the 0.6% reported by VanderPlas and Hofmann (2017). This difference may reflect our practice trials, which explicitly demonstrated both target types, and our requirement to make multiple selections before moving on to the next trial, which could reduce the "masking" effect observed in the earlier study. When comparing participants working in STEM fields to those who did not, accuracy rates were similar with 69.7% and 67.2%, respectively, selected at least one target. Among STEM participants, 11.1% identified both targets, 37.5% identified only the trend, and 18.7% identified only the cluster. Among non-STEM participants, 12.2% identified both, 37.5% identified only the trend, and 20.1% identified only the cluster.

When evaluating the accuracy of each lineup plot, as shown in Figure 5, participants selected the trend target more often for smaller $\lambda = 0.25$ values (surrounding null plots appear to be more clustered) and (although less drastic) the cluster target for larger $\lambda = 0.75$ values (surrounding null plots appear to follow a more linear trend). This result makes sense, as it would correspond to the cluster or trend plot being the "most different" from the null plots. When $\lambda = 0.5$, participants tended to select the linear trend more often, but not as often as $\lambda = 0.25$. Due to

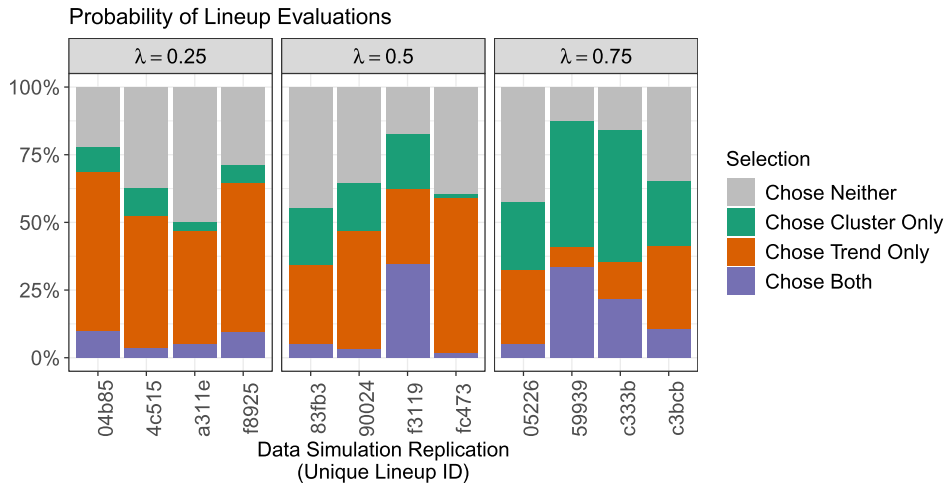


Figure 5: Proportion of evaluations selecting each outcome (cluster only, trend only, both, neither) by lineup and mixture proportion (λ). Bars sum to 100% within each lineup. At ($\lambda = 0.25$) (nulls more clustered), trend selections are more common; at ($\lambda = 0.75$) (nulls more linear), cluster selections are more common.

variability in data generation, participants were able to select both target plots more often for some of the lineups. The overall accuracy (at least one target) for each unique lineup plot varied between 50% and 80%.

3.2.1 Cluster vs. Trend Accuracy

Similar to VanderPlas and Hofmann (2017), we examined whether participants selected the linear or cluster target more often (Figure 6). Overall, participants favored the linear target, but most did not consistently choose one target type across all trials. Only 59 participants exclusively selected the linear target throughout the study, and a different 59 individuals exclusively selected the cluster target, with the majority alternating between target types across the six study trials. Only 12 participants failed to select any target in any study trial.

Starting Features The results of the GLMM including the mixture proportion (λ), starting features, and their interaction as fixed effects indicate that λ is the only significant effect in the model ($\chi^2 = 20.74$; df = 2, $p < 0.0001$). This means that we do not have enough evidence to conclude the starting features have a significant effect on a participant selecting the trend target over the cluster target, and that the choice of λ significantly changes the odds of selecting the trend target over the cluster target. Appendix C provides the full ANOVA Type III tests of fixed effects for each term included in the model.

Figure 7 shows the estimated (log) odds of successfully identifying a trend target in a lineup, given that a participant selected one target with the Tukey compact letter displays indicating discernible differences in the odds of target detection at an $\alpha = 0.05$ level and colors indicating whether the feature(s) support clusters (e.g., C, El, C + El), trends (e.g., L, Er, L + Er), or the combination is conflicting (e.g., C + L, C + El). As previously mentioned, the odds of selecting the linear trend decrease as λ increases (nulls become more linear in appearance). While we expected λ to have a significant effect on the odds of detecting the trend target over the cluster

