Supplement to: Clusters beat Trend!? Testing feature hierarchy in statistical graphics

Susan VanderPlas*
Department of Statistics and Statistical Laboratory, Iowa State University
and
Heike Hofmann
Department of Statistics and Statistical Laboratory, Iowa State University

February 13, 2016

Supplement A Simulations of the Parameter Space

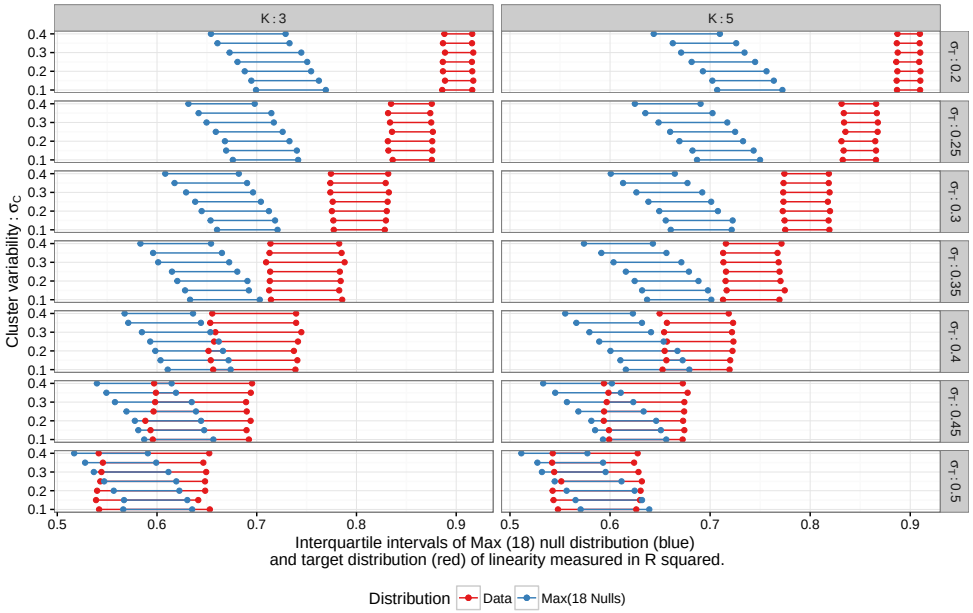
Using 1000 simulations for each of the 98 combinations of parameters ($K = \{3, 5\}$, $\sigma_C = \{.1, .15, .2, .25, .3, .35, .4\}$, $\sigma_T = \{.2, .25, .3, .35, .4, .45, .5\}$), we explored the effect of parameter value on the distribution of summary statistics describing the strength of the linear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

Figures 1a and 1b show the 25th and 75th percentiles of the distribution of R^2 and cluster strength summary statistics for each set of parameter values. These plots guide our evaluation of "easy", "medium" and "hard" parameter values for trend 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 width of the blue intervals (bottom figure) changes between different levels of K and 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 data set). Each data set 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.

*The authors gratefully acknowledge funding from the National Science Foundation Grant # DMS 1007697. All data collection has been conducted with approval from the Institutional Review Board IRB 10-347

(a) R^2 values for target and null data distributions



(b) Cluster cohesion statistics C^2.

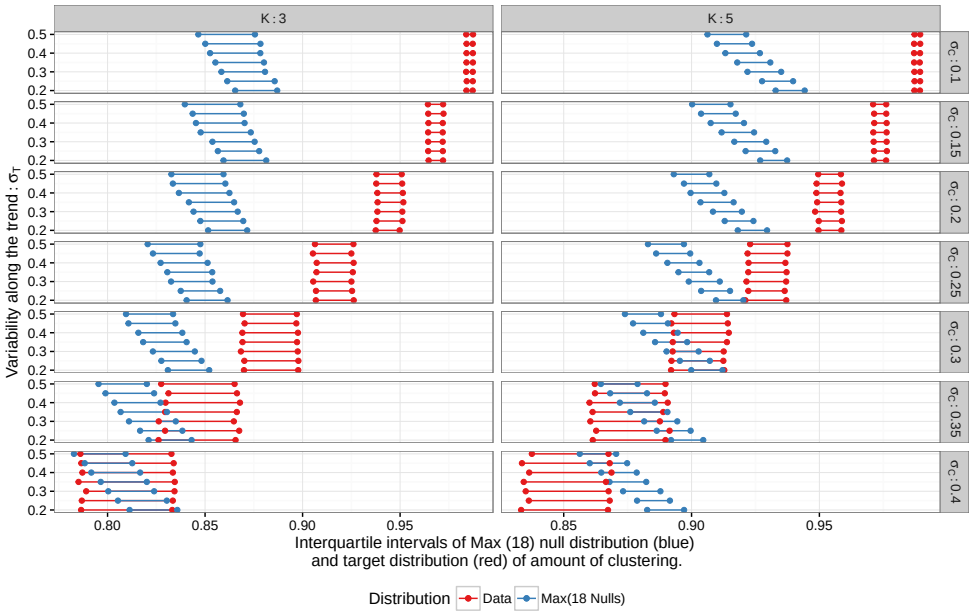


Figure 1: Simulated interquartile ranges between target and most extreme statistic from one of the 18 null plots.

