Group beats Trend!?
Testing feature hierarchy in statistical graphics

Susan VanderPlas, Heike Hofmann*

March 4, 2015

Abstract

abstract goes here

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?

Soft introduction

Our understanding of the perception of statistical charts is informed by several levels of research. Cognitive psychologists and neuroscientists often focus on pre-attentive perception, which occurs automatically in the first 200 ms of exposure to a visual stimulus. Gestalt psychologists focus instead on perception as a holistic experience; they consider the heuristics used to transform visual stimuli into useful, coherent information. Finally, statistical graphics researchers apply low-level perceptual research and gestalt ideas to statistical charts, using tools such as lineups (Buja et al., 2009; Majumder et al., 2013; Wickham et al., 2010; Hofmann et al., 2012) to determine which graphics are accurately perceived and communicate information effectively.

Research into the preattentive stage of visual 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). Healey and Enns (1999) extended this work, demonstrating that certain features of three-dimensional data displays are processed preattentively, but that neither target identification nor three-dimensional data processing always translate into faster or more accurate inference about the data displayed, particularly when participants must integrate several preattentive features to understand the data.

add citations and text describing preattentive interference.

add a bit of transition between preattentive stages of perception and gestalt rules.

Gestalt rules of perception also impact statistical graphics. These rules describe the way we organize visual input, focusing on the holistic experience rather than the individual perceptual

*Department of Statistics and Statistical Laboratory, Iowa State University

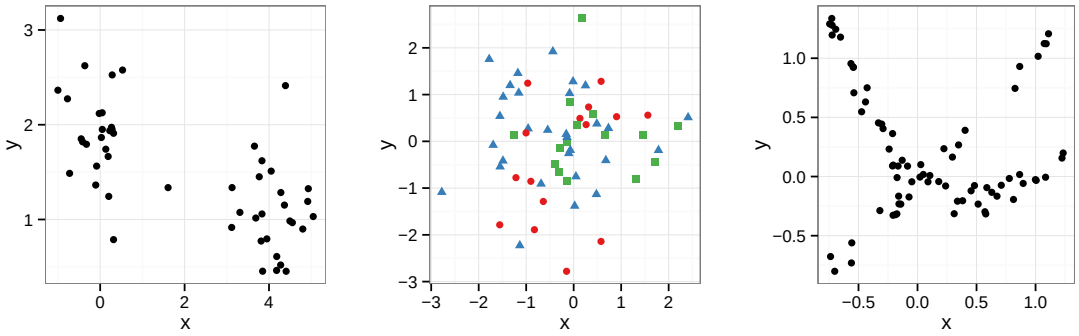


Figure 1: *Proximity* renders the fifty points of the first scatterplot as two distinct (and equal-sized) groups. Shapes and colors create different groups of points in the middle scatterplot, invoking the Gestalt principle of *Similarity*. *Good Continuation* renders the points in the scatterplot on 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.

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.

We should also look at the time to response – it would be interesting, to see if the conflicting stimuli need more time to come to a decision. It's obviously not milliseconds that we measure, but it might still be informative. (We would need to exclude everybody's first attempt).

This study is designed to understand 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.

1.1 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.

We should add that we allow users to select multiple targets, so that we don't get them into the position of having to guess between targets.

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 data and generating model; we are measuring the "power" (in a statistical sense) of the human perceptual system.

Using this testing framework, we can 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.

beautiful!!

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. 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.

2 Experimental 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: $x_i = \tilde{x}_i + \eta_i$, $\eta_i \sim$ Unif$(-z, z)$, $z = 1/5 \cdot 2/(N-1)$*

3. Generate y_i: $y_i = ax_i + e_i$, $e_i \sim N(0, \sigma_T^2)$

4. Center and scale x_i, y_i

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

$$r = 1 - RSS/TSS,$$

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 correlation coefficient ρ in this scenario is