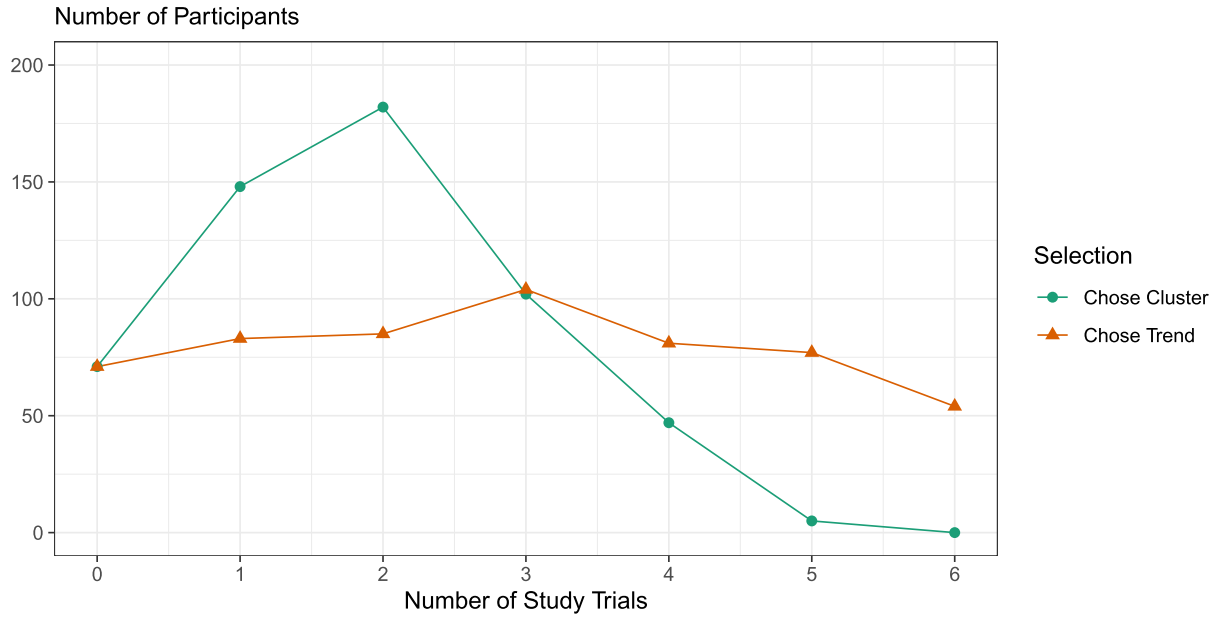


Figure 6: The number of participants by the number of study trials (0–6) in which they selected each target type. Counts include trials where both targets were selected (i.e., a participant contributes to both series in those cases).

target, the main objective of the study was to understand the effect of the four aesthetics and combinations. Although, we did not find evidence of discernible differences in the odds of trend target detection between starting features, we observe the line and error band (L + Er) starting feature combination has one of the greatest odds of selecting the trend target within each λ facet, indicating that this feature combination tends to lead participants to select the linear trend, compared to other starting feature combinations. Conversely, the color and trend line (C + L) feature combination and color (C) combination have lower odds of selecting the trend within each λ grouping, showing that these feature combinations are may be associated with selected the cluster target more often. Interestingly, for larger λ mixture proportions, the line (L) feature has one of the lowest odds, indicating we observed that participants were more likely to identify the cluster when the trendline aesthetic was shown, providing insight into the identification strategy.

Ending Features We previously defined the ending features as the aesthetics enabled by the user before submitting their selections. The results of the GLMM including the mixture proportion (λ), ending features (and their two-way interactions), and the λ interactions with ending features as fixed effects indicates error band $\times$ ellipse $\times$ λ (χ^2s = 9.38, df = 2, p = 0.009), ellipse $\times$ λ ($\chi^2 = 12.17$, df = 2, p = 0.002) are significant interactions on the odds of selecting the trend target. A significant three-way interaction indicates a change in λ significantly affects the relationship between the odds of selecting the target and ending with the error band feature and ellipses.

Particularly, Figure 7 shows the estimated (log) odds of successfully identifying the linear trend decrease as λ increases, given they selected a target. Within a $\lambda = 0.25$, we did not find evidence of a discernible difference in the odds of detecting a trend between ending features.