Supplement B Simulation based inference in a two-target lineup scenario

Assume that there are two targets embedded in a lineup of overall size m, where m in our experiment is taken to be $m = 20$. Let A be the event that one of these targets is chosen. Under the null hypothesis that both targets are consistent with being created based on data from the null model, we can assume that under the null hypothesis the expected value of the probability that an observer picks one of these plots from the lineup is $2/m = E[P(A \mid H_o)]$. For the distribution of $A \mid H_o$ we employ a simulation-based strategy: Under the null hypothesis, we can assume, that the p-value corresponding to a hypothesis test 'the presented data is consistent with the null model' has a standard uniform distribution, i.e. $p_i \sim U[0, 1]$ i.i.d. for all $1 \leq i \leq m$. We assume that the choice observers make can be modeled using a multinomial distribution, where the probability π_i to pick panel i is inversely linear to p_i, with $\sum_{i=1}^{m} \pi_i = 1$.

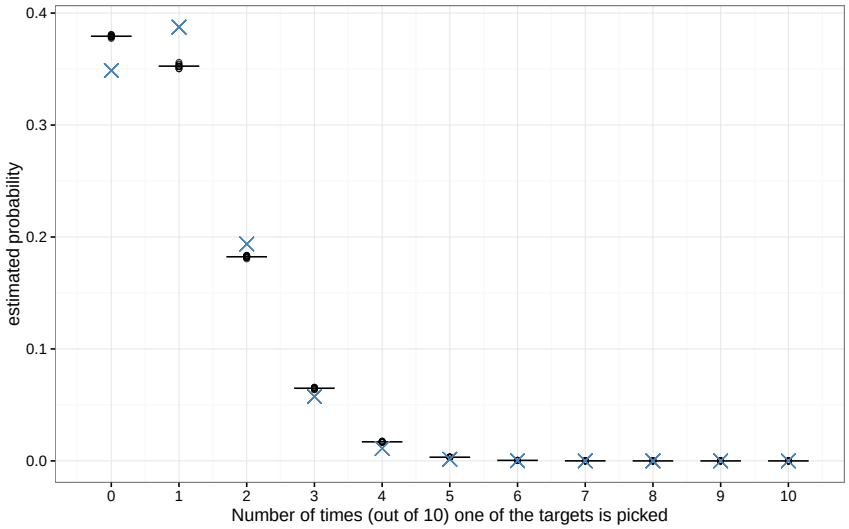


Figure 2: Ten simulations of size $b_2 = 1,000$ and $b_1 = 100$ for lineups of size $m = 20$ assuming $K = 10$ evaluations. The averages of the ten simulation runs are shown as lines. The crosses are probabilities from Binomial $B_{2/20,10}$.

W.l.o.g. we can assume that the two target plots are in positions 1 and 2. Given that a lineup was evaluated by K individuals, the simulation process for the conditional probability of identifying one of the targets given that both are consistent with the null model, $P(A|H_o)$, is then as follows:

1. Pick two values $p_i \sim U[0, 1], i = 1, 2$.
2. Repeat b_1 times:
   (a) Pick $m - 2$ values $p_i \sim U[0, 1], i = 3, ..., m$.
   (b) Pick K values from a Multinomial distribution with $\pi = \frac{1-p}{||1-p||}$, i.e. $x_j \sim M_\pi, i = 1, ..., K$

(c) Return the number of times that x_j is 1 or 2.

Repeat the above process b_2 times, and average results for a distribution of $A \mid H_o$. The choice of b_1 and b_2 decides on the number of decimal places to which the estimated distribution can be used reliably.

Figure 2 shows the result of this simulation approach for a lineup of size 20 assuming $K = 10$ evaluation. The density of $A \mid H_o$ is plotted for ten runs (open circles). The variability in the results is relatively small - for comparison, the density of a Binomial distribution $B_{2/20,10}$ is shown using crosses. The main difference between the densities is the probability of zero or only one identification, while the tail probabilities are very similar.

Supplement C Model Discussion and Results

In this supplement, we discuss four primary models used to examine different aspects of the data collected in this experiment. According to its dependent variables, these models are:

Accuracy (in section C.1): a lineup is considered to be evaluated 'accurately', if at least one of its two targets is identified in an evaluation. We model this probability, i.e. $P(C_{ijk} \cup T_{ijk})$, where C_{ijk} and T_{ijk} are binary variables describing the events that participant k identifies the cluster/trend target in data set j presented with design i.

Response time (in section C.2): response time is measured as the time between which a lineup is presented to a participant and the time point at which an answer is submitted. On average, participants need 40 seconds to respond to a lineup. Factors that affect this average (including accuracy, confidence level, and user justification) are also investigated.

