

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

Anna E. Hughes^a, Anna Nowakowska^b, Alasdair D. F. Clarke¹

^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 propose running a new fully within-subjects experiment to attempt to replicate the key original findings, with some changes to help distinguish between theories.

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

6 has been between efficient (also referred to as parallel or pop-out) and inefficient
7 (serial) search. These are often studied in the context of the regression slope be-
8 tween the number of distractors and mean reaction time, which has been termed
9 the *search slope*. When the search slope is shallow (usually positive, but occasion-
10 ally negative e.g. (Rangelov et al., 2017)), the search is called efficient or parallel,
11 and the addition of more non-target distractors has little impact on an observers
12 difficulty in finding a target. When the slope is steeper, each additional distrac-
13 tor has a noticeable impact on increasing difficulty, and the search is described
14 as inefficient or serial. However, the distinction between these types of search is
15 often less clear in real experimental data, with a range of different search slopes
16 being seen for different types of targets and distractors (Duncan and Humphreys,
17 1989; Cave and Wolfe, 1990; Wolfe, 1998; Liesefeld et al., 2016). Recent work
18 has also attempted to model the variation in search slopes at the boundary between
19 inefficient and efficient search (Liesefeld et al., 2016).

20 In the current study, we are interested in what has traditionally been termed
21 efficient or parallel search, and the factors that affect search slope in these condi-
22 tions. Recent work has suggested that for efficient search, there is a logarithmic
23 relationship between distractor set size and reaction time, and that this relation-
24 ship can be modified by target-distractor similarity (Buetti et al., 2016), providing
25 evidence that search behaviour in parallel search is more complex than has pre-
26 viously been assumed. This observation has formed the basis of the ‘Target Con-
27 trast Signal (TCS) Theory’ (Lleras et al., 2020), which aims to provide a means
28 of predicting observer search slopes for new search arrays by quantifying target-
29 distractor differences. For example, by measuring search slopes for conditions in
30 which the distractors differ from the target along a *single feature* (e.g. colour *or*
31 shape), it has been shown that you can predict search times for arrays in which
32 the target differs from the distractors along two features (e.g., colour *and* shape)
33 which we refer to here as *double feature* search (Buetti et al., 2019) (but simi-
34 lar paradigms have been known by other names e.g. ‘redundant feature search’
35 (Krummenacher and Müller, 2012; Mordkoff and Yantis, 1991)). Here, we aim
36 to replicate and extend this work both theoretically and empirically, to test the
37 generalisability of the TCS model, and to suggest ways in which the TCS model
38 could be modified to generate better predictions.

39 1.1. Previous Work

40 Many different forms of visual search models have been proposed. One well
41 developed class of models are the saliency models, which aim to predict eye move-
42 ments during scene viewing, including visual search. They rest on the assump-

tion 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:

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

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

121 1.2. Rationale for proposed work

122 While many aspects of the TCS framework have been tested, with extremely
 123 promising results, there remains a great deal of scope for verification of some of
 124 the key findings to date, and extensions of aspects of the model. In all implementa-
 125 tions of TCS so far, predictions of search efficiency (e.g. in heterogeneous scenes)
 126 have been made on the average of a group of participants, using data from a dif-
 127 ferent group performing a different task (e.g. searching in homogeneous scenes).
 128 Thus, we know that TCS can replicate group-level averages between subjects in
 129 search well, but we do not know to what extent it is also able to make predictions
 130 at the individual level. This is particularly important given that conclusions based
 131 on aggregate data can be different from those that take individual differences into
 132 account; in one study where participants searched for a target in an array of ran-
 133 domly oriented line segments, aggregating the data suggested that participants
 134 were using a stochastic search model (Nowakowska et al., 2017). However, when
 135 considering each participant individually, it became clear that there was a high
 136 level of heterogeneity in responses, with some participants performing close to
 137 optimally, and others actually performing worse than chance (Nowakowska et al.,
 138 2017). Similarly striking variability has also been reported in other search studies
 139 (Irons and Leber, 2016, 2018; Clarke et al., 2022a).

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

146 We also extend the TCS model into a Bayesian framework, where we begin
 147 with existing 'prior' beliefs that are updated with data to give 'posterior' beliefs
 148 that can be used for inference (McElreath, 2020). We think this has a number
 149 of advantages over frequentist approaches. Perhaps most importantly, Bayesian
 150 models are highly flexible. We demonstrate how we are able to specify a model
 151 that is able to more accurately represent the distribution of responses (for exam-
 152 ple, by specifying a response distribution that avoids predicting negative reaction

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.

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).

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 successfully replicate the original analysis

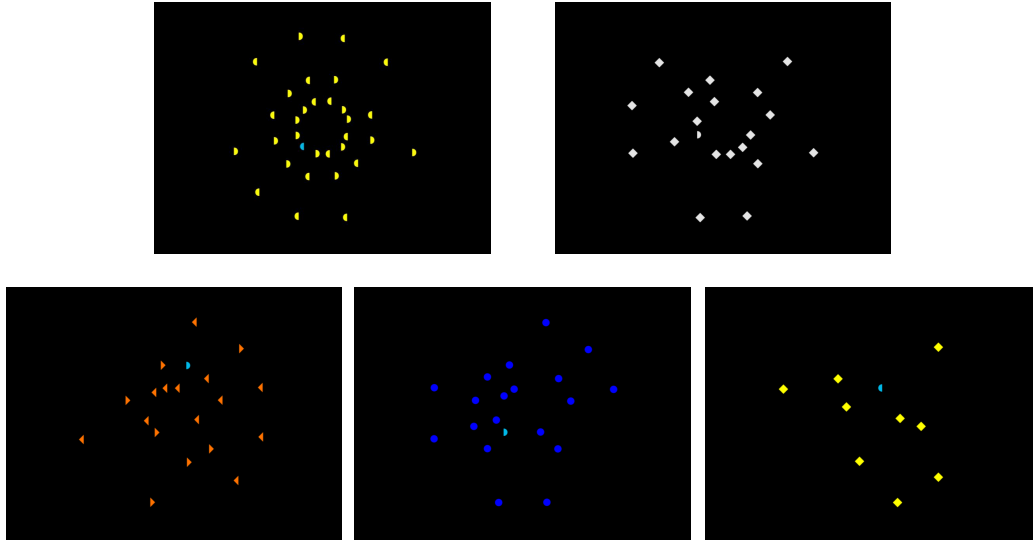


Figure 1: Example stimuli from Buetti et al. (2019) Top left: Expt 1A. Here, the target is a blue semicircle within a set of homogeneous (yellow semicircle) distractors. Top right: Expt 1B. The target is a grey semicircle in circular grey distractors. Bottom left: Expt 2A. The target is a blue semicircle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