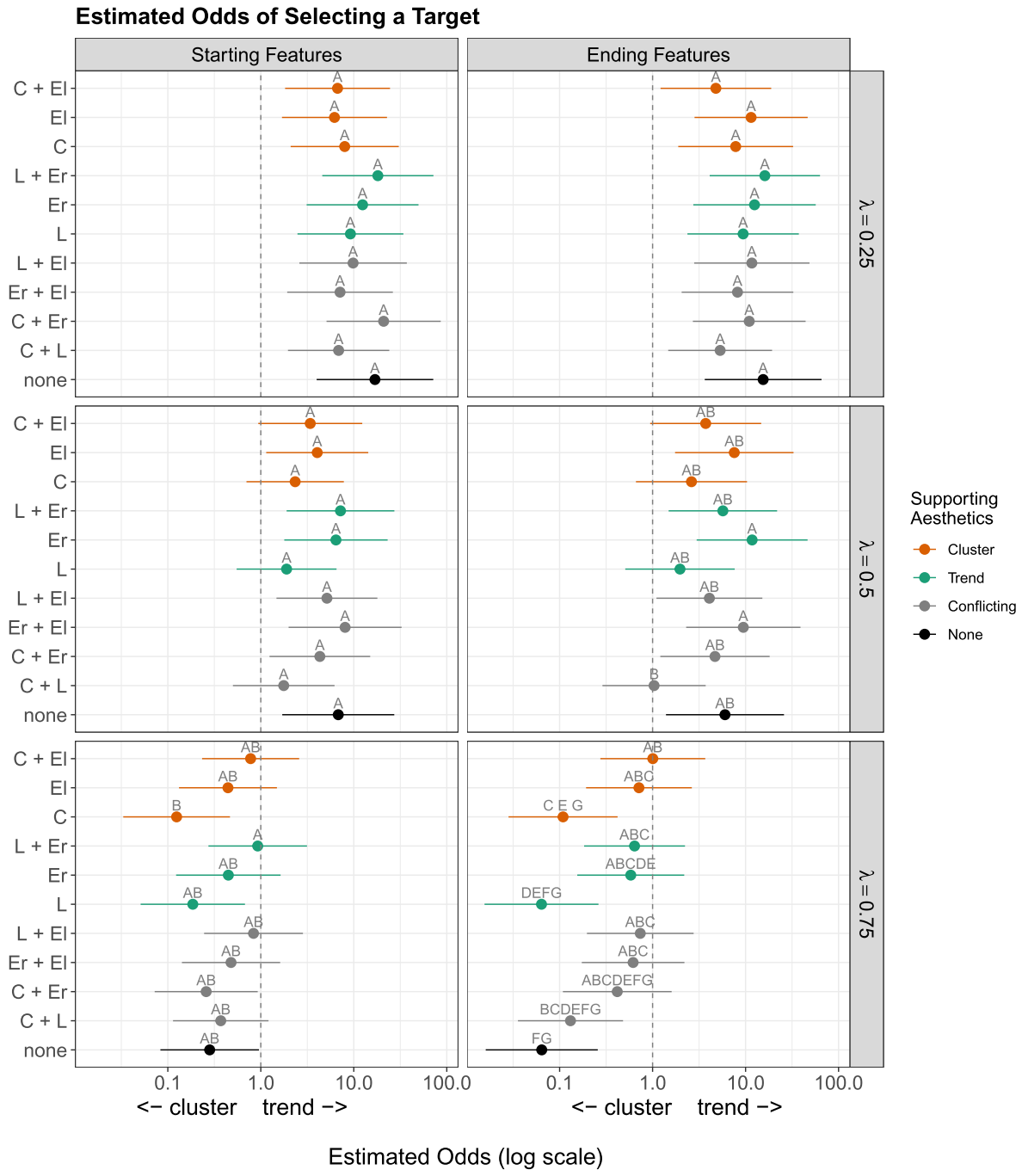


Figure 7: Estimated odds (log scale) of selecting the trend target, given that a participant selected exactly one target, by aesthetic feature combination and mixture proportion for both the Starting Feature and Ending Feature GLMMs. The points represent model estimates from the GLMM; horizontal lines show 95% confidence intervals; colors indicate whether feature(s) support clusters, trends, conflicting combination, or none. Values greater than 1 indicate higher odds of selecting the trend target, and values less than 1 indicate higher odds of selecting the cluster target. A vertical reference line at 1 indicates equal odds.

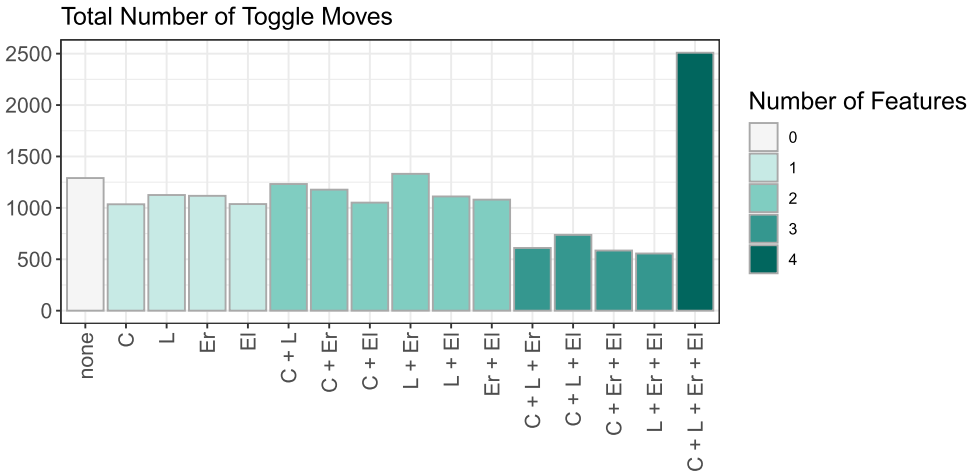


Figure 8: Counts of feature combinations toggled by participants across all study trial evaluations. The bars show the total number of times each combination was used; fill indicates the number of features enabled (0–4).

However, when $\lambda = 0.5$, we find error bars have a higher odds of detecting the trend target over the cluster target compared to the color and line (conflicting) aesthetics. Interestingly, when $\lambda = 0.75$, we found both color and line alone supported greater odds of detecting a cluster target compared to a trend target.

3.3 Interactive Toggle Behavior

A noticeable pattern in the interactive toggle data was the tendency for participants to gravitate toward two extremes. The first group, which we call maximalists, turned on all four aesthetic features and kept them on. The second group, minimalists, used few or no features throughout the trial. Across all participants and evaluations, the "all features on" combination was toggled on 2,517 times which was nearly twice as often as the next most common state, "no features," at 1,298 times (Figure 8). The maximalist preference was also evident in feature usage duration where all four features remained enabled for a total of 578 minutes across all participant evaluations, 2.7 times longer than the second longest-duration combination (Line + Error).

