Inter-trial effects in visual search: Factorial comparison of Bayesian updating models

Fredrik Allenmark, Hermann J. Müller, Zhuanghua Shi

General and Experimental Psychology, Psychology Department, LMU Munich

1. Experimental Psychology, Department of Psychology, LMU Munich, Germany
2. Department of Psychological Science, Birkbeck College (University of London), London, UK

1 Acknowledgements

This work was supported by German research foundation DFG MU773/16-1 to HJM and DFG SH166/3-1 to ZS. The authors would like to thank Lena Schröder and Kerstin Wenzel for recruiting and testing participants and Uta Noppeney for helpful discussions.

Correspondence concerning this article should be addressed to Zhuanghua Shi, General and Experimental Psychology, Department of Psychology, Ludwig-Maximilians-Universität München, 80802, Munich, Germany. Email: strongway@psy.lmu.de.

2 Abstract

Many previous studies on visual search have reported inter-trial effects, that is, observers respond faster when some target property, such as a defining feature or dimension, or the response associate with the target repeats versus changes across consecutive trial episodes. However, what processes drive these inter-trial effects is still controversial. Here, we investigated this question using a combination of Bayesian modelling of belief updating and evidence accumulation modelling in perceptual decision-making. In three visual singleton ('pop-out') search experiments, we explored how the probability of the response-critical states of the search display (e.g., target presence/absence) and the repetition/switch of the target-defining dimension (color/orientation) affect reaction time distributions. The results replicated the mean reaction time (RT) inter-trial and dimension repetition/switch effects that have been reported in previous studies. Going beyond this, to uncover the underlying mechanisms, we used the Drift-Diffusion Model (DDM) and the Linear Approach to Threshold with Ergodic Rate model (LATER) to explain the RT distributions in terms of decision bias (starting point) and information processing speed (evidence accumulation rate). We further investigated how these different aspects of the decision-making process are affected by different properties of stimulus history, giving rise to dissociable inter-trial effects. We approached this question by (i) combining each perceptual decision making model (DDM or LATER) with different updating models, each specifying a plausible rule for updating of either the starting point or the rate, based on stimulus history, and (ii) comparing every possible combination of trial-wise updating mechanism and perceptual decision model in a factorial model comparison. Consistently across experiments, we found that the (recent) history of the response-critical property influences the initial decision bias, while repetition/switch of the target-defining dimension affects the accumulation rate, reflecting the top-down modulation process. This provides strong evidence of a disassociation between response- and dimension-based inter-trial effects.

3 Results

Experiments 1 and 2 both consisted of three equally long blocks. The frequency of pop-out target presence (or absence) was varied across blocks in Experiment 1. In Experiment 2, a target was always present, and the