Confidence (in section C.3): for each lineup, participants are asked to provide a subjective evaluation of the level of confidence they have in the choice they made, in the form of an integer value between 0 (least) to 5 (highest).

Balance between targets (in section C.4): because each lineup includes two targets, we can compare the frequency of identification of each target type. The objective here is to model the conditional probability of a cluster target identification given one of the two targets was identified, i.e. we want to model $P(C_{ijk} \mid C_{ijk} \cup T_{ijk})$.

The parameter estimation for the first three models is based on all 12010 available lineup evaluations, while the last model is based on only those lineups where at least one of the targets was identified. An overview of all of the fixed effects considered for each one of these models is given in Table 1. All models are fitted using the same random effects structure: we assumed one random effect for data set specific characteristics, one random effect to account for individuals' different skills, and a random error term. We are assuming that all random effects are normally distributed and pairwise orthogonal.

The general model structure we are using for all four main models is the following:

$$g\left(E[Y]\right) = \mathbf{W}\alpha + \mathbf{X}\beta + \mathbf{J}\gamma + \mathbf{K}\eta, \tag{1}$$

where

$g(.)$, the left hand side, is made up of a link function g that connects the (conditional) expected value of dependent variable Y to the structural right hand side. We further assume that $Y = g^{-1}(.) + \epsilon$, with $\epsilon \overset{\text{approx}}{\sim} N\left(0, \sigma^2 I_{n\times n}\right)$.

α, β are vectors of fixed effects, where β is a vector corresponding to the ten designs investigated, and α encompasses all 'other' fixed effects that might have an effect on the dependent variable besides the design of a plot,

		Model			
Dependent Variable		Face-off C.4	Accuracy C.1	Response Time C.2	Confidence C.3
Transformation		(logit)	(logit)	(log)	—
Effect	**Parameter**				
Design	β	✓	✓	✓	✓
Within cluster variability	α_C	✓	✓	✓	✓
Variability around trendline	α_T	✓	✓	✓	✓
Number of clusters	α_K	✓		✓	✓
First trial	α_1			✓	✓
Lineup order	α_O		✓	✓	✓
(log) Response time	τ		✓	—	✓
Gini impurity	γ		✓		
Single missing ellipse	ν	✓			
Target type (cluster, trend, neither, both)	$\omega_{\{C,T,N,B\}}$	—	—	✓	✓
Data		identified targets	all evaluations	all evaluations	all evaluations

Table 1: Overview of all models and their respective fixed effects. The same random effects were used for each model. All effects were investigated if applicable (otherwise marked by '—'). Checkmarks indicate significance at the 0.05 level.

W, X are the design matrices corresponding to the fixed effects,

$\gamma_j \overset{iid}{\sim} N(0, \sigma^2_{\text{data}})$, random effect for data set specific characteristics,

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

We also assume that random effects for data set and participant are orthogonal.

C.1 Accuracy Model

The lineup protocol provides an easy way of measuring accuracy of evaluations by assessing the number of participants who identified the data plot. In the modified version, we can use this as well by regarding any lineup evaluation resulting in an identification of at least one of the two targets as 'accurate'. We therefore want to model the probability that participant k identifies (at least) one of the targets on the lineup (using aesthetics set i) of data set j, $P(C_{ijk} \cup T_{ijk})$. We use the logit function as the link function $g(.)$.

The vector α of fixed effects consists of $(\mu, \alpha_T, \alpha_C, \alpha_K, \alpha_O)$, where

μ is an average baseline accuracy (and should not be interpreted, because all other effects are assumed to be zero, which is not practically possible),

α_C, α_T are the average effects of the standard error around trend lines $s_T \in \{0.25, 0.35, 0.45\}$ and clusters $s_C \in \{0.2, 0.25, 0.3, 0.35\}$ on accuracy,

α_K is the effect of the number of clusters $K \in \{3, 5\}$, and

α_O is an order effect, i.e. an effect on accuracy for successive lineups.

Table 2 shows an overview of the parameters of the accuracy model and their estimates. Both α_T and α_C have large negative effects that are highly significant. This indicates that

as the signal in the target plots weakens (by an increase in variability around the trend line or within cluster variability), accuracy of participants decreases on average. α_K has a small negative effect, i.e. participants are on average answering lineups with three clusters with more accuracy than lineups with five clusters. However, this effect is not significant. The order effect α_O is small, but significant; as participants answer more lineups, their accuracy decreases on average by about 2% for each additional evaluation α_O may represent a fatigue effect or an experience effect. Fatigue effects have been documented (Abdul-Rahman et al., 2014) in studies where participants complete several trials which are similar in a short period of time. It is also possible that as participants completed additional trials, they learned to look for certain cues (e.g. number of groups) which led to the selection of plots which had large differences in group size. While this approach is reasonable using inductive logic, participants using this strategy would be less likely to select the cluster or trend targets as a result.