Case studies for the maximalist and minimalist strategies are illustrated in Figure 9 and Figure 10. The maximalist participant consistently enabled all features early in each lineup and made quick selections, rarely switching features off. They consistently selected the trend while in lineup 5 they selected the cluster but never selected both the cluster and target trends at the same time. The minimalist participant rarely had more than two features on at once, explored combinations more in early trials, and later settled on a preferred set (Line + Error), spending less time evaluating each lineup. They selected both the trend and cluster targets during the study and were able to select both target plots during their third evaluation.

While many participants fell somewhere between these two extremes, toggle use was not universal. In 44.8% of trials, participants made no toggle moves at all. When looking at when toggle moves were made, most occurred before the first selection (mean = 1.24, median = 0 toggles). Fewer toggles happened between the first and second selection (mean = 0.48) or after the second selection (mean = 0.36), indicating that interaction typically decreased once an initial

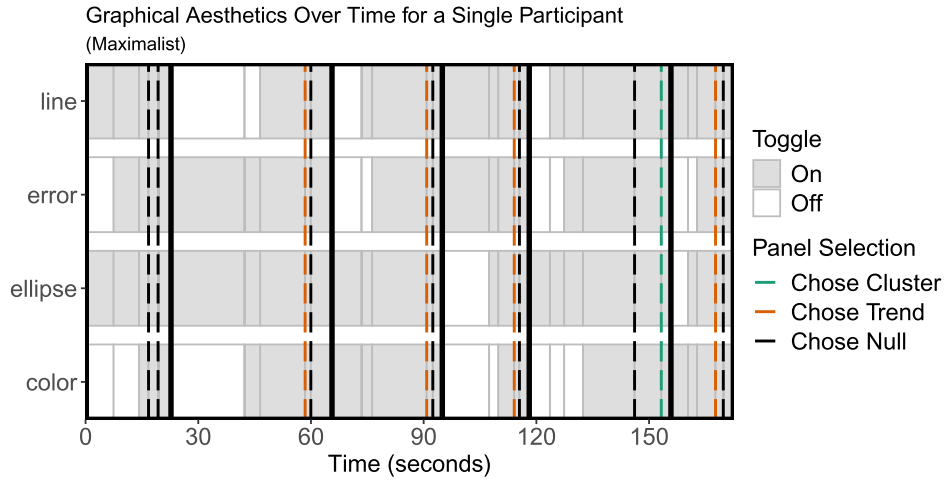


Figure 9: Maximalist workflow: This participant enabled all four aesthetic features in every lineup and rarely turned any off. Most toggles occurred early in the trial to turn features on, after which they made rapid selections. Across the study, they chose the cluster target only once, never selected both targets in the same lineup, and tended to make both selections in quick succession. In the plot, dashed vertical lines indicate when a panel was selected: green for the cluster target, orange for the trend target, and black for a null plot. Thick black vertical lines mark when the participant submitted their selections and moved to the next lineup.

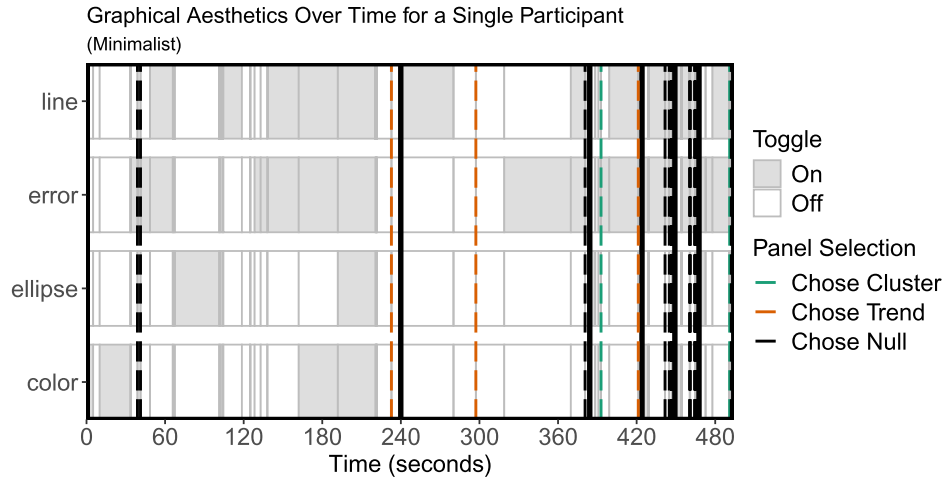


Figure 10: Minimalist workflow: This participant rarely had more than two features enabled at a time. They explored more feature combinations during their first two trials, then settled on Line + Error as their preferred set after the third lineup. They selected both the trend and cluster targets during the study, including both in the same lineup during their third trial. In the plot, dashed vertical lines indicate when a panel was selected: green for the cluster target, orange for the trend target, and black for a null plot. Thick black vertical lines mark when the participant submitted their selections and moved to the next lineup.