$$\rho = \frac{\frac{1}{3}}{\frac{1}{3} + \sigma_T^2},$$

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

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

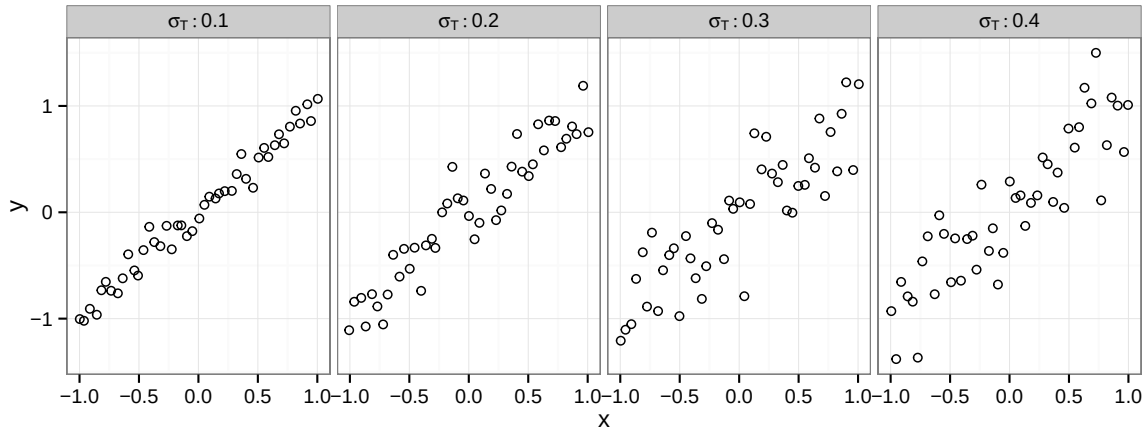


Figure 2: Set of scatterplots showing one draw each from the trend model M_T for parameter values of $\sigma_T \in \{0.1, 0.2, 0.3, 0.4\}$.

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*

(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)*:*

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

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 are therefore determined as a draw from a multinomial distribution: determine group sizes* $g = (g_1, ..., g_K)$*, with* $N = \sum_{i=1}^K g_i$*, for clusters* $1, ..., K$ *as a random draw*

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

4. *Generate points around cluster centers:*

(a) $x_i = \tilde{c}_{g_i}^x + e_i^x$*, where* $e_i^x \sim N(0, \sigma_C^2)$

(b) $y_i = \tilde{c}_{g_i}^y + e_i^y$*, where* $e_i^y \sim N(0, \sigma_C^2)$

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 the study we used $a = 1$, right?

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).

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.

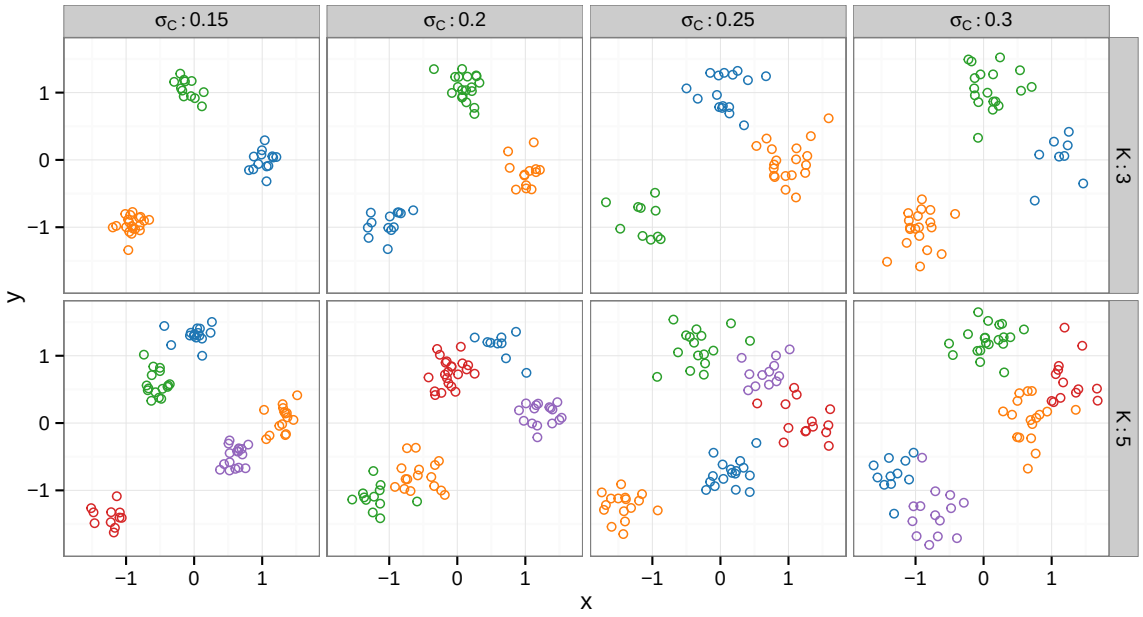


Figure 3: Scatterplots of clustering output for different inner cluster spread σ_C (left to right) and different number of clusters K (top and bottom).

the simulation is so nice, maybe you could include one more sentence on the idea. Niladri did something similar in his Where's Waldo paper. Need to ask Di for a reference.

The parameter values were chosen after examining the full parameter space through simulation; results are provided in appendix 4. In accordance with 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.