As the lineups in this study were randomly ordered, either hypothesized effect is unlikely to bias the estimates for data set and aesthetic effects. In future experiments, it may be advantageous to utilize a two-part experimental design or add a short break to mitigate both potential fatigue and learning effects.

	Parameter	Estimate	Std. Error	z-value	Pr(>|z|)
	μ	6.47	1.08	6.01	< 0.0001
	α_T	-2.55	1.20	-2.12	0.0339
	α_C	-8.59	2.40	-3.58	0.0003
	α_K	-0.11	0.11	-0.93	0.3523
	α_O	-0.02	0.01	-2.33	0.0199
Design β	Plain	0.00	—	—	—
	Trend + Error	-0.08	0.14	-0.57	0.5701
	Shape	-0.08	0.14	-0.57	0.5694
	Trend	-0.35	0.13	-2.61	0.0090
	Color + Shape	-0.51	0.13	-3.95	0.0001
	Color	-0.60	0.13	-4.64	< 0.0001
	Color + Trend	-0.70	0.13	-5.49	< 0.0001
	Color + Ellipse + Trend + Error	-1.12	0.12	-9.05	< 0.0001
	Color + Shape + Ellipse	-1.40	0.12	-11.45	< 0.0001
	Color + Ellipse	-1.47	0.12	-12.09	< 0.0001

Table 2: Parameters and estimates of the accuracy model.

We additionally investigate two more effects: (log) response time and the effect of imbalances in the group allocation on accuracy.

C.1.1 Effect of response time

The effect of (log) response times on accuracy is highly significant (χ^2_1=80.7, P-value$<$ 0.0001). With each unit increase in (log) response time the probability for a target identification is reduced on average by about 1/3. However, in the long run, a secondary effect

takes place, and response time has a positive effect on accuracy again. Fitting an additional quadratic term in the model is also highly significant (χ^2_1=13.3, P-value=3×10^{-4}), and leads to an overall minimum accuracy over time at a response time of about 150 seconds.

This speed/accuracy relationship is more complicated than the typical speed/accuracy trade off, which assumes that accuracy increases with decreasing speed. Instead, when evaluating lineups, two effects seem to compete: Initially, participants may be able to fairly quickly (and accurately) select the target plot using high-bandwidth visual differences between the plots. If this initial strategy fails, participants may begin making pairwise comparisons in an attempt to select between one of several potential "targets". These comparisons are slow, but eventually participants may succeed in using inductive reasoning to identify the target plot, producing a slow increase in accuracy after about 150 seconds.

The accuracy of initial impressions made with relatively little information has been documented in the popular press (Gladwell, 2007) and in scholarly literature (Curhan and Pentland, 2007; Ambady and Rosenthal, 1993; Hogarth, 2014). Participants who select a target quickly are likely utilizing intuition as well as conscious reasoning; when combined with lineups which have a clear signal, this approach is fairly accurate. When intuition and initial approaches fail, however, participants must resort to a slower, more deliberate approach which may include attempting to determine both the evaluation metric to utilize and the most different plot when utilizing that metric. This process would take more time and have a higher error rate, particularly if the participant uses a strategy which is not aligned with the data-generating model (for instance, evaluating the size of the groups rather than the spatial clustering of points). The combination of these two effects could explain the quadratic speed/accuracy relationship seen in this model.

C.1.2 Effect of group imbalances

Gini impurity measures the homogeneity of group allocations. Let n_i be the number of elements in the ith cluster, $i = 1, ..., K$, with $n = \sum_{i=1}^{K} n_i$ and let $p_i = n_i/n$ be the frequency of cluster i. Then the gini impurity is calculated as

$$G(p_1, ..., p_K) = \frac{K}{K-1} \sum_{i=1}^{K} p_i(1 - p_i).$$

G is an index between 0 and 1, where 0 is maximum diversity - in the sense, that there is only one group present, i.e. $p_i = 0$ for all but one of the groups. A Gini impurity of 1 indicates perfect homogeneity, i.e. $p_i = 1/K$ for all i.

The probability of picking at least one of the targets significantly increases with an increase in Gini impurity, that is, with more equal group sizes (χ^2_1=5.9, P-value=0.0153). In other words, a small Gini index indicates the presence of a null plot with large differences in the number of elements in each group. This served as a(n unintended) distractor away from the intended targets and therefore led to a drop in accuracy. Different designs are affected differently strong (two-way interaction: χ^2_9=34.7, P-value < 0.0001): Plots making use of color are all affected more strongly by deviations from homogeneity than plots without color. If ellipses are drawn, the effect becomes even more pronounced. Figure 3 gives an overview of the relationship between predicted accuracy and designs for levels

