Noname manuscript No.
(will be inserted by the editor)

Bayesian multi-level modelling for predicting single and double feature visual search

Replication and Generalisation of the Target Contrast Signal Theory

Anna E. Hughes · **Anna Nowakowska** ·
Alasdair D. F. Clarke

Received: date / Accepted: date

Abstract Performance in visual search tasks is frequently summarised by "search slopes" - the additional cost in reaction time for each additional distractor. While search tasks with a shallow search slopes are termed efficient (pop-out, parallel, feature), there is no clear dichotomy between efficient and inefficient (serial, conjunction) search. Indeed, a range of search slopes are observed in empirical data. The Target Contrast Signal (TCS) Theory is a rare example of quantitative model that attempts to predict search slopes for efficient visual search. In particular, the search slope in a double-feature search (where the target differs in both colour and shape from the distractors) can be estimated from the slopes of the associated single-feature searches. This estimating is done be using a contrast combination model, and previous work shows that a collinear contrast integration model outperformed other options. In our work, we extend the TCS to a Bayesian multi-level framework. We demonstrate that moving from using a normal distribution to a shifted-lognormal allows for a better fit of previously published data, and in doing so, we show that previous datasets are not sufficient to distinguish between the various contrast combination models that had been considered. We will conduct experiments that attempt to replicate the key original findings, with some changes to help distinguish between theories.

Keywords Visual search · Efficient search · Parallel processing

ESRC grant?

F. Author
first address
Tel.: +123-45-678910
Fax: +123-45-678910
E-mail: fauthor@example.com

S. Author
second address

1 Introduction

Visual search, where participants are asked to find a target within a cluttered scene, has been extensively studied within psychology. Several models have been developed that can generate testable predictions about how different types of distractors and targets affect search efficiency. One of the key distinctions in the field has been between efficient (also referred to as parallel or pop-out) and inefficient (serial) search. These are often studied in the context of the regression slope between the number of distractors and mean reaction time, which has been termed the *search slope*. When the search slope is shallow (usually positive, but sometimes negative), the search is called efficient or parallel, and the addition of more non-target distractors has little impact on an observers difficulty in finding a target. When the slope is steeper, each additional distractor has a noticeable impact on increasing difficulty, and the search is described as inefficient or serial. However, the distinction between these types of search is often less clear in real experimental data, with a range of different search slopes being seen for different types of targets and distractors (Cave and Wolfe 1990).

In the current study, we are interested in what has traditionally been termed efficient or parallel search, and the factors that affect search slope in these conditions. Recent work has suggested that for efficient search, there is a logarithmic relationship between distractor set size and reaction time, and that this relationship can be modified by target-distractor similarity (Buetti et al. 2016), providing evidence that search behaviour in parallel search is more complex than has previously been assumed. This observation has formed the basis of the 'Target Contrast Signal (TCS) Theory' (Lleras et al. 2020), which aims to provide a means of predicting observer search slopes for new search arrays by quantifying target-distractor differences. For example, by measuring search slopes for conditions with single feature distractors (e.g. colour or shape), it has been shown that you can predict search times for arrays with double feature distractors that differ from the target in both colour and shape (Buetti et al. 2019). Here, we aim to replicate and extend this work both theoretically and empirically, to test the generalisability of the TCS model, and to suggest ways in which the TCS model could be improved to generate better predictions.

1.1 Previous Work

Many different forms of visual search models have been proposed. One well developed class of models are the saliency models, which aim to predict eye movements during scene viewing, including visual search. They rest on the assumption that fixations are directed to objects or locations that are most dissimilar to the background or other objects in the visual display (Itti and Koch 2000; Itti et al. 1998; Koch and Ullman 1987). While the original saliency model was able to predict fixation allocation in a visual search task above chance (Parkhurst et al. 2002), further research demonstrated that a comparable level of performance could be achieved using a simple central fixation bias heuristic (Tatler 2007). The saliency models have since been extended and improved (see for example Zhang et al. (2008)): however, the main issue with this family of models remains their limited usability in complex real-life

search arrays (Tatler et al. 2011; Koehler et al. 2014). In addition, in most instances of visual search, the target is clearly defined (i.e. the goal is to find a specific object) and inspecting the most salient areas of the display may in these cases be inefficient. Finally, by focusing on eye movements, these models do not provide a theoretical framework for the cognitive processes underlying visual search.

Perhaps the most established class of models of visual search are based around Feature Integration Theory (Treisman and Gelade 1980), which has been modified and extended by Wolfe and colleagues in the Guided Search Model (Wolfe et al. 1989; Wolfe 2014). These theories have been developed using data from visual search tasks with discrete sets of abstract items. These models combine top-down influences (how closely an item resembles the observer's goal) with bottom-up image properties. For example, if one's goal (top-down processing) is to find a red horizontal bar, all the red and horizontal items in a visual search display will be given greater weight than distractors (e.g. vertical and blue items) in the model. The salience of a given object in the display (how distinctive it is from the surrounding objects) also activates bottom-up processing. For instance, a blue item among red items is ranked higher than red among orange items. In such cases, a salient item can capture attention even without resembling the target. Combining bottom-up and top-down sources of activation generates an activation map which generates a prediction of the order in which stimuli are processed in visual search. Thus, these models aim to produce a representation of the visual properties of the distractors at each location in the visual field. However, these are predominantly verbal models, and thus it is difficult to use them to make specific quantitative predictions.

TCS falls under a class of models that take a different approach, in that they focus solely on representing the difference between targets and distractors. For example, in work on eye movement patterns, it has been proposed that performance in inefficient (serial) visual search is mostly determined by the size of the 'functional viewing field', whose size varies as a function of target-distractor similarity (Hulleman and Olivers 2017). Similarly, work on attention has proposed the notion of 'relative features', where attention is tuned to feature relationships i.e. the appearance of the target relative to distractors in the environment (Becker et al. 2014; Becker 2010). However, TCS (Lleras et al. 2020) aims to provide a unifying framework that can make quantitative behavioural predictions for visual search based on this general assumption. As such, it is an attractive candidate model for a formal registered replication.