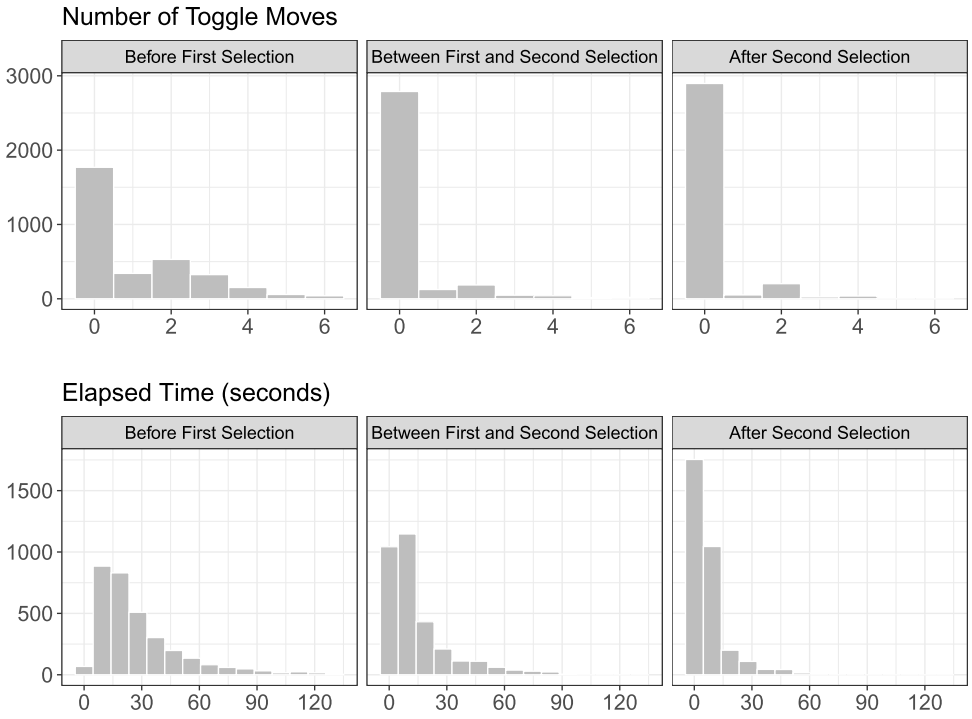


Figure 11: Distribution of (top) the number of toggle moves and (bottom) elapsed time per lineup evaluation, shown for before the first selection, between the first and second selections, and after the second selection. Most toggling occurs before the first selection, with substantially fewer changes later in the evaluation. Participants also spend the most time before the first selection, with shorter times between selections and after the second selection.

target was identified (Figure 11). This trend was mirrored in elapsed times where participants spent the most time before the first selection (mean = 32.3 seconds), less between the first and second (17.0 seconds), and the least after the second selection (9.2 seconds).

4 Discussion & Conclusion

In this study, we conducted a graphical testing experiment using statistical lineups to understand how participants' perceptions of patterns in data are influenced by various graphical aesthetics in an interactive setting. By recording toggle actions and panel selections, we captured how participants explored and interpreted visualizations when given control over the display.

We found that participants selected the linear trend target more often than the cluster target, particularly for lower λ mixture values. In VanderPlas and Hofmann (2017), participants tended to select the cluster target more often, suggesting that task context and data generation parameters can influence which patterns are more prominent. In our case, the standard deviations for both the trend and cluster data generation models were held constant at the lowest level, whereas VanderPlas and Hofmann (2017) varied them and used a $\lambda = 0.5$ only.

We did not find enough evidence to conclude that starting aesthetic features significantly affected whether participants selected the cluster or trend target. This suggests there is no single "best" default combination of features for identifying patterns in interactive exploratory graphics. Targets were easier to identify when they contrasted strongly with surrounding null plots.

For example, cluster targets were easier to detect when nulls appeared more linear (corresponding to $\lambda = 0.75$), and trend targets were easier to detect when nulls appeared more clustered (corresponding to $\lambda = 0.25$).

The most commonly used feature combination overall was having all four aesthetics enabled, reflecting a "maximalist" approach in which participants may assume that more visual cues provide more "help" in finding patterns. Toggle activity was most frequent before the first selection and decreased substantially thereafter, indicating that participants often committed to an initial impression. Many participants did not toggle features at all and relied entirely on the pre-filled starting features, further emphasizing the importance of careful default choices in interactive visualizations. Among those who toggled regularly, some individuals developed a preferred feature combination and applied it consistently across trials. Preferences varied widely, with some users favoring minimal features and others enabling everything available.

4.1 Limitations

As with most human-subject experiments, results are subject to variability in participant behavior. Some participants may not have followed instructions closely or may have disengaged during the task. To focus on meaningful toggle usage, we excluded participants who toggled in fewer than half of their trials, which reduced the dataset but improved interpretability. Non-use of toggles could be due to misunderstanding the controls (despite practice trials) or assuming that the starting features were optimal.

The sample was not fully representative of the general population, skewing younger. Prior research has found that demographic factors have minimal effect on accuracy for static lineup plots (Majumder et al., 2014). Exploratory data analysis comparing accuracy between STEM and non-STEM individuals as well as including various demographics including age, gender, STEM, and education into the GLMM analysis revealed these minimal effects of demographics on accuracy may also hold for interactive lineup studies.