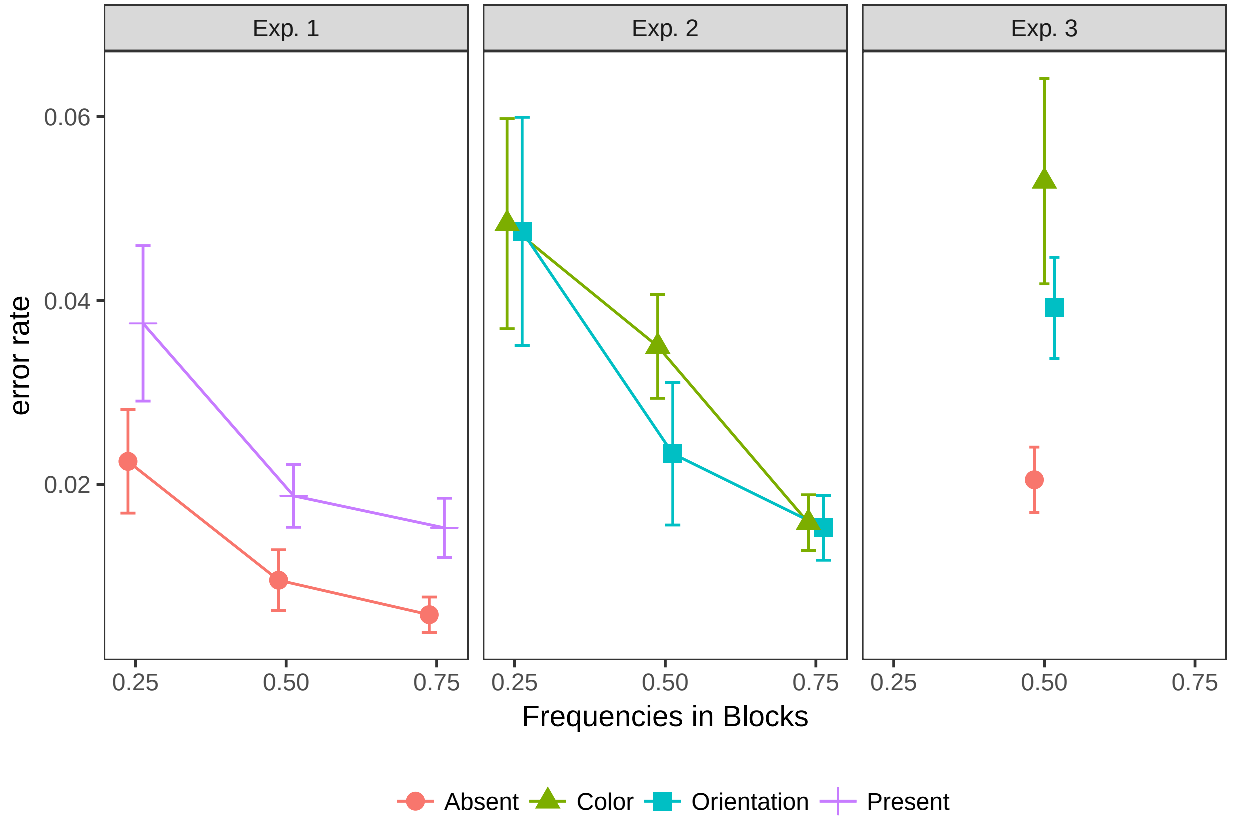

Figure 1: Error rates from Experiments 1, 2, and 3 for all combinations of target frequency. Target frequency is defined relative to the target condition, as the frequency with which that target condition occurred within a given block. Of note, this means that, for a given frequency, the data from the different target conditions do not necessarily come from the same block of the experiment. Error bars show the standard error of the mean.

frequency of the target being a color-defined or, alternatively, an orientation-defined singleton was varied across blocks. In Experiment 3, target presence and absence were kept equal frequent, as were trials with color- and orientation-defined singleton targets. One implication of this design is that the high-frequency condition for one target condition (present/absent, color/orientation) was implemented in the same block as the low-frequency condition for the other target condition. So, in all figures and analyses of the effects of frequency, the high- and low-frequency conditions are based on data collected in different blocks for each target condition, while the data for the medium-frequency condition comes from the same block for each target condition.

3.1 Error rates

The singleton search was quite easy, with participants making few errors overall: mean error rates were 1.5%, 2.5%, and 3.3% in Experiments 1, 2, and 3 respectively (Figure 1). Despite the low average error rates, error rates differed significantly between blocks in both Experiments 1 and 2 [$F(1.34, 14.78) = 11.50$, Huynh-Feldt-corrected degrees of freedom, $p < 0.01, \eta_p^2 = 0.51$, and $F(2, 22) = 12.20$,, $p < 0.001, \eta_p^2 = 0.53$, respectively]: as indicated by posthoc comparisons, error rates were higher in the low-frequency blocks. In particular, in Experiment 1 (target-present vs. -absent), error rates were significantly higher in the low-frequency compared to the medium- and high-frequency blocks [$t(11) = 3.67, p < 0.01$, $t(11) = 4.51, p < 0.001$, Bonferoni-corrected p-values], with no difference between the latter [$t(11) = 0.84, p > 0.9$]. In Experiment 2, error rates were also significantly higher in the low-frequency compared to the medium- and high-frequency blocks [$t(11) = 2.85, p < 0.05$; $t(11) = 4.92, p < 0.001$, Bonferoni-corrected p-values], without a significant

