Group beats Trend!?
Testing feature hierarchy in statistical graphics

Susan VanderPlas, Heike Hofmann*

March 8, 2015

Abstract

abstract goes here

Contents

*Department of Statistics and Statistical Laboratory, Iowa State University

1 Introduction and background

Discussion of pre-attentive visual features (Healey and Enns, 2012) - with a focus on hierarchy of pre-attentive features: color trumps shape - do we also see this in our results, and if so, by how much?

Numerical information can be difficult to communicate effectively in raw form, due to limits on attention span, short term memory, and information storage mechanisms within the human brain. Graphics are much more effective for communicating numerical information, as (well-designed) graphics order the numerical information spatially and utilize the higher-bandwidth visual system. Visual data displays serve as a form of external cognition **??**, ordering and visually summarizing data which would be hopelessly confusing in tabular format. One fantastic example of this phenomenon is the Hertzsprung-Russell (HR) diagram, which was described as "one of the greatest observational syntheses in astronomy and astrophysics" because it allowed astronomers to clearly relate the absolute magnitude of a star to its' spectral classification; facilitating greater understanding of stellar evolution (Spence and Garrison, 1993). The data it displayed was previously available in several different tables; when plotted on the same chart, information that was invisible in a tabular representation became immediately clear (Lewandowsky and Spence, 1989b). Graphical displays more efficiently utilize cognitive resources by reducing the burden of storing, ordering, and summarizing raw data; this frees bandwidth for higher levels of information synthesis, allowing observers to note outliers, understand relationships between variables, and form new hypotheses.

Graphical displays are powerful because they efficiently and effectively convey numerical information, but there exists relatively sparse empirical information about how the human perceptual system processes these displays. Our understanding of the perception of statistical graphics is informed by general psychological and psychophysics research as well as more specific research into the perception of data displays (Cleveland and McGill, 1984).

One relevant focus of psychological research is pre-attentive perception, that is, perception which occurs automatically in the first 200 ms of exposure to a visual stimulus (Treisman, 1985).

Research into preattentive perception provides us with some information about the temporal hierarchy of graphical feature processing. Color, line orientation, and shape are processed preattentively; that is, within 200 ms, it is possible to identify a single target in a field of distractors, if the target differs with respect to color or shape (Goldstein, 2009). Research by Healey and Enns (1999) extends this work, demonstrating that certain features of three-dimensional data displays are also processed preattentively. However, neither target identification nor three-dimensional data processing always translate into faster or more accurate inference about the data displayed, particularly when participants have to integrate several preattentive features to understand the data.

Feature detection at the attentive stage of perception has also been examined in the context of statistical graphics; researchers have evaluated the perceptual implications of utilizing color, fill, shapes, and letters to denote categorical or stratified data in scatterplots. Cleveland and McGill (1984) ranked the optimality of these plot aesthetics based on response accuracy, preferring colors, amount of fill, shapes, and finally letters to indicate category membership. Lewandowsky and Spence (1989a) examined both accuracy and response time, finding that color is faster and more accurately perceived (except by individuals with color deficiency). Shape, fill, and discriminable letters (letters which do not share visual features, such as HQX) were identified as less accurate than color, while confusable letters (such as HEF) result in significantly decreased accuracy.

Another area of psychological research, Gestalt psychology, examines perception as a holistic experience, establishing and evaluating mental heuristics used to transform visual stimuli into useful,

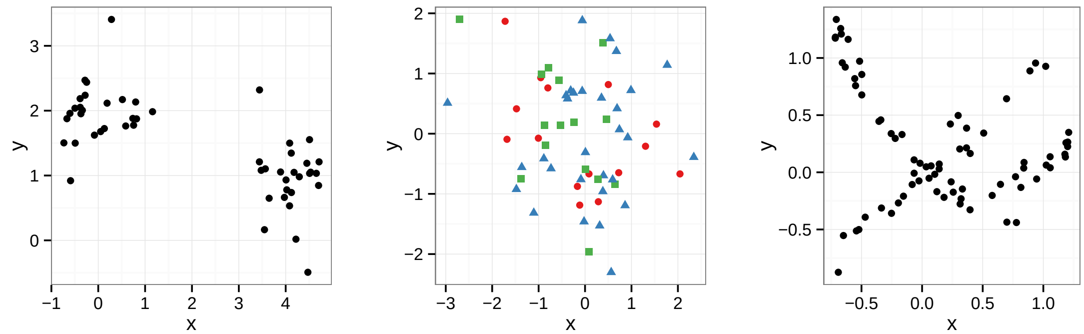

Figure 1: *Proximity* renders the fifty points of the first scatterplot as two distinct (and equal-sized) groups. Shapes and colors create different groups of points in the middle scatterplot, invoking the Gestalt principle of *Similarity*. *Good Continuation* renders the points in the scatterplot on the right hand side into two groups of points on curves: one a straight line with an upward slope, the other a curve that initially decreases and at the end of the range shows an uptick.

coherent information. Gestalt rules of perception can be easily applied to statistical graphics, as they describe the way we organize visual input, focusing on the holistic experience rather than the individual perceptual features.

For example, rather than perceiving four legs, a tail, two eyes, two ears, and a nose, we perceive a dog. The rules of perceptual grouping or organization, as stated in Goldstein (2009) are:

- Proximity: two elements which are close together are more likely to belong to a single unit.
- Similarity: the more similar two elements are, the more likely they belong to a single unit.
- Common fate: two elements moving together likely belong to a single unit.
- Good continuation: two elements which blend together smoothly likely belong to one unit.
- Closure: elements which can be assembled into closed or convex objects likely belong together.
- Common region: elements contained within a common region likely belong together.
- Connectedness: elements physically connected to each other are more likely to belong together.

The plots in figure 1 demonstrate several of the gestalt principles which combine to order our perceptual experience from the top down. These laws help to order our perception of charts as well: points which are colored or shaped the same are perceived as belonging to a group (similarity), points within a bounding interval or ellipse are perceived as belonging to the same group (common region), and regression lines with confidence intervals are perceived as single units (connectedness, closure, and/or common region).

clarify next sentence

The use of physical location, color, and shape to organize graphical units mentally utilizes both preattentive processing and higher-order gestalt schemas, identifying and grouping similar graphical features and simultaneously directing attention to graphical features which stand alone.

Research on preattentive perception is important because features that are perceived preattentively do not require as much mental effort to process from raw visual stimuli; theoretically, subsequent top-down gestalt heuristics can be applied to such stimuli more quickly.

This paper describes the results of a user study designed to explore the hierarchy of gestalt principles in perception of statistical graphics. We utilize information from previous studies (Çağatay Demiralp et al., 2014; Robinson, 2003) concerning the hierarchy of preattentive feature perception in order to maximize the effect of preattentive feature differences.

might be useful to have a small diagram describing the perceptual process (with preattentive processing way at the top and gestalt heuristic processing in the middle, with "cognitive effort" at the bottom). Not sure if it's necessary, though. HH: good idea, let's see how much space we'll have.

Statistical graphics can be difficult to examine experimentally; qualitative studies rely on descriptions of the plot by participants who may not be able to articulate their observations precisely, while quantitative studies may only be able to examine whether the viewer can accurately read numerical information from the chart, instead of exploring the overall utility of the data display holistically. Statistical lineups, described in the next section, are an important experimental tool for evaluating the perceptual utility of graphical displays. Lineups fuse commonly used psychological tests (target identification, visual search)

XXX cite the ACM Transact submission here

with statistical hypothesis tests to facilitate formal experimental evaluation of statistical graphics.

Statistical Lineups

Intro to lineups (**?**Majumder et al., 2013; **?**; **?**).

Describe the lineup protocol, including basic statistics. Link to the psychological "target and distractors" approach, which can be used to justify the addition of a second target, even with the PITA of the statistical complications.

In this study, we modify the lineup protocol by introducing a second target to each lineup. The two targets represent two different, competing signals; the participant's choice then demonstrates empirically which signal is more salient. If both targets exhibit similar signal, participants may identify both targets, removing any forced-choice scenario which might skew results (few participants exercised this option).

By tracking the proportion of observers choosing either target plot (a measure of overall lineup difficulty) as well as which proportion of observers choose one target over the other target, we can determine the relative strength of the two competing signals amid a field of distractors. At this level, signal strength is determined by the experimental data and the generating model; we are measuring the "power" (in a statistical sense) of the human perceptual system, rather than raw numerical signal.

Using this testing framework, we apply different aesthetics, such as color and shape, as well as plot objects which display statistical calculations, such as trend lines and bounding ellipses. These additional plot layers, discussed in more detail in the next section, are designed to emphasize one of the two competing targets and affect the overall visual signal of the target plot relative to the null plots. We expect that in a situation similar to the third plot of figure 1, the addition of two trend lines would emphasize the "good continuation" of points in the plot, producing a stronger visual signal, even though the underlying data has not changed. Similarly, the grouping effect in the first plot in the figure would be enhanced if the points in each group were colored differently, as the proximity heuristic would be supplemented by similarity. In plots that are ambiguous, containing some clustering of points as well as a linear relationship between x and y, additional aesthetic cues may "tip the balance" in favor of recognizing one type of signal.