Each of the generated datasets is then plotted in lineups, 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 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.

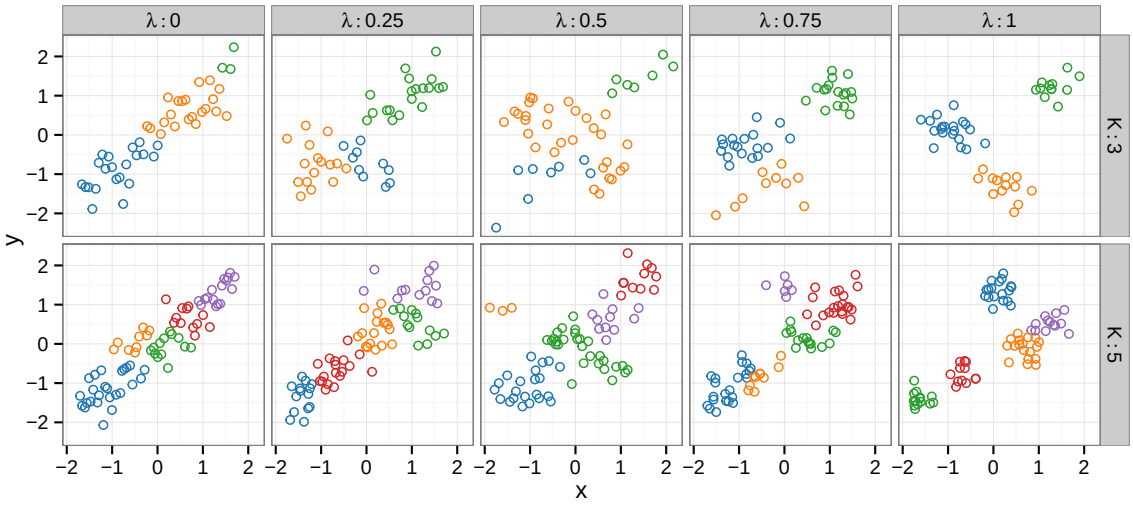


Figure 4: Scatterplots of data generated from M_0 using different values of λ.

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.

Figure 5: Each of the 10 plot feature combinations tested in this study, with $K = 3$, $\sigma_T = .25$ and $\sigma_C = .2$.

<table>
<tr><td></td><td></td><td colspan="3">Line Emphasis</td></tr>
<tr><td></td><td>Strength</td><td>0</td><td>1</td><td>2</td></tr>
<tr><td rowspan="5">Cluster Emphasis</td><td>0</td><td>None</td><td>Line</td><td>Line + Prediction</td></tr>
<tr><td rowspan="2">1</td><td>Color</td><td rowspan="2">Color + Line</td><td rowspan="2"></td></tr>
<tr><td>Shape</td></tr>
<tr><td rowspan="2">2</td><td>Color + Shape</td><td rowspan="2"></td><td rowspan="2">Color + Ellipse + Line + Prediction</td></tr>
<tr><td>Color + Ellipse</td></tr>
<tr><td>3</td><td>Color + Shape + Ellipse</td><td></td><td></td></tr>
</table>

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.

2.3 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.

2.4 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 grey from the palette to make the experiment more inclusive of participants with colorblindness.

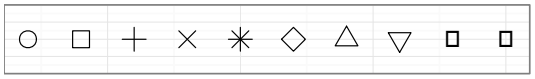


Figure 7: Shapes in Çağatay Demiralp et al. (2014). In order to control for varying point size due to Unicode vs. non-Unicode characters, the last two shapes were removed.

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 grey hue from the palette. This modification produced maximally different color combinations which did not include red-green combinations, while also removing a color (grey) 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.

2.5 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.

2.6 Participant Recruitment

describe amazon turk, participant instructions, screening procedures, etc.

3 Results

3.1 General results

demographic information, N people participated, completing on average M plots, etc., overall accuracy rates by individual and plot type

In our last paper we got blasted for includign the table with Z statistics and p-values, and the reviewer wanted to have confidence ratios instead - at that point, there is no real difference to the picture. I'd suggest to move the model tables into the appendix and discuss the results based on the charts...

We will 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.

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 $k = 1, ..., P$,

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

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

where β_i describe the effect of specific plot aesthetics

$$\gamma_j \overset{iid}{\sim} N\left(0, \sigma^2_{\text{data}}\right)$$

$$\eta_k \overset{iid}{\sim} N\left(0, \sigma^2_{\text{participant}}\right)$$

and $\epsilon_{ijk} \overset{iid}{\sim} N\left(0, \sigma^2\right)$

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.