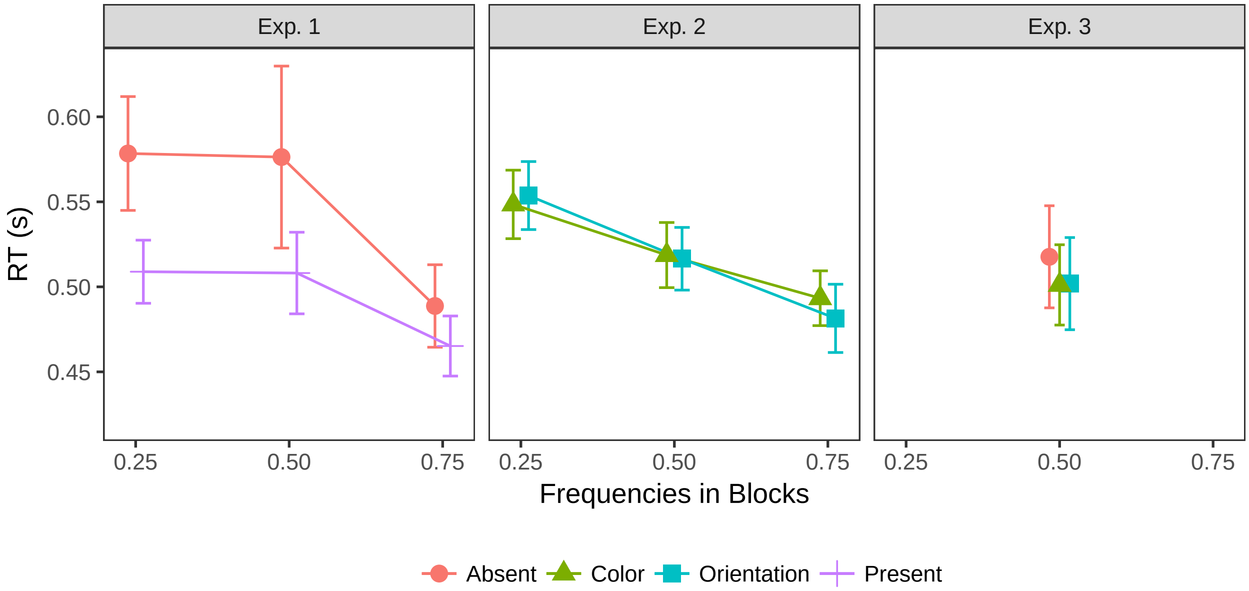

Figure 2: Mean RTs from Experiments 1, 2, and 3 for all combinations of target condition and target frequency. Target frequency is defined relative to the target condition, as the frequency with which that target condition occurred within a given block. Of note, this means that for a given frequency, the data from the different target conditions do not necessarily come from the same block of the experiment. Error bars show the standard error of the mean.

difference between the latter $[t(11) = 2.07, p = 0.15]$.

In addition, in Experiment 1, error rates were overall higher for the target-present than for target-absent trials, that is, there were more misses than false alarms, $F(1, 11) = 11.43, p < 0.01, \eta_p^2 = 0.51$. In contrast, there was no difference in error rates between color and orientation targets in Experiment 2, $F(1, 11) = 0.70, p = 0.42, BF_{10} = 0.31$. Interestingly, there was no interaction between target condition and frequency in either Experiment 1 or Experiment 2 $[F(2, 22) = 0.83, p = 0.45, BF = 0.24$, and, respectively, $F(1.28, 14.04) = 0.76$, Huynh-Feldt Corrected degrees of freedom, $p = 0.43, BF = 0.27$; Bayes factors compare the model with both main effects and an interaction term to the model with both main effects but no interaction] - suggesting the effect of the frequency of a condition within a block is independent of the target stimuli.

In Experiment 3, there was no manipulation of target (or dimension) frequency, but like in Experiment 1, error rates were higher on target-present compared to target-absent trials, $t(11) = 3.71, p < 0.001$; and similar to Experiment 2, there was no significant difference in error rates between color and orientation targets, $t(11) = 1.51, p = 0.16, BF_{10} = 0.71$.

3.2 Mean Reaction times (RTs)

Given that the error rates were low, we analyzed only RTs from trials with a correct response, though excluding outliers which were defined as trials on which the reciprocal RT was more than three standard deviations from the mean for any individual participant or the RT was shorter than 40 ms. Figure 2 shows the pattern of mean RTs from all three experiments. In both Experiments 1 and 2, there were significant main effects of frequency on RTs $[F(2, 22) = 10.25, p < 0.001, \eta_p^2 = 0.48$, and, respectively, $F(1.27, 13.96) = 29.83$, Huynh-Feldt-corrected degrees of freedom, $p < 0.01, \eta_p^2 = 0.73]$. Post-hoc comparisons indicated that RTs were faster in high-frequency compared to low-frequency blocks, suggesting participants were adapting to the statistics of stimuli in a way that allowed a faster response to the most frequent type of trial within a given block. In particular, in Experiment 1, RTs were significantly faster in the high-frequency compared to both the low- and medium-frequency block $[t(11) = 3.96, p < 0.01$, and, respectively, $t(11) = 3.88, p < 0.01$; Bonferroni-corrected p-values], but there was no significant difference between the medium- and low-frequency blocks, $t(11) = 0.086, p > 0.9$. Similarly, in Experiment 2, RTs were significantly faster in the high-frequency compared to the low- and medium-frequency blocks [$t(11) = 7.72, p < 0.001$, and, respectively, $t(11) = 3.66, p < 0.01$; Bonferroni-corrected p-values], and they were also significantly faster in the medium- compared to the low-frequency block [$t(11) = 4.06, p < 0.01$; Bonferroni-corrected p-value].