4.2 Future Work

The dataset we collected during this study is rich, encompassing a wide range of variables and participant behaviors. This paper offers only an initial exploration of how participants used the interactive features. Given the lack of a defined "grammar of interactive graphics," it remains challenging to isolate and test the effects of individual components on visual perception. Nevertheless, several of our analyses, such as visualizing participant workflows, provide a strong foundation for understanding how users interact with graphical interfaces.

Future work will focus on generalizing the visual workflows across all participants to identify common toggle behaviors and frequently used feature combinations. Since participants were required to select two targets per evaluation, we also plan to use a distance metric for comparing their data plot selections within and across trials. This method of comparison was used by Chowdhury et al. (2018) to compare participant lineup selections to various measures of distance meant to quantify the similarity between plots. These distance metrics included dividing the graphical space into a grid and counting the number of points within each division, comparing regression lines within binned data, and measuring the distance between cluster means. By assigning quantitative values to assess the similarity of plots within a single lineup, we can assess whether individuals tend to focus on similar patterns over time or shift their attention toward different patterns.

A large-scale qualitative analysis will complement this quantitative work. We will code and synthesize open-ended responses alongside generalized workflow data to better understand why participants favored particular feature combinations and how they perceived their role in finding patterns.

Supplementary Material

- **Shiny App Code:** The code used to replicate the study Shiny app can be accessed at https://github.com/earobinson95/interactive-lineup-study-applet.
- **`accuracy_clean.csv`:** De-identified participant accuracy data collected in the study and used for accuracy analyses.
- **`toggles_clean.csv`:** De-identified participant toggle moves data collected in the study and used for understanding toggle workflow analyses.
- **`analysis.qmd`** The code used to replicate the analyses presented in this paper.

Appendix

A Data Simulation Algorithms

All data-generating models were simulated following the approach in VanderPlas and Hofmann (2017). Three models were used for each lineup plot: a linear trend model, a cluster model, and a mixture model. For each lineup, 20 plots were generated: one from the linear trend model, one from the cluster model, and 18 from the mixture model.

The linear trend model generates $N = 45$ points with a positive linear relationship between x and y.

The cluster model generates $N = 45$ points jittered around $K = 3$ cluster centers, with a correlation r between 0.25 and 0.75 to avoid negative correlations.

The mixture model combines points from the linear and cluster models. For each null plot, points are sampled from each model according to the mixture proportion λ. A lower λ produces more clustered points, making the linear trend more visible; a higher λ produces more linear points, making the trend stronger. $\lambda = 0$ yields a pure cluster model, $\lambda = 1$ a pure linear model, and $\lambda = 0.5$ an even mix.

Algorithm A.1 GenerateLinearTrend(N, σ_T).

Input: N – number of points; σ_T – standard deviation for the linear trend
Output: A data frame with N points $\{(x_1, y_1, k_1), (x_2, y_2, k_2), \ldots, (x_N, y_N, k_N)\}$ that follow a linear trend

1: Generate N points between $[-1, 1]$ that are evenly spaced
2: Jitter the x values: $x_i = x_i + \eta_i$, with $\eta_i \sim \text{Unif}\left(-\frac{2}{5(N-1)}, \frac{2}{5(N-1)}\right)$
3: Generate y values as a linear function of x: $y_i = x_i + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \sigma_T)$
4: Center and scale x_i and y_i
5: Use k-means clustering to compute and assign simulated points to K = 3 clusters for color and ellipse aesthetics.

return $(x_1, y_1, k_1), (x_2, y_2, k_2), \ldots, (x_N, y_N, k_N)$

Algorithm A.2 GENERATECLUSTERTREND(N, K, σ_C).

Input: N: number of points, K: the number of clusters, σ_C: the standard deviation for the clusters

Output: A data frame with N points $\{(x_1, y_1, k_1), (x_2, y_2, k_2), \ldots, (x_N, y_N, k_N)\}$, that create a cluster trend

1: Generate K cluster centers $(c_1^x, c_1^y), \ldots, (c_K^x, c_K^y)$, such that $r \in (0.25, 0.75)$.
2: Determine the number of points in each cluster, where $g = (g_1, \ldots, g_K)$ are the cluster sizes and $N = \sum_{k=1}^{K} g_k$. Select $g \sim \text{Multinomial}(N, p)$ with $p = \frac{\widetilde{p_k}}{\sum_{k=1}^{K} \widetilde{p_k}}$ and $\widetilde{p_k} \sim \text{N}\left(\frac{1}{K}, \frac{1}{2K^2}\right)$
3: Set g points to the value of its respective cluster centers, and jitter the values around the cluster centers, where $x_i = c_{g_k}^x + \epsilon_i$, where $\epsilon_i \sim \text{N}(0, \sigma_C)$ and $y_i = c_{g_k}^y + \epsilon_i$, where $\epsilon_i \sim \text{N}(0, \sigma_C)$
4: Center and scale x_i, y_i
5: Use k-means clustering to compute and assign simulated points to K = 3 clusters for color and ellipse aesthetics.

return $(x_1, y_1, k_1), (x_2, y_2, k_2), \ldots, (x_N, y_N, k_N)$

Algorithm A.3 GENERATEMIXEDMODEL(N, λ, K, σ_T, σ_C).