Commentary goes here... too tired tonight

Discuss variance/covariance and random effects

Plot Aesthetic	Log Odds	Std. Error	Z	P value
Trend + Error	0.6081	0.1014	6.00	0.0000
Color + Ellipse + Trend + Error	0.1425	0.1028	1.39	0.1658
Trend	-0.1436	0.1036	-1.39	0.1658
Shape	-0.2387	0.1046	-2.28	0.0225
Color + Shape	-0.3877	0.1060	-3.66	0.0003
Color	-0.4546	0.1070	-4.25	0.0000
Color + Trend	-0.5448	0.1071	-5.09	0.0000
Color + Ellipse	-0.8861	0.1111	-7.98	0.0000
Color + Shape + Ellipse	-0.9763	0.1124	-8.69	0.0000

Table 3: Fitted values of fixed effects for the model described in (1). 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.

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). For plot type $i = 1, ..., 10$ displaying dataset $j = 1, ..., 54$ by participant $k = 1, ..., P$,

$$P(\text{success}) = \left(e^{\theta}\right) / \left(1 + e^{\theta}\right) \tag{2}$$

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

where β_i describe the effect of specific plot aesthetics

$$\gamma_j \overset{iid}{\sim} N\left(0, \sigma^2_{\text{data}}\right)$$

$$\eta_k \overset{iid}{\sim} N\left(0, \sigma^2_{\text{participant}}\right)$$

$$\text{and } \epsilon_{ijk} \overset{iid}{\sim} N\left(0, \sigma^2\right)$$

As before, 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.

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

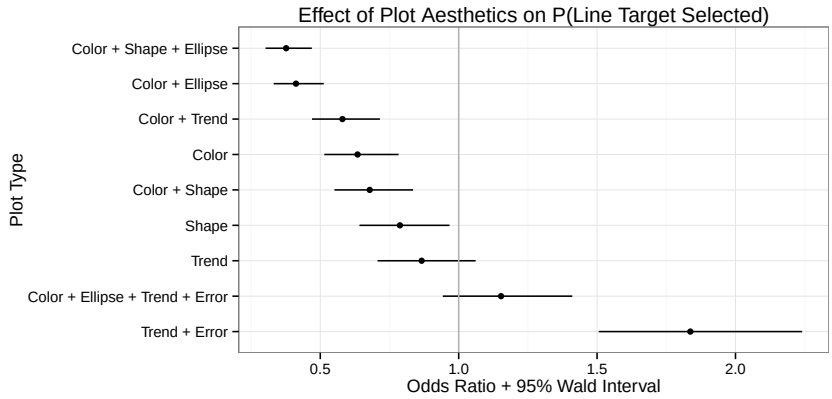


Figure 8: Odds ratios describing the odds of detecting the linear target plot for each aesthetic, relative to a plain scatterplot. Only the combination of Trend + Error significantly increases the odds of linear target plot detection relative to the control plot.

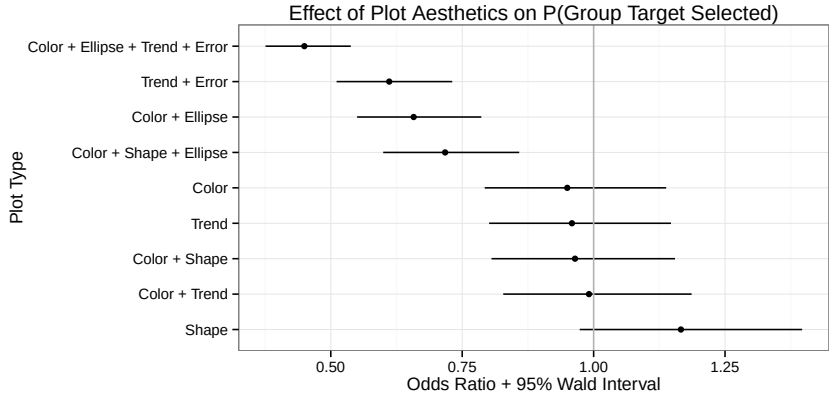


Figure 9: Odds ratios describing the odds of detecting the cluster target plot for each aesthetic, relative to a plain scatterplot. The presence of error lines or bounding ellipses significantly decreases the probability of correct target detection, and no aesthetic successfully increases the probability 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.