This study is designed to inform our understanding of the perceptual implications of these additional aesthetics, in order to provide guidelines for the creation of data displays which provide visual cues consistent with gestalt heuristics and preattentive perceptual preferences.

We will discuss the experimental design, model considerations, and graphical features compared experimentally in section 2. The next section discusses the particulars of the experimental design, including the data generation model, plot aesthetics, selection of color and shape palettes, and other important considerations. Experimental results are presented in section 3, and implications and conclusions are discussed in section 4.

2 Experimental Setup and Design

In this section, we discuss the generating data models for the two types of signal plots and the null plots, the selection of plot aesthetic combinations and aesthetic values, and the design and execution of the experiment.

I know this will have to be rearranged, expanded, and transitions between sections will need to be added, but I want to get the paragraphs out.

2.1 Data Generation

Lineups require a single "target" data set (which we are expanding to two competing "target" data sets), and a method for generating null plots. When utilizing real data for target plots, null plots are often generated through bootstrap sampling, but this introduces some dependencies between target and null plots which complicate the statistical analysis of the results.

add citations

When possible, it is desireable to generate true null plots, which are generated from the null model and do not depend on the data used in the target plot. This experiment will measure two competing gestalt heuristics, proximity and good continuation, using two data-generating models: M_C, which generates data with K clusters, and M_T, which generates data with a positive correlation between x and y. True null datasets are created using a mixture model M_0 which combines M_C and M_T. Both M_C and M_T generate data in the same range of values. Additionally, M_C generates clustered data with linear correlations that are within $\rho = (0.25, 0.75)$, similar to the linear relationship between datasets generated by M_0, and M_T generates data with clustering similar to M_0. These constraints provide some assurance that participants who select a plot with data generated from M_T are doing so because of visual cues indicating a linear trend (rather than a lack of clustering compared to plots with data generated from M_0), and participants who select a plot with data generated from M_C are doing so because of visual cues indicating clustering, rather than a lack of a linear relationship relative to plots with data generated from M_0.

2.1.1 Regression Model M_T

This model has the parameter σ_T to reflect the amount of scatter around the trend line. It generates N points $(x_i, y_i), i = 1, ..., N$ where x and y have a positive linear relationship. The data generation mechanism is as follows:

Algorithm 2.1
Input Parameters: sample size N, σ_T standard deviation around the line
Output: N points, in form of vectors x and y.

1. *Generate $\tilde{x}_i$, $i = 1, ..., N$, as a sequence of evenly spaced points from $[-1, 1]$.*
2. *Jitter $\tilde{x}_i$ by adding small uniformly distributed perturbations to each of the values: $x_i = \tilde{x}_i + \eta_i$, where $\eta_i \sim$ Unif$(-z, z)$, $z = \frac{2}{5(N-1)}$.*
3. *Generate y_i as a linear regressand of x_i: $y_i = x_i + e_i$, $e_i \sim N(0, \sigma_T^2)$.*
4. *Center and scale x_i, y_i.*

We compute the coefficient of determination for all of the plots to assess the amount of linearity in each panel, computed as

$$R^2 = 1 - \frac{RSS}{TSS}, \tag{1}$$

where TSS is the total sum of squares, $TSS = \sum_{i=1}^{N} (y_i - \bar{y})^2$ and $RSS = \sum_{i=1}^{N} e_i^2$, the residual sum of squares. The expected value of the coefficient of determination $E\left[R^2\right]$ in this scenario is

$$E\left[R^2\right] = \frac{1}{1 + 3\sigma_T^2},$$

because $E[RSS] = N\sigma_T^2$ and $E[TSS] = \sum_{i=1}^{N} E\left[y_i^2\right]$ (as $E[Y] = 0$), where

$$E\left[y_i^2\right] = E\left[x_i^2 + e_i^2 + 2x_i e_i\right] = \frac{1}{3} + \sigma_T^2.$$

The use of R^2 to assess the strength of the linear relationship (rather than the correlation) is indicated because human perception of correlation strength more closely aligns with R^2 (Bobko and Karren, 1979; Lewandowsky and Spence, 1989b).

2.1.2 Cluster Model M_C

We begin by generating K cluster centers on a $K \times K$ grid, then we generate points around selected cluster centers.

Algorithm 2.2
Input Parameters: N points, K clusters, σ_C cluster standard deviation
Output: N points, in form of vectors x and y.

1. *Generate cluster centers (c_i^x, c_i^y) for each of the K clusters, $i = 1, ..., K$:*
   (a) *in form of two vectors c^x and c^y of permutations of $\{1, ..., K\}$, such that*

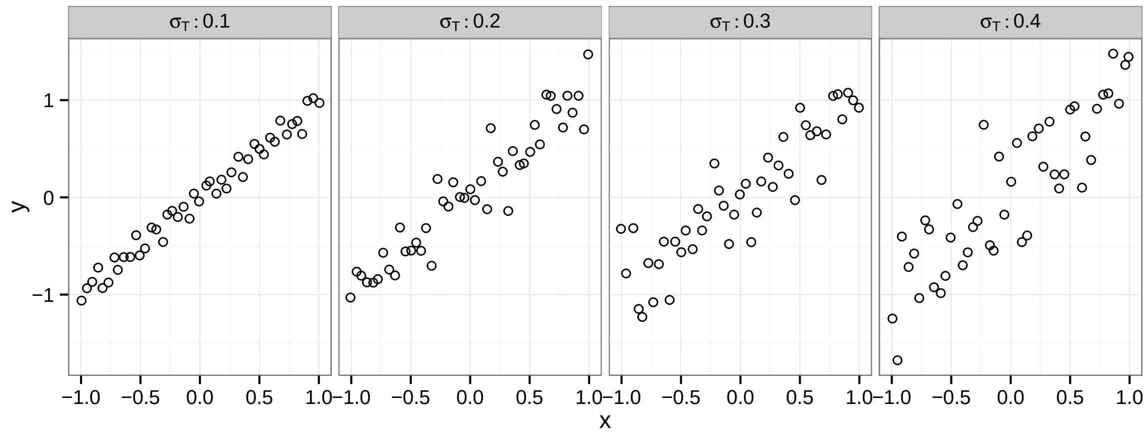

Figure 2: Set of scatterplots showing one draw each from the trend model M_T for parameter values of $\sigma_T \in \{0.1, 0.2, 0.3, 0.4\}$.

(b) the correlation between cluster centers $cor(c^x, c^y)$ falls into a range of $[.25, .75]$.

2. *Center and standardize cluster centers (c^x, c^y):*

$$\tilde{c}_i^x = \frac{c_i^x - \bar{c}}{s_c} \quad and \quad \tilde{c}_i^y = \frac{c_i^y - \bar{c}}{s_c},$$

where $\bar{c} = (K+1)/2$ and $s_c^2 = \frac{K(K+1)}{12}$ for all $i = 1, ..., K$.

3. *For the K clusters, we want to have nearly equal sized groups, but allow some variability. Group sizes $g = (g_1, ..., g_K)$ with $N = \sum_{i=1}^K g_i$, for clusters $1, ..., K$ are therefore determined as a draw from a multinomial distribution:*

$$g \sim Multinomial\ (K, p)\ where\ p = \tilde{p} / \sum_{i=1}^{K} \tilde{p}_i,\ for\ \tilde{p} \sim N\left(\frac{1}{K}, \frac{1}{2K^2}\right).$$

4. *Generate points around cluster centers by adding small normal perturbations:*

$$\begin{aligned} x_i &= \tilde{c}_{g_i}^x + e_i^x,\ where\ e_i^x \sim N(0, \sigma_C^2), \\ y_i &= \tilde{c}_{g_i}^y + e_i^y,\ where\ e_i^y \sim N(0, \sigma_C^2). \end{aligned}$$

5. *Center and scale x_i, y_i.*

As a measure of clustering we use a coefficient to assess the amount of variability within groups, compared to total variability. Note that for the purpose of clustering, variability is measured as the variability in both x and y from a common mean, i.e. we implicitly assume that the values in x and y are on the same scale (which we achieve by scaling in the final step of the generation algorithm).