A key assumption of the TCS model is that behaviour is determined by comparing the target template (held in memory) with every element present in the scene in parallel. This allows the visual system to reject peripheral non-targets quickly; the speed at which items are evaluated is determined by how different the item is from the template through an evidence accumulation process (formally, the slope of the logarithmic function is assumed to be inversely proportional to the overall magnitude of the contrast signal between the target and distractor). The model thus focuses on an initial, efficient processing stage of search; if sufficient evidence is not accumulated during this process, the model posits that a second stage is entered, requiring a sequence of eye movements to search for the target in a serial manner. TCS has been successful in predicting a number of empirical results, including search performance in heterogeneous scenes based on parameters estimated in homogeneous

scenes, both with artificial stimuli (Buetti et al. 2016; Lleras et al. 2019) and with real-world objects visualised on a computer display (Wang et al. 2017). Table 1 provides an overview of studies investigating the TCS framework to date.

The original version of the TCS model is given below:

$$\hat{RT} = a + \sum_{j=1}^{L} (D_j - D_{j-1}) \log \left(N_T - \left(\sum_{i=1}^{j-1} N_i \right) I_{[2,\infty]}(L) + 1 \right) \quad (1)$$

This is essentially log-linear model in the number of distractors. The intercept, a, corresponds to stimuli in which only the target is present and there are no distractors. L represents the number of different types of distractors present in the display, while N_i represents the number of individual distractors present of type i. N_T is the total number of distractors, i.e., $N_T = \sum_{i=1}^{L} N_i$. Finally, $I_{[2,\infty]}$ is an index function that equals 0 when there $L < 2$, and 1 otherwise[1]. In our paper, we will follow Buetti et al. (2019) and only consider the specific case of $L = 1$, of a target among a homogeneous set of distractors As such, the Equation 1 simplifies to:

$$\hat{RT} = a + D \log(N_T + 1) \quad (2)$$

1.2 Rationale for proposed work

While results to date for TCS appear promising, they remain relatively preliminary, being tested by only one research group so far. There remains a great deal of scope for extending beyond the parameters tested to date (such as the number of distractors) in order to test the robustness and generalisability of the model. In addition, in all implementations of TCS so far, predictions of search efficiency (e.g. in heterogeneous scenes) have been made on the average of a group of participants, using data from a different group performing a different task (e.g. searching in homogeneous scenes). Thus, we know that TCS can replicate group-level averages between subjects in search well, but we do not know whether it is also able to make predictions at the individual level. This is particularly important given that conclusions based on aggregate data can be different from those that take individual differences into account; in one study where participants searched for a target in an array of randomly oriented line segments, aggregating the data suggested that participants were using a stochastic search model. However, when considering each participant individually, it became clear that there was a high level of heterogeneity in responses, with some participants performing close to optimally, and others actually performing worse than chance (Nowakowska et al. 2017). Similarly striking variability has also been reported in other search studies (Irons and Leber 2016, 2018).

In the current manuscript, we focus on replicating and extending findings from Buetti et al. (2019). In this study, participants searched for a target in a scene of homogeneous distractors. First, parallel search efficiency (measured by the logarithmic search slope) was estimated for cases where the distractors varied from the target in

[1] Note: Equation 1 in Lleras et al. (2020) had j here rather than L. The current version is correct.

Reference	Overview
Buetti et al. (2016)	For efficient search with a specific target, there is a logarithmic relationship between distractor set size and reaction time. The steepness of this relationship is modulated by distractor-target similarity, with steeper slopes for more similar distractors.
Wang et al. (2017)	Data from homogeneous search arrays can be used to predict search reaction times in heterogeneous displays, using an equation assuming parallel, unlimited capacity, exhaustive processing, and independence of inter-item processing.
Madison et al. (2018)	Logarithmic efficiency in efficient search cannot be explained by crowding in peripheral vision.
Ng et al. (2018)	Logarithmic efficiency in efficient search cannot be explained by eye movements.
Lleras et al. (2019)	Validation of previous results showing data from homogeneous search arrays can be used to predict reaction times in heterogeneous displays. Distractor-distractor interactions can also facilitate processing.
Buetti et al. (2019)	Data from search arrays where the distractors are distinguished from the target by one feature can be used to predict search reaction times in displays with compound stimuli, defined by two features. Reaction times can be predicted using a collinear contrast integration model, which assumes that the overall target-distractor contrast is the sum of the contrasts from the two feature vectors separately.
Lleras et al. (2020)	Full proposal of the Target Contrast Signal Theory, proposing that the initial stage of processing computes a difference signal between each item in the scene and the target template, using this to determine which items are in the scene are unlikely to be the target.
Ng et al. (2020)	Attention works in a two stage process, first discarding target-dissimilar distractors in a distributed, parallel way. Focused spatial attention then visits target-similar items at random.

Table 1 An overview of work on the Target Contrast Signal Theory. The key paper for our replication is highlighted.

one dimension: either colour (e.g. a cyan target being searched for in either yellow, blue or orange distractors) or shape (e.g. a semicircle target in either circle, diamond or triangle distractors). New participants then searched for the same targets in displays where the distractors were compounds, differing from the target in both colour and shape (e.g. searching for a cyan semicircle in either blue circles, orange diamonds or yellow triangles). Figure 1 shows example stimuli from their paper. The logarithmic search slopes in the initial experiments were then used to predict the logarithmic slopes and reaction times using a number of models. The authors found that the best model was a 'collinear contrast integration model' where the distinctiveness scores were summed along each attribute in the unidimensional experiments, creating an overall contrast score that was used for compound stimuli predictions.

We first run a replication of Buetti et al. (2019), in an online, within-subjects study. This design allows us to extend the modelling, both incorporating a multilevel design to predict within-subjects effects and by utilising a Bayesian generalised linear model framework to better represent the distribution of responses (e.g. avoiding predicting negative reaction times, accounting for uncertainty in model predictions). We also carry out a direct analytical replication using the same methods as in Buetti et al. (2019) allowing us to ask whether the choice of analysis affects the results.