Plot Aesthetic	Log Odds	Std. Error	Z	P value
Trend + Error	0.6081	0.1014	6.00	0.0000
Color + Ellipse + Trend + Error	0.1425	0.1028	1.39	0.1658
Trend	-0.1436	0.1036	-1.39	0.1658
Shape	-0.2387	0.1046	-2.28	0.0225
Color + Shape	-0.3877	0.1060	-3.66	0.0003
Color	-0.4546	0.1070	-4.25	0.0000
Color + Trend	-0.5448	0.1071	-5.09	0.0000
Color + Ellipse	-0.8861	0.1111	-7.98	0.0000
Color + Shape + Ellipse	-0.9763	0.1124	-8.69	0.0000

Table 4: Fitted values of fixed effects for the model described in (2).

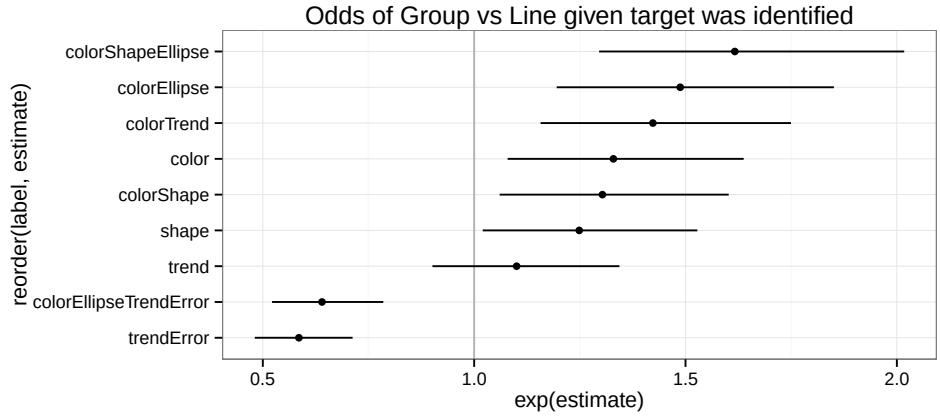


Figure 10: Estimated odds of decision for group versus line target based on evaluations that resulted in the identification of one of these targets.

3.5 Signal Strength

4 Discussion

References

Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E.-K., Swayne, D. F., and 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, 4361–4383.

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

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

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.

Hofmann, H., Follett, L., Majumder, M., and Cook, D. (2012), "Graphical Tests for Power Comparison of Competing Designs," *IEEE Transactions on Visualization and Computer Graphics*, 18, 2441–2448.

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.

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

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

Wickham, H., Cook, D., Hofmann, H., and Buja, A. (2010), "Graphical inference for infovis," *Visualization and Computer Graphics, IEEE Transactions on*, 16, 973–979.

Simulation Studies of Parameter Space

4.1 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.

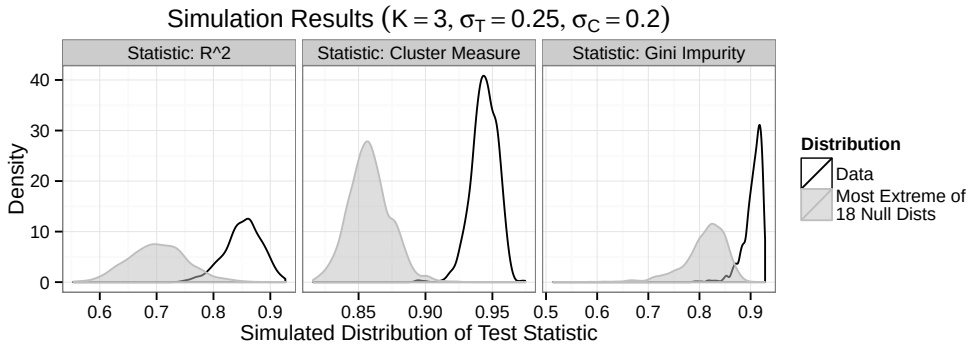
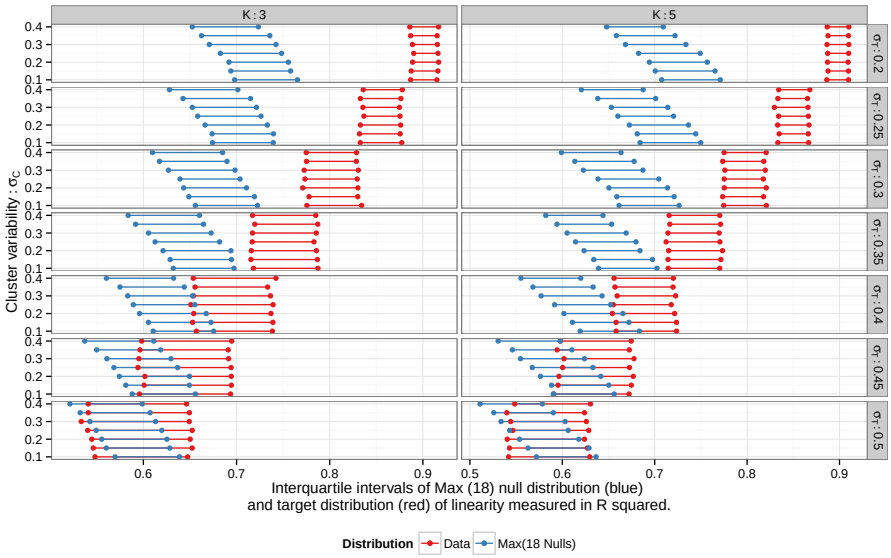