For two numeric variables x and y and grouping variable g with $g_i \in \{1, ..., K\}, i = 1, ..., n$, we compute the *cluster index* as follows: let $j(i)$ be the function that maps index $i = 1, ..., n$ to one of the clusters $1, ..., K$ given by the grouping variable g. Then for each level of g, we find a cluster center as $\bar{x}_{j(i)}$ and $\bar{y}_{j(i)}$, and we determine the strength of the clustering by comparing the within cluster variability with the overall variability:

$$\begin{aligned} \gamma^2 &= 1 - \frac{CSS}{TSS}, \\ CSS &= \sum_{i=1}^{n} \left(x_{j(i)} - \overline{x}_{j(i)}\right)^2 + \left(y_{j(i)} - \overline{y}_{j(i)}\right)^2, \\ TSS &= \sum_{i=1}^{n} (x_i - \bar{x})^2 + (y_i - \bar{y})^2 . \end{aligned} \tag{2}$$

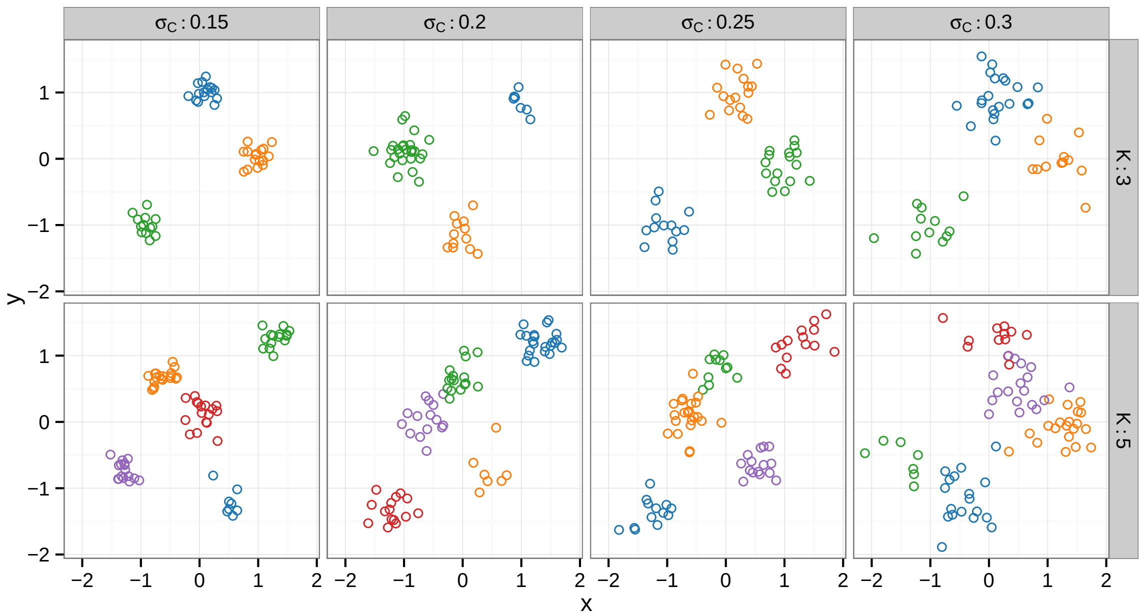

Figure 3: Scatterplots of clustering output for different inner cluster spread σ_C (left to right) and different number of clusters K (top and bottom).

Could you include the same colors and shapes as in our lineups?

2.1.3 Null Model M_0

The generative model for null data is a mixture model M_0 that draws $n_c \sim \text{Binomial}(N, \lambda)$ observations from the cluster model, and $n_T = N - n_c$ from the regression model M_T. Observations are assigned groups using hierarchical clustering, which creates groups consistent with any structure present in the generated data. This provides a plausible grouping for use in aesthetic and statistics requiring categorical data (color, shape, bounding ellipses).

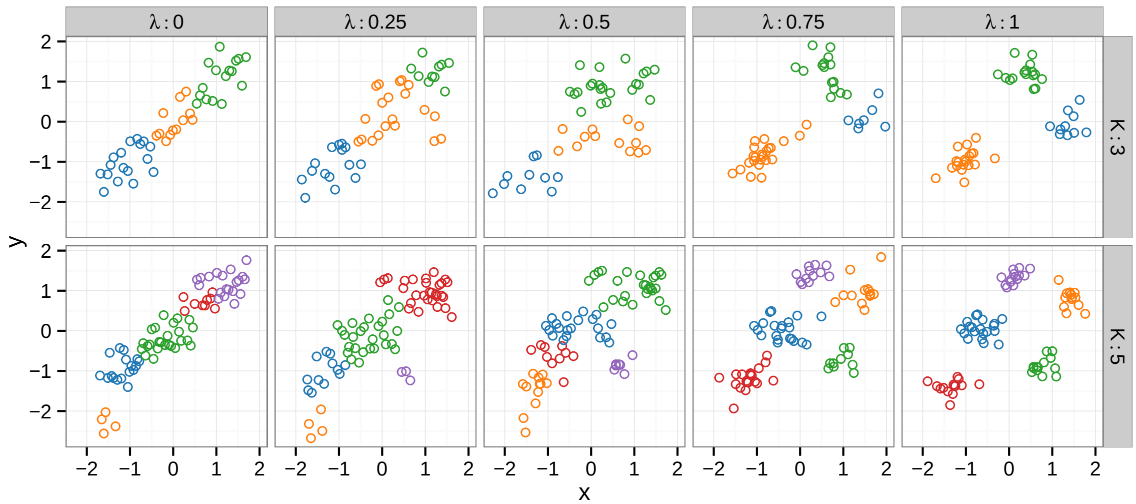

Figure 4: Scatterplots of data generated from M_0 using different values of λ.

Null data in this experiment is generated using $\lambda = 0.5$, that is, each point in a null data set is equally likely to have been generated from M_C and M_T.

2.1.4 Parameters used in Data Generation

These models provide the foundation for this experiment; by manipulating cluster standard deviation σ_C and regression standard deviation σ_T (directly related to correlation strength) for varying numbers of clusters $K = 3, 5$, we can systematically control the statistical signal present in the target plots and generate corresponding null plots that are mixtures of the two distributions. For each parameter set $\{K, N, \sigma_C, \sigma_T\}$, as described in table 1, we generate a lineup dataset consisting of one set drawn from M_C, one set drawn from M_T, and 18 sets drawn from M_0.

Parameter	Description	Choices
K	# Clusters	3, 5
N	# Points	$15 \cdot K$
σ_T	Scatter around trend line	.15, .25, .35
σ_C	Scatter around cluster centers	.15, .20, .25 ($K = 3$) .20, .25, .30 ($K = 5$)

Table 1: Parameter settings for generation of lineup datasets.

The parameter values were chosen after examining the full parameter space through simulation of 1000 lineup datasets for each combination of $\sigma_T \in \{0.2, 0.25, ..., 0.5\}$, $\sigma_C \in \{0.1, 0.15, ..., 0.4\}$, and $K \in \{3, 5\}$; for each data set generated, the previously described statistics for trend and cluster strength were computed. We compared the statistics for the relevant target plot to the most extreme value for the 18 null plots. These evaluations provide an estimate of the difficulty of identifying the target plot numerically; a target plot with $R^2 = 0.95$ would be very easy to identify when surrounded by null plots with $R^2 = 0.5$, while null plots with $R^2 = 0.9$ would make the target a lot more difficult. Graphical summaries of simulation results are provided in appendix A.

Using information from the simulation, we identified values of σ_T and σ_C corresponding to "easy", "medium" and "hard" numerical comparisons between corresponding target data sets and null data sets. It is important to note that the numerical measures we have described in equations (1) and (2) only provide information on the numerical discriminability of the target datasets from the null datasets; the simulation cannot provide us with information on the perceptual discriminability, and it has been established that human perception of scatterplots does not replicate statistical measures exactly (Bobko and Karren, 1979; Mosteller et al., 1981; Lewandowsky and Spence, 1989b).

Each of the generated datasets is then plotted as a lineup, where we apply aesthetics which emphasize clusters and/or linear relationships, to experimentally determine how these aesthetics change participants' ability to identify each target plot. The next section describes the aesthetic combinations and their anticipated effect on participant responses.

2.2 Lineup Rendering

2.2.1 Plot Aesthetics

Gestalt perceptual theory suggests that perceptual features such as shape, color, trend lines, and boundary regions modify the perception of ambiguous graphs, emphasizing clustering in the data (in the case of shape, color, and bounding ellipses) or linear relationships (in the case of trend lines and prediction intervals), as demonstrated in figure 1. For each dataset we examine the effect of plot aesthetics (color, shape) and statistical layers (trend line, boundary ellipses, prediction intervals) shown in table 2 on target identification. Examples of these plot aesthetics are shown in figure 5.