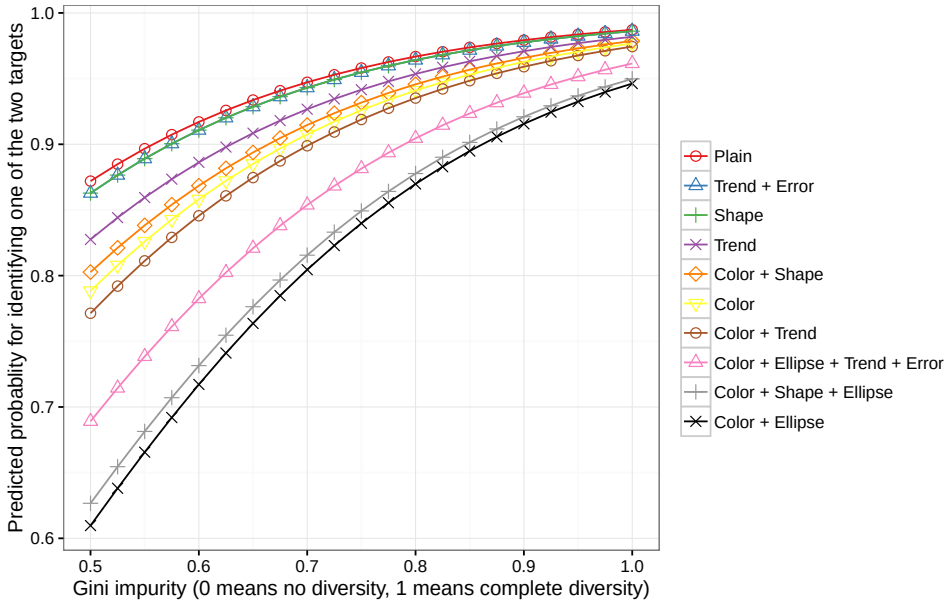


Figure 3: Predicted accuracy values for designs given different values of the gini impurity measure. Values of gini impurity between 0.59 and 1 were actually observed in the panels of the lineups. Accuracy of all designs with color are lower than their non-colored counterparts, and the addition of ellipses further decreases accuracy. With a large deviation from equi-distributed groups, accuracy is disproportionately affected when color or ellipse aesthetics are present.

of homogeneity between 0.5 and 1 (similar to the range of Gini impurity observed in the study).

Other features measuring the imbalance within a lineup that we considered besides Gini impurity are

- the difference between the maximum and the minimum number of elements in each of the groups of a lineup, and
- the number of ellipses missing from the lineup.

The range of group sizes does not have a significant effect on the probability to pick at least one of the targets, even if different designs are taken into account. Similarly, a single absent ellipse does not lead to a significant change in the probability of detecting one of the target plots. Neither the number of missing ellipses nor the absence of at least one ellipse have a significant effect on this probability, not even when we consider the impact of individual designs.

These features have also been included in the face-off model of section C.4 to investigate whether group size imbalance or missing ellipses affects the probability of selecting one target over the other.

C.2 Modelling response times

While we do not have the same amount of control in an AMT study that we would have in a lab setting, we can accurately capture the time between presenting a lineup to a

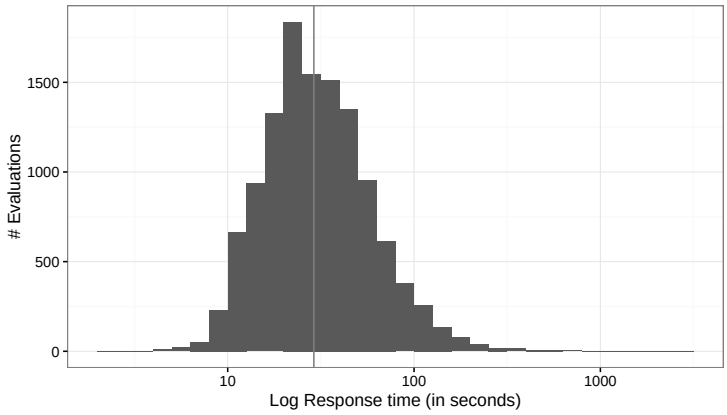


Figure 4: Histogram of (log) response times. The median evaluation time (vertical line) is 29 seconds.

participant and the time at which results are submitted. A histogram of these times is given in Figure 4. Response times are extremely skew. In the model we therefore use the log of response times $T = (t_{ijk})_{n\times 1}$ as the dependent variable.

For the vector of fixed effects, α, we estimate, as before, the effects of variability (α_C and α_T) on response times, as well as the number of clusters shown (α_K). Aside from this, we also consider the average effect of the first trial, α_1, and the effect of successive trials, α_O, on the response time. $\omega_C, \omega_T, \omega_B$, and ω_N is a set of parameters assessing the effect of the outcome of the lineup evaluation on response time (the letters in the subscript stand for 'C'luster, 'T'rend, 'B'oth, and 'N'one).

Table 3 gives an overview of all parameters of the response time model and their estimates. Similar to what has been found in other lineup studies (Majumder et al., 2014; Hofmann et al., 2012), participants take on average 25% longer to respond to the first lineup than to subsequent lineups. Aside from this, we see that as the difficulty of lineups increases (controlled by an increase in the parameters s_C and s_T), the average amount of time participants spend on each evaluation significantly increases. On average, participants respond to each successive lineup about 3% faster.