189 (both using frequentist modelling and Bayesian modelling; see *Supplementary*
190 *Materials*).

191 2.1. TCS modelling overview

192 In Experiment 1a of Buetti et al. (2019), participants searched for a cyan
193 semicircle target among blue, yellow or orange semicircular distractors i.e. they
194 searched for a target that differed from the distractors by a *single feature* (colour).
195 The experiment was then repeated (1b) using a different single feature (shape,
196 with participants searching for the semicircular target within triangle, circle or di-
197 amond distractors). In Experiments 2a, 2b and 2c, participants again searched for
198 a cyan semicircle, but this time, the distractors differed in both shape and colour.
199 We will refer to these conditions as *double features*. Note, unlike in standard con-
200 junction searches, in this paradigm, the distractors are all identical with respect to
201 these features (i.e. orange triangles). Examples of all these stimuli are shown in
202 Figure 1. Buetti et al. (2019) also carried out a replication of their basic results
203 using slightly different target and distractor stimuli (Experiments 3 and 4).

204 The *Target Signal Contrast* theory is built around a linear model for predicting
 205 mean reaction times from the logarithm of the number of distractors (see Equation
 206 1). In particular, the TCS theory allows us to predict the value of the logarithmic
 207 slope, $D_{c,s}$, in this condition based on the corresponding D_i in the single feature
 208 search experiments.

209 2.1.1. Calculating the intercept, a , and the logarithmic slope parameter, D_i

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

218 Fitting the model specified in Equation 1 to the data, we obtain the values for
 219 D_c and D_s given in Table 2. As can be seen, the more similar the distractors are to
 220 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* for full values for all experiments.

221 2.1.2. Estimating $D_{c,s}$, the logarithmic slope parameter for compound features

222 In the context of the current experiments, the core idea of TCS theory is that
 223 we can estimate the (natural) logarithmic slope parameter for a double feature
 224 visual search from the slopes parameters in the two independent single feature
 225 searches i.e., $D_{c,s} = f(D_c, D_s)$. Buetti et al. (2019) tested three different models
 226 for predicting D for compound colour-shape stimuli. The best feature guidance
 227 model (Equation 2) suggests that when the target and lures differ in two dimen-
 228 sions, participants will choose to attend to whichever feature dimension is the
 229 most discriminable (i.e. has the smallest D value):

$$D_{c,s} = \min(D_c, D_s) \quad (2)$$

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

$$D_{c,s} = \frac{1}{\sqrt{(\frac{1}{D_c})^2 + (\frac{1}{D_s})^2}} \quad (3)$$

233 Finally, the collinear contrast integration model also assumes independence of
 234 feature dimensions, but assumes that while the visual features create a multidimensional space, the contrast between them is unidimensional. As D is assumed
 235 to be inversely proportional to contrast, the equation can be written as follows:
 236

$$\frac{1}{D_{c,s}} = \frac{1}{D_c} + \frac{1}{D_s} \quad (4)$$

237 Buetti et al. (2019) found that with their dataset, the collinear contrast inte-
 238 gration model was best able to predict $D_{c,s}$ from D_c and D_s , with $R^2 = 0.915$.
 239 We verified we were able to replicate this result using the dataset available on
 240 OSF (<https://osf.io/f3m24/>)¹ and using the exclusion criteria originally applied;
 241 see Figure 2 (left panel) and *Supplementary Materials* for details. We show that
 242 we are able to do this using both the frequentist modelling approaches used in the
 243 original paper, and using Bayesian modelling.

244 2.1.3. Estimating a , the intercept parameter for compound features

245 As a is the intercept of the model, it represents how long observers take to find
 246 a target when $N_T = 0$, i.e., there are no distractors. As such, it should be inde-
 247 pendent of both shape and colour, and can be thought of as the role of non-search
 248 processes (such as motivation, motor preparation etc.) that influence reaction time.
 249 In Buetti et al. (2019), a was calculated for each sub-experiment. Here, we follow
 250 that method in order to replicate their results exactly.

251 2.1.4. Estimating mean reaction times

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

¹downloaded on 28th August 2020

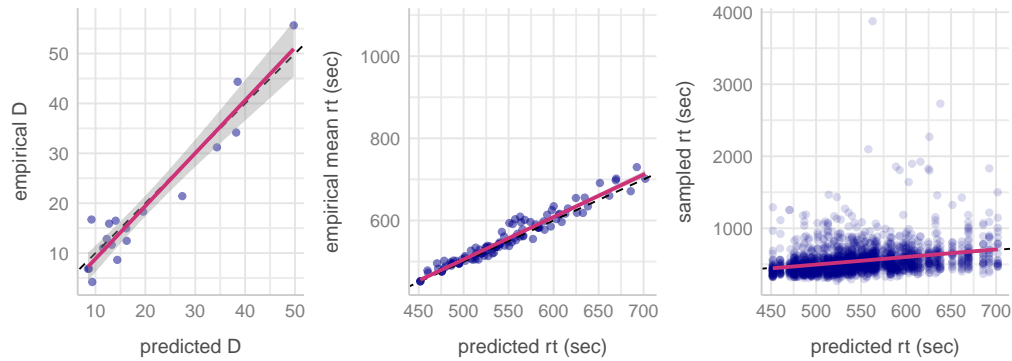


Figure 2: (left) The collinear method for calculating D offers a good prediction. (centre) Using the TCS to predict reaction times. (right) Each dot now represents a randomly sampled reaction time from an observer. Note that there is greater spread in the data points here, due to the fact that there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

2.1.5. 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.