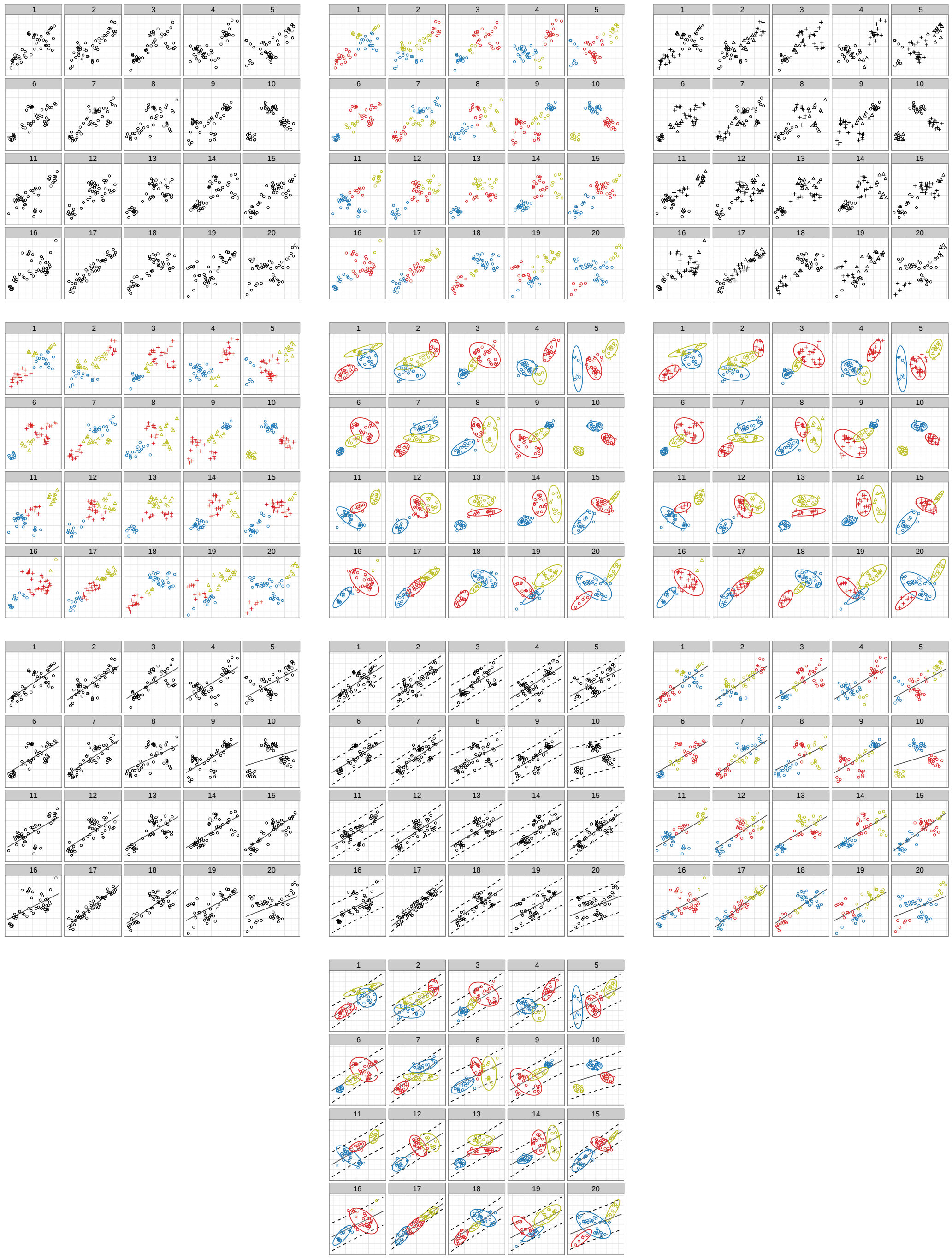

Figure 5: Each of the 10 plot feature combinations tested in this study, with $K = 3$, $\sigma_T = 0.25$ and $\sigma_C = 0.20$.

		Line Emphasis		
	Strength	0	1	2
Cluster Emphasis	0	None	Line	Line + Prediction
	1	Color Shape	Color + Line	
	2	Color + Shape Color + Ellipse		Color + Ellipse + Line + Prediction
	3	Color + Shape + Ellipse		

Table 2: Plot aesthetics and statistical layers which impact perception of statistical plots, according to gestalt theory.

We expect that relative to a plot with no extra aesthetics or statistical layers, the addition of color, shape, and 95% boundary ellipses increases the probability of a participant selecting the target plot with data generated from M_C, the cluster model, and that the addition of these aesthetics decreases the probability of a participant selecting the target plot with data generated from M_T, the linear model.

Similarly, we expect that relative to a plot with no extra aesthetics or statistical layers, the addition of a trend line and prediction interval increases the probability of a participant selecting the target plot with data generated from M_T, the linear model, and decreases the probability of a participant selecting the target plot with data generated from M_C, the cluster model.

2.2.2 Experimental Design

The study is designed hierarchically, as a factorial experiment for combinations of σ_C, σ_T, and K, with three replicates at each parameter combination. These parameters are used to generate lineup datasets which serve as blocks for the plot aesthetic level of the experiment; each dataset is rendered with every combination of aesthetics described in table 2. Participants are assigned to generated plots according to an augmented balanced incomplete block scheme: each participant is asked to evaluate 10 plots, which consist of one plot at each combination of σ_C and σ_T, randomized across levels of K, with one additional plot providing replication of one level of $\sigma_C \times \sigma_T$. Each of a participant's 10 plots will present a different aesthetic combination.

Need to find some graphic/table which makes this a bit more clear.

2.2.3 Color and Shape Palettes

Colors and shapes used in this study were selected in order to maximize preattentive feature differentiation. Çağatay Demiralp et al. (2014) provide sets of 10 colors and 10 shapes, with corresponding distance matrices, determined by user studies. Using these perceptual kernels for shape and color, we identified sets of 3 and 5 colors and shapes which maximize the sum of pairwise differences, subject to certain constraints imposed by software and accessibility concerns.

The color palette used in Çağatay Demiralp et al. (2014) and shown in figure 6 is derived from colors available in Tableau visualization software.

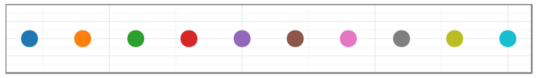

Figure 6: Colors in Çağatay Demiralp et al. (2014). This study removed gray from the palette to make the experiment more inclusive of participants with colorblindness.

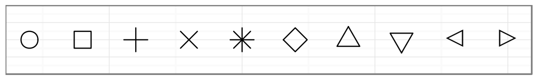

Figure 7: Shapes in Çağatay Demiralp et al. (2014). In order to control for varying point size due to Unicode vs. non-Unicode characters, the last two shapes were removed.

citation for Tableau?

In order to produce experimental stimuli accessible to the approximately 4% of the population with red-green color deficiency (Gegenfurtner and Sharpe, 2001), we removed the gray hue from the palette. This modification produced maximally different color combinations which did not include red-green combinations, while also removing a color (gray) which is difficult to distinguish for those with color deficiency.

Software compatibility issues led us to exclude two shapes used in Çağatay Demiralp et al. (2014) and shown in figure 7. The left and right triangle shapes (available only in unicode within R) were excluded due to size differences between unicode and non-unicode shapes. After optimization over the sum of all pairwise distances, the maximally different shape sequences for the 3 and 5 group datasets also conform to the guidelines in Robinson (2003): for $K = 3$ the shapes are from Robinson's group 1, 2, and 9, for $K = 5$ the shapes are from groups 1, 2, 3, 9, and 10. Robinson's groups are designed so that shapes in different groups show differences in preattentive properties; that is, they are easily distinguishable. In addition, all shapes are non-filled shapes, which means that they are consistent with one of the simplest solutions to overplotting of points in the tradition of Tukey (1977); Cleveland (1994) and Few (2009). For this reason we abstained from the additional use of alpha-blending of points to diminish the effect of overplotting in the plots.

2.3 Hypotheses

The primary purpose of this study is to understand how visual aesthetics affect signal detection in the presence of competing signals. We expect that plot modifications which emphasize similarity and proximity, such as color, shape, and 95% bounding ellipses, will increase the probability of detecting the clustering relationship, while plot modifications which emphasize good continuation, such as trend lines and prediction intervals, will increase the probability of detecting the linear relationship.

A secondary purpose of the study is to relate signal strength (as determined by dataset parameters σ_C, σ_T, and K) to signal detection in a visualization by a human observer.

2.4 Participant Recruitment

Participants were recruited using Amazon's Mechanical Turk service, which connects interested workers with "Human Intelligence Tasks" (HITs), which are (typically) short tasks which cannot be easily automated. Only workers with at least 100 previous HITs at a 95% successful completion rate were allowed to sign up for completing the task. These restrictions reduce the amount of data cleaning required by ensuring that participants have experience with the Mechanical Turk system.

Participants were asked to complete an example task similar to the task in the experiment before deciding whether or not to complete the HIT. The lineups used as examples contained only one target (5 trend and 5 cluster trials were provided), and participants had to correctly identify target plots in at least two lineups before being allowed into the HIT and proceeding to the experimental phase. The webpage used to collect data from Amazon Turk participants is available at `http://www.mlcape.com:8080/mahbub/turk16/index.html`. No data was recorded from the example task because participants had not yet provided informed consent.

Once participants completed the example task and provided informed consent, they could accept the HIT through Amazon and were directed to the main experimental task. Participants were asked to complete 10 lineups, answering "Which plot is the most different from the others?". Participants were also asked to provide a short summary of their reasoning, such as "Strong linear trend" or "Groups of points", and to rate their confidence in their selection from 1 (least confident) to 5 (most confident). After the first question, participants were asked to provide basic demographic information: age range, gender, and highest level of education. Upon completion of 10 lineups, participants were provided with a text code to enter into Amazon Turk in order to receive the advertised payment of $1.00. Participants took (on average) 8 minutes to complete the HIT (not including the required example task).

3 Results

3.1 General results

It would also be interesting to see, if individuals picked mostly lines or mostly groups or whether there was a mix of both.

Data collection was conducted over a 24 hour period, during which time 1356 individuals completed 13519 unique lineup evaluations. Participants who completed fewer than 10 lineups were removed from the study (159 participants, 1060 evaluations), and lineup evaluations in excess of 10 for each participant were also removed from the study (421 evaluations).