Input: N: number of points, K: the number of clusters, σ_C: the standard deviation for the clusters, σ_T: the standard deviation for the linear trend, λ: proportion of points from the cluster model

Output: A data frame with N points $\{(x_1, y_1, k_1), (x_2, y_2, k_2), \ldots, (x_N, y_N, k_N)\}$

1: Generate a set of data from the linear model $GenerateLinearModel(N, \sigma_T)$ and a set of data from the cluster model, $GenerateClusterModel(N, K, \sigma_C)$
2: Sample n_T points from the linear trend model, where $n_c \sim Bin(N, \lambda)$
3: Sample n_C points from the cluster trend model, where $n_C = N - n_T$
4: Combine sampled points into one dataset.
5: Use k-means clustering to compute and assign simulated points to K = 3 clusters for color and ellipse aesthetics.

return $(x_1, y_1, k_1), (x_2, y_2, k_2), \ldots, (x_N, y_N, k_N)$

B Practice Trials

We used two practice lineups, one with *similar* targets (both cluster or both trend) and one with *competing* targets (one cluster, one trend). For training salience, we reduced variability

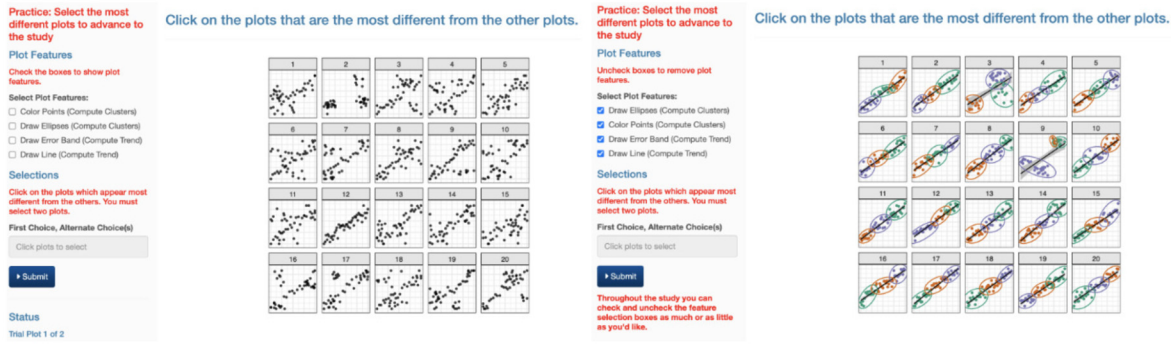


Figure B.1: Practice trial lineups: competing vs. similar signals.

($\sigma_C = 0.15$, $\sigma_T = 0.20$) and generated nulls with $\lambda = 0.5$. Participants saw notifications if they chose non-target panels during practice.

C Statistical Models (GLMMs)

To assess the statistical significance of each aesthetic visual feature, λ mixture value, and their interactions on participant accuracy of selecting a trend target over the cluster, we conducted Type III Wald chi-square tests for both the starting feature and ending feature models. The table below reports the chi-square test statistic, degrees of freedom, and associated p-values for all main effects and interactions, organized by feature combinations and their interactions with significant effects reported at an $\alpha = 0.05$ highlighted in bold. This appendix table is provided to document the full results from the models presented in the main text.

Table 1: Type III Wald χ^2 tests for starting and ending feature models. Bold entries indicate $p < .05$.

Effect	Starting model			Ending model		
	χ^2	df	p	χ^2	df	p
C	1.17	1	.280	2.19	1	.139
L	0.80	1	.372	0.05	1	.828
Er	0.18	1	.669	0.13	1	.722
El	2.22	1	.136	0.65	1	.420
$C \times L$	0.26	1	.612	0.03	1	.858
$C \times Er$	1.67	1	.196	0.80	1	.371
$C \times El$	0.81	1	.368	0.09	1	.761
$L \times Er$	1.05	1	.306	1.50	1	.221
$L \times El$	1.37	1	.242	0.68	1	.410
$Er \times El$	0.23	1	.635	0.04	1	.839
λ	**20.74**	**2**	**< .001**	3.10	2	.213
$C \times \lambda$	0.14	2	.930	2.96	2	.228
$L \times \lambda$	1.15	2	.562	0.93	2	.629
$Er \times \lambda$	0.81	2	.668	3.13	2	.209
$El \times \lambda$	3.30	2	.192	**12.17**	**2**	**.002**
$C \times L \times \lambda$	0.78	2	.676	0.01	2	.994
$C \times Er \times \lambda$	0.61	2	.737	3.16	2	.206
$C \times El \times \lambda$	0.28	2	.869	0.20	2	.904
$L \times Er \times \lambda$	0.10	2	.950	0.67	2	.715
$L \times El \times \lambda$	0.21	2	.900	0.48	2	.786
$Er \times El \times \lambda$	1.04	2	.595	**9.38**	**2**	**.009**

Note. C = color, L = line, Er = error, El = ellipse, λ = trend strength. Boldface indicates $p < .05$.

