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

Anna E. Hughes[a], Anna Nowakowska[b], Alasdair D. F. Clarke[1]

[a]Department of Psychology, University of Essex, Colchester, CO4 3SQ, UK
[b]School of Psychology, University of Aberdeen, Aberdeen, AB24 3FX, UK

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. One study using the TCS framework has shown that 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 estimation is done using a contrast combination model, and a collinear contrast integration model was shown to outperform other options. In our work, we extend TCS to a Bayesian multi-level framework. We investigate modelling using normal and shifted-lognormal distributions, and show that the latter allows for a better fit to previously published data. We run a new fully within-subjects experiment to attempt to replicate the key original findings, and show that overall, TCS does a good job of predicting the data. However, we do not replicate the finding that the collinear combination model outperforms the other contrast combination models, instead finding that it may be difficult to conclusively distinguish between them.

Keywords: Visual search, Efficient search, Parallel processing

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 occasionally negative e.g. (Rangelov et al., 2017)), 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 (Duncan and Humphreys, 1989; Cave and Wolfe, 1990; Wolfe, 1998; Liesefeld et al., 2016). Recent work has also attempted to model the variation in search slopes at the boundary between inefficient and efficient search (Liesefeld et al., 2016).

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 in which the distractors differ from the target along a *single feature* (e.g. colour *or* shape), it has been shown that you can predict search times for arrays in which the target differs from the distractors along two features (e.g., colour *and* shape) which we refer to here as *double feature* search (Buetti et al., 2019) (but similar paradigms have been known by other names e.g. 'redundant feature search' (Krummenacher and Müller, 2012; Mordkoff and Yantis, 1991)). 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 modified 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), and even in abstract laboratory search arrays (Kotseruba et al., 2020). 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 necessarily 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. Other extensions to these models have been proposed, such as the Dimension Weighting Account, in which saliency weightings are assigned to different target 'dimensions' (e.g. colour or shape), helping to explain results where varying the target dimension within blocks of trials leads to longer reaction times than where the dimension remains consistent within a block (Krummenacher and Müller, 2012). 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 qualitative 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). TCS also has features in common with other models that propose parallel identification of all items in a scene, with diffusion based mechanisms for identifying targets from distractors (Moran et al., 2013, 2016). 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 essentially a (natural) log-linear model in the number of distractors. The full model contains a variable L, which represents the number of different types of distractors present in the display. However, 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. In this case, the TCS model can be represented in the following way:

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

The intercept, a, corresponds to search arrays in which only the target is present and there are no distractors. N_T is the total number of distractors.

1.2. Rationale for proposed work

While many aspects of the TCS framework have been tested, with extremely promising results, there remains a great deal of scope for verification of some of the key findings to date, and extensions of aspects of the model. 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 to what extent 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 (Nowakowska et al., 2017). 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; Clarke et al., 2022a).

Taking search time distributions into account is also important for constraining theories of visual search (Wolfe et al., 2010; Liesefeld and Müller, 2020): for example, they have been used to help distinguish between models that make similar predictions at the level of average reaction times (Moran et al., 2016, 2017). Including subject and trial level data into our implementation of the TCS will therefore further aid model development and assumption testing.

We also extend the TCS model into a Bayesian framework, where we begin with existing 'prior' beliefs that are updated with data to give 'posterior' beliefs that can be used for inference (McElreath, 2020). We think this has a number

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 containing images of real-world objects, 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 when nearby items are similar to each other.
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 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.
Xu et al. (2021)	Extension of Buetti et al. (2019) to new features (shape and texture), which combine according to a Euclidean metric (orthogonal contrast integration model).

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

of advantages over frequentist approaches. Perhaps most importantly, Bayesian models are highly flexible. We demonstrate how we are able to specify a model that is able to more accurately represent the distribution of responses (for example, by specifying a response distribution that avoids predicting negative reaction times) with a relatively complex model structure, that can be fit to a relatively small amount of pilot data: something that would be challenging within a frequentist framework. We also believe that Bayesian models offer very intuitive methods for model testing and comparison and straightforward interpretation of results, and we hope that this manuscript can act as a demonstration of these benefits, showing how they can be applied to real scientific questions beyond the simplified examples often found in textbooks or tutorials.

In the current manuscript, we focus on replicating and extending findings from Buetti et al. (2019). In their study, participants searched for a target in a scene of homogeneous distractors (see Figure 1). First, parallel search efficiency (measured by the logarithmic search slope) was estimated for cases where the distractors varied from the target in 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). 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. In our registered replication, we will attempt to verify the conclusions of Buetti et al. (2019), that the collinear contrast integration model does indeed offer the best characterisation of contrast signal combinations in visual search within the TCS framework.

We begin by verifying the analysis of Buetti et al. (2019). We then describe our proposed replication study, showing with pilot data how we are able to extend their model of how multi-dimensional contrasts are calculated, both by incorporating a multi-level 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).

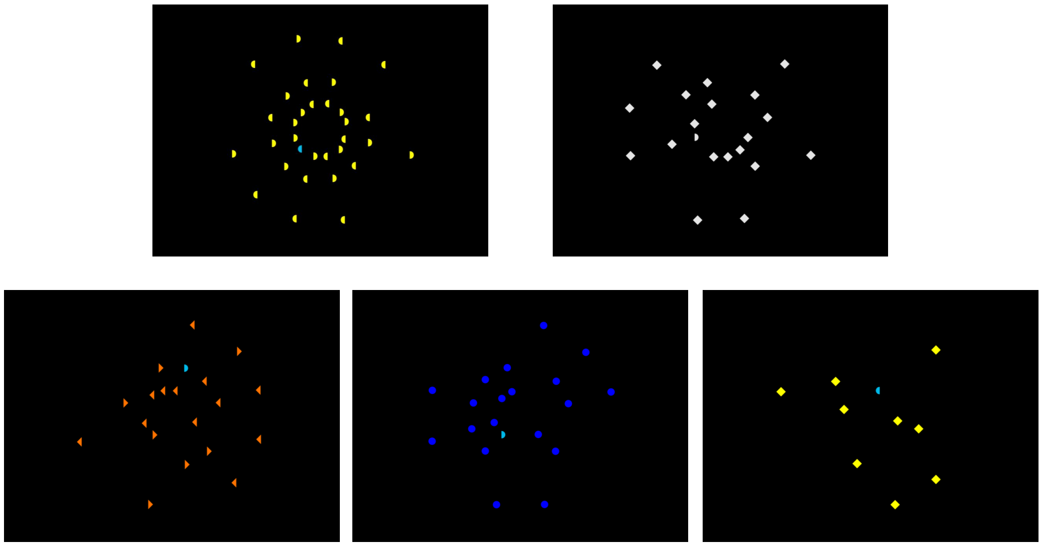

Figure 1: Example stimuli from Buetti et al. (2019) Top left: Expt 1A. Here, the target is a blue semicircle within a set of homogeneous (yellow semicircle) distractors. Top right: Expt 1B. The target is a grey semicircle in circular grey distractors. Bottom left: Expt 2A. The target is a blue semicircle 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 (both using frequentist modelling and Bayesian modelling; see *Supplementary Materials - Computational Verification)*.