After these data filtration steps, our data consist of 12010 trials completed by 1200 participants.

$12010 \neq 1200 \times 10$ something went wrong.

Of the participants who completed at least 10 lineup evaluations, 61% were male, and 52 % were between 18 and 30 years of age.

Participants were fairly well-educated: 62% had at least some undergraduate education, and 30% had at least some graduate education.

Each plot was evaluated by between 11 and 37 individuals (Mean: 22.24, SD= 4.62). 82.7% of the participant evaluations identified at least one of the two target plots successfully (Trend:

26.6%, Cluster: 56.7%). Only 2.9% of participant evaluations identified more than one target plot, and of these multiple identifications, 22.7% identified both targets correctly.

We first consider the effect of plot aesthetics on target selection for each target type (separately), and then we will analyze the effect of parameter values on participant performance.

The two models (1) and (2) have the exact same shape, so just discuss the structure of the model first, and then report the results for the two different dependent variables in the subsections.

3.2 Linear Target Model

We will model the probability of selecting the linear target plot as a function of plot type, with random effects for dataset (which encompasses parameter effects) and participant (accounting for variation in individual skill level). For plot type $i = 1, ..., 10$ displaying dataset $j = 1, ..., 54$ by participant $k = 1, ..., P$,

We model the probability of selecting the linear target plot as a function of plot type, with random effects for dataset (which encompasses parameter effects) and participant (accounting for variation in individual skill level). For plot type $i = 1, ..., 10$ displaying dataset $j = 1, ..., 54$ by participant $k = 1, ..., P$,

$$P(\text{success}) = \left(e^{\theta}\right) / \left(1 + e^{\theta}\right) \tag{3}$$

$$\theta = \mathbf{X}\beta + \mathbf{J}\gamma + \mathbf{K}\eta + \epsilon \tag{a}$$

$$\begin{aligned} \text{where } \beta_i \quad & \text{describe the effect of specific plot aesthetics} \\ \gamma_j \overset{iid}{\sim} & N\left(0, \sigma^2_{\text{data}}\right), \text{ the random effect for dataset specific characteristics} \\ \eta_k \overset{iid}{\sim} & N\left(0, \sigma^2_{\text{participant}}\right), \text{ the random effect for participant characteristics} \\ \text{and } \epsilon_{ijk} \overset{iid}{\sim} & N\left(0, \sigma^2_{e}\right), \text{ the error associated with a single trial evaluation} \end{aligned}$$

We note that any variance due to parameters K, σ_T, and σ_C is contained within σ^2_{data} and can be examined using a subsequent model.

We examine the probability of selecting the group target plot as a function of plot type, with random effects for dataset (which encompasses parameter effects) and participant (accounting for variation in individual skill level). For plot type $i = 1, ..., 10$ displaying dataset $j = 1, ..., 54$ by participant $k = 1, ..., P$,

3.3 Group Target Selection

We now examine the probability of selecting the group target plot as a function of plot type, with random effects for dataset (which encompasses parameter effects) and participant (accounting for variation in individual skill level). The model fit here is the same as that shown in equation (3), except that success in this model is defined as identification of the cluster target plot.

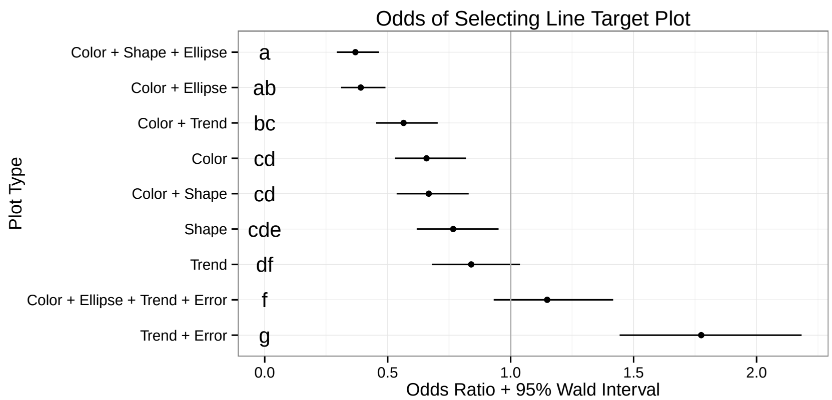

Figure 8: Odds ratios describing the odds of detecting the linear target plot for each aesthetic, relative to a plain scatterplot. Only the combination of Trend + Error significantly increases the odds of linear target plot detection relative to the control plot.

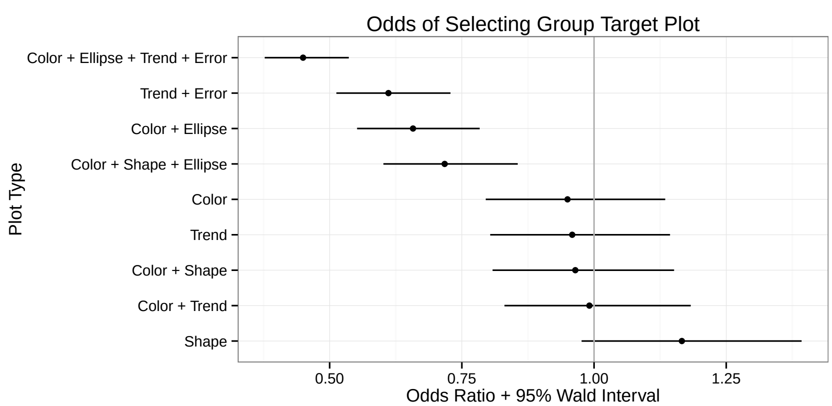

Figure 9: Odds ratios describing the odds of detecting the cluster target plot for each aesthetic, relative to a plain scatterplot. The presence of error lines or bounding ellipses significantly decreases the probability of correct target detection, and no aesthetic successfully increases the probability of correct target detection. This may be due to differences in group size for null plots, with data generated under M_0 compared with the group target plot displaying data generated under M_C.

3.4 Face-Off: Group versus Line

Just another idea of evaluating this data set: For each data set we only consider those evaluations that correctly identify one of the targets. This reveals that participants overall favored groups to lines at a ratio of about 2:1. We remove this overall effect using an intercept, and model group vs line decisions using a logistic regression with a random effect for each dataset to account for different difficulty levels in the generated data. The estimated odds of a decision in favor of group over line target are shown in figure 10. From left to right the odds of selecting the group target over the line target increase. As hypothesized, the strongest signal for identifying groups, is color + shape + ellipse, while trend + error results in the strongest signal in favor of trends. Most of the effects are not significantly different (see the letter values Piepho (2004) on the left hand side of the figure, representing pairwise comparisons of all of the designs, adjusted for multiple comparison). Trend+error plots and ellipse+trend plots are significantly different from all of the other designs. Apart from that, the only significant difference between designs is between color+shape+ellipse plots and trend plots. Surprisingly(?), the two designs sending mixed signals (color + trend, color+ ellipse+trend) end on opposite ends of the scale.

My take on this: HH: I like it, let's keep this!
The similarity/proximity effect (as indicated by clustering and color/shape/etc.) dominates the equation, including dominating the color+trend (good continuation) condition.
The addition of common region (ellipses, error bars) modifies this effect somewhat, reinforcing the clustering and trend when present (as those are at the extreme ends of the plot).
When (trend + error) are present in the same plot, you get additional gestalt ordering principle(s): common region + good continuation **+ connectedness + closure**:
since the error band would be perceived as containing the points and connected to the trendline - this does somewhat lessen with the lines for error bands we have shown here, but I think the point still holds.
Plus, you could argue for closure, since you have fairly symmetric lines on either side of a central object.
(color + ellipse) + clustering = (similarity + common region) + proximity
is not as strong as
(trend + error) + correlation = (good continuation (trendline) + common region + connectedness + closure) + good continuation (points)

3.5 Signal Strength

4 Discussion

References

Bobko, P. and Karren, R. (1979), "The perception of Pearson product moment correlations from bivariate scatterplots," *Personnel Psychology*, 32, 313–325.

Çağatay Demiralp, Bernstein, M., and Heer, J. (2014), "Learning Perceptual Kernels for Visualization Design," *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis).*

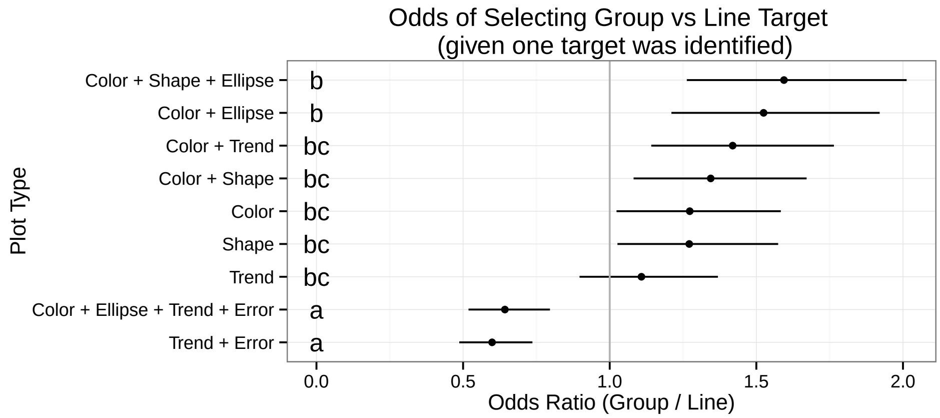

Figure 10: Estimated odds of decision for group versus line target based on evaluations that resulted in the identification of one of these targets. Plot types are significantly different, if they do not share a letter as given on the left hand side of the plot.