Acknowledgements

All participants read the study instructions and provided informed consent under IRB protocol #2024-222.

References

Bates D, Mächler M, Bolker B, Walker S (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1): 1–48. https://doi.org/10.18637/jss.v067.i01

Buja A, Cook D, Hofmann H, Lawrence M, Lee EK, . . . , Wickham H (2009). 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. https://doi.org/10.1098/rsta.2009.0120

Chang W (2024). Shiny - Shiny Assistant.

Chang W, Cheng J, Allaire J, Sievert C, Schloerke B, . . . , Borges B (2024). shiny: Web Application Framework for R. R package version 1.9.1.

Chowdhury NR, Cook D, Hofmann H, Majumder M (2018). Measuring lineup difficulty by matching distance metrics with subject choices in crowd-sourced data. *Journal of Computational and Graphical Statistics*, 27(1): 132–145. https://doi.org/10.1080/10618600.2017.1356323

Cleveland WS, McGill R (1987). Graphical perception: The visual decoding of quantitative information on graphical displays of data. *Journal of the Royal Statistical Society: Series A (General)*, 150(3): 192–210. https://doi.org/10.2307/2981473

Correll M, Li M, Kindlmann G, Scheidegger C (2019). Looks good to me: Visualizations as sanity checks. *IEEE Transactions on Visualization and Computer Graphics*, 25(1): 830–839. https://doi.org/10.1109/TVCG.2018.2864907

Glicksohn A, Cohen A (2011). The role of gestalt grouping principles in visual statistical learning. *Attention, Perception, & Psychophysics*, 73(3): 708–713. https://doi.org/10.3758/s13414-010-0084-4

Hartig F (2024). DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models. R package version 0.4.7.

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. Conference Name: IEEE Transactions on Visualization and Computer Graphics. https://doi.org/10.1109/TVCG.2012.230

Hullman J, Gelman A (2021). Designing for interactive exploratory data analysis requires theories of graphical inference. *Harvard Data Science Review*, 3(3): 10–1162. https://doi.org/10.1162/99608f92.3ab8a587

Komorowski M, Marshall D, Salciccioli J, Crutain Y (2016). Exploratory Data Analysis. Chapter 15 in *Secondary Analysis of Electronic Health Records*. Springer, Cham. https://doi.org/10.1007/978-3-319-43742-2_15

Lewandowsky S, Spence I (1989). The perception of statistical graphs. *Sociological Methods & Research*, 18(2–3): 200–242. https://doi.org/10.1177/0049124189018002002

Li NT, Brossard D, Scheufele DA, Wilson PH, Rose KM (2018). Communicating data: Interactive infographics, scientific data and credibility. *Journal of Science Communication*, 17. A06.

Liu S, Maljovec D, Wang B, Bremer PT, Pascucci V (2016). Visualizing high-dimensional data: Advances in the past decade. *IEEE Transactions on Visualization and Computer Graphics*, 23(3): 1249–1268. https://doi.org/10.1109/TVCG.2016.2640960

Loy A, Hofmann H, Cook D (2017). Model choice and diagnostics for linear mixed-effects models using statistics on street corners. *Journal of Computational and Graphical Statistics*, 26(3): 478–492. https://doi.org/10.1080/10618600.2017.1330207

Majumder M, Hofmann H, Cook D (2014). Human factors influencing visual statistical inference. arXiv preprint: https://arxiv.org/abs/1408.1974.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reda K, Szafir DA (2021). Rainbows revisited: Modeling effective colormap design for graphical inference. *IEEE Transactions on Visualization and Computer Graphics*, 27(2): 1032–1042. https://doi.org/10.1109/TVCG.2020.3030439

Rutter H, Parker S, Stahl-Timmins W, Noakes C, Smyth A , . . . , Freeman AL (2021). Visualising SARS-CoV-2 transmission routes and mitigations. *BMJ*, 375. e065312. https://doi.org/10.1136/bmj-2021-065312

SAS Institute Inc (2023). *JMP®, Version 18.0*. SAS Institute Inc., Cary, NC. Computer software.

Shah P, Miyake A (2005). *The Cambridge Handbook of Visuospatial Thinking*. Cambridge University Press.

Spence I (1990). Visual psychophysics of simple graphical elements. *Journal of Experimental Psychology: Human Perception and Performance*, 16(4): 683–692.

Swayne DF, Buja A (2004). Exploratory visual analysis of graphs in GGOBI. In: *COMPSTAT 2004—Proceedings in Computational Statistics: 16th Symposium Held in* Prague, Czech Republic. 2004, 477–488. Springer.

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

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

Ward M, Grinstein GG, Keim D (2021). *Interactive Data Visualization: Foundations, Techniques, and Applications*. CRC Press.

Weissgerber TL, Garovic VD, Savic M, Winham SJ, Milic NM (2016). From static to interactive: Transforming data visualization to improve transparency. *PLoS Biology*, 14(6): e1002484. https://doi.org/10.1371/journal.pbio.1002484

Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York.

Zeileis A, Hornik K, Murrell P (2009). Escaping RGBland: Selecting colors for statistical graphics. *Computational Statistics & Data Analysis*, 53(9): 3259–3270. https://doi.org/10.1016/j.csda.2008.11.033