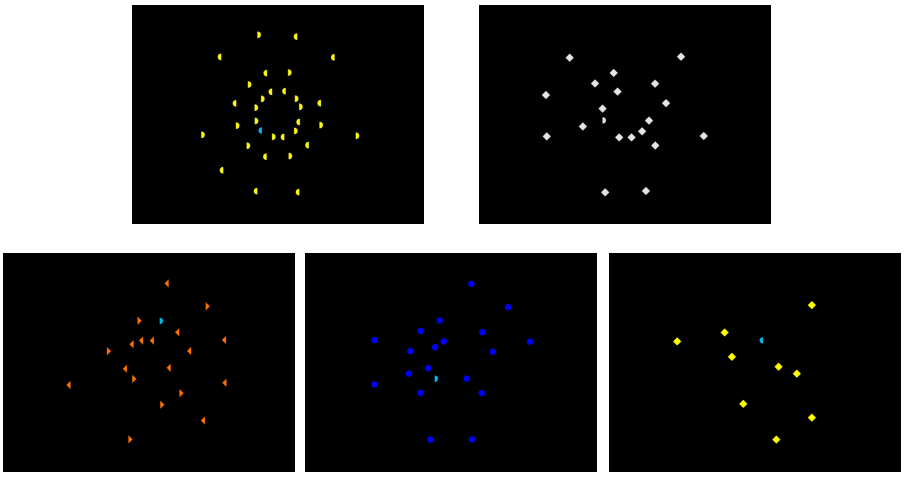

Fig. 1 Example stimuli from Buetti et al. (2019) Top left: Expt 1A. Here, the target is a blue semicircle within a set of homogeneous (yellow semicircle) distractors. Top right: Expt 1B. The target is a grey semicircle in circular grey distractors. Bottom left: Expt 2A. The target is a blue semicircle in orange diamond distractors. Bottom middle: Expt 2B. The target is a blue semicircle in dark blue triangle distractors. Bottom right: Expt 2C. The target is a blue semicircle in yellow circular distractors.

2 The Target Contrast Model

We first describe the original Target Contrast Model, as presented in Buetti et al. (2019) and verify that we can succesfully replicate the original analysis (see Supplementary Materials). This is followed by Section 2.2 in which we detail our proposed modifications and show how they change the interpretation of the data collected by Buetti et al. (2019).

2.1 TCS modelling overview

In Experiment 1a of Buetti et al. (2019), participants searched for a cyan semicircle target among blue, yellow or orange semicircular distractors i.e. they searched for a target that differed from the distractors by a *single feature* (colour). The experiment was then repeated (1b) using a different single feature (shape, with participants searching for the semicircular target within triangle, circle or diamond distractors). In Experiments 2a, 2b and 2c, participants again searched for a cyan semicircle, but this time, the distractors differed in both shape and colour. We will refer to these conditions as *double features*. Note, unlike in standard conjunction searches, in this paradigm, the distractors are all identical with respect to these features (i.e, orange triangles). Examples of all these stimuli are shown in Figure 1.

The *Target Signal Contrast* theory is built around a linear model for predicting mean reaction times from the logarithm of the number of distractors (see Equation 2). In particular, the TCS theory allows us to predict the value of the logarithmic slope,

$D_{c,s}$, in this condition based on the corresponding D_j in the single feature search experiments.

2.1.1 Calculating the intercept, a, and the logarithmic slope parameter, D_i

Experiments 1a and 1b (referred to jointly as Experiment 1) were used to calculate the logarithmic slope parameter D_i. In both experiments, the number of distractors varied, allowing the data to be used to fit a log-linear model for reaction times, where reaction times increase logarithmically with N_T, the number of distractors. (See Equation 2.) In the original model the error distribution was assumed to be normal. Thus the results of Experiment 1 were used to calculate D_i, for each type of distractor. When colour varied, we will refer to D_c, for $c = 1,2,3$. Similarly for shape we will denote this (D_s), and the compound features are denoted as ($D_{c,s}$).

Fitting the model specified in Equation 2 to the data, we obtain the values for D_c and D_s given in Table 2. As can be seen, the more similar the distractors are to the target, the steeper the slope parameter is.

feature	D_c	a	feature	D_s	a
blue	76.8	575.1	triangle	141.1	551.6
yellow	16.0	575.1	diamond	77.2	537.7
orange	9.8	575.1	circle	62.1	555.8

Table 2 A table of D_i values for Experiment 1a and 1b

As a is the intercept in our model, it represents how long observers take to find a target when $N_T = 0$, i.e., there are no distractors. As such, it should be independent of both shape and colour. However, this leads to some ambiguity in how a should be defined; in Buetti et al. (2019), a was calculated for each feature in each sub-experiment (e.g. Experiments 1a, 1b, 2a, 2b, 2c; see Table 2). Here, we follow that method in order to replicate their results exactly. Unfortunately, this leads to a situation in which the model's prediction of how long participants take to find a target when N_T depends on the features of distractors that are not present in the stimulus! As such, in our proposed new version of the TCS theory, we will estimate a at the level of an experiment (e.g. Experiments 1 and 2).

2.1.2 Estimating $D_{c,s}$, the logarithmic slope parameter for compound features

In the context of the current experiments, the core idea of TCS theory is that we can estimate the logarithmic slope parameter for a double feature visual search from the slopes parameters in the two independent single feature searches. I.e., $D_{c,s} = f(D_c, D_s)$. Buetti et al. (2019) tested three different models for predicting D for compound colour-shape stimuli. The best feature guidance model (Equation 3) suggests that when the target and lures differ in two dimensions, participants will choose to attend to whichever feature dimension is the most discriminable (i.e. has the smallest D value):

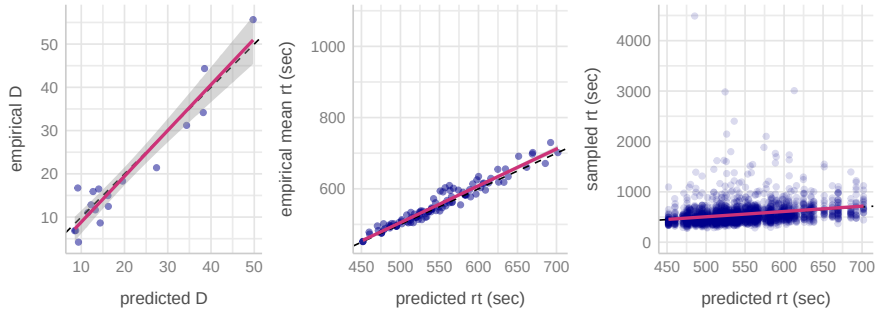

Fig. 2 (left) The collinear method for calculating D offers a good prediction. (centre) Using the TCS to predict reaction times. (right) Each dot now represents a randomly sampled reaction time from an observer.