CHowdhury, N. R., Cook, D., Hofmann, H., Majumder, M., and Zhao, Y. (2014), "Utilizing Distance Metrics on Lineups to Examine What People Read From Data Plots," *arXiv.org*.

Cleveland, W. S. (1994), *The Elements of Graphing Data*, Hobart Press, 1st ed.

Cleveland, W. S. and McGill, R. (1984), "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods," *Journal of the American Statistical Association*, 79, pp. 531–554.

Few, S. (2009), *Now You See It: Simple Visualization Techniques for Quantitative Analysis*, Burlingame, CA: Analytics Press, 1st ed.

Gegenfurtner, K. R. and Sharpe, L. T. (2001), *Color vision: From genes to perception*, Cambridge University Press.

Goldstein, E. B. (2009), *Encyclopedia of perception*, Sage Publications.

Healey, C. G. and Enns, J. T. (1999), "Large datasets at a glance: Combining textures and colors in scientific visualization," *Visualization and Computer Graphics, IEEE Transactions on*, 5, 145–167.

— (2012), "Attention and visual memory in visualization and computer graphics," *Visualization and Computer Graphics, IEEE Transactions on*, 18, 1170–1188.

Lewandowsky, S. and Spence, I. (1989a), "Discriminating strata in scatterplots," *Journal of the American Statistical Association*, 84, 682–688.

— (1989b), "The perception of statistical graphs," *Sociological Methods & Research*, 18, 200–242.

Majumder, M., Hofmann, H., and Cook, D. (2013), "Validation of Visual Statistical Inference, Applied to Linear Models," *Journal of the American Statistical Association*, 108, 942–956.

Mosteller, F., Siegel, A. F., Trapido, E., and Youtz, C. (1981), "Eye fitting straight lines," *The American Statistician*, 35, 150–152.

Piepho, H.-P. (2004), "An algorithm for a letter-based representation of all-pairwise comparisons," *Journal of Computational and Graphical Statistics*, 13, 456–466.

Robinson, H. (2003), "Usability of Scatter Plot Symbols," *ASA Statistical Computing & Graphics Newsletter*, 14, 9–14.

Spence, I. and Garrison, R. F. (1993), "A remarkable scatterplot," *The American Statistician*, 47, 12–19.

Treisman, A. (1985), "Preattentive processing in vision," *Computer Vision, Graphics, and Image Processing*, 31, 156 – 177.

Tukey, J. W. (1977), *Exploratory Data Analysis*, Lebanon, IN: Addison Wesley.

A Simulation Studies of Parameter Space

A.1 Distribution of Test Statistics

Simulating lineup data sets, we can compare test statistics measuring trend strength, cluster strength, and cluster size inequality for the null plots and target plots. These distributions allow us to objectively assess the difficulty of detecting the target datasets computationally (without relying on human perception).

Add equations for test statistics

Figure 11 show computed densities of the maximum null distribution measure compared with the measure in the signal plot. There is some overlap in the distribution of R^2 for the null plots compared with the target plot displaying data drawn from M_T. We have two measures comparing data drawn from M_C and M_0; the cluster measure examines the variance in x and y described by the cluster center; the gini coefficient examines the inequality in group sizes. These simulations indicate that it may be possible to differentiate M_C based on two different features in clustered data. In future experiments, it may be beneficial to control cluster size more tightly to remove this additional feature.

The distribution of the cluster statistic values are more easily separated from the null plots than the distribution of the line statistic, indicating that $\sigma_C = 0.20$ is producing target plots that are a bit easier to spot than trend targets with a parameter value of $\sigma_T = 0.25$, however, the inequality of group sizes may distract participants from the intended target signal of cluster cohesion.

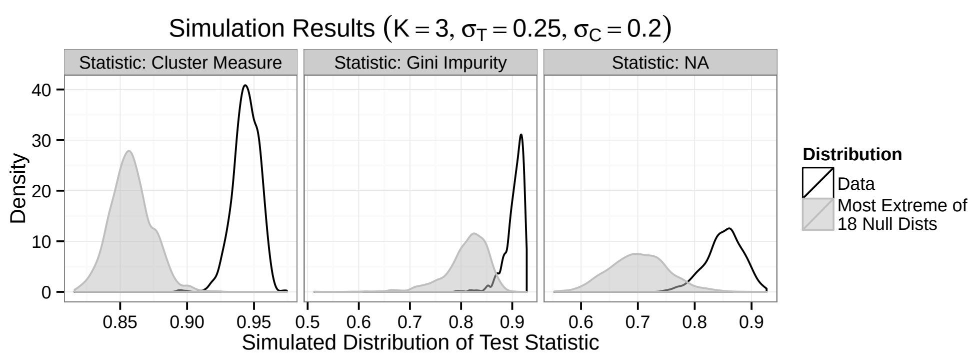

Figure 11: Density of test statistics measuring trend strength, cluster strength, and cluster inequality for target distributions and null plots.

A.2 Full Parameter Space Simulation Study

Using 1000 simulations for each of the 98 combinations of parameters ($K = \{3, 5\}$, $\sigma_C = \{.1, .15, .2, .25, .3, .35, .4\}$, $\sigma_T = \{.2, .25, .3, .35, .4, .45, .5\}$), we explored the effect of parameter value on the distribution of summary statistics describing the line strength (R^2) and cluster strength for null and target plots.

The setup for this simulation is similar to what Niladri is doing in CHowdhury et al. (2014).

Describe statistics

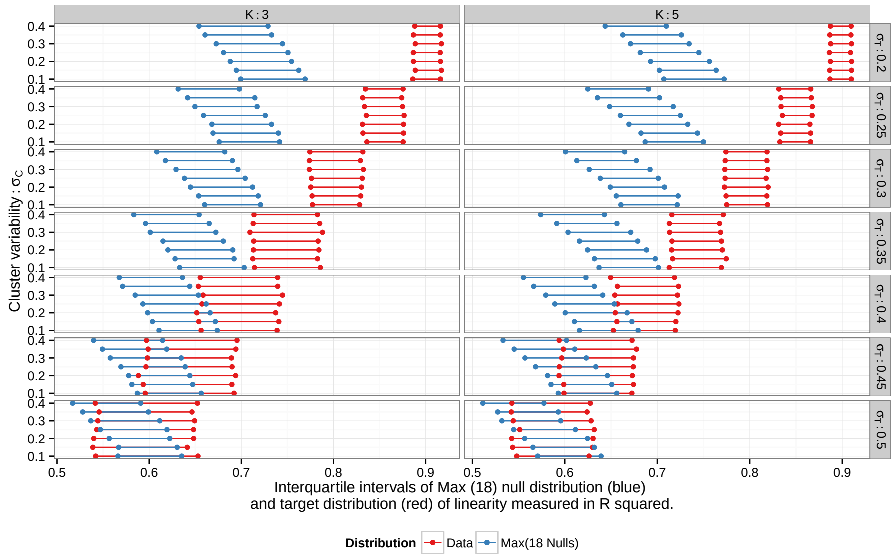

Figure 12: Simulated interquartile range of R^2 values for target and null data distributions.

Figures 12 and 13 show the 25th and 75th percentiles of the distribution of line and cluster summary statistics for each set of parameter values. These plots guided our evaluation of "easy", "medium" and "hard" parameter values for line and cluster tasks.

Additionally, we note that there is an interaction between σ_C and σ_T: the distinction between target and null on a fixed setting of clustering becomes increasingly difficult as the standard deviation for the linear trend is increased, and vice versa. There may additionally be a three-way interaction between σ_C, σ_T, and K: the size of the blue intervals (bottom figure) changes in size between different levels of K, it changes for different levels of σ_C and σ_T. These interactions suggest that in order to examine differences in aesthetics, we must block by parameter settings (this can be accomplished through blocking by dataset). Each dataset is non-deterministic, because we have a random process generating from different parameter settings, not a deterministic run setting as in an engineering setting. It is thus important to use replicates of each parameter setting to ensure that we can separate data-level effects from parameter-level effects.

Additionally, after the experiment was complete, we examined the distribution of group size (as measured by gini impurity) to establish whether there were any systematic differences in group size inequality between data generated from M_0 (null data) and data generated from M_C (cluster data). Figure 14 demonstrates that the cluster plots have significantly lower group size differences than null plots at all parameter combinations. It is therefore possible that some participants will identify extraordinarily unequal group sizes present in null plots as significantly different from the other lineup plots, ignoring any cluster signal. Future studies should more tightly control group size in order to reduce this effect.