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. Buetti et al. (2019) also carried out a replication of their basic results using slightly different target and distractor stimuli (Experiments 3 and 4).

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 1). In particular, the TCS theory allows us to predict the value of the logarithmic slope, $D_{\mathrm{c,s}}$, in this condition based on the corresponding D_i in the single feature search experiments.

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

Experiments 1a and 1b and 3a and 3b were used to calculate the logarithmic slope parameter D_i. In all 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 1). In the original model the error distribution was assumed to be normal. Thus the results of Experiments 1 and 3 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 1 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	feature	D_s
blue	76.8	triangle	141.1
yellow	16.0	diamond	77.2
orange	9.8	circle	62.1

Table 2: A table of D_i values for Experiment 1a and 1b. See *Supplementary Materials - Computational Verification* for full values for all experiments.

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 (natural) 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 2) 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):

$$D_{\mathrm{c,s}} = \min\left(D_{\mathrm{c}}, D_{\mathrm{s}}\right) \tag{2}$$

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 $\mathrm{D}_{\mathrm{c,s}}$ can be represented as:

$$D_{\mathrm{c,s}} = \frac{1}{\sqrt{(\frac{1}{D_{\mathrm{c}}})^2 + (\frac{1}{D_{\mathrm{s}}})^2}} \tag{3}$$

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:

$$\frac{1}{D_{\mathrm{c,s}}} = \frac{1}{D_{\mathrm{c}}} + \frac{1}{D_{\mathrm{s}}} \tag{4}$$

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/)[1] and using the exclusion criteria originally applied; see Figure 2 (left panel) and *Supplementary Materials - Computational Verification* for details. We show that we are able to do this using both the frequentist modelling approaches used in the original paper, and using Bayesian modelling.

2.1.3. Estimating a, the intercept parameter for compound features

As a is the intercept of the 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, and can be thought of as the role of non-search processes (such as motivation, motor preparation etc.) that influence reaction time. In Buetti et al. (2019), a was calculated for each sub-experiment. Here, we follow that method in order to replicate their results exactly.

[1] downloaded on 28th August 2020

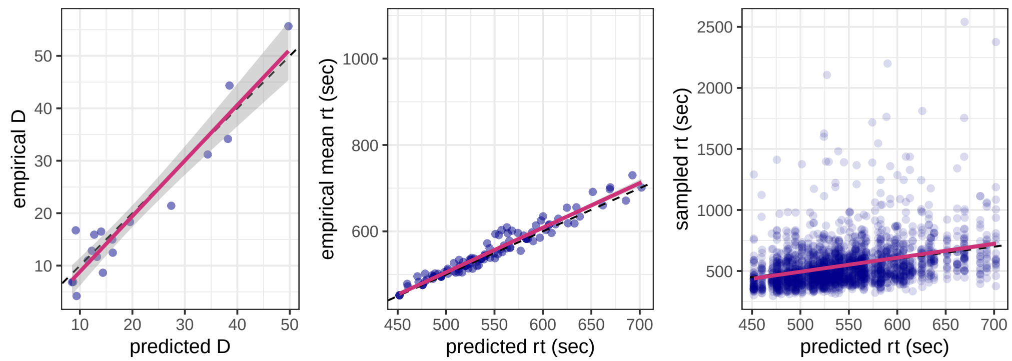

Figure 2: (left) The collinear method for calculating D offers a good prediction. (centre) Using the TCS to predict reaction times. (right) Each dot now represents a randomly sampled reaction time from an observer. Note that there is greater spread in the data points here, due to the fact that there will be trial-to-trial variability due to target position, inter-item distances, observer differences and so on.

2.1.4. Estimating mean reaction times

Finally, we can use Equation 1 to predict mean reaction times. As can be seen in Figure 2 (centre panel), these predictions are essentially identical to the empirical RT results: $R^2 = 0.93\%$.

2.1.5. 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 adapting TCS to make use of multi-level modelling techniques. Multi-level models allow us to take into account the hierarchical structure of the data (i.e. that each participant completes multiple trials) in a way that does not require averaging, meaning that we are able to model participant variability as well as group-level effects (Gelman and Hill, 2006).

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 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. 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. Note that a Wald, or inverse Gaussian distribution, would also be a reasonable distribution choice for this data given that TCS is based on a diffusion process e.g. (Moran et al., 2013), and this distribution has been argued to be psychologically more plausible (e.g. Kieffaber et al. (2006), though see Matzke and Wagenmakers (2009)): we chose not to use this distribution as it often leads to computational issues, which would make it harder for others to reproduce or build on our approach later.

3. Hypotheses

We plan an experiment to test the extent to which the original results in Buetti et al. (2019) replicate and generalise, using our new modelling approach.

3.1. Proposed Modifications to Experimental Design

In order to better test the above, and increase 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.** If the distractors are a very different colour from the target, they may not distinguish well between different contrast models. We will therefore run a version of the experiment where the target is a red semicircle, with distractors being either orange, purple or pink.

3.2. Registered Hypotheses

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. We expect that this will confirm the results previously seen in papers testing TCS i.e. that the log-linear approach will be best.

3. **Contrast model comparisons.** We will test the hypothesis proposed by (Buetti et al., 2019): specifically, that the *collinear contrast integration model* outperforms the *best feature guidance*, and *orthogonal contrast combination models* for the calculation of D, by calculating and comparing the mean absolute prediction error for each model.

4. **Reaction time predictions.** We will further test the hypothesis proposed by (Buetti et al., 2019) by testing which model gives the best prediction at the trial-by-trial RT level.

We will test each of these hypotheses by calculating the marginal likelihood of the relevant models, and then calculating the posterior 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%. This approach is most appropriate for our model: other measures of model fit, such as AIC, require an assumption of flat priors (which is not valid for multi-level models) and are based on point estimates (which is not valid for Bayesian models) (McElreath, 2020).

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.

One of the benefits of using a multi-level modelling approach is that it is relatively easy to extend to incorporate other factors that may contribute to reaction times, such as eccentricity and inter-item distance, which may help to explain behaviour further. To demonstrate this, we will also run exploratory analyses including a factor for which ring the target is in to assess whether this improves model fit or affects any of the conclusions that can be drawn from the model.

3.4. Pilot Experiment

Full details of a pilot experiment with $n = 4$ participants (960 trials each) using our proposed analyses can be found in *Supplementary Materials - Pilot Analysis*. This suggests that even with a small sample, we can convincingly demonstrate H1 and H2. However, more data will be required to discriminate between the models, particularly for H4. Given that our methods are within-subject, we have reduced the number of trials per condition compared to Buetti et al. (2019) (12 in our pilot study, 20 in our proposed, compared to 40 in theirs). It is therefore possible that the increased noise in our estimated D single-feature parameters will make it more difficult to predict double-feature Ds accurately. However, we think this is unlikely to be the case as we can see that even in a small amount of pilot data, we can verify H3, with the collinear model having the lowest mean absolute prediction error.