Depending on the outcome of the evaluation, there are differences in the amount of time: if either one of the targets is identified, the amount of time taken to answer is significantly shorter than if neither of the targets is found. Answers take on average the longest, if both targets are identified (however, this only happens in 0.6% of the responses). Plot aesthetics have a significant impact on the amount of time for responses, with increasing plot complexity associated with increased evaluation time. This may be a function of increased cognitive load, as participants must examine more features in order to identify which plot has the strongest signal. For instance, when color, ellipses, trend lines, and error bands are present, participants have to compare the allocation of color to points, the size, shape, and distance between each set of ellipses, the slope of each trend line, and the width of the error bands. While each participant almost certainly does not complete a full pairwise comparison of all 20 lineup plots across each feature set, the increased complexity of each additional feature does increase the space which must be examined using perceptual

heuristics in order to identify the target plot correctly. This is consistent with Borgo et al. (2012), who found that visual embellishments increase the time required to perform visual search tasks using data displays.

	Parameter	Estimate	Std. Error	z value	P-value
	μ	2.664	0.108	24.559	< 0.0001
	$\alpha_{K=3}$	0.000	—	—	—
	$\alpha_{K=5}$	0.124	0.029	4.297	< 0.0001
	α_T	0.481	0.153	3.152	< 0.0001
	α_C	1.664	0.302	5.503	< 0.0001
	α_1	0.084	0.016	5.133	< 0.0001
	α_O	-0.030	0.002	-17.273	< 0.0001
Outcome ω	Trend	0.000	—	—	—
	Cluster	0.027	0.013	2.117	0.0342
	Neither	0.181	0.015	11.736	< 0.0001
	Both	0.347	0.057	6.079	< 0.0001
Design β	Plain	0.000	—	—	—
	Shape	0.110	0.019	5.903	< 0.0001
	Color	0.130	0.019	6.961	< 0.0001
	Trend	0.144	0.019	7.714	< 0.0001
	Trend + Error	0.163	0.019	8.725	< 0.0001
	Color + Ellipse	0.206	0.019	10.895	< 0.0001
	Color + Shape	0.209	0.019	11.194	< 0.0001
	Color + Trend	0.215	0.019	11.469	< 0.0001
	Color + Shape + Ellipse	0.203	0.019	10.752	< 0.0001
	Color + Ellipse + Trend + Error	0.248	0.019	13.187	< 0.0001

Table 3: Model parameters and estimates for (log) response time in seconds. The P-values are based on a normal approximation of the t statistics.

C.3 Model of confidence levels

With each lineup evaluation, participants were asked to give feedback on their level of confidence from 0 (least) to 5 (most). As an approximation, we can fit a mixed effects model with this variable as the dependent, and investigate its relationship with the parameters controlling difficulty of a lineup, the time taken to evaluate the lineup and its outcome.

The approximation of confidence level (which is a bounded, discrete variable) by a normal distribution is far from perfect, but the results are very interpretable.

The vector of fixed effects, α, includes besides the previously described effects of the control parameters $\alpha_C, \alpha_T, \alpha_K$, the effects of lineup order α_1, α_O, the effect of evaluation outcome $\omega_C, \omega_T, \omega_B, \omega_N$, also parameters that control for the amount of time taken by participants to evaluate a lineup: τ is the effect of log response time on the level of confidence reported by a participant.

Table 4 gives an overview of the parameters and estimates of the model. Over the course of the study, confidence deteriorates with each additional lineup by about 0.03 (while significant, this effect on confidence is fairly small, as confidence is measured on a scale of 0 to 5). Beyond this order effect, the first trial does not have a significant effect on the reported confidence. The longer a participant needs to evaluate a lineup, the lower on average will be the value of confidence reported along with it. Similarly, an increase in lineup difficulty (as controlled by increased values of s_C and s_T) goes hand in hand with a significant decrease in confidence. If neither one or both of the two targets were identified, the reported confidence level is significantly lower than if one of the two targets was identified[1]. Aesthetics in general did not have a significant effect on confidence levels. However, individual aesthetics did lead to a significant increase in confidence: any plot showing ellipses increases the level of confidence on average by about 0.1. These results suggest that the speed of evaluation is not significantly contributing to shifting the balance between selecting one target over the other.