In addition, in Experiment 1, RTs were faster for the target-present than for target-absent trials, $F(1, 11) = 5.94, p < 0.05, \eta_p^2 = 0.35$, consistent with the visual search literature. In contrast, there was no difference between color- and orientation-defined target trials in Experiment 2, $F(1, 11) = 0.45, p = 0.52, BF_{10} = 0.25$. Interestingly, there was no interaction between target condition and frequency in either Experiment 1 or 2 [$F(2, 22) = 2.44, p = 0.11, BF = 0.38$, and, respectively, $F(2, 22) = 0.87, p = 0.43, BF = 0.26$; Bayes factors compare the model with both main effects and an interaction term to the model with both main effects but no interaction] - suggesting that the effect of the frequency is independent of the target stimuli.

Comparing the error rates depicted in Figure 1 and the mean RTs in Figure 2, error rates tended to be lower in those frequency conditions in which RTs were faster. While this suggests that there were no speed-accuracy trade-offs, it favors the view that participants were adapting to the statistics of stimuli in a way that permitted faster and more accurate responding to the most frequent type of trial within a given block, at the cost of slower and less accurate responding on the less frequent type of trial. A possible explanation of these effects is a shift of the starting point of a drift-diffusion model towards the boundary associated with the response required on the most frequent type of trial; as will be seen below (see modeling section), the shapes of the RT distributions were consistent with this interpretation.

Without the manipulation of frequency, Experiment 3 yielded a standard outcome: all three types of trials had similar mean RTs, $F(2, 22) = 2.15, p = 0.14, BF_{10} = 0.72$. This is different from Experiment 1, in which target-absent RTs were significantly slower that target-present RTs. This difference likely occurred because the target dimension was kept constant within short mini-blocks in Experiment 1, but varied randomly across trials in Experiment 3, yielding a dimension switch cost and therefore slower average RTs on target-present trials (see modeling section for further confirmation of this interpretation).

3.3 Inter-trial effects

As we are interested in inter-trial dynamic changes in response times, we compared trials on which the target condition was switched to trials on which it was repeated from the previous trial. Figure 3 illustrates the inter-trial effects on RTs for all three experiments. Target-repeat trials were significantly faster than target-switch trial in Experiment 1 [$F(1, 11) = 6.13, p < 0.05, \eta_p^2 = 0.48$], Experiment 2 [$F(1, 11) = 71.29, p < 0.001, \eta_p^2 = 0.87$], and Experiment 3 [$F(1, 11) = 32.68, p < 0.001, \eta_p^2 = 0.75$]. This is consistent with trial-wise updating of an internal model (see the modeling section). In addition, we found the target repetition/switch effect to be larger for target-absent responses (i.e., comparing repetition of target absence to a switch from target presence to target absence) compared to target-present responses in Experiment 3 (interaction between inter-trial condition and target condition, $F(1, 11) = 14.80, p < 0.01, \eta_p^2 = 0.57$), while there was no such a difference in Experiment 1, ($F(1, 11) = 2.55, p = 0.14, BF = 0.42$, Bayes factor compares the model with both main effects and an interaction term to the model with both main effects but no interaction). These findings suggest that the target repetition/switch effect is as such stable across experiments, though its magnitude may fluctuate across different conditions. The interaction between target condition and inter-trial condition found in Experiment 3, but not in Experiment 1, might be a consequence of the fact that color and orientation targets were randomly interleaved in Experiment 3 and the target-present repetitions include trials on which the target dimensions may either repeat or change - whereas the target dimension was always repeated in Experiment 1. The effects of repeating/switching the target dimension are considered further below.