4. General Methods

4.1. Sample Size: Participants and Trials

We plan to test 40 participants during the experiment. Our pilot experiment shows that H1 and H2 are easily demonstrated with 10 times less data, and Buetti et al. (2019) used 20 participants per experiment. Our sample size will therefore be in line with previous work testing H3 and H4. Ethical approval for the study was granted by the University of Aberdeen (application number PEC/4677/2021/2).

Our pilot study above suggests that just 12 trials per condition may be sufficient to fit our models. To be conservative, we propose using 20 in our experiment. We have demonstrated that using just half the data (20/40 trials per condition) from Buetti et al. (2019) makes no difference to our computational verification (see *Supplementary Materials - Computational Verification*).

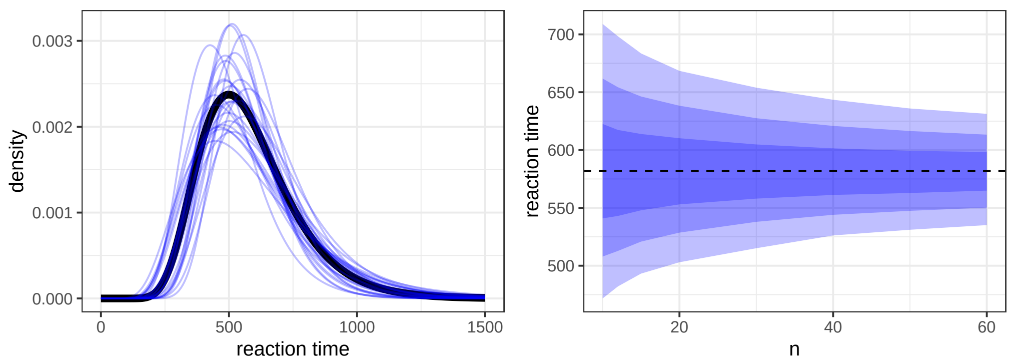

Figure 3: (left) The dark line shows the distribution we sampled from. The blue lines show distributions fitted to different samples of 20 data points. (right) Plot showing how the distribution of sample means vary with n. Shaded regions indicate the 50%, 80% and 95% confidence intervals.

Finally, we have carried out a simulation experiment to estimate the confidence intervals on the mean when sampling from a log-normal distribution. We defined our distribution to have a mean-log of 6.135 and a standard deviation of 0.32. These values were loosely based on the distributions of reaction times in Buetti et al. (2019). The results are shown in Figure 3. Based on these simulations, we find that a sample of $n = 20$ leads to a 95% confidence interval that is approximately 1.4 times larger than $n = 40$. We feel this is a suitable compromise given we will be collecting our data within-subjects.

4.2. Stimuli

The targets and distractors are randomly assigned to the display based on an invisible grid. Within each quadrant of the screen, there are three 'spokes' each with four possible target positions (starting from the centre of the screen and moving outwards), creating 36 different target positions in total, in three concentric circles. A small amount of jitter is added to each possible position to make the target locations less predictable.

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

Set sizes: we will run all the distractor set sizes used in Buetti et al. (2019) (1, 4, 9, 19 and 31). We will also run target-only 'zero distractor' trials (60 in total, with 12 being the white semicircle target and the remainder the red semicircle target).

The experiments were programmed in PsychoPy and Pavlovia (Peirce et al., 2019). Stimuli were pre-made to generate search array images with 1920×1080 resolution.

4.3. Procedure

Participants will complete the experiment in the laboratory, sitting at a viewing distance of 45cm from the screen (viewing distance will be fixed by using a chin rest). They will view a fixation cross before viewing a search array: they will press the space bar to continue to the trial. Participants will be told to search for the target among distractors (either a red semicircle or a white semicircle, depending on the block) and report if the semicircle target points to the left or right, by pressing either the left or right button on a button box (Cedrus RB-540). They will first complete 16 practice trials where they will receive feedback immediately after completing each trial. In the real experimental trials, participants will receive feedback on their average accuracy and reaction time after each block of 320 trials. Participants will complete 5 blocks of trials (1600 trials overall i.e. 320 trials in each of 5 experiments, consisting of 5 set sizes x 3 distractor conditions x 20 repeats + 20 zero distractor trials). The trials where the distractors are single-feature shapes (i.e. the target is a white semicircle - Experiment 1b in Buetti et al. (2019)) will all appear in one block (which will appear at a randomly selected position within the experiment). All other trials (where the target is red semicircle) will be fully randomised i.e. all different conditions will be completely intermixed. This approach will be taken as TCS requires the participant to have a well-defined target template in mind in order to compare this to the stimuli in the display. Thus, participants will be cued to search for the relevant target at the beginning of each block.

In both the practice and experimental trials, the search display will always remain on screen until a response is made, or until 5 seconds had passed.

4.4. Data Pre-processing

Only participants who complete the full experiment will be considered candidates for inclusion in the data analysis. We will apply the same inclusion criteria as the original paper: participants will only be included if their search accuracy is over 90% and their average response time is not smaller or larger than two standard deviations from the group average response time.

For participants included in the analysis, we will apply the data cleaning used in the pilot data analysis i.e. removing the top and bottom 1% of their data. [2]

4.5. Analysis Plan

All analysis will be carried out using R (v4.2.0), brms (v2.17.0) and rStan (v2.26.11) As discussed above, we will use mixed-effect models with either normal, lognormal or shifted lognormal distributions.

Please see the analysis of our pilot data for a full implementation of our analysis pipeline, including all code (available on Github at `https://github.com/Riadsala/single_double_feature_search`).

4.6. Registered Report

The original Stage 1 registered report for this manuscript is available at `https://osf.io/f9sua/`. All study data, materials and analysis code for both Stage 1 and Stage 2 are available at `https://github.com/Riadsala/single_double_feature_search`.

We report how we determined our sample size (see Section 4.1), all data exclusions (if any), all inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis (see Section 4.4) all manipulations, and all measures in the study (see Section 4.2).

5. Results

All 40 participants had accuracy over 90% (minimum 93.1%). One participant had an average response time (1100ms) over two standard deviations from the group average response time (781ms) and was removed. Incorrect trials were then removed, and the data was trimmed (only including response times between the 1% and 99% quantiles) leaving us with 39 participants completing a total of 59,587 trials.

[2]See *Supplementary Materials - Pilot Analysis* and *Supplementary Materials - Registered Analysis* for full details of data cleaning.

All Bayesian models were fit to the new data using exactly the same procedure[3] as the pilot data presented in the Stage One review process. We checked for convergence of our models by visually inspecting the chains as well as verifying that the $\hat{R}$ was close to 1 for all parameters of all the fitted models (see *Supplementary Material - Main Analysis* for full model fit information).

5.1. Hypothesis 1: Shifted-lognormal model

Our first hypothesis concerns which distribution best fits the single feature response time data. We fit multi-level models with a i) normal, ii) lognormal, and iii) shifted-lognormal distribution. The models all used the same model formula that estimated search slopes in terms of $\log N_t$ for each feature. Maximal random effect structures were used.