- 310 **2. Log-linear effect of N_T .** We will test the TCS model assumption that N_T
 311 has a log-linear effect by testing models with and without the log of this
 312 term. We expect that this will confirm the results previously seen in papers
 313 testing TCS i.e. that the log-linear approach will be best.
 314
- 315 **3. Contrast model comparisons.** We will test the hypothesis proposed by
 316 (Buetti et al., 2019): specifically, that the *collinear contrast integration*
 317 *model* outperforms the *best feature guidance*, and *orthogonal contrast com-*
 318 *bination models* for the calculation of D , by calculating and comparing the
 319 mean absolute prediction error for each model.
 320
- 321 **4. Reaction time predictions.** We will further test the hypothesis proposed by
 322 (Buetti et al., 2019) by testing which model gives the best prediction at the
 323 trial-by-trial RT level.

324 We will test each of these hypotheses by calculating the marginal likelihood
 325 of the relevant models, and then calculating the posterior probabilities. This will
 326 give us a probability for each model that represents the likelihood that the model
 327 gives the best prediction. We will consider there to be evidence for one model over
 328 the others if a given model has a probability above 90%. We will consider there
 329 to be strong evidence for one model over the others if that model has a posterior
 330 probability above 99%. This approach is most appropriate for our model: other
 331 measures of model fit, such as AIC, require an assumption of flat priors (which is
 332 not valid for multi-level models) and are based on point estimates (which is not
 333 valid for Bayesian models) (McElreath, 2020).

334 3.3. *Planned Explorations*

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

344 One of the benefits of using a multi-level modelling approach is that it is rel-
 345 atively easy to extend to incorporate other factors that may contribute to reaction

346 times, such as eccentricity and inter-item distance, which may help to explain
347 behaviour further. To demonstrate this, we will also run exploratory analyses in-
348 cluding a factor for which ring the target is in to assess whether this improves
349 model fit or affects any of the conclusions that can be drawn from the model.

350 3.4. Pilot Experiment

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

362 4. General Methods

363 4.1. Sample Size: Participants and Trials

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

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

374 Finally, we have carried out a simulation experiment to estimate the confi-
375 dence intervals on the mean when sampling from a log-normal distribution. We
376 defined our distribution to have a mean-log of 6.135 and a standard deviation of
377 0.32. These values were loosely based on the distributions of reaction times in
378 Buetti et al. (2019). The results are shown in Figure 3. Based on these simula-
379 tions, we find that a sample of $n = 20$ leads to a 95% confidence interval that is



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.

380 approximately 1.4 times larger than $n = 40$. We feel this is a suitable compromise
 381 given we will be collecting our data within-subjects.

382 4.2. Stimuli

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

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

398 **Set sizes:** we will run all the distractor set sizes used in Buetti et al. (2019) (1,
 399 4, 9, 19 and 31). We will also run target-only 'zero distractor' trials (60 in total,
 400 with 12 being the white semicircle target and the remainder the red semicircle
 401 target).

402 The experiments were programmed in PsychoPy and Pavlovia (Peirce et al.,
403 2019). Stimuli were pre-made to generate search array images with 1920×1080
404 resolution.

405 4.3. Procedure

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

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

429 4.4. Data Pre-processing

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

435 For participants included in the analysis, we will apply the data cleaning used
436 in the pilot data analysis i.e. *removing incorrect trials* and removing the top and
437 bottom 1% of their data.

438 4.5. Analysis Plan

439 All analysis will be carried out using R (v4.2.0), brms (v2.17.0) and rStan
440 (v2.26.11) As discussed above, we will use mixed-effect models with either nor-
441 mal, lognormal or shifted lognormal distributions.

442 Please see the analysis of our pilot data for a full implementation of our anal-
443 ysis pipeline, including all code (available on Github at https://github.com/scienceanna/TCS_Bayesian).
444

445 5. Results

446 All 40 participants had accuracy over 90% (minimum 93.1%). One participant
447 had an average response time (1100ms) over two standard deviations from the
448 group average response time (781ms) and was removed. Incorrect trials were
449 then removed, and the data was trimmed (only including response times between
450 the 1% and 99% quantiles) leaving us with 39 participants completing a total of
451 59,587 trials.

452 All Bayesian models were fit to the new data using exactly the same proce-
453 dure² as the pilot data presented in the Stage One review process. We checked
454 for convergence of our models by visually inspecting the chains as well as ver-
455 ifying that the \hat{R} was close to 1 for all parameters of all the fitted models (see
456 Supplementary Material for full model fit information).

457 5.1. Hypothesis 1: Shifted-lognormal model

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

463 After each of these models had been fit to the data, leave-one-out (LOO) model
464 comparison was used to calculate posterior probabilities for each. The results of
465 this procedure allocated $\sim 100\%$ of the weight to the shifted-lognormal model, so
466 we can conclude that it is the best distribution (out of the three we tested³) to use

²The 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.

³See discussion for Wald, Weibull, etc.

467 for modelling response times in this paradigm. This model is shown in Figure 2.2
 468 of the supplementary materials.

469 5.2. Hypothesis 2: log-linear effect of N_T

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

474 5.3. Hypothesis 3: Contrast Model Comparison

475 Now that we have confirmed that the shifted-lognormal multilevel model (with
 476 a log-linear effect of N_T) is indeed the best fit to the data we will extract the search
 477 slopes for each feature. These are summarised in Table 3. We can see that we have
 478 successfully obtained a range of values for both D_c and D_s . As with Buetti et al.
 479 (2019) we find that the values for D_s are larger than D_c (see Table 2), meaning
 480 that search slopes for colour features are shallower than shape.