$$D_{\mathrm{c,s}} = \min\left(D_{\mathrm{c}}, D_{\mathrm{s}}\right) \tag{3}$$

The orthogonal contrast combination model instead suggests that independent feature dimensions comprise a multidimensional space, where an object can be described by the overall vector in this space, and thus $\mathrm{D_{c,s}}$ can be represented as[2]:

$$D_{\mathrm{c,s}} = \frac{1}{\sqrt{\frac{1}{(D_{\mathrm{c}})^2+(D_{\mathrm{s}})^2}}} \tag{4}$$

Finally, the collinear contrast integration model also assumes independence of feature dimensions, but assumes that while the visual features create a multidimensional space, the contrast between them is unidimensional. As D is assumed to be inversely proportional to contrast, the equation can be written as follows:

$$\frac{1}{D_{\mathrm{c,s}}} = \frac{1}{D_{\mathrm{c}}} + \frac{1}{D_{\mathrm{s}}} \tag{5}$$

Buetti et al. (2019) found that with their dataset, the collinear contrast integration model was best able to predict $D_{c,s}$ from D_c and D_s, with $R^2 = 0.915$. We verified we were able to replicate this result using the dataset available on OSF (https://osf.io/f3m24/)[3] and using the exclusion criteria originally applied; see Figure 2 (left panel) and *Supplementary Materials* for details.

2.1.3 Estimating mean reaction times

Finally, we can use Equation 2 to predict mean reaction times. As can be seen in Figure 2 (centre panel), these predictions are essentially identical to the empirical RT results: $R^2 = 0.93\%$.

[2] Note: there is a small mistake in Equation 4 in Buetti et al. (2019). The version given here is correct.

[3] downloaded on 28th August 2020

2.1.4 Discussion

While TCS theory offers a good prediction of search slopes and corresponding mean reaction times for double feature search, there are two related limitations. Firstly, it is unable to account for individual differences between observers, only the changes to the sample average. Secondly, it cannot account for the distribution of reaction times over multiple trials. Figure 2 (right panel) shows clearly that these factors generate high levels of variability within the individual trial-level data. To address these issues, we propose a second version of the TCS that make use of multi-level modelling techniques.

It is also worth noting that relying simply on R^2 values may give a misleading view of the accuracy of the TCS model. Figure 3 (top) shows the predicted RTs (pink line) for three conditions from Experiment 2 (a, b and c) of Buetti et al. (2019), plotted next to the mean of means for each condition. It is clear that for at least some conditions, the TCS model is doing a poor job of predicting the empirical data, even when allowing quite rigid constraints on the model; for example, the intercept (a) is set at the level of each sub-experiment and thus would be expected to be perfect. To some extent, this could reflect the fact that the R^2 measure of goodness-of-fit simply measures how tightly clustered the points are around a line with arbitrary intercept and slope. It is possible therefore for the R^2 value to be high, but for the line to deviate significantly from an intercept of 0 and a slope of 1, which is the unity line required for the predictions to be a good match to the empirical data. For our reimplementation of the TCS model, we will therefore assess model fit by computing the log marginal likelihoods via bridge sampling. This directly assesses how well our values of $D_{c,s}$ are able to predict RTs, avoiding the issues with the R^2 measure.

Finally, if we look at the different predictions made by the three proposed feature combination models, we can see that in most of the parameter space explored in Buetti et al. (2019), the models all make very similar predictions (Figure 3, bottom). In Experiment 2, we are only able to discriminate between models in the *blue* conditions, whereas in the *yellow* and *orange* conditions, they make essentially identical predictions. As such, if we want to distinguish between these three models, we will have to combinations of features that lead to larger target-distractor combinations. We will therefore use yellow, green and blue distractors in our replication study where participants are required to find a cyan target, hopefully generating a wider range of target-distractor differences.

2.2 A multi-level TCS

Switching from a linear regression model to a multi-level model will allow us to compute D for each participant, while simultaneously estimating the trial-to-trial variance. We also switch from a frequentist to Bayesian framework, as this allows us to naturally account for the uncertainty in the model's predictions.

However, switching from linear regression to a multi-level model raises the problem of which distribution to use for modelling reaction times. Using a normal distribution is unlikely to be satisfactory, as it is unable to account for the skew frequently

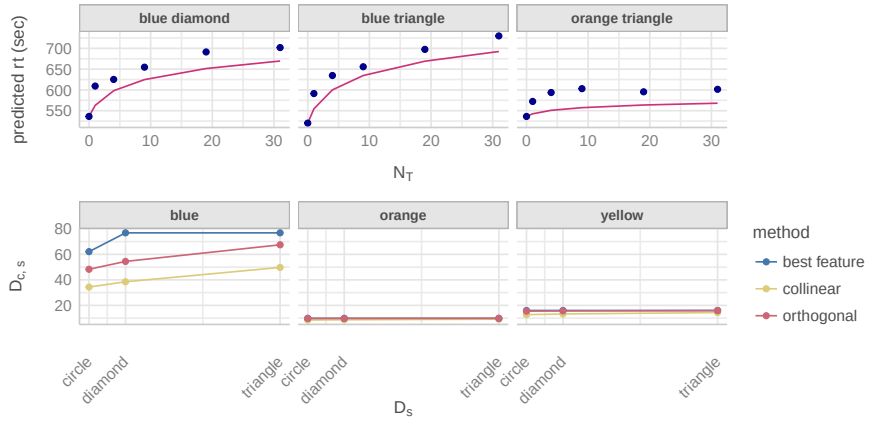

Fig. 3 (top) Three conditions from Experiment 2. Blue dots represent the mean of means for each condition, while the line shows the prediction of the TCS theory. (bottom) Estimates of $D_{c,s}$ for the different colours and shape from Buetti et al. (2019). As can be seen, in two of the three colour conditions, we are unable to distinguish between best feature, collinear or orthogonal contrast models.

seen in reaction time distributions, and also allows the possibility of negative reaction times. We can account for both of these problems by using a log-normal distribution, rt lognormal(μ, σ)). We will also test whether a slightly more complex extension of this model, the *shifted lognormal* model (which allows the distribution to be offset to the right i.e. mimicking the patterns seen in reaction time data, where valid responses begin at around 100ms) offers any improvement in model fit.

Throughout the paper, we will make use of the 53% and 97% *highest posterior density intervals* (HPDI) to summarise probability distributions. These can be thought of as the smallest interval that, given our data and assumptions, contain 53% and 97% of the probability mass.