After each of these models had been fit to the data, leave-one-out (LOO) model comparison was used to calculate posterior probabilities for each. The results of this procedure allocated $\sim 100\%$ of the weight to the shifted-lognormal model, so we can conclude that, in accordance with our registered hypothesis, it is the best distribution (out of the three we tested[4]) to use for modelling response times in this paradigm. This model is shown in Figure 2.2 of the *Supplementary Materials - Registered Analysis*.

5.2. Hypothesis 2: log-linear effect of N_T

We then used the same methods to verify that using $\log N_T$ for the search slope does indeed give a better fit to the data than simply using N_T. The results are again conclusive with $\sim 100\%$ of the model weight being assigned to the model that is log-linear in N_T, again in accordance with our original hypothesis.

5.3. Hypothesis 3: Contrast Model Comparison

Now that we have confirmed that the shifted-lognormal multilevel model (with a log-linear effect of N_T) is indeed the best fit to the data we will extract the search slopes for each feature. These are summarised in Table 3. We can see that we have successfully obtained a range of values for both D_c and D_s. As with Buetti et al. (2019) we find that the values for D_s are larger than D_c (see Table 2), meaning that search slopes for colour features are shallower than shape.

[3]The only departure was an increase in iterations from 5000 to 80000 for the model predicting reaction times, based on advice given in the Stan forums, to enable the bridge sampling process to work properly.

[4]See discussion for Wald, Weibull, etc.

feature	D_c	95%HDCI	feature	D_s	95%HDCI
orange	0.156	[0.139 , 0.173]	triangle	0.253	[0.230 , 0.275]
pink	0.042	[0.028 , 0.057]	diamond	0.187	[0.171 , 0.205]
purple	0.015	[0.002 , 0.030]	circle	0.191	[0.175 , 0.204]

Table 3: A summary of the posterior estimates of D_c and D_s values from our Experiment. Note that our values are reported in seconds, in contrast to Table 2, which follows (Buetti et al., 2019) and reports the slopes in milliseconds.

We now combine the *single-feature* search slopes, D_c and D_s, to predict the *double-feature* conditions ($D_{c,s}$) using Equations 2, 3 and 4 and above. The results are summarised in Figure 4. We find that while the collinear contrast model has the highest R^2 (0.922, compared to $R^2 = 0.884$ for best feature, and $R^2 = 0.916$ for orthogonal contrast), the orthogonal contrast model is the most accurate, both in terms of mean absolute error (0.165, compared to 0.185 for best feature and 0.271 for collinear) and having a regression slope closest to 1 (1 compared to 0.753 and 1.48). Therefore, Hypothesis 3 does not hold: orthogonal contrast rather than collinear contrast offers the best prediction of search slopes in the double-feature condition.

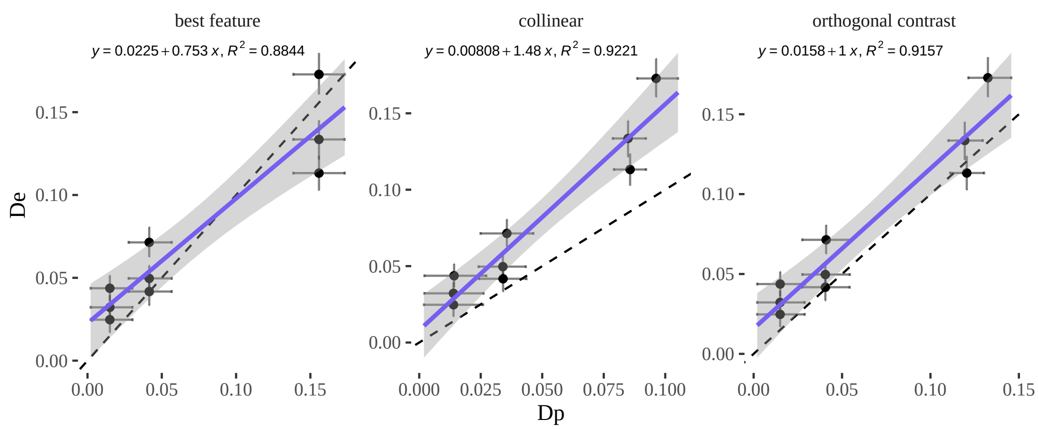

Figure 4: Predicting $D_{c,s}$ from D_c and D_s. The x-axis shows our predictions, D_p, using the best feature, collinear contrast, and orthogonal contrast models.

5.4. *Hypothesis 4: Reaction Time Predictions*

Upon reflection, the approach to model comparison we outlined in our registered analysis was limited in a number of ways. Our original plan was to use the posterior predictions from a model trained on the single-feature data to act as a prior for the double-feature data. While we initially thought this would be an elegant approach, there are a large number of parameters that are outside the main focus of this paper yet still require priors (intercepts, group level variance and residual variance). Furthermore, while the methods for estimating $D_{c,s}$ presented above give good predictions in terms of the mean value, it is not clear that the standard deviation for these distributions will be accurate. As such, we have developed a new, simpler method for this final comparison. To maintain full transparency, we present both methods here.

5.4.1. Registered Method

Our final hypothesis concerns how well the different feature combination models perform when predicting reaction times. We find very little difference between the three methods in terms of LOO model weights: 0.318 for best feature, 0.346 for collinear and 0.336 for orthogonal contrast. Thus, according to this analysis, we find no conclusive answer to hypothesis 4: all models give similar predictions at the trial-by-trial RT level.

5.4.2. Updated Method

Our new method for exploring this hypothesis involves taking $n = 100$ samples of the fixed effects from both the model fitted to the single-feature data and the model fitted to the double-feature data. Each of these samples includes an intercept (a), slope (D), non-decision time (ndt), and residual variance (σ). We then take the parameters from the double-feature model, but replace the D values with our predicted D using the single-feature model. Finally the predicted mean $\log(rt)$ is calculated for each feature and number of distractors. These are then compared to the empirical reaction times and we compute the absolute error.

We can also calculate an upper-bound by carrying out the above process, but without replacing the fitted $D_{c,s}$ with the predicted. This allows us to report 'relative absolute error'. As all of the methods under consideration make identical predictions for trials with no distractors, these are omitted from this calculation.

The results of this procedure are in-line with the registered analysis presented above: all three methods perform well relative to our baseline (see Table 4), and thus we cannot make any strong conclusions related to hypothesis 4. All three contrast combination methods do a good job of accounting for the reaction time data collected.

metric	abs error		
	lower	median	upper
orthogonal	0.994	1.00	1.02
collinear	0.990	1.01	1.05
best feature	0.999	1.01	1.02

Table 4: How well can we predict RTs using D_p (collinear, best feature or orthogonal contrast) comped to using D_e? A value of 1 means that our estimates of D derived from the single-feature trials does an equally good job at predicting the double-feature trials as using the D fit to the data.

6. Planned Explorations

Our interpretation of the null/neutral results for Hypothesis 4 (the prediction of reaction times) is that the differences in predictions from the three contrast combination methods are small relative to the (i) individual differences between participants and (ii) trial-to-trial variability due to target eccentricity. Thus, in our exploratory analysis, we investigate how incorporating these factors affects our conclusions.