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.

481 We now combine the *single-feature* search slopes, D_c and D_s , to predict the
 482 *double-feature* conditions ($D_{c,s}$) using Equations 2, 3 and 4 and above. The results
 483 are summarised in Figure 4. We find that while the collinear contrast model has
 484 the highest R^2 (0.922, compared to $R^2 = 0.884$ for best feature, and $R^2 = 0.916$ for
 485 orthogonal contrast), the orthogonal contrast model is the most accurate, both in
 486 terms of mean absolute error (0.165, compared to 0.185 for best feature and 0.271
 487 for collinear) and having a regression slope closest to 1 (1 compared to 0.753 and
 488 1.48). Therefore, Hypothesis 3 does not hold: orthogonal contrast rather than
 489 collinear contrast offers the best prediction of search slopes in the double-feature
 490 condition.

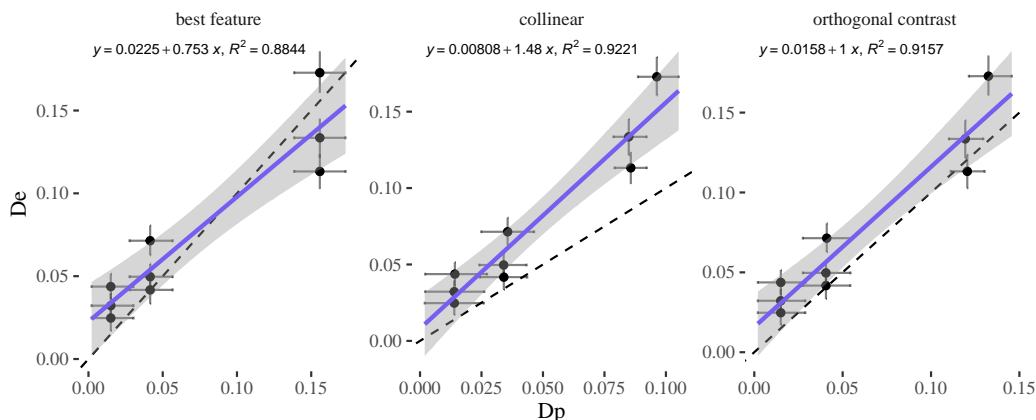


Figure 4: Predicting $D_{c,s}$ from D_c and D_s . The x-axis shows our predictions, D_p , using the best feature, collinear contrast, and orthogonal contrast models.

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.

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).

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

536 Pearson's r for each sample from our posterior, which gives us a full posterior dis-
 537 tribution for the correlations. We can see in Figure 5 (*right*) that while both the D_c
 538 and D_s are correlated within feature classes (~ 0.75), there is no correlation of any
 539 of the colour features with any of the shape features. The individual differences
 540 for the *double-feature* conditions are much less pronounced - these conditions are
 541 easy and the search slopes are quite close to flat. Hence, the correlations are all
 542 much weaker, presumably due to range restriction.

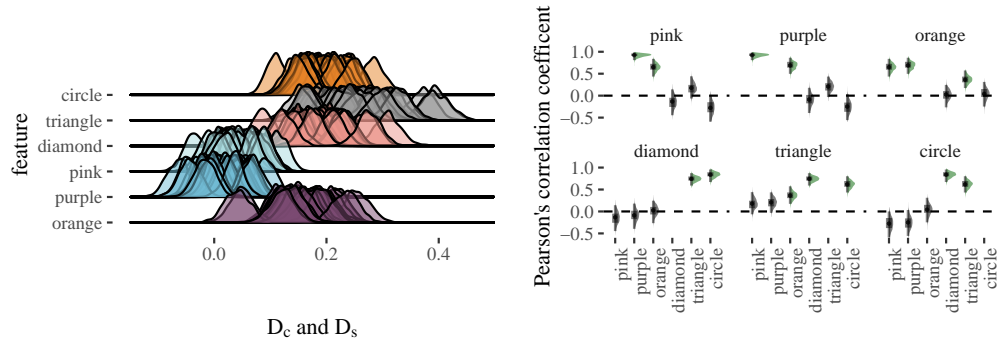


Figure 5: Individual differences in D_c and D_s . (*left*) Posterior probability distributions for D_c and D_s for each individual. (*right*) Estimated correlations between each of the D_c and D_s .

543 Given these results, it is perhaps unsurprising that our analysis for Hypothesis
 544 4 leads to an inconclusive result for distinguishing between the three contrast com-
 545 bination methods. Perhaps taking these individual differences into account when
 546 we predict reaction times will lead to improved power to discriminate between
 547 the models. However, before we do so, we will also investigate incorporating
 548 information about target eccentricity into the model.

549 6.2. Target Eccentricity

550 It is well known that there are eccentricity effects in visual search, with reac-
 551 tion times being longer for targets that are further away from fixation (Carrasco
 552 et al., 1995). To investigate this in our dataset, we will use the same methods as
 553 above (fitting a multi-level shifted-lognormal model) but now including an addi-
 554 tional factor that represents how far the target was from the fixation cross. This
 555 is coded as a three-level categorical factor representing which ring contained the
 556 target (see stimulus details, above). Allowing for interactions with the *feature* and
 557 $\log N_T$ increases the number of fixed effect parameters in the model from XX to
 558 XX.

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

559 We experimented with including r in the random effect structure, but this
 560 proved difficult to fit. We also had to revise the priors used in our registered analy-
 561 sis, in order to lower the intercept. Full details can be found in the supplementary
 562 materials.