	Parameter	Estimate	Std. Error	z-value	P-value
	μ	5.898	0.147	40.044	< 0.0001
	α_O	-0.033	0.003	-10.738	< 0.0001
	α_1	0.002	0.029	0.083	0.9336
	τ	-0.295	0.016	-18.507	< 0.0001
	α_T	-0.538	0.198	-2.719	0.0066
	α_C	-1.899	0.389	-4.879	< 0.0001
	$\alpha_{K=3}$	0.000	—	—	—
	$\alpha_{K=5}$	-0.131	0.037	-3.559	0.0004
Outcome ω	Trend	0.000	—	—	—
	Cluster	-0.028	0.022	-1.263	0.2065
	Both	-0.213	0.102	-2.084	0.0372
	Neither	-0.220	0.028	-7.971	< 0.0001
Design β	Shape	-0.006	0.034	-0.187	0.8518
	Plain	0.000	—	—	—
	Color + Shape	0.024	0.034	0.698	0.4855
	Color	0.040	0.034	1.178	0.2390
	Trend	0.050	0.034	1.497	0.1343
	Trend + Error	0.052	0.034	1.533	0.1252
	Color + Trend	0.058	0.034	1.707	0.0878
	Color + Ellipse	0.068	0.034	1.997	0.0459
	Color + Ellipse + Trend + Error	0.106	0.034	3.132	0.0017
	Color + Shape + Ellipse	0.107	0.034	3.156	0.0016

Table 4: Parameters and estimates for the model of participants' confidence.

[1]The decrease in confidence when both targets are identified may be due to the additional complexity of dual-target search (Fleck et al., 2010; Cain et al., 2011; Adamo et al., 2015)

C.4 Face-off Model

Figures 5 and 6 show the proportion of outcomes for either the cluster target, the trend target, both or none of them. Overall, cluster targets are picked more often than trend targets. For very small residual errors around the line fit and large within-cluster errors, the number of line target picks are highest. As the standard error around the trend line increases, the number of times the corresponding target is picked decreases. Similarly, an increase in within-cluster error is associated with a decrease of the number of cluster target picks. The effect of the different designs is consistent across different parameter settings (the order of plot designs is given by the marginal effects as estimated in the face-off model. Numerical estimates can be found in Table 5). The effect of designs is most pronounced, when the ambiguity between the two targets is strong, i.e. close to a 50:50 decision between the targets. In those cases the additional aesthetics tip the balance in favor of one target over the other.

	Parameter	**Log Odds Ratio**	**95% Lower**	**95% Upper**
	Intercept	1.018	-1.615	3.651
	α_T	16.254	13.276	19.231
	α_C	-16.038	-21.935	-10.140
	α_K	-0.281	-0.563	0.001
Design β	Trend + Error	-0.650	-0.877	-0.423
	Color + Ellipse + Trend + Error	-0.515	-0.751	-0.279
	Plain	0.000	—	—
	Trend	0.130	-0.101	0.361
	Color	0.271	0.032	0.509
	Shape	0.277	0.043	0.510
	Color + Shape	0.314	0.076	0.551
	Color + Ellipse	0.459	0.207	0.711
	Color + Trend	0.459	0.219	0.700
	Color + Shape + Ellipse	0.573	0.317	0.830

Table 5: Odds Ratios of picking the cluster target over the trend target (with the plain design as a baseline). The last two columns are 95% confidence intervals. Within the plain plots, the odds of choosing the cluster target over the trend target is about 2:1.

Response time (composed of log(response time) and effect of first trial) does not have a significant effect on the decision between cluster and trend target (χ^2_2=4.5, P-value=0.1067). Nor does the confidence level of participants (χ^2_5=4.6, P-value=0.4716).

What does have a significant effect on the balance between cluster target and trend target is the absence of one of the ellipses in one of the panels of the lineup: a single missing ellipse (that is, a group size of less than 4) cuts the probability that the cluster target is selected by more than half (44.6%; χ^2_1=9.6, P-value=0.002). We also find a significant effect if we additionally take the two-way interaction between a single missing ellipse and individual designs into account (χ^2_9=20.4, P-value=0.0155). These effects are summarized in Figure 7.