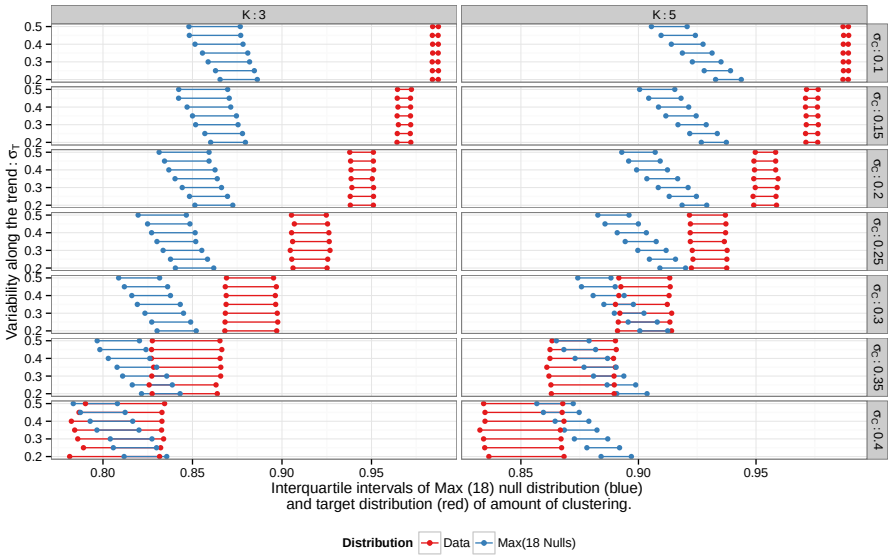
Figure 11: Density of test statistics measuring trend strength, cluster strength, and cluster inequality for target distributions and null plots.

4.2 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 (short description here) for null and target plots. The plot below shows the 25th and 75th percentiles of the distribution of these 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.

What we also see from these plots, is that we do have a $\sigma_C\sigma_T$ interaction: 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. We might also have 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. I am not sure whether that is an actual three-way interaction or just all three two-way interactions, but it doesn't matter, at this point we are just talking about potentially saving one parameter. What is clear, is that we need to block by parameter setting. We do so, by blocking on each dataset. Each dataset is non-deterministic, though, because we have a random process generating from different parameter settings, not a deterministic run setting as in an engineering setting. We therefore need repetitions of the data generation to be able to separate the variability coming from within the parameter setting from the additional variability introduced by the subjects' evaluations of the lineups.





Experimental Design

Initially, assume a fully factorial, balanced design, with r unique datasets per parameter set (replicates) and P evaluations per (aesthetic|dataset). The experiment is conducted at three levels: parameter sets (with replication, so EUs are data sets), plot types (i.e. a certain set of aesthetics), and participant evaluations. At the first level, there are three parameters: $K = 3$, $\sigma_T \in \{.15, .25, .35\}$, and $\sigma_C \in \{.15, .25, .35\}$. At the second level, there are blocks (by data set), and then 4 aesthetic

combinations.

We'll have to use contrasts to measure the effect of color individually, etc., for now let's just consider the ANOVA evaluation

Finally, at the lowest level, there are participant effects.

At the participant level, we need to decide if we're going to fully randomize, try to block, etc. - are participants going to get 10 different data sets? 5? Not sure how to conceptualize that, and I would imagine it will affect how we organize model evaluation. Grr, I hate mixed models.
HH: yes, I would assume that participants get ten plots each, one from each of the designs in a random order. (We have the data base set up that way).

Modified from Table 10.6 (pg 181) of Design of Experiments by Dr. Morris. The table in the book has a four-factor split plot design with three levels (randomized, block, block).