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

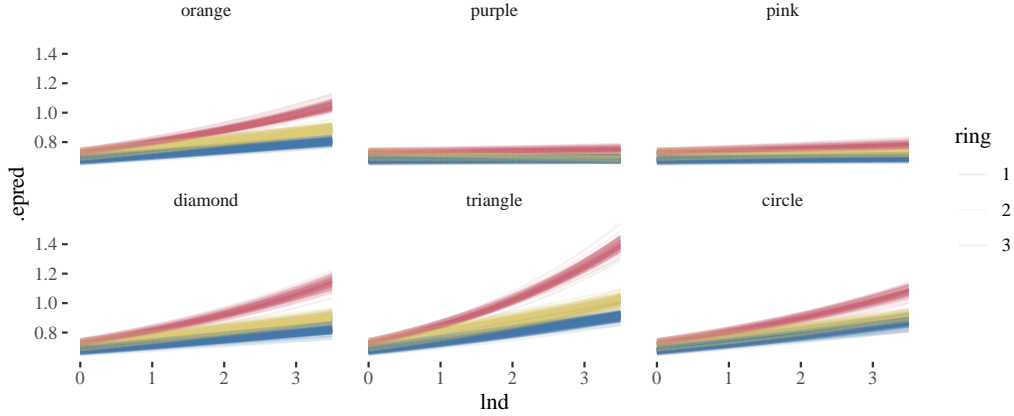


Figure 6: Fixed effects for predicting the effect of ring, feature and number of distractors on response times. SWITCH TO HPDI PLOT TO BE CONSISTENT WITH WHAT WE USED BEFORE. Shaded regions represent 97% HPDI. We can see that *ring* has an effect on search slopes, and that this effect is more pronounced for some features (i.e., triangles) than others.

570 We can now compute our predictions (D_p) for $D_{c,s}$ taking the *ring* into ac-
 571 count. Doing so leads us to a similar result as before with orthogonal contrast
 572 outperforming the best feature and collinear measures in terms of absolute error
 573 (0.023 compared to 0.025 (best feature) and 0.034 (collinear)). However, the re-
 574 gression slopes are all relatively similar (0.90 for best feature, 1.58 for collinear
 575 and 1.15 for orthogonal contrast).

576 6.3. Predicting Response Times

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

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.

582 6.4. Issues with the Collinear Contrast method

583 The collinear contrast method performs very poorly in this test. To explain
 584 this, we can look back at Equation 4 and realise that empirical data is not required
 585 to reject this hypothesis as a suitable method for predicting search slopes in feature
 586 search. This is because when search slopes are close to 0, it is possible that we will
 587 observe negative values in the empirical data. Breaking down our data to compute
 588 search slopes for each person and each target eccentricity increases the chances
 589 of this being observed. Looking at Equation 4 we can see that $D_1 \sim -D_2 \Rightarrow$
 590 $1/D_1 + 1/D_2 \sim 0$. This leads to our estimated $D = \frac{1}{1/D_1 + 1/D_2} \gg D_1, D_2$ which
 591 is clearly incorrect.

592 7. General Discussion

593 In this paper, we aimed to test the extent to which the results of Buetti et al.
 594 (2019) replicate and generalise, using a new modelling approach. Our results
 595 allow us to confirm our pre-registered hypotheses 1 and 2. Firstly, a shifted-
 596 lognormal distribution of response times outperforms normal and lognormal dis-
 597 tributions, demonstrating that reaction time data are best modelled by a skewed
 598 distribution with an offset. Similarly, we confirmed that the number of distractors
 599 has a log-linear effect in this model, in line with the predictions of TCS theory.

600 We also replicated other aspects of the original Buetti et al. (2019) paper with
601 a different experimental set up, such as to observing shallower search slopes for
602 colour features compared to shape features.

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

614 7.1. *Modelling of reaction times*

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

624 However, there have been concerted efforts within the literature to model re-
625 action time distributions more effectively: indeed, Buetti et al. (2019) use $\log N_T$
626 when computing their search slopes. In terms of reaction time distributions, log
627 transformations are frequently taught as a way to normalise reaction time data
628 (although often with caveats regarding how this can change the interpretation of
629 the results (Osborne, 2002)) and are frequently used in analysing reaction time
630 data e.g. (Clarke et al., 2022b). Researchers have also looked at other distribu-
631 tions to assess which offer the best fit to empirical response times in visual search.
632 For example, Palmer et al. (2011) compared ex-Gaussian, ex-Wald, Gamma, and
633 Weibull distributions and found that the distributions with exponential compo-
634 nents offer a better fit to the data. Our results are in line with this. However,
635 we opted to use a shifted-lognormal distribution in our analysis above for mostly
636 pragmatic reasons, as these more complex distributions are often computationally

difficult to fit ⁴. 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, 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 ?? 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 more sophisticated approaches that better account for the data distribution can be taken with relative ease.

7.2. *Discriminating between combination methods*

CHECK ALL THIS; AL THINKS THAT THIER ORIG DATA + OUR ANALYSIS METHODS LEADS TO SAME CONCLUSION AS OUR DATA + OUR 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 assumptions 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 was better. This does not seem to be something inherent to our slightly altered methods, as we were able to replicate the original Buetti et al. (2019) findings using our methods on their data, and our pilot analysis also suggested the collinear contrast model had the lowest mean error for that (small) dataset.

One possibility is that our small modifications to the experimental stimuli

⁴See: <https://discourse.mc-stan.org/t/model-fails-to-converge-when-using-brms/9062>

671 changed the strategy that participants used. However, this seems unlikely given
672 that we only made changes to the colour of the stimuli, a manipulation that Buetti
673 et al. (2019) also used, with no changes to their overall conclusions. The orthog-
674 onal contrast method would arguably be more likely for feature dimensions that
675 are not independent of each other, but the colour and shape dimensions used in
676 our stimuli are similar to those used previously (and our exploratory correlation
677 analyses suggest that they can be considered independent feature dimensions).

678 Another possibility is that because some participants had negative search slopes,
679 the collinear contrast model predicts implausibly large reaction times, due to the
680 mathematical formulation of this model, leading to worse predictions. Given that
681 negative search slopes do occur in some situations (Utochkin, 2013), we suggest
682 that a future improvement for the collinear contrast integration model would be to
683 modify it to be able to give sensible predictions in these situations.