Finally, when comparing models, we discuss that R^2 value on its own is not an adequate measure of the goodness-of-fit of the model, and that the match to the unity line also needs to be considered. We suggest adjustments to the original models proposed by Buetti et al. (2019) to improve the predictive power of the TCS model: namely that we should assess goodness-of-fit of the model, using bridge sampling to calculate the log marginal likelihoods.

2.2.1 Calculating the logarithmic slope parameter, D_i

We began in a similar fashion to section 2.1.1, although this time, we fitted each model three times using a (i) normal, (ii) lognormal, and (iii) shifted lognormal distribution (please see Supplementary Materials for full details of our modelling approach, including prior predictive checks and model fit diagnostics). The three models were compared using bridge sampling to calculate the log marginal likelihoods, and the results showed that the shifted-lognormal model (illustrated in Figure 4) was

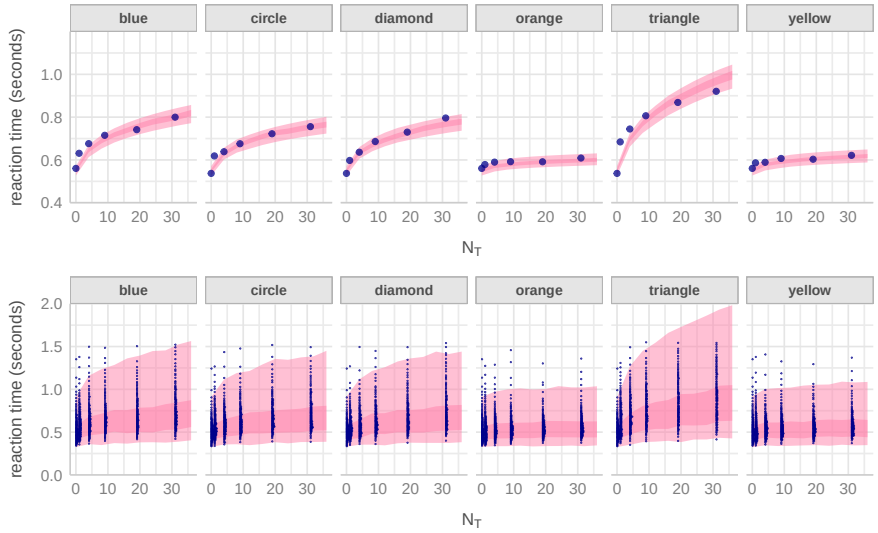

Fig. 4 (*top*) The shaded regions show the model's estimate (53% and 97% intervals) of the average participant's mean reaction time, while the points show the empircal mean reaction time. (*bottom*): The shaded regions now indicate the distribution of reaction times (over a new simulated group of participants) generated by the model. The points now represent the 100 quantiles from the empirical data.

given 100% of the weight[4]. Therefore, we used only this model for the rest of the analysis.

2.2.2 Estimating $D_{c,s}$, the logarithmic slope parameter for compound features

The first step in predicting the values for $D_{c,s}$ is to extract the logarithmic slope parameters D_c and D_c from our model outlined above (Figure 5 (top)). This is the same as the original TCS, except this time we use 1000[5] samples from the posterior distributions, rather than just the maximum likelihood fit. The then combine D_c and D_c using the best feature, orthogonal contrast, and collinear contrast integeration models to get three different predictions for $D_{c,s}$. These are illustrated in Figure 5 (bottom).

The R^2 values for the three approaches are shown in Table 3. We can see that as with the original model, we still find the *best feature* is the worst of the three methods, at least in terms of the raw R^2 value. However, it is less clear which of the other two methods gives the best performance. While the collinear contrast approach offers the highest R^2 (0.824-0.963), it consistently predicts values that are too low. The orthogonal contrast method gives more accurate predictions, albeit with a slightly lower R^2 (0.739-0.944). As such, we propose a different method, (see below) to distinguish between these approaches.

[4] we also tested a version of the shifted lognormal model that did not use the logarithm of the number of distractors i.e. testing this assumption of the TCS model, but found no support for this model being better: see Supplementary Materials for further details.

[5] check!

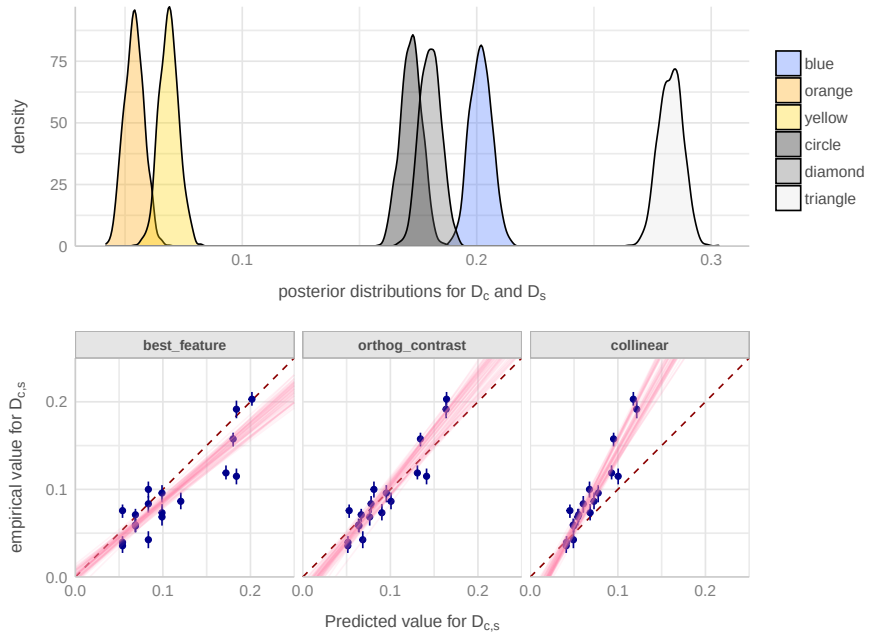

Fig. 5 (*top*) Posterior probability density functions for each D_c and D_s in Experiment 1. As with the original TCS model, when searching for a blue semicircle, triangular distractors are the hardest, while yellow and orange distractors are the eaiest.) (*bottom*) Predicting $D_{c,s}$ from D_c and D_s. Crosshairs indicate 97.5 HPDI for estimates and predictions, while the pink region shows the HPDI for the best fit line.