6.1. Individual Differences

We start this exploratory analysis looking at how the D_c and D_s values vary from participant to participant. From Figure 5 (*left*) we can see that there is considerable variation between observers - in fact, the variation from one observer to the next is often larger than the variation across features. To investigate this further we calculated the correlations between each of the features, by calculating Pearson's r for each sample from our posterior, which gives us a full posterior distribution for the correlations. We can see in Figure 5 (*right*) that while both the D_c and D_s are correlated within feature classes (~ 0.75), there is no correlation of any of the colour features with any of the shape features. The individual differences for the *double-feature* conditions are much less pronounced - these conditions are easy and the search slopes are quite close to flat. Hence, the correlations are all much weaker, presumably due to range restriction.

Given these results, it is perhaps unsurprising that our analysis for Hypothesis 4 leads to an inconclusive result for distinguishing between the three contrast combination methods. Perhaps taking these individual differences into account when we predict reaction times will lead to improved power to discriminate between the models. However, before we do so, we will also investigate incorporating information about target eccentricity into the model.

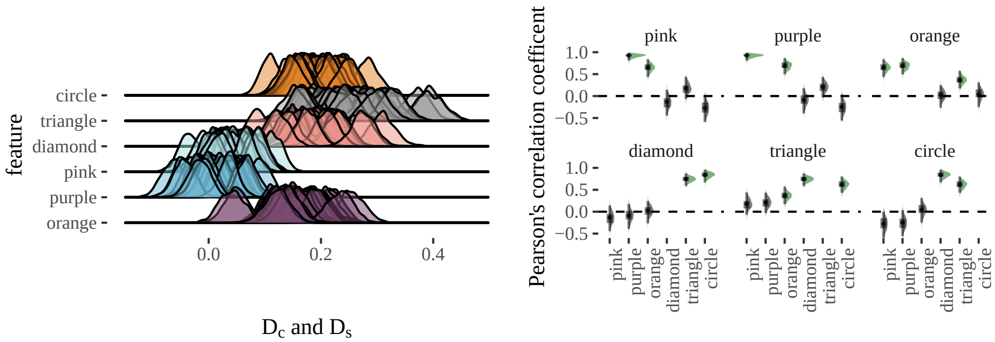

Figure 5: Individual differences in D_c and D_s. (*left*) Posterior probability distributions for D_c and D_s for each individual. (*right*) Estimated correlations between each of the D_c and D_s.

6.2. *Target Eccentricity*

It is well known that there are eccentricity effects in visual search, with reaction times being longer for targets that are further away from fixation (Carrasco et al., 1995; Wang et al., 2017). To investigate this in our dataset, we will use the same methods as above (fitting a multi-level shifted-lognormal model) but now including an additional factor that represents how far the target was from the fixation cross. This is coded as a three-level categorical factor representing which ring contained the target (see stimulus details, above). Allowing for interactions with the *feature* and $\log N_T$ increases the number of fixed effect parameters in the model from 8 to 22, with the model equation becoming the following:

$$y \sim 0 + r + r : f : \log(N_T) + (1|id) \tag{5}$$

We experimented with including r in the random effect structure, but this proved difficult to fit. We also had to revise the priors used in our registered analysis, in order to lower the intercept. Full details can be found in *Supplementary Materials - Planned Explorations*.

After obtaining a model that passed all convergence checks, we examined the posterior distribution for the effect of *ring*. Figure 6 paints an interesting and complex picture in which some features (e.g. some colours, particularly those that are more distinct from the target colour) are clearly leading to ‘pre-attentive search‘ in which response times are unaffected by either the number of distractors or target eccentricity. However, shape features seem to be strongly affected by eccentricity, particularly when there are multiple distractors in the stimulus.

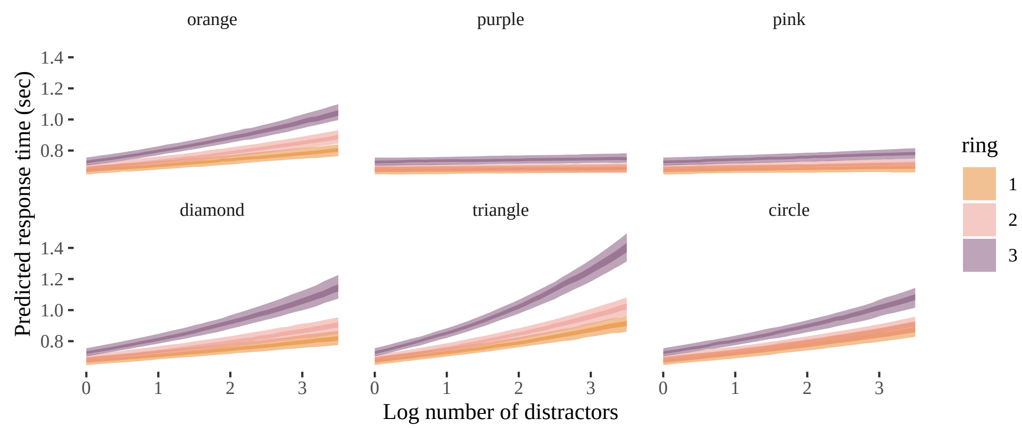

Figure 6: Fixed effects for predicting the effect of ring, feature and number of distracters on response times. Shaded regions represent the 53% and 97% HDCIs. We can see that *ring* has an effect on search slopes, and that this effect is more pronounced for some features (i.e., triangles) than others.

We can now compute our predictions (D_p) for $D_{c,s}$ taking the *ring* into account. Doing so leads us to a similar result as before with orthogonal contrast outperforming the best feature and collinear measures in terms of absolute error (0.023 compared to 0.025 (best feature) and 0.034 (collinear)). However, the regression slopes are all relatively similar (0.90 for best feature, 1.58 for collinear and 1.15 for orthogonal contrast). Thus, adding ring into the model does not drastically change our overall conclusions, with the orthogonal contrast model still giving the best prediction of search slopes in the double-feature condition.

6.3. Predicting Response Times

We will now test to see if we can discriminate between the three contrast combination methods when we take target eccentricity (ring) and individual-level slopes into account. We use the same model comparison as before (see *Supplementary Materials - Planned Explorations* for full code) and find orthogonal contrast performs best, closely followed by best feature.

6.3.1. Issues with the Collinear Contrast method

In the previous model, the upper bound on the error in the collinear contrast method is high (see Table 5). To explain this, we can look back at Equation 4: when search slopes are close to 0, it is possible that we will observe negative values in the empirical data. Breaking down our data to compute search slopes

metric	abs error		
	lower	median	upper
orthogonal	1.01	1.00	1.04
collinear	1.03	1.00	5.15
best feature	1.03	1.00	1.07

Table 5: How well can we predict RTs using D_p (collinear, best feature or orthogonal contrast) comped to using D_e when using a model containing the ring of the target? A value of 1 means that our estimates of D derived from the single-feature trials does an equally good job at predicting the double-feature trials as using the D fit to the data.

for each person and each target eccentricity increases the chances of this being observed. Looking at Equation 4 we can see that in the case where both D_1 and D_2 are small but one is negative (i.e. $D_1 \sim -D_2$), then $1/D_1 + 1/D_2 \sim 0$. This leads to our estimated $D = \frac{1}{1/D_1+1/D_2} >>> D_1, D_2$ i.e. our estimated D is much larger than the slopes that were used to generate it, which is clearly incorrect. However, we do note that the main conclusions of our analyses still hold even if we remove these negative slopes (by restricting our analyses only to certain colours and rings of the data - see *Supplementary Materials: Suggestions from Reviewers* for more details), suggesting that addressing this mathematical issue may not necessarily lead to the collinear contrast method being preferred.