684 Finally, we would argue that it is difficult from these results to definitively
685 make a decision about which model is best: all three models give very similar
686 predictive weights during our model evaluation process. One challenge is that
687 in general, the double feature searches are easy, and therefore the search slopes
688 are fairly flat and there is not much variability to allow different models to make
689 different predictions. For the current paradigm, a fruitful approach for future re-
690 search could be to consider only trials in which there are a large number of dis-
691 tractors: the small differences in predicted search slope diverge as D_T increases,
692 leading to larger differences in predicted reaction times.

693 7.2.1. Individual differences

694 Our (planned) exploratory analysis of the individual differences in search slopes
695 suggests that there are large differences from one observer to the next. Indeed
696 in some cases, these are larger than the differences from one feature to another.
697 The difference between the steepest and shallowest search slopes (fixed effects) is
698 0.238 ($D_{triangle} = 0.253$ 3 while $D_{purple} = 0.015$). If we compare this to the range
699 of observer search slopes within a feature, we find this varies from 0.242 ($D_{triangle}$
700 per-observer ranges from 0.395 to 0.152) to 0.149 (D_{pink} ranges from -0.065 to
701 0.092). This suggests a challenge for modelling based on average performance:
702 can we be sure that averages represent a meaningful summary of the data, given
703 that we see very clear individual differences? It could certainly be argued that ob-
704 servers might be using different strategies, and thus some members of the sample
705 population might use (for example) a collinear combination strategy, while others
706 use an orthogonal contrast strategy. Variable strategies have been found for other
707 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. Understanding whether these correlations reflect something about an observer’s behaviour with different features, or instead how an observer’s behaviour changes over time, would be an interesting question for future research.

7.2.2. *Eccentricity*

(Buetti et al., 2019) argues that the “odd one out” 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. However, the experimental design did not preclude slower, serial search (eye movements were not monitored, and participants had up to five seconds to make a response). The relatively strong eccentricity effects we found during our exploratory analyses (a model with target ring number included was a better predictor of the data than one without) also suggests the possibility that, at least in some cases, peripheral information was insufficient (or at least, observers felt it was insufficient) to make judgements. This supports the idea that it may not be helpful to distinguish between serial and parallel search, and many models now assume that one mechanism underlies all different types of search (Wolfe, 1998). DO WE WANT TO ALSO CITE HULLMAN’S DEMISE OF ITEM SEARCH PAPER HERE???? Given that the experiment was not designed to test eccentricity effects and this analysis was exploratory, we make no strong claims about what this finding might mean for Target Contrast Signal Theory Lleras et al. (2020). Instead, our claim is simply that this modelling methodology is useful precisely because it can incorporate a wider range of factors easily, giving us a better insight into what is driving behaviour in visual search, and helping to test model assumptions.

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 make sensible predictions with negative search slopes b) focusing on trials with a larger number of distractors 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

780 (Moshontz et al., 2018) to our scientific questions. As such, we would like to
781 conclude by encouraging other researchers to critique, build on and improve the
782 approach we have taken in this manuscript, in order to further improve our ability
783 to model performance in visual search tasks.

784 **Conflict of interest**

785 The authors declare that they have no conflict of interest.

786 **Acknowledgements**

787 This work was supported by an Economic and Social Research Council grant
788 (ES/S016120/1) to ADFC and employing AN.

789 **References**

- 790 Stefanie I Becker. The role of target–distractor relationships in guiding attention
791 and the eyes in visual search. *Journal of Experimental Psychology: General*,
792 139(2):247, 2010.
- 793 Stefanie I Becker, Christian Valuch, and Ulrich Ansorge. Color priming in pop-
794 out search depends on the relative color of the target. *Frontiers in psychology*,
795 5:289, 2014.
- 796 Aline Bompas, Petroc Sumner, and Craig Hedge. Non-decision time: the higg’s
797 boson of decision. *bioRxiv*, pages 2023–02, 2023.
- 798 Simona Buetti, Deborah A Cronin, Anna M Madison, Zhiyuan Wang, and Ale-
799 jandro Lleras. Towards a better understanding of parallel visual processing in
800 human vision: Evidence for exhaustive analysis of visual information. *Journal*
801 *of Experimental Psychology: General*, 145(6):672, 2016.
- 802 Simona Buetti, Jing Xu, and Alejandro Lleras. Predicting how color and shape
803 combine in the human visual system to direct attention. *Scientific reports*, 9(1):
804 1–11, 2019.
- 805 Marisa Carrasco, Denise L Evert, Irene Chang, and Svetlana M Katz. The eccen-
806 tricity effect: Target eccentricity affects performance on conjunction searches.
807 *Perception & psychophysics*, 57:1241–1261, 1995.

- 808 Kyle R Cave and Jeremy M Wolfe. Modeling the role of parallel processing in
809 visual search. *Cognitive psychology*, 22(2):225–271, 1990.
- 810 Alasdair DF Clarke, Jessica L Irons, Warren James, Andrew B Leber, and
811 Amelia R Hunt. Stable individual differences in strategies within, but not be-
812 tween, visual search tasks. *Quarterly Journal of Experimental Psychology*, 75
813 (2):289–296, 2022a.
- 814 Alasdair DF Clarke, Anna Nowakowska, and Amelia R Hunt. Visual search habits
815 and the spatial structure of scenes. *Attention, Perception, & Psychophysics*, 84
816 (6):1874–1885, 2022b.
- 817 Elaine A Corbett and Philip L Smith. A diffusion model analysis of target detec-
818 tion in near-threshold visual search. *Cognitive Psychology*, 120:101289, 2020.
- 819 John Duncan and Glyn W Humphreys. Visual search and stimulus similarity.
820 *Psychological review*, 96(3):433, 1989.
- 821 Nathan J Evans and Eric-Jan Wagenmakers. Evidence accumulation models: Cur-
822 rent limitations and future directions. 2019.
- 823 Andrew Gelman and Jennifer Hill. *Data analysis using regression and multi-
824 level/hierarchical models*. Cambridge university press, 2006.
- 825 Olivia Guest and Andrea E Martin. How computational modeling can force theory
826 building in psychological science. *Perspectives on Psychological Science*, 16
827 (4):789–802, 2021.
- 828 Anna E Hughes, Rosy V Southwell, Iain D Gilchrist, and David J Tolhurst. Quan-
829 tifying peripheral and foveal perceived differences in natural image patches to
830 predict visual search performance. *Journal of vision*, 16(10):18–18, 2016.
- 831 Johan Hulleman and Christian NL Olivers. On the brink: The demise of the item
832 in visual search moves closer. *Behavioral and Brain Sciences*, 40, 2017.
- 833 Jessica L Irons and Andrew B Leber. Choosing attentional control settings in a
834 dynamically changing environment. *Attention, Perception, & Psychophysics*,
835 78(7):2031–2048, 2016.
- 836 Jessica L Irons and Andrew B Leber. Characterizing individual variation in the
837 strategic use of attentional control. *Journal of Experimental Psychology: Hu-
838 man Perception and Performance*, 44(10):1637, 2018.