Method	Intercept	Slope	R^2	R^2 Lower	R^2 Upper	Model Weight
best feature	0.003	0.831	0.752	0.545	0.927	34.0%
orthogonal contrast	-0.019	1.206	0.856	0.739	0.944	33.6%
collinear	-0.034	1.814	0.901	0.824	0.963	32.4%

Table 3 A table of R^2 values

2.2.3 Estimating other parameters and predicting reaction times

Unlike the original TCS, before we can use the predicted values of $D_{c,s}$ to generate reaction times, we first have to estimate some additional parameters:

- α - the intercept for the shift[6] parameter.
- ϕ_a - the random intercepts for the linear predictor for μ.
- ϕ_α - the random intercepts for a.
- σ - the residual (trial-to-trial) variance.

For the implementation presented here, in order to predict the data from Experiment 2 we will simply use the values for the above parameters obtained from Experiment 1. Once these have been set, we can use the specified model to generate

[6] ndt

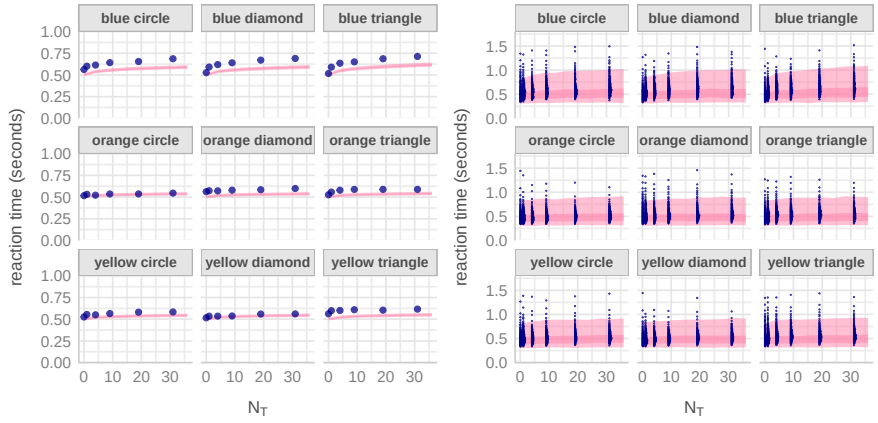

Fig. 6 (*left*) HPDI and empirical mean reaction times for each condition in Experiment 2, using the collinear contrast integration model to predict D values. (*right*) Similar, for the full distribution of reaction times over trials.

the HPDI intervals for the expected average performance and the full distribution of reaction times over a number of simulated new observers (Figure 6).

We can summarise how well the model predicts the mean reaction times (Figure 6, left) by repeatedly i) sampling from the posterior, ii) computing the predicted mean reaction times, iii) computing the R^2 statistic for the correlation between predicted and empirical mean reaction times (as done in Buetti et al. (2019)) and iv) calculating a measure of goodness-of-fit of the model predicting RTs from the $D_{c,s}$ values.

At first glance, our model appears to be offering a poorer fit with the the data than the original TCS. This is due to our decision to estimate a (the intercept) from the Experiment 1 data, whereas Buetti et al. (2019) estimated three values for a, one for each sub-experiment, and these values were estimating from the Experiment 2 empirical data. With the within-subject design we propose for our replication study, this issue should not arise. Our new model gives us a HPDI of $[0.78, 0.83]$ for the R^2 value when fitting the best fit model using the $D_{c,s}$ predicted values from the collinear contrast integration method (see Table 3), which compares with the value of $R^2 = 0.93$ obtained for the original TCS.

However, our goodness-of-fit measures showed that there was little difference between our models: the posterior model probability of the collinear contrast model was 32.4%, compared to 33.6% for the orthogonal contrast model and 34.0% for the best feature model. There is therefore no strong evidence for one model over another. This contrasts with Buetti et al. (2019), who concluded that the collinear contrast integration model offered the best fit to their data. They reasoned that this was due to the independence of the two features in their study, and they would not expect this result to generalize to integral dimensions[7], in which case they expect the *orthogonal* contrast combination model to offer the best predictions. However, our

[7] cite Garner for odd terminology?

analysis suggests that there is no strong evidence that any of the models tested is superior to the others. As we can see in Table 3, the R^2 values are very similar across methods.

Overall, while our new model appears to fit slightly more poorly than the than the original method, as discussed above, it offers substantial improvements over the original model in terms of predicting the full distribution of reaction times over multiple trials and participants. In addition, as we discuss previously, the R^2 value may not capture goodness-of-fit optimally, and more sensitive goodness-of-fit metrics suggest that the models have equivalent predictive value.

2.3 Discussion

Our extension to a multi-level TCS model allows us to go beyond the original version and predict the full distribution of reaction times over samples of known, or new, observers. Moving from a normal to a shifted-lognormal distribution allows us to accurately model the skew seen in the empirical data, and avoid predicting negative response times. Our modelling has also highlighted other issues that we will address during our replication study: firstly, using a within-subjects design allows a to be calculated in a more principled manner, and secondly, we will choose distractor features to give the highest possible D sensitivity, and thus give us greater power to detect true differences between models.

3 Hypotheses

We plan an experiment to test the extent to which the original results in Buetti et al. (2019) replicate and generalise.

3.1 Proposed Modifications to Experimental Design

In order to better test the above, and increases sensitivity, we propose to make the following changes to the experiment described in Buetti et al. (2019):

1. **Within-subjects design** This modification should give us greater power to detect differences between different models, as well as allowing us to investigate how individual differences in the single-feature task might explain differences in the double-feature task.

2. **Increase target-distractor similarity** Our reanalysis of Buetti et al. (2019) indicates that the orange distractor condition does not distinguish well between different contrast models. We will therefore replace the orange distractor condition with a green distractor condition, which is more similar to the light blue target.

3. **Increase number of distractors, and more peripheral distractors** The TCS model, if it is to be a useful predictor of human behaviour, should generalise beyond the conditions originally tested. We therefore will add an extra condition, with 43 distractors. This requires the addition of another ring of possible distractor positions compared to Buetti et al. (2019), allowing the distractors in all conditions to be placed in more peripheral locations.

4. **Online data collection** This again tests the generalisability of the results obtained so far in laboratory conditions.