Level	Factor	Source	DF	Sum of Squares
Dataset	K	α	1	$\sum_i (4)(5)(r)(10P)(\overline{y}_{i\cdots\cdots} - \overline{y}_{\cdots\cdots})^2$
	σ_T^2	β	3	$\sum_j (2)(5)(r)(10P)(\overline{y}_{\cdot j\cdots\cdots} - \overline{y}_{\cdots\cdots})^2$
	σ_C^2	γ	4	$\sum_i (2)(4)(r)(10P)(\overline{y}_{\cdot\cdot k\cdots} - \overline{y}_{\cdots\cdots})^2$
		$(\alpha\beta)$	3	$\sum_{ij} (5)(r)(10P)(\overline{y}_{ij\cdots\cdots} - \overline{y}_{i\cdots\cdots} - \overline{y}_{\cdot j\cdots\cdots} + \overline{y}_{\cdots\cdots})^2$
		$(\alpha\gamma)$	4	$\sum_{ik} (4)(r)(10P)(\overline{y}_{i\cdot k\cdots} - \overline{y}_{i\cdots\cdots} - \overline{y}_{\cdot\cdot k\cdots} + \overline{y}_{\cdots\cdots})^2$
		$(\beta\gamma)$	12	$\sum_{jk} (2)(r)(10P)(\overline{y}_{\cdot jk\cdots} - \overline{y}_{\cdot j\cdots\cdots} - \overline{y}_{\cdot\cdot k\cdots} + \overline{y}_{\cdots\cdots})^2$
		$(\alpha\beta\gamma)$	12	$\sum_{ijk} (r)(10P)(\overline{y}_{ijk\cdots} - \overline{y}_{i\cdots\cdots} - \overline{y}_{\cdot j\cdots\cdots} - \overline{y}_{\cdot\cdot k\cdots} + \overline{y}_{ij\cdots\cdots} + \overline{y}_{\cdot jk\cdots} + \overline{y}_{i\cdot k\cdots} - \overline{y}_{\cdots\cdots})^2$
		Resid.	$(2)(4)(5)(r-1)$	$\sum_{ijkl} (10P)(\overline{y}_{ijkl\cdot\cdot} - \overline{y}_{i\cdots\cdots} - \overline{y}_{\cdot j\cdots\cdots} - \overline{y}_{\cdot\cdot k\cdots} + \overline{y}_{ij\cdots\cdots} + \overline{y}_{\cdot jk\cdots} + \overline{y}_{i\cdot k\cdots} - \overline{y}_{ijk\cdots} - \overline{y}_{ij\cdot l\cdot\cdot} - \overline{y}_{i\cdot kl\cdots} - \overline{y}_{\cdot jkl\cdots} + \overline{y}_{\cdots\cdots})^2$
	Total		$(2)(4)(5)(r)-1$	$\sum_{ijkl} (10P)(\overline{y}_{ijkl\cdot\cdot} - \overline{y}_{\cdots\cdots})^2$
Plot	Dataset	blocks	$(2)(4)(5)(r)-1$	$\sum_{ijkl} (10P)(\overline{y}_{ijkl\cdot\cdot} - \overline{y}_{\cdots\cdots})^2$
	Aes.	δ	9	$\sum_m (2)(4)(5)(P)(\overline{y}_{\cdots\cdots m\cdot} - \overline{y}_{\cdots\cdots})^2$
	Aes x K	$(\alpha\delta)$	9	$\sum_{im} (4)(5)(P)(\overline{y}_{i\cdots m\cdot} - \overline{y}_{i\cdots\cdots} - \overline{y}_{\cdots\cdots m\cdot} + \overline{y}_{\cdots\cdots})^2$
	Aes x σ_T	$(\beta\delta)$	27	$\sum_{jm} (2)(5)(P)(\overline{y}_{\cdot j\cdots m\cdot} - \overline{y}_{\cdot j\cdots\cdots} - \overline{y}_{\cdots\cdots m\cdot} + \overline{y}_{\cdots\cdots})^2$
	Aes x σ_C	$(\gamma\delta)$	36	$\sum_{km} (2)(4)(P)(\overline{y}_{\cdot\cdot k\cdot m\cdot} - \overline{y}_{\cdot\cdot k\cdots} - \overline{y}_{\cdots\cdots m\cdot} + \overline{y}_{\cdots\cdots})^2$
	Others		9(31)	
	Resid		40(rP-1)-(40r-1)	difference
	Total		$400r-1$	$\sum_{ijklm} (P)(\overline{y}_{ijklm\cdot} - \overline{y}_{ijkl\cdot\cdot})^2$
Trial	Picture	Sub-blocks	$400r-1$	$\sum_{ijklm} (P)(\overline{y}_{ijklm\cdot} - \overline{y}_{ijkl\cdot\cdot})^2$
	Participants	τ	$P-1$	$\sum_n (2)(4)(5)(r)(10)(\overline{y}_{\cdots\cdots n} - \overline{y}_{\cdots\cdots})^2$
	Resid		$(400r-1)(P)$	difference
	Total		$400(r)(P)-1$	$\sum_{ijklmn} (y_{ijklmn} - \overline{y}_{\cdots\cdots})^2$

Table 5: Evaluation of sources of error in a full factorial version of the experiment, with r replicates of each parameter combination and P participant evaluations of each plot(data/aesthetic combination).