- 839 Laurent Itti and Christof Koch. A saliency-based search mechanism for overt and
840 covert shifts of visual attention. *Vision research*, 40(10-12):1489–1506, 2000.
- 841 Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual
842 attention for rapid scene analysis. *IEEE Transactions on pattern analysis and*
843 *machine intelligence*, 20(11):1254–1259, 1998.
- 844 Paul D Kieffaber, Emily S Kappenman, Misty Bodkins, Anantha Shekhar, Brian F
845 O’Donnell, and William P Hetrick. Switch and maintenance of task set in
846 schizophrenia. *Schizophrenia research*, 84(2-3):345–358, 2006.
- 847 Christof Koch and Shimon Ullman. Shifts in selective visual attention: towards
848 the underlying neural circuitry. In *Matters of intelligence*, pages 115–141.
849 Springer, 1987.
- 850 Kathryn Koehler, Fei Guo, Sheng Zhang, and Miguel P Eckstein. What do
851 saliency models predict? *Journal of vision*, 14(3):14–14, 2014.
- 852 Iuliia Kotseruba, Calden Wloka, Amir Rasouli, and John K Tsotsos. Do saliency
853 models detect odd-one-out targets? new datasets and evaluations. *arXiv*
854 *preprint arXiv:2005.06583*, 2020.
- 855 Árni Kristjánsson, Ómar I Jóhannesson, and Ian M Thornton. Common attentional
856 constraints in visual foraging. *PloS one*, 9(6):e100752, 2014.
- 857 Joseph Krummenacher and Hermann J Müller. Dynamic weighting of feature di-
858 mensions in visual search: behavioral and psychophysiological evidence. *Front-*
859 *iers in psychology*, 3:221, 2012.
- 860 Walden Y Li, Molly R McKinney, Jessica L Irons, and Andrew B Leber. As-
861 sessing the generality of strategy optimization across distinct attentional tasks.
862 *Journal of Experimental Psychology: Human Perception and Performance*, 48
863 (6):582, 2022.
- 864 Heinrich René Liesefeld and Hermann J Müller. A theoretical attempt to revive
865 the serial/parallel-search dichotomy. *Attention, Perception, & Psychophysics*,
866 82(1):228–245, 2020.
- 867 Heinrich René Liesefeld, Rani Moran, Marius Usher, Hermann J Müller, and
868 Michael Zehetleitner. Search efficiency as a function of target saliency: The
869 transition from inefficient to efficient search and beyond. *Journal of Experi-*
870 *mental Psychology: Human Perception and Performance*, 42(6):821, 2016.

- 871 Alejandro Lleras, Zhiyuan Wang, Anna Madison, and Simona Buetti. Predicting
872 search performance in heterogeneous scenes: Quantifying the impact of homo-
873 geneity effects in efficient search. *Collabra: Psychology*, 5(1), 2019.
- 874 Alejandro Lleras, Zhiyuan Wang, Gavin Jun Peng Ng, Kirk Ballew, Jing Xu, and
875 Simona Buetti. A target contrast signal theory of parallel processing in goal-
876 directed search. *Attention, Perception, & Psychophysics*, pages 1–32, 2020.
- 877 Anna Madison, Alejandro Lleras, and Simona Buetti. The role of crowding in par-
878 allel search: Peripheral pooling is not responsible for logarithmic efficiency in
879 parallel search. *Attention, Perception, & Psychophysics*, 80(2):352–373, 2018.
- 880 Dora Matzke and Eric-Jan Wagenmakers. Psychological interpretation of the ex-
881 gaussian and shifted wald parameters: A diffusion model analysis. *Psycho-*
882 *nomic bulletin & review*, 16(5):798–817, 2009.
- 883 Richard McElreath. *Statistical rethinking: A Bayesian course with examples in R*
884 *and Stan*. Chapman and Hall/CRC, 2020.
- 885 Rani Moran, Michael Zehetleitner, Hermann J Mueller, and Marius Usher. Com-
886 petitive guided search: Meeting the challenge of benchmark rt distributions.
887 *Journal of Vision*, 13(8):24–24, 2013.
- 888 Rani Moran, Michael Zehetleitner, Heinrich René Liesefeld, Hermann J Müller,
889 and Marius Usher. Serial vs. parallel models of attention in visual search: ac-
890 counting for benchmark rt-distributions. *Psychonomic bulletin & review*, 23(5):
891 1300–1315, 2016.
- 892 Rani Moran, Heinrich René Liesefeld, Marius Usher, and Hermann J Muller. An
893 appeal against the item’s death sentence: Accounting for diagnostic data pat-
894 terns with an item-based model of visual search. *Behavioral and Brain Sci-*
895 *ences*, 40:e148, 2017.
- 896 J Toby Mordkoff and Steven Yantis. An interactive race model of divided at-
897 tention. *Journal of Experimental Psychology: Human Perception and Perform-*
898 *ance*, 17(2):520, 1991.
- 899 Hannah Moshontz, Lorne Campbell, Charles R Ebersole, Hans IJzerman,
900 Heather L Urry, Patrick S Forscher, Jon E Grahe, Randy J McCarthy, Erica D
901 Musser, Jan Antfolk, et al. The psychological science accelerator: Advancing