3.2 Registered Hypothesis

1. **Shifted lognormal model** We hypothesise that a shifted lognormal model will give the best fit to our single-feature data, when compared to a lognormal and a normal model.

2. **Log-linear effect of N_T.** We will test the TCS model assumption that N_T has a log-linear effect by testing models with and without the log of this term.

3. **Contrast model comparisons** We will test the hypothesis proposed by (Buetti et al. 2019): specifically, that the *collinear contrast ingratiation model* outperforms the *best feature guidance*, and *orthogonal contrast combination models*.

We will test each of these hypotheses by calculating the marginal likelihood of the relevant models, and then calculating the poster probabilities. This will give us a probability for each model that represents the likelihood that the model gives the best prediction. We will consider there to be evidence for one model over the others if a given model has a probability above 90%. We will consider there to be strong evidence for one model over the others if that model has a posterior probability above 99%.

3.3 Planned Explorations

We plan to investigate the effect of individual differences in this paradigm: to what extent performance in the single-feature task can predict performance in the double-feature task for a given individual (Buetti et al. (2019) were not able to investigate this due to the between-subjects design of their study). We plan to do this by specifying a more complex random effects structure for the model, that allows for individual differences across different slopes for different features. This allows us to then study the random effect correlation structure. However, given these models can be challenging to fit, we will do this in an exploratory manner after carrying out our formally registered analysis.

4 General Methods

4.1 Sample Size: Participants and Trials

To determine our sample size, we carried out a simulation study to estimate the effect of reducing the number of trials and increasing the number of participants on our ability to accurately measure the D_i. While our reanalysis of Buetti et al. (2019)'s data is unable to distinguish between the three models of contrast combination, we believe that this is due to the feature values used in their study, rather than low statistical power from sampling (see above and Figure 3). Therefore, we will base our sensitivity analysis around the width of the 97% HPDI when estimating D_{blue}. We have chosen to look at this single feature level as it is clear from above that the blue distractor trials offer greater discrimination between the three contrast models.

We carried out our simulation sensitivity analysis by generating synthetic data from the shifted lognormal model outlined in Section 2.2, Figure 4, for the blue distractor conditions. This was done for several different sample sizes, both in terms of number of trials and number of participants. For each synthetic dataset, we then refit the model and calculated the width of the 97% HPDI (see Figure 7). F

As we are moving from the within-subjects design used by Buetti et al. (2019) to a within-subjects design in which each participant sees all combinations of features, we prefer to decrease the number of trials per person $\times$ condition from 40 to six. Increasing the number of participants from 20 to 80 will allow us to maintain sufficient accurcay in the estimating of logarithmic search slopes.

Ethics?

4.2 Stimuli

The targets and distractors were randomly assigned to the display based on an invisible grid. Within each quadrant of the screen, there were three 'spokes' each with four possible target positions (starting from the centre of the screen and moving outwards), creating 48 different target positions in total, in four concentric circles. A small amount of jitter was added to each possible position to make the target locations less predictable.

Distractor and target types: we replicated the distractor types used in Buetti et al. (2019), apart from that we changed one distractor colour (from orange to green) to allow us to discriminate better between different models of the data (see above). There were six single-feature conditions (yellow, green and blue distractors and triangle, circle and diamond distractors) and nine double-feature conditions (all possible pairings of the single-feature conditions). The target was always a light blue semicircle, except in the trials where the distractors were single-feature shapes (triangles, circles and diamonds) in which case the target was a white semicircle.

Set sizes: we ran all the distractor set sizes used in Buetti et al. (2019) (1, 4, 9, 19 and 31) along with an extra condition with more distractors (43). We also ran target-only 'zero distractor' trials (30 in total, with 6 being the white semicircle target and the remainder the light blue semicircle target).

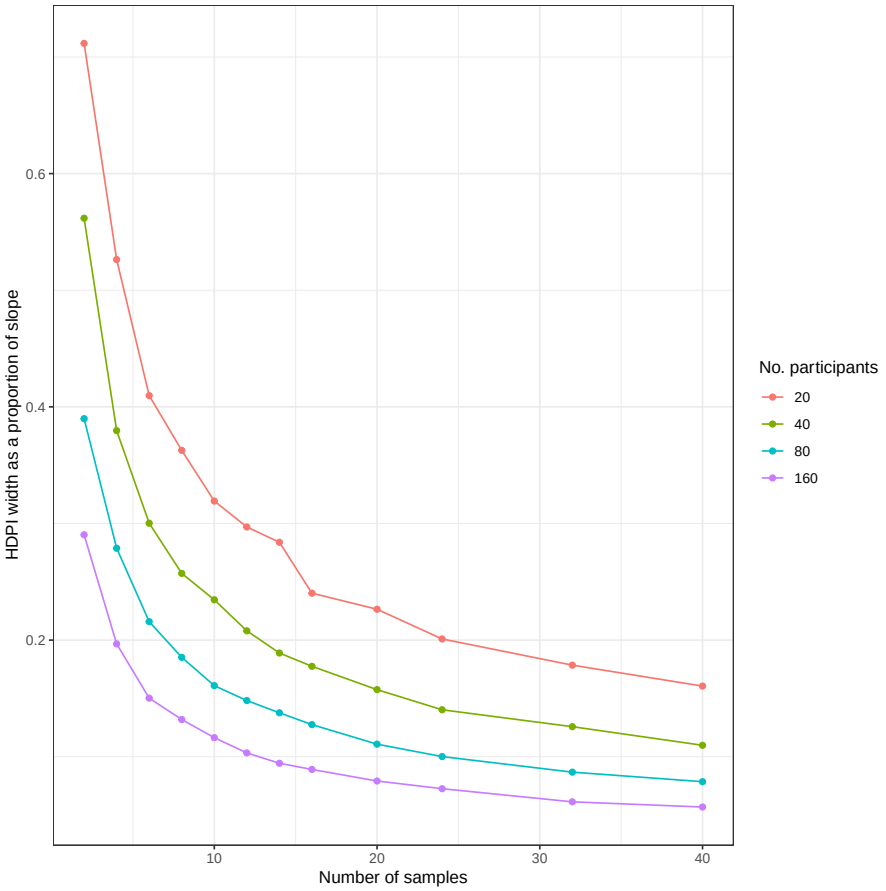

Fig. 7 Simulation sensitivity analysis, showing the 97% HPDI width for the blue distractor condition in a shifted lognormal model, across different numbers of trials and participants.