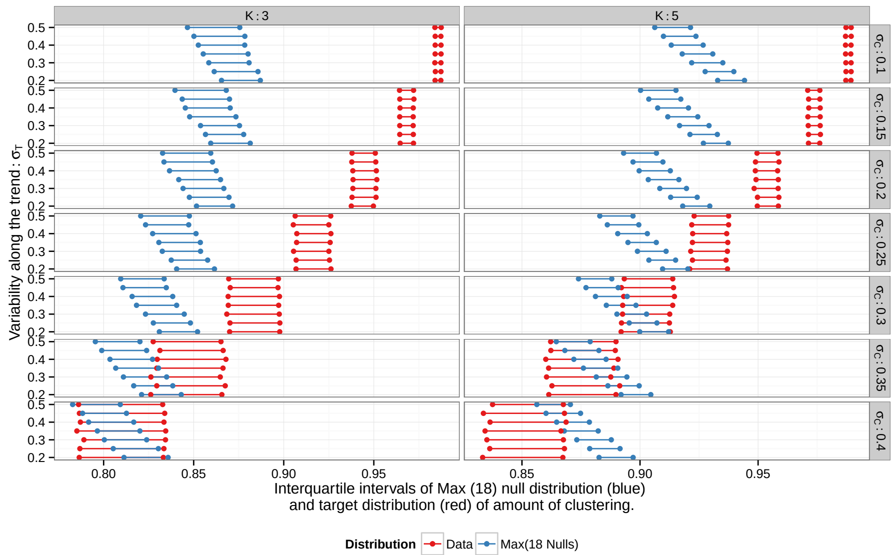

Figure 13: Simulated interquartile range of cluster cohesion statistic values for target and null data distributions.

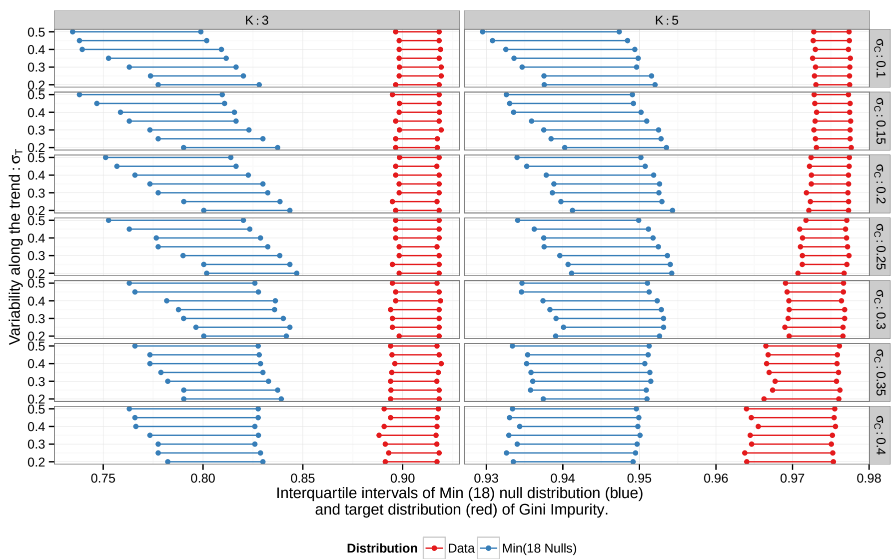

Figure 14: Simulated interquartile range of group size inequality statistic values for cluster and null data distributions.

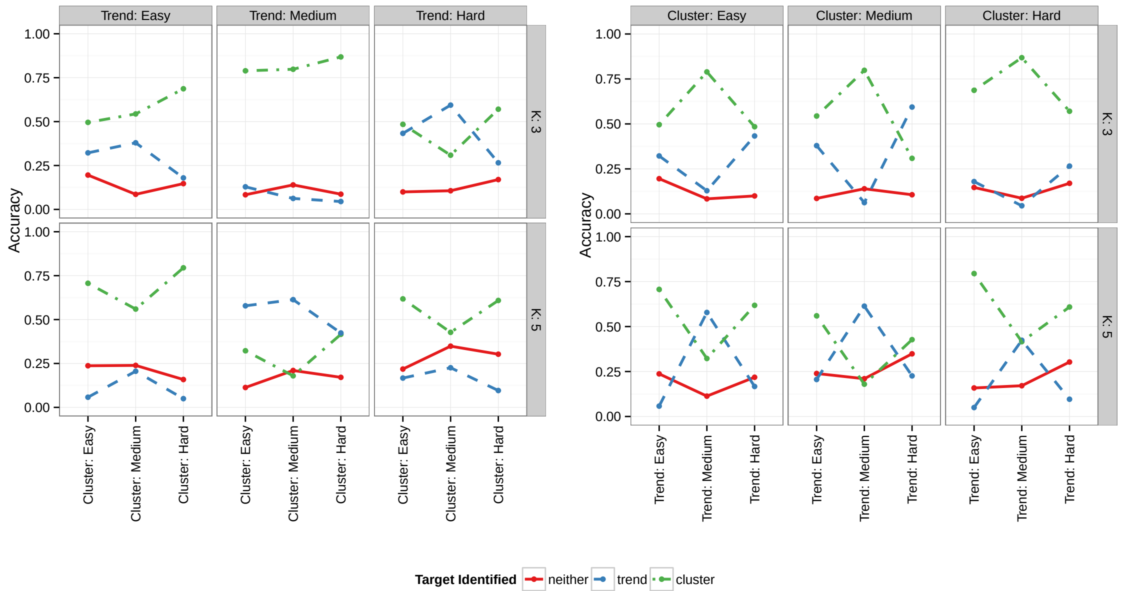

Figure 15: Identified targets for each level of parameter settings (across plot types).

B Graphical Exploration

Figure 16 shows aggregate accuracy rates for each plot aesthetic combination. It is again apparent that the cluster targets were overall more likely to be identified than line targets across all aesthetic combinations, however, it is also evident that plot aesthetics influence the identified target.

In addition to participant identification of target plots, we also asked that participants to rate their confidence in their answer. Figure 17 shows aggregate participant confidence rating as a function of trial outcome. Participants who did not identify either target plot were less likely to be "extremely confident" in their answer, while participants who identified either the trend or the cluster target correctly were highly confident that their answer was correct. Overall, though, participants seem to have some degree of confidence in their answer, regardless of whether the answer was correct.

As data collection was conducted entirely online, we cannot measure responses in the millisecond range characteristic of many psychometric studies, however, the data server does record the time between initial lineup presentation (trial start) and answer submission (trial end). Examining differences in average response times across trials provides us with an additional measure of trial difficulty or perceptual complexity. We can also explore whether participants spent more time on certain types of plots and whether that additional time increased accurate target identification. In order to remove the "novelty" effect of an unfamiliar task, we excluded every participant's first trial from this portion of the analysis.

Additionally, all trials which took greater than 3 minutes to complete have also been removed; we have no guarantee that a participant was not multitasking during their trials, and 3 minutes is an excessive amount of time for lineup completion (only 1% of trials were more than 3 minutes

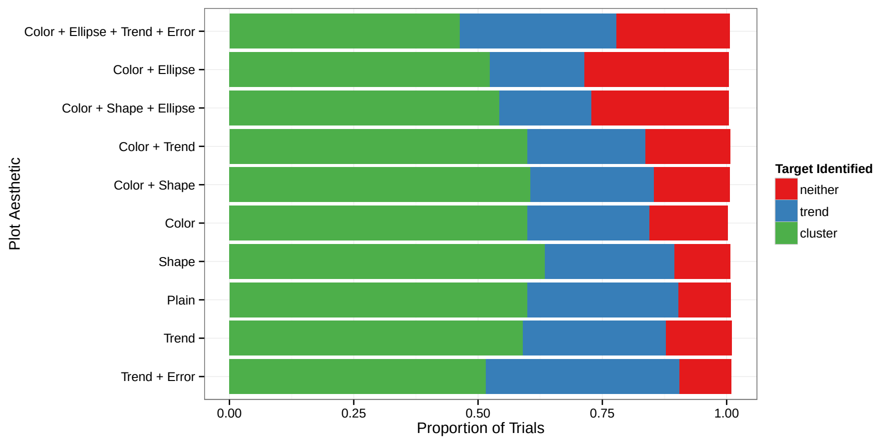

Figure 16: Proportion of trials identifying each target for each plot type (across parameter settings).

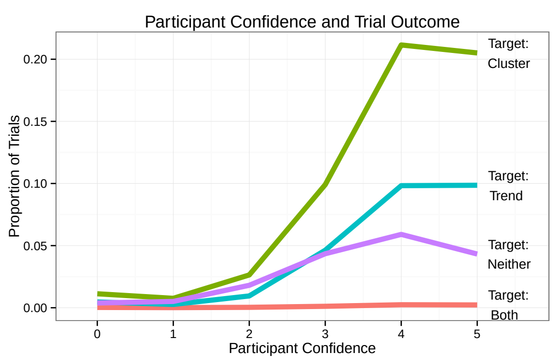

Figure 17: Participant confidence levels compared with trial results.