We have a couple of options:

- keep the full factorial experiment, use one (at most two) replicates, and use higher level factorial effects to beef up any error variance terms.
- Do a full factorial experiment for $K = 3$ and use a subset of the factorial experiment for $K = 5$ (either using a subset of cases for σ_T and σ_C, or a subset of combinations of the two cases/fractional factorial.)

The fractional factorial option will be a pain to explain when we write things up; it will be simpler to explain using a subset of cases. Given that we don't particularly care about the third-order effects (and possibly not even the second-order effects) for the parameters, I'm inclined to say that the single-replicate option is the easiest way to go (and lets us keep the simple SSQ in the table, which is a huge bonus in my opinion). Even if we just use the third-order interaction effect as error, we still have 12 degrees of freedom; that should be plenty - we'd only need F=2.69 to get a significant result for even the $(\sigma_T \sigma_C)$ test.
HH: We might not care about interpreting the two-way interactions, but unfortunately they will be there (see comment at the back). So I would suggest to go with a full factorial design in σ_C, σ_T, and K, with three replications each (we need the replicates, also explained in the back). This gives us 18 parameter settings, and $18 \cdot 3 = 54$ data sets. In case you still want to consider the effect of the number of datapoints N, we could switch from fully factorial to fractional factorial and replace the three-way interaction of σ_C, σ_T, and K by the settings of N. That way we will keep the 18 settings.

Table 6: ANOVA table - only one replicate. Evaluation of sources of error in a full factorial version of the experiment, with one replicate of each parameter combination and P participant evaluations of each plot(data/aesthetic combination).

Level	Factor	Source	DF	Sum of Squares
Dataset	K	α	1	$\sum_i (4)(5)(10P)(\overline{y}_{i\ldots.} - \overline{y}_{\ldots..})^2$
	σ_T^2	β	3	$\sum_j (2)(5)(10P)(\overline{y}_{.j\ldots} - \overline{y}_{\ldots..})^2$
	σ_C^2	γ	4	$\sum_i (2)(4)(10P)(\overline{y}_{..k.} - \overline{y}_{\ldots..})^2$
		Resid.	22	difference
	Total		39	$\sum_{ijk} (10P)(\overline{y}_{ijk.} - \overline{y}_{\ldots..})^2$
Plot	Dataset	blocks	39	$\sum_{ijk} (10P)(\overline{y}_{ijk.} - \overline{y}_{\ldots..})^2$
	Aes.	δ	9	$\sum_m (2)(4)(5)(P)(\overline{y}_{\ldots m.} - \overline{y}_{\ldots..})^2$
	Aes x K	$(\alpha\delta)$	9	$\sum_{im} (4)(5)(P)(\overline{y}_{i\cdot\cdot m.} - \overline{y}_{i\ldots.} - \overline{y}_{\ldots m.} + \overline{y}_{\ldots..})^2$
	Aes x σ_T	$(\beta\delta)$	27	$\sum_{jm} (2)(5)(P)(\overline{y}_{.j\cdot m.} - \overline{y}_{.j\ldots} - \overline{y}_{\ldots m.} + \overline{y}_{\ldots..})^2$
	Aes x σ_C	$(\gamma\delta)$	36	$\sum_{km} (2)(4)(P)(\overline{y}_{\cdot\cdot km.} - \overline{y}_{\cdot\cdot k.} - \overline{y}_{\ldots m.} + \overline{y}_{\ldots..})^2$
	Resid		9(31)	difference
	Total		399	$\sum_{ijkm} (P)(\overline{y}_{ijkm.} - \overline{y}_{ijk.})^2$
Trial	Picture	Sub-blocks	399	$\sum_{ijkm} (P)(\overline{y}_{ijkm.} - \overline{y}_{ijk.})^2$
	Participants	τ	$P-1$	$\sum_n (2)(4)(5)(10)(\overline{y}_{\ldots.n} - \overline{y}_{\ldots..})^2$
	Resid		$399(P-1)$	difference
	Total		$400P-1$	$\sum_{ijkmn} (y_{ijkmn} - \overline{y}_{\ldots.})^2$