The experiments were programmed in PsychoPy and Pavlovia Peirce et al. (2019). Stimuli were pre-made offline to generate search array images with 1920 x 1080 resolution. [Need to add something here about target/distractor sizes at this resolution? Obviously potentially not really that useful for online experiments though...]

4.3 Procedure

Participants initially viewed a fixation cross for Xms before viewing a search array. Participants were told to search for the target among distractors and report if the semicircle target pointed to the left or right, by pressing either the '1' or '2' key respectively on their keyboard. They first completed 20 practice trials where they received feedback immediately after completing each trial. In the real experimental trials, participants received feedback on their average accuracy and reaction time after

each block of 30 trials. Participants completed 19 blocks of trials (570 trials overall) with the order of the stimuli being fully randomised.

In both the practice and experimental trials, the search display always remained on screen until a response was made, or until 5 seconds had passed. The inter-trial interval was Xms.

4.4 Data Pre-processing

In the original paper, there were inclusion criteria: search accuracy over 90% and individual average response times smaller than two standard deviations from the group average response time. Even in a lab set up, this normally led to at least one or two participants being excluded. Do we need something similar? More stringent?

incorrect trials? Poorly behaved participants? RTs that are far too short? Or far too long?

4.5 Analysis Plan

We will follow the analysis given in Section 2.2.

5 Results

– blank –

6 General Discussion

Is discriminating between the different models one of our aims? Or is this a discussion point i.e. it's quite hard to do? And therefore maybe a follow up paper?

Acknowledgements Thank you to AL for help and encouragement!

Conflict of interest

The authors declare that they have no conflict of interest.

References

Stefanie I Becker. The role of target–distractor relationships in guiding attention and the eyes in visual search. *Journal of Experimental Psychology: General*, 139(2):247, 2010.

Stefanie I Becker, Christian Valuch, and Ulrich Ansorge. Color priming in pop-out search depends on the relative color of the target. *Frontiers in psychology*, 5:289, 2014.

Simona Buetti, Deborah A Cronin, Anna M Madison, Zhiyuan Wang, and Alejandro Lleras. Towards a better understanding of parallel visual processing in human vision: Evidence for exhaustive analysis of visual information. *Journal of Experimental Psychology: General*, 145(6):672, 2016.

Simona Buetti, Jing Xu, and Alejandro Lleras. Predicting how color and shape combine in the human visual system to direct attention. *Scientific reports*, 9(1):1–11, 2019.

Kyle R Cave and Jeremy M Wolfe. Modeling the role of parallel processing in visual search. *Cognitive psychology*, 22(2):225–271, 1990.

Johan Hulleman and Christian NL Olivers. On the brink: The demise of the item in visual search moves closer. *Behavioral and Brain Sciences*, 40, 2017.

Jessica L Irons and Andrew B Leber. Choosing attentional control settings in a dynamically changing environment. *Attention, Perception, & Psychophysics*, 78(7):2031–2048, 2016.

Jessica L Irons and Andrew B Leber. Characterizing individual variation in the strategic use of attentional control. *Journal of Experimental Psychology: Human Perception and Performance*, 44(10):1637, 2018.

Laurent Itti and Christof Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research*, 40(10-12):1489–1506, 2000.

Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259, 1998.

Christof Koch and Shimon Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of intelligence*, pages 115–141. Springer, 1987.

Kathryn Koehler, Fei Guo, Sheng Zhang, and Miguel P Eckstein. What do saliency models predict? *Journal of vision*, 14(3):14–14, 2014.

Alejandro Lleras, Zhiyuan Wang, Anna Madison, and Simona Buetti. Predicting search performance in heterogeneous scenes: Quantifying the impact of homogeneity effects in efficient search. *Collabra: Psychology*, 5(1), 2019.

Alejandro Lleras, Zhiyuan Wang, Gavin Jun Peng Ng, Kirk Ballew, Jing Xu, and Simona Buetti. A target contrast signal theory of parallel processing in goal-directed search. *Attention, Perception, & Psychophysics*, pages 1–32, 2020.

Anna Madison, Alejandro Lleras, and Simona Buetti. The role of crowding in parallel search: Peripheral pooling is not responsible for logarithmic efficiency in parallel search. *Attention, Perception, & Psychophysics*, 80(2):352–373, 2018.

Gavin JP Ng, Simona Buetti, Trisha N Patel, and Alejandro Lleras. Prioritization in visual attention does not work the way you think it does. *Journal of Experimental Psychology: Human Perception and Performance*, 2020.

Gavin Jun Peng Ng, Alejandro Lleras, and Simona Buetti. Fixed-target efficient search has logarithmic efficiency with and without eye movements. *Attention, Perception, & Psychophysics*, 80(7):1752–1762, 2018.

Anna Nowakowska, Alasdair DF Clarke, and Amelia R Hunt. Human visual search behaviour is far from ideal. *Proceedings of the Royal Society B: Biological Sciences*, 284(1849):20162767, 2017.

Derrick Parkhurst, Klinton Law, and Ernst Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision research*, 42(1):107–123, 2002.

Jonathan Peirce, Jeremy R Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyuki Sogo, Erik Kastman, and Jonas Kristoffer Lindeløv. Psychopy2: Experiments in behavior made easy. *Behavior research methods*, 51(1):195–203, 2019.

Benjamin W Tatler. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of vision*, 7(14):4–4, 2007.

Benjamin W Tatler, Mary M Hayhoe, Michael F Land, and Dana H Ballard. Eye guidance in natural vision: Reinterpreting salience. *Journal of vision*, 11(5):5–5, 2011.

Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12 (1):97–136, 1980.

Zhiyuan Wang, Simona Buetti, and Alejandro Lleras. Predicting search performance in heterogeneous visual search scenes with real-world objects. *Collabra: Psychology*, 3(1), 2017.

Jeremy M Wolfe. Approaches to visual search: Feature integration theory and guided search. *Oxford handbook of attention*, pages 11–55, 2014.

Jeremy M Wolfe, Kyle R Cave, and Susan L Franzel. Guided search: an alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human perception and performance*, 15(3):419, 1989.

Lingyun Zhang, Matthew H Tong, Tim K Marks, Honghao Shan, and Garrison W Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of vision*, 8(7):32–32, 2008.