7. General Discussion

In this paper, we aimed to test the extent to which the results of Buetti et al. (2019) replicate and generalise, using a new modelling approach. Our results allow us to confirm our pre-registered hypotheses 1 and 2. Firstly, a shifted-lognormal distribution of response times outperforms normal and lognormal distributions, demonstrating that reaction time data are best modelled by a skewed distribution with an offset. Similarly, we confirmed that the number of distractors has a log-linear effect in this model, in line with the predictions of TCS theory. We also replicated other aspects of the original Buetti et al. (2019) paper with a different experimental set up, such as to observing shallower search slopes for colour features compared to shape features.

We do not find support for our pre-registered hypotheses 3 and 4. For predicting D in the double-feature conditions, our analyses found that the orthogonal contrast model was favoured over collinear, which is not in line with the registered hypothesis, which predicted that the collinear contrast model would be best (in line with Buetti et al. (2019)). Similarly, for hypothesis 4, we found that there was relatively little difference between the three combination methods for prediction of trial-by-trial reaction times. Our exploratory analyses suggest that incorporating additional factors (e.g. individual differences in participant D_c and D_s values, and the eccentricity of the target) allows better discrimination between models, but again suggests that the orthogonal contrast combination method gives the best predictions.

7.1. Modelling of reaction times

In much of the literature on visual search, mean reaction times are modelled using a simple linear model $\bar{y} = bN_T + a$ (e.g. Treisman and Gormican (1988); Rosenholtz et al. (2012); Hughes et al. (2016)). The b coefficients are often referred to as "search slopes" and are often treated as measurements of theoretical importance. Our results indicate that a shifted-lognormal model that is loglinear in N_T offers a much better fit to the data ($\log(y) - ndt = b\log(N_T) + a$), which is perhaps not surprising, given the properties of reaction time data, where valid responses normally begin at around 100ms, and the distribution often has a long "tail" of slower responses.

However, there have been concerted efforts within the literature to model reaction time distributions more effectively: indeed, Buetti et al. (2019) use $\log N_T$ when computing their search slopes. In terms of reaction time distributions, log transformations are frequently taught as a way to normalise reaction time data (although often with caveats regarding how this can change the interpretation of the results e.g. Osborne (2002)) and are frequently used in analysing reaction time data e.g. (Clarke et al., 2022b). Researchers have also looked at other distributions to assess which offer the best fit to empirical response times in visual search. For example, Palmer et al. (2011) compared ex-Gaussian, ex-Wald, Gamma, and Weibull distributions and found that the distributions with exponential components offer a better fit to the data. Our results are in line with this. However, we opted to use a shifted-lognormal distribution in our analysis above for mostly pragmatic reasons, as these more complex distributions are often computationally difficult to fit[5]. It has also been argued that trying to select a "correct" distribution is likely to be problematic for empirical data, which is probably a mixture of multiple components (Wolfe et al., 2010). Similarly, some recent approaches make use of drift-diffusion methods (e.g. Wolfe and Van Wert (2010); Yu et al. (2022); Corbett and Smith (2020)), though again these models can be challenging,

[5] See: https://discourse.mc-stan.org/t/model-fails-to-converge-when-using-brms/9062

particularly when considering how to interpret the parameters (Evans and Wagenmakers, 2019; Bompas et al., 2023). While important, these debates are outside the scope of the present Registered Report.

Despite these previous findings, the use of linear search slopes is still prevalent in the visual search literature. Our work shows that these choices of distribution can influence results and conclusions (see section 7.2 below), and therefore we recommend that other researchers consider carefully how they want to model their data. Even in the case where the search slopes are the primary outcome measure of interest (as opposed to the potentially more 'cognitive' parameters of e.g. Wald distributions, or drift diffusion models), we demonstrate that approaches that better account for the data distribution can be taken with relative ease.

7.2. Discriminating between combination methods

In Buetti et al. (2019), the collinear contrast integration model was found to provide the best fit for their data, providing a more precise prediction than the orthogonal contrast combination model (as measured by both the closeness of the slope of the regression line to one, and the mean average prediction error). Accepting this model of how the combination process works has theoretical implications e.g. it implies that colour and shape contrasts independently contribute to attentional guidance. However, we did not find strong support for this model, instead finding that the orthogonal contrast combination model provides the best fit with the data.

One possibility is that our small modifications to the experimental stimuli changed the strategy that participants used. However, this seems unlikely given that we only made changes to the colour of the stimuli, a manipulation that Buetti et al. (2019) also used, with no changes to their overall conclusions. In addition, our reanalysis of the original Buetti et al. (2019) data using our new methods also suggested that the orthogonal contrast model was best supported. Thus, we suggest that the choice of modelling distribution (e.g. shifted-lognormal v.s. lognormal) affects the conclusions drawn, and thus we should aim to use the models that best align with the data in order to better understand the theoretical implications of our findings.

We also modified the way stimuli were presented in our experiment compared to Buetti et al. (2019): rather than running each experiment separately, we (mostly) intermixed conditions. Models of attention presume that we hold a target template in our memory (Duncan and Humphreys, 1989), and thus we ensured that the trials where the target was the white semicircle were blocked separately from the trials where the target was the red semicircle, to try to avoid conflict between maintaining multiple target templates in memory. However, the distractors could change from trial to trial in all other blocks, unlike in Buetti et al. (2019). In their experiments, it is possible participants used strategies such as shifting the target representation away from the distractors, or generally using relational strategies (Navalpakkam and Itti, 2007; Becker, 2010; Yu et al., 2023), which would be more challenging in our experimental set up. In relation to the models, this type of target representation shift could occur more strongly for one feature dimension (e.g. colour) than the other, perhaps changing the relationship between the contrasts for different feature dimensions and therefore the preferred model. If future work were to confirm this hypothesis, it would suggest that observers are able to cognitively shift their strategy based on the information available in the task.

Another possibility is that because some participants had negative search slopes, the collinear contrast model predicts implausibly large reaction times, due to the mathematical formulation of this model, leading to worse predictions. Despite the fact that our exploratory analyses suggested that removing these negative slopes would not change our conclusions, we suggest that a future improvement for the collinear contrast integration model would be to modify it to be able to give sensible predictions in these situations, given that negative search slopes do occur in some situations (Utochkin, 2013).

Finally, we would argue that it is difficult from these results to definitively make a decision about which model is best: all three models give very similar predictive weights during our model evaluation process. One challenge is that in general, the double feature searches are easy, and therefore the search slopes are fairly flat and there is not much variability to allow different models to make different predictions. For the current paradigm, a fruitful approach for future research could be to consider using different feature sets, and in particular, moving away from colour as feature, which may be a particularly salient cue (see Section 7.2.3 below).