- 902 psychology through a distributed collaborative network. *Advances in Methods*
903 *and Practices in Psychological Science*, 1(4):501–515, 2018.
- 904 Gavin JP Ng, Simona Buetti, Trisha N Patel, and Alejandro Lleras. Prioritiza-
905 tion in visual attention does not work the way you think it does. *Journal of*
906 *Experimental Psychology: Human Perception and Performance*, 2020.
- 907 Gavin Jun Peng Ng, Alejandro Lleras, and Simona Buetti. Fixed-target efficient
908 search has logarithmic efficiency with and without eye movements. *Attention,*
909 *Perception, & Psychophysics*, 80(7):1752–1762, 2018.
- 910 Anna Nowakowska, Alasdair DF Clarke, and Amelia R Hunt. Human visual
911 search behaviour is far from ideal. *Proceedings of the Royal Society B: Biolog-*
912 *ical Sciences*, 284(1849):20162767, 2017.
- 913 Klaus Oberauer and Stephan Lewandowsky. Addressing the theory crisis in psy-
914 chology. *Psychonomic bulletin & review*, 26:1596–1618, 2019.
- 915 Jason Osborne. Notes on the use of data transformations. *Practical assessment,*
916 *research, and evaluation*, 8(1):6, 2002.
- 917 Evan M Palmer, Todd S Horowitz, Antonio Torralba, and Jeremy M Wolfe. What
918 are the shapes of response time distributions in visual search? *Journal of ex-*
919 *perimental psychology: human perception and performance*, 37(1):58, 2011.
- 920 Derrick Parkhurst, Klintan Law, and Ernst Niebur. Modeling the role of salience
921 in the allocation of overt visual attention. *Vision research*, 42(1):107–123, 2002.
- 922 Jonathan Peirce, Jeremy R Gray, Sol Simpson, Michael MacAskill, Richard
923 Höchenberger, Hiroyuki Sogo, Erik Kastman, and Jonas Kristoffer Lindeløv.
924 Psychopy2: Experiments in behavior made easy. *Behavior research methods*,
925 51(1):195–203, 2019.
- 926 Michael J Proulx. Individual differences and metacognitive knowledge of visual
927 search strategy. *PLoS One*, 6(10):e27043, 2011.
- 928 Dragan Rangelov, Hermann J Müller, and Michael Zehetleitner. Failure to pop
929 out: Feature singletons do not capture attention under low signal-to-noise ratio
930 conditions. *Journal of Experimental Psychology: General*, 146(5):651, 2017.

- 931 Ruth Rosenholtz, Jie Huang, Alvin Raj, Benjamin J Balas, and Livia Ilie. A sum-
932 mary statistic representation in peripheral vision explains visual search. *Journal*
933 *of vision*, 12(4):14–14, 2012.
- 934 Benjamin W Tatler. The central fixation bias in scene viewing: Selecting an op-
935 timal viewing position independently of motor biases and image feature distri-
936 butions. *Journal of vision*, 7(14):4–4, 2007.
- 937 Benjamin W Tatler, Mary M Hayhoe, Michael F Land, and Dana H Ballard. Eye
938 guidance in natural vision: Reinterpreting salience. *Journal of vision*, 11(5):
939 5–5, 2011.
- 940 Anne Treisman and Stephen Gormican. Feature analysis in early vision: evidence
941 from search asymmetries. *Psychological review*, 95(1):15, 1988.
- 942 Anne M Treisman and Garry Gelade. A feature-integration theory of attention.
943 *Cognitive psychology*, 12(1):97–136, 1980.
- 944 Igor S Utochkin. Visual search with negative slopes: The statistical power of
945 numerosity guides attention. *Journal of vision*, 13(3):18–18, 2013.
- 946 Zhiyuan Wang, Simona Buetti, and Alejandro Lleras. Predicting search perfor-
947 mance in heterogeneous visual search scenes with real-world objects. *Collabra:*
948 *Psychology*, 3(1), 2017.
- 949 Jeremy M Wolfe. What can 1 million trials tell us about visual search? *Psycho-*
950 *logical Science*, 9(1):33–39, 1998.
- 951 Jeremy M Wolfe. Approaches to visual search: Feature integration theory and
952 guided search. *Oxford handbook of attention*, pages 11–55, 2014.
- 953 Jeremy M Wolfe and Michael J Van Wert. Varying target prevalence reveals two
954 dissociable decision criteria in visual search. *Current biology*, 20(2):121–124,
955 2010.
- 956 Jeremy M Wolfe, Kyle R Cave, and Susan L Franzel. Guided search: an alterna-
957 tive to the feature integration model for visual search. *Journal of Experimental*
958 *Psychology: Human perception and performance*, 15(3):419, 1989.
- 959 Jeremy M Wolfe, Evan M Palmer, and Todd S Horowitz. Reaction time distri-
960 butions constrain models of visual search. *Vision research*, 50(14):1304–1311,
961 2010.

- 962 Zoe Jing Xu, Alejandro Lleras, and Simona Buetti. Predicting how surface texture
963 and shape combine in the human visual system to direct attention. *Scientific*
964 *reports*, 11(1):1–13, 2021.
- 965 Xinger Yu, Timothy D Hanks, and Joy J Geng. Attentional guidance and match
966 decisions rely on different template information during visual search. *Psycho-*
967 *logical science*, 33(1):105–120, 2022.
- 968 Lingyun Zhang, Matthew H Tong, Tim K Marks, Honghao Shan, and Garrison W
969 Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Jour-*
970 *nal of vision*, 8(7):32–32, 2008.