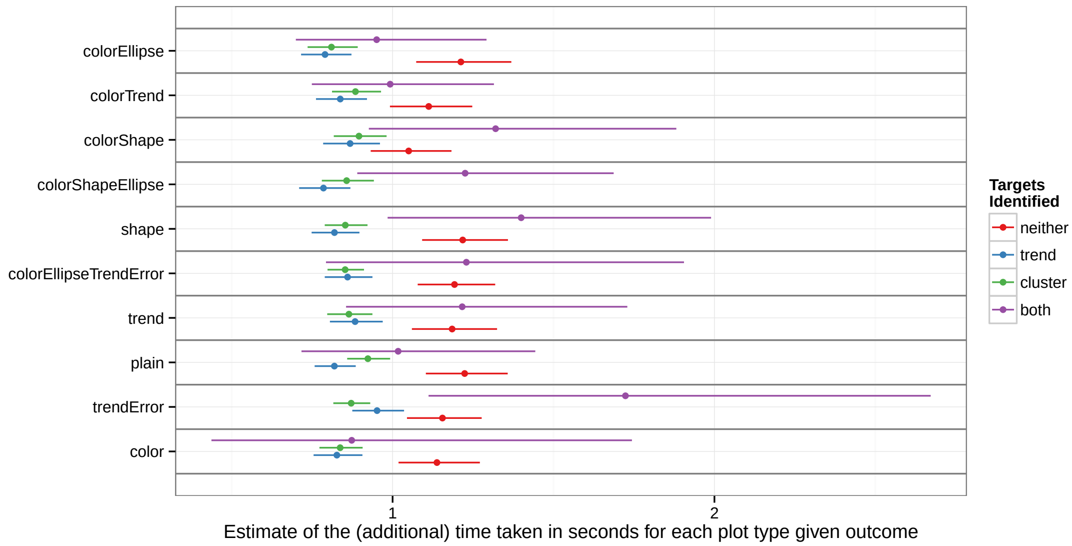

Figure 18: XXX figure would need some additional work, in case we want to keep it.

long).

No, don't select based on the value of the dependent variable. Use a log transform instead. Most likely, the time taken by participants is close to a Gamma distribution. This is what we have used in the past to model this kind of data - have a look at Mahbub's paper (the arxiv one on Human Factors). Instead of average trial ties, use a robust approach, such as the median - that will give you almost the same answers, see figure 18. There is no indication that the plot type matters for the time taken, even when we take the outcome of that viewing into account. What matters, is whether the lineup is the participants first lineup.

Average trial times and accuracy rates for each lineup are shown in figure 19 facetted by plot type, and in figure 20 facetted by parameter difficulty.

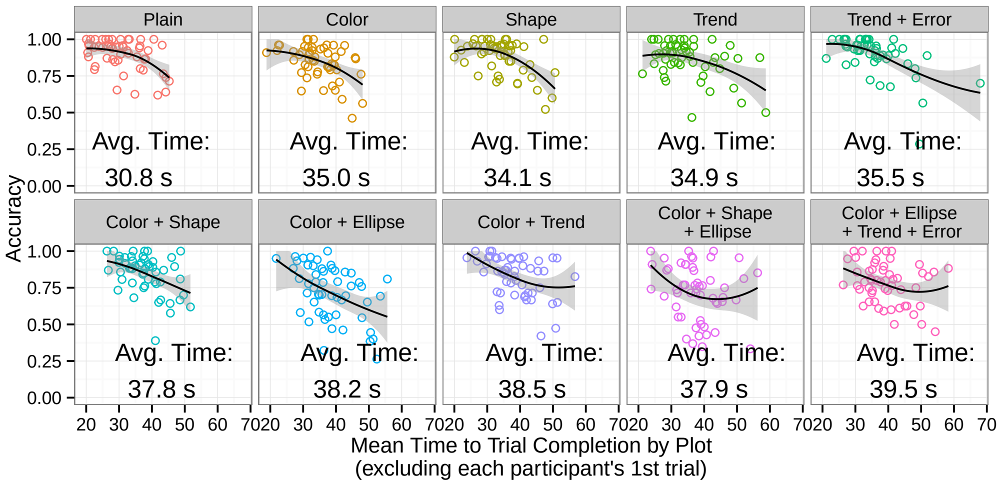

Figure 19: Accuracy (identifying either target plot) compared with mean trial time, by plot aesthetic.

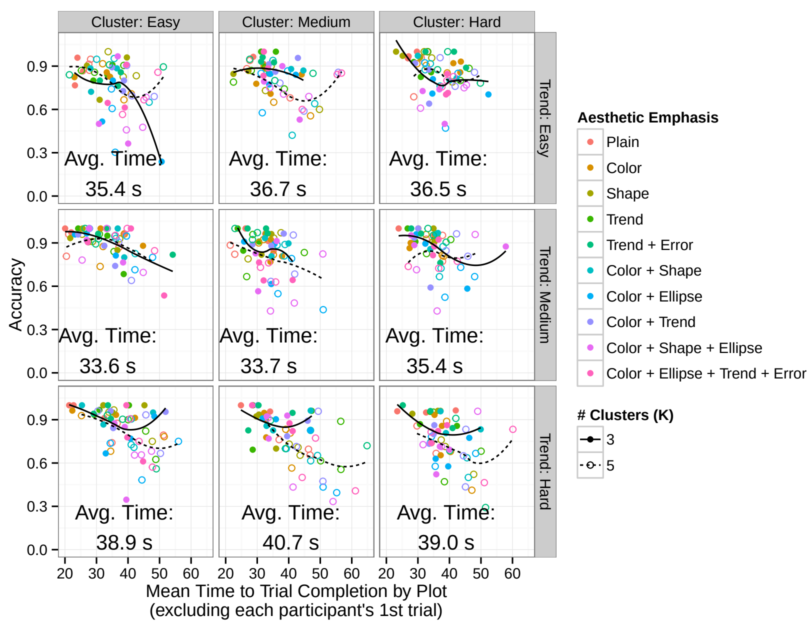

Figure 20: Accuracy (identifying either target plot) compared with mean trial time, by parameter settings.

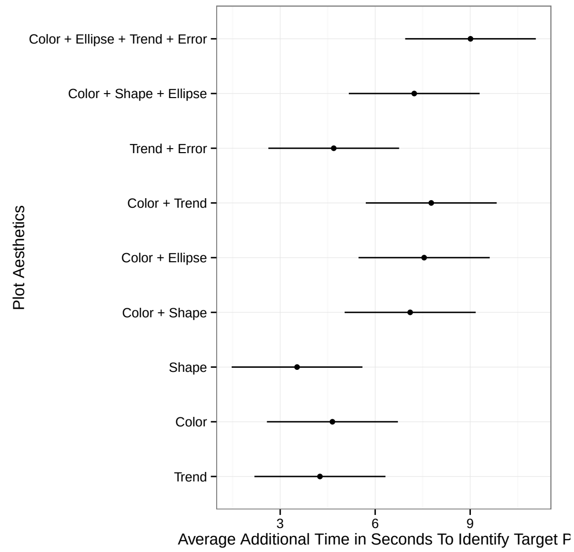

Figure 20 shows a reasonably clear effect of increasing trial time in response to increasing difficulty in cluster or trend variation parameters. While medium difficulty trend plots seem to be somewhat faster, on average, than easy or hard trend plots, there is a notable increase in time from either easy or medium trend plots to hard trend plots.

Nice finding!

C Model Results

Plot Aesthetic	Log Odds	Std. Error	Z	P value	Tukey Post Hoc Differences
Trend + Error	0.5738	0.1056	5.43	0.0000	g
Color + Ellipse + Trend + Error	0.1386	0.1072	1.29	0.1960	f
Trend	-0.1746	0.1083	-1.61	0.1067	df
Shape	-0.2658	0.1099	-2.42	0.0156	cde
Color + Shape	-0.4050	0.1111	-3.64	0.0003	cd
Color	-0.4186	0.1115	-3.75	0.0002	cd
Color + Trend	-0.5715	0.1120	-5.10	0.0000	bc
Color + Ellipse	-0.9401	0.1168	-8.05	0.0000	ab
Color + Shape + Ellipse	-0.9975	0.1178	-8.46	0.0000	a

Table 3: Fitted values of fixed effects for the model described in (3). Only Trend+Error plots significantly increase the probability of detecting the linear target plot (with data generated from M_T), while most other aesthetic combinations decrease the probability of detecting the linear target plot.

Plot Aesthetic	Log Odds	Std. Error	Z	P value	Tukey Post Hoc Differences
Shape	0.1727	0.0961	1.80	0.0724	d
Color + Shape	0.0274	0.0959	0.29	0.7752	d
Color + Trend	0.0054	0.0955	0.06	0.9546	d
Trend	-0.0445	0.0953	-0.47	0.6401	cd
Color	-0.0595	0.0958	-0.62	0.5345	cd
Color + Shape + Ellipse	-0.3065	0.0950	-3.23	0.0013	bc
Color + Ellipse	-0.4023	0.0945	-4.26	0.0000	b
Trend + Error	-0.4766	0.0947	-5.03	0.0000	b
Color + Ellipse + Trend + Error	-0.7867	0.0947	-8.31	0.0000	a

Table 4: Fitted values of fixed effects for the model described in (**??**).