7.2.1. Individual differences

Our (planned) exploratory analysis of the individual differences in search slopes suggests that there are large differences from one observer to the next. Indeed in some cases, these are larger than the differences from one feature to another. The difference between the steepest and shallowest search slopes (fixed effects) is 0.238 ($D_{triangle} = 0.253$, while $D_{purple} = 0.015$). If we compare this to the range of observer search slopes within a feature, we find this varies from 0.242 ($D_{triangle}$ per-observer ranges from 0.395 to 0.152) to 0.149 (D_{pink} ranges from -0.065 to 0.092). This suggests a challenge for modelling based on average performance:

can we be sure that averages represent a meaningful summary of the data, given that we see very clear individual differences? It could certainly be argued that observers might be using different strategies, and thus some members of the sample population might use (for example) a collinear combination strategy, while others use an orthogonal contrast strategy. Variable strategies have been found for other search behaviour (Clarke et al., 2022a; Kristjánsson et al., 2014; Proulx, 2011; Li et al., 2022), highlighting the importance of considering individual differences when understanding behaviour.

We also found that search slopes were correlated within feature, but not between: i.e, knowing that an observer's search slope for a colour condition allows us to predict their search slopes for the other colour conditions, but not any of the shape conditions. However, given the block design of our experiment, it is possible that this reflects a type of priming effect: knowing the search slope for a feature in the first block tells allows us to predict the search slopes of the other features in that block, but tells us nothing of the observer's behaviour in the second block. Post-hoc analyses looking at correlations within the colour condition by block suggest that this seems unlikely to explain our results fully, as we still observe good correlations between different colour search slopes across blocks (see *Supplementary Materials - Suggestions from Reviewers* for further details). However, to test this fully we would need to design the experiment differently in order to avoid block confounds, allowing us to disentangle whether these correlations reflect something about an observer's behaviour with different features, or instead how an observer's behaviour changes over time.

7.2.2. Eccentricity

Buetti et al. (2019) argues that the processing undertaken in this type of task can be done in parallel, with observers using peripheral vision to distinguish between target and distractors, and that there is systematic variation in reaction times as a function of set size associated with parallel processing. Target Contrast Signal Theory incorporates eccentricity effects into this type of parallel processing via a time-out parameter (T_0) (Lleras et al., 2020; Wang et al., 2018; Ng et al., 2018). Here, we confirm in our exploratory analyses that we are able to detect relatively strong eccentricity effects, as a model with target ring number included was a better predictor of the data than one without. However, including this factor did not change our overall conclusions about which model best predicted D in the double-feature condition, or which model best predicted reaction times.

In our experiment we followed Buetti et al. (2019)'s original methods, with participants freely viewing the displays. It is therefore likely that in some cases, observers felt that peripheral information was insufficient to make judgements, and thus made eye movements, moving into a more serial, focused-attention processing stage. Future work could more exclusively investigate peripheral effects in parallel processing by ensuring fixation when viewing the displays.

7.2.3. Limitations

One limitation of the experimental approach may be the feature dimensions chosen. We kept these the same as in Buetti et al. (2019) (colour and shape), but there is good evidence that colour may in some ways be a 'basic' feature dimension that is particularly salient, especially in peripheral vision, whereas guidance of attention by shape may be more complex (Wolfe, 2021). Mathematically, it would be better to have features where the slope values across the two dimensions are more similar, as all of the contrast combination formulae essentially consist of sums of inverse values, and if the slope values are highly dissimilar, the inverse sums will be disproportionately determined by one feature. This may indeed reflect how participants are approaching this task, as it may be the case that they preferentially attend to the more discriminating feature (colour) and the contribution of shape to their behaviour in the double-feature condition may be negligible. However, for the purposes of discriminating between the models, it would be beneficial in future experiments to adjust the target set, perhaps by making the shape dimension more salient (e.g. by increasing the size of the targets), or by selecting a different pair of features (e.g. shape and orientation).

7.3. Conclusions and future directions

In the current paper, we have independently reproduced the findings of Buetti et al. (2019) by extending their modelling to a multi-level framework. We have used a Bayesian approach, but note that this is in many ways entirely arbitrary: all of the modelling decisions we have taken would be possible within a frequentist framework as well. We also aimed to replicate the previous findings by running a within-subjects experiment, and broadly find that the Target Contrast Signal Theory does a good job of predicting the data. When using single-feature search slopes to predict double-feature search slopes, we do not replicate the previous finding that the collinear contrast integration method outperforms other options, but instead find that all combination methods do reasonably well, and in this particular experimental design, it may be difficult to conclusively distinguish between them.

One of the clear benefits of Target Contrast Signal Theory (Lleras et al., 2020) is its quantitative nature, allowing it to be empirically tested in a straightforward manner. Here, we demonstrate that we can independently replicate many aspects of TCS, while also offering extensions to the model that we hope will stimulate more research and refinement of this theory. Some suggestions for possible future directions and hypotheses that could be tested include:

1. It is relatively straightforward to make predictions about the mean reaction time per participant in the double-feature search condition: however, we have not attempted to predict an individual's trial-to-trial variance for different features, which could improve the model fit further.

2. We find correlations within feature classes (i.e. D_c and D_s) but not between: however, these may be a side-effect of the block design of the experiment. A future experiment could randomise trial type in order to more fully understand the nature of these correlations.

3. To more fully explore which combination model best predicts the data, we suggest a) modifying the collinear contrast model to accommodate negative search slopes b) attempting to find experimental conditions that best differentiate between the models, perhaps by using feature dimensions other than colour and c) modifying the experimental design to enforce parallel processing e.g. by making the display gaze contingent.

Computational modelling approaches alongside detailed, quantitative theory building has been argued to be one way to improve the reliability of psychological research (Oberauer and Lewandowsky, 2019; Guest and Martin, 2021). By combining this approach with fully open datasets and analysis scripts, we can hopefully begin to take a more "distributed collaborative network" approach (Moshontz et al., 2018) to our scientific questions. As such, we would like to conclude by encouraging other researchers to critique, build on and improve the approach we have taken in this manuscript, in order to further improve our ability to model performance in visual search tasks.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgements

This work was supported by an Economic and Social Research Council grant (ES/S016120/1) to ADFC and employing AN. We thank the reviewers for their

helpful comments which greatly improved the manuscript.

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.

Aline Bompas, Petroc Sumner, and Craig Hedge. Non-decision time: the higg's boson of decision. *bioRxiv*, pages 2023–02, 2023.

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.

Marisa Carrasco, Denise L Evert, Irene Chang, and Svetlana M Katz. The eccentricity effect: Target eccentricity affects performance on conjunction searches. *Perception & psychophysics*, 57:1241–1261, 1995.

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

Alasdair DF Clarke, Jessica L Irons, Warren James, Andrew B Leber, and Amelia R Hunt. Stable individual differences in strategies within, but not between, visual search tasks. *Quarterly Journal of Experimental Psychology*, 75 (2):289–296, 2022a.