Note that in all experiments, we mapped two alternative target conditions to two fixed alternative responses. The repetition and switch effects observed above may be partly due to response repetitions and switches. To further analyze dimension repetition and switch effects when both dimensions were mapped to the same response, we extracted those target-present trials from Experiment 3 on which a target was also present

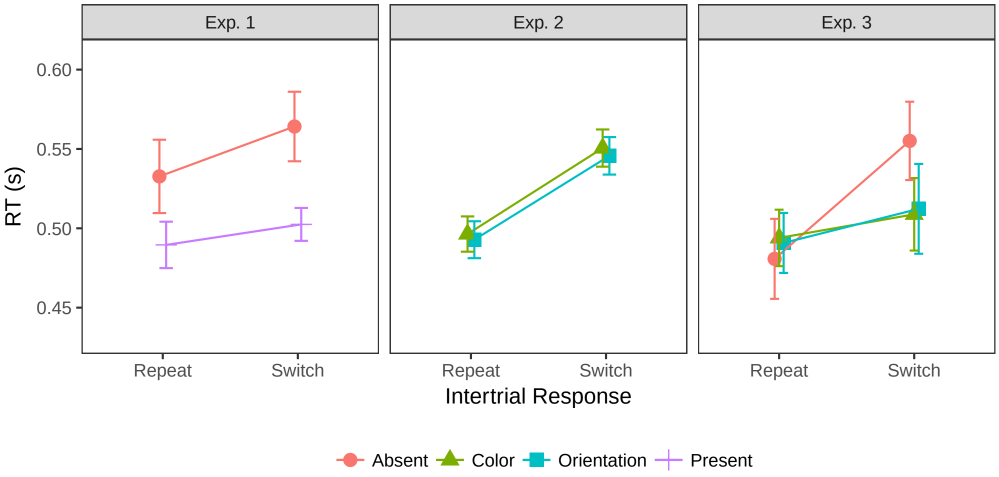

Figure 3: Intertrial effects for all three experiments. Error bars show the standard error of the mean.

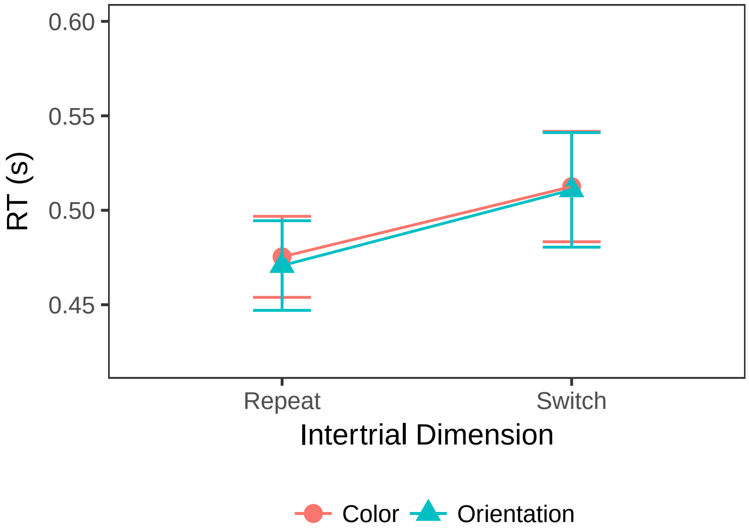

Figure 4: Dimension repetition/switch effect in Experiment 3. RTs were signicantly faster when the target-defining dimension was repeated. Error bars show the standard error of the mean.

on the previous trial. Figure 4 depicts the mean RTs for the dimension-repeat vs. -switch trials. Mean RTs were faster when the target dimension was repeated, compared to when it was switched, $F(1, 11) = 25.06, p < 0.001, \eta_p^2 = 0.70$. There were no differences between the color and orientation dimensions, $F(1, 11) = 0.16, p = 0.69, BF_{10} = 0.30$], and no interaction between the type of dimension and dimension repetition, $F(1, 11) = 0.04, p = 0.84, BF = 0.36$. This pattern is consistent with the prediction of the dimension-weighting account (Müller, Heller, and Ziegler 1995).

In addition to intter-trial effects from reptetion and switching of the target dimension, there may also be effects of repeating/switching the individual features. To address this question, we extracted those trials on which a target was present and where the same target dimension was repeated from the previous trial. Figure 5 shows the mean RTs for feature switch vs. repeat trials. In Experiment 1 and Experiment 3 there was no significant effect of feature repetition/switch [Exp. 1: $F(1, 11) = 0.30, p = 0.593, BF_{10} = 0.284$, Exp. 3: $F(1, 11) = 3.77, p = 0.078, BF_{10} = 0.748$], nor was there any significant interaction with the target dimension [Exp. 1: $F(1, 11) = 2.122, p = 0.17, BF = 0.463$, Exp. 3: $F(1, 11) = 0.007, p = 0.93, BF = 0.364$]. However, in Experiment 2 RTs were significantly faster when the same feature was repeated compared to when the feature changed between trials, $F(1, 11) = 35.535, p < 0.001, \eta_p^2 = 0.764$, and this effect did not interact with