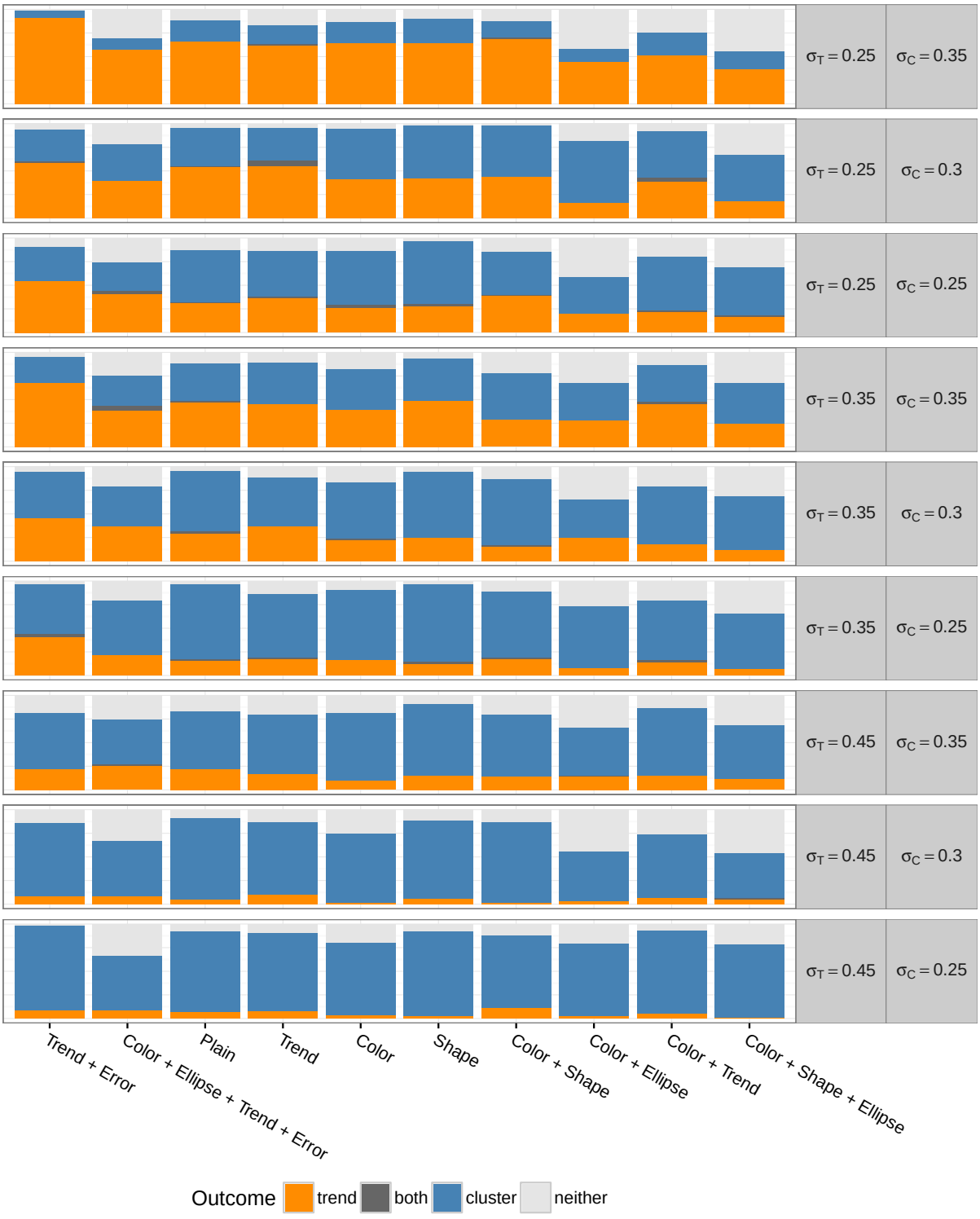


Figure 5: Outcome by design and parameter setting for lineups with trend and cluster targets. The cluster target consists of $K = 3$ clusters.

Figure 6: Outcome by design and parameter setting for lineups with trend and cluster targets. The cluster target consists of $K = 5$ clusters.

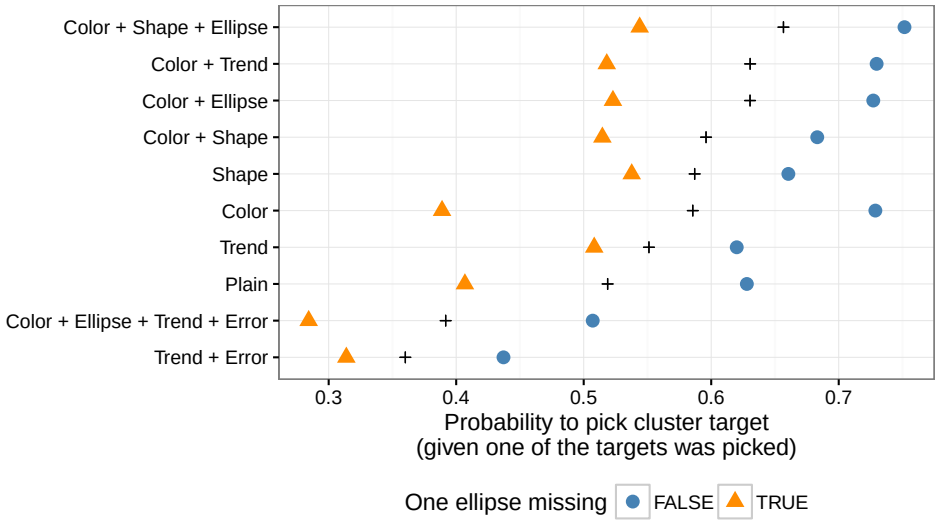


Figure 7: Overview of the probability to pick the cluster target given the different designs. s_C and s_T are set to 0.25 and 0.275, respectively, and $K = 3$ is assumed. The plus symbols indicate probabilities from the base model, filled triangles and circles represent predicted probabilities under a model including the two-way interaction between a single missing ellipse and designs. Plots with trend and shape aesthetics are the least affected by the imbalance in groups, while plots with color aesthetics show huge differences in the predicted probability.