Alasdair DF Clarke, Anna Nowakowska, and Amelia R Hunt. Visual search habits and the spatial structure of scenes. *Attention, Perception, & Psychophysics*, 84 (6):1874–1885, 2022b.

Elaine A Corbett and Philip L Smith. A diffusion model analysis of target detection in near-threshold visual search. *Cognitive Psychology*, 120:101289, 2020.

John Duncan and Glyn W Humphreys. Visual search and stimulus similarity. *Psychological review*, 96(3):433, 1989.

Nathan J Evans and Eric-Jan Wagenmakers. Evidence accumulation models: Current limitations and future directions. 2019.

Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.

Olivia Guest and Andrea E Martin. How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, 16 (4):789–802, 2021.

Anna E Hughes, Rosy V Southwell, Iain D Gilchrist, and David J Tolhurst. Quantifying peripheral and foveal perceived differences in natural image patches to predict visual search performance. *Journal of vision*, 16(10):18–18, 2016.

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.

Paul D Kieffaber, Emily S Kappenman, Misty Bodkins, Anantha Shekhar, Brian F O'Donnell, and William P Hetrick. Switch and maintenance of task set in schizophrenia. *Schizophrenia research*, 84(2-3):345–358, 2006.

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.

Iuliia Kotseruba, Calden Wloka, Amir Rasouli, and John K Tsotsos. Do saliency models detect odd-one-out targets? new datasets and evaluations. *arXiv preprint arXiv:2005.06583*, 2020.

Árni Kristjánsson, Ómar I Jóhannesson, and Ian M Thornton. Common attentional constraints in visual foraging. *PloS one*, 9(6):e100752, 2014.

Joseph Krummenacher and Hermann J Müller. Dynamic weighting of feature dimensions in visual search: behavioral and psychophysiological evidence. *Frontiers in psychology*, 3:221, 2012.

Walden Y Li, Molly R McKinney, Jessica L Irons, and Andrew B Leber. Assessing the generality of strategy optimization across distinct attentional tasks. *Journal of Experimental Psychology: Human Perception and Performance*, 48 (6):582, 2022.

Heinrich René Liesefeld and Hermann J Müller. A theoretical attempt to revive the serial/parallel-search dichotomy. *Attention, Perception, & Psychophysics*, 82(1):228–245, 2020.

Heinrich René Liesefeld, Rani Moran, Marius Usher, Hermann J Müller, and Michael Zehetleitner. Search efficiency as a function of target saliency: The transition from inefficient to efficient search and beyond. *Journal of Experimental Psychology: Human Perception and Performance*, 42(6):821, 2016.

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.

Dora Matzke and Eric-Jan Wagenmakers. Psychological interpretation of the ex-gaussian and shifted wald parameters: A diffusion model analysis. *Psychonomic bulletin & review*, 16(5):798–817, 2009.

Richard McElreath. *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC, 2020.

Rani Moran, Michael Zehetleitner, Hermann J Mueller, and Marius Usher. Competitive guided search: Meeting the challenge of benchmark rt distributions. *Journal of Vision*, 13(8):24–24, 2013.

Rani Moran, Michael Zehetleitner, Heinrich René Liesefeld, Hermann J Müller, and Marius Usher. Serial vs. parallel models of attention in visual search: accounting for benchmark rt-distributions. *Psychonomic bulletin & review*, 23(5): 1300–1315, 2016.

Rani Moran, Heinrich René Liesefeld, Marius Usher, and Hermann J Muller. An appeal against the item's death sentence: Accounting for diagnostic data patterns with an item-based model of visual search. *Behavioral and Brain Sciences*, 40:e148, 2017.

J Toby Mordkoff and Steven Yantis. An interactive race model of divided attention. *Journal of Experimental Psychology: Human Perception and Performance*, 17(2):520, 1991.

Hannah Moshontz, Lorne Campbell, Charles R Ebersole, Hans IJzerman, Heather L Urry, Patrick S Forscher, Jon E Grahe, Randy J McCarthy, Erica D Musser, Jan Antfolk, et al. The psychological science accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science*, 1(4):501–515, 2018.

Vidhya Navalpakkam and Laurent Itti. Search goal tunes visual features optimally. *Neuron*, 53(4):605–617, 2007.

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.

Klaus Oberauer and Stephan Lewandowsky. Addressing the theory crisis in psychology. *Psychonomic bulletin & review*, 26:1596–1618, 2019.

Jason Osborne. Notes on the use of data transformations. *Practical assessment, research, and evaluation*, 8(1):6, 2002.

Evan M Palmer, Todd S Horowitz, Antonio Torralba, and Jeremy M Wolfe. What are the shapes of response time distributions in visual search? *Journal of experimental psychology: human perception and performance*, 37(1):58, 2011.

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.

Michael J Proulx. Individual differences and metacognitive knowledge of visual search strategy. *PLoS One*, 6(10):e27043, 2011.

Dragan Rangelov, Hermann J Müller, and Michael Zehetleitner. Failure to pop out: Feature singletons do not capture attention under low signal-to-noise ratio conditions. *Journal of Experimental Psychology: General*, 146(5):651, 2017.

Ruth Rosenholtz, Jie Huang, Alvin Raj, Benjamin J Balas, and Livia Ilie. A summary statistic representation in peripheral vision explains visual search. *Journal of vision*, 12(4):14–14, 2012.

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 Treisman and Stephen Gormican. Feature analysis in early vision: evidence from search asymmetries. *Psychological review*, 95(1):15, 1988.

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

Igor S Utochkin. Visual search with negative slopes: The statistical power of numerosity guides attention. *Journal of vision*, 13(3):18–18, 2013.

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

Zhiyuan Wang, Alejandro Lleras, and Simona Buetti. Parallel, exhaustive processing underlies logarithmic search functions: Visual search with cortical magnification. *Psychonomic bulletin & review*, 25:1343–1350, 2018.

Jeremy M Wolfe. What can 1 million trials tell us about visual search? *Psychological Science*, 9(1):33–39, 1998.

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

Jeremy M Wolfe. Guided search 6.0: An updated model of visual search. *Psychonomic Bulletin & Review*, 28(4):1060–1092, 2021.

Jeremy M Wolfe and Michael J Van Wert. Varying target prevalence reveals two dissociable decision criteria in visual search. *Current biology*, 20(2):121–124, 2010.

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.

Jeremy M Wolfe, Evan M Palmer, and Todd S Horowitz. Reaction time distributions constrain models of visual search. *Vision research*, 50(14):1304–1311, 2010.

Zoe Jing Xu, Alejandro Lleras, and Simona Buetti. Predicting how surface texture and shape combine in the human visual system to direct attention. *Scientific reports*, 11(1):1–13, 2021.

Xinger Yu, Timothy D Hanks, and Joy J Geng. Attentional guidance and match decisions rely on different template information during visual search. *Psychological science*, 33(1):105–120, 2022.

Xinger Yu, Zhiheng Zhou, Stefanie I Becker, Sage EP Boettcher, and Joy J Geng. Good-enough attentional guidance. *Trends in Cognitive Sciences*, 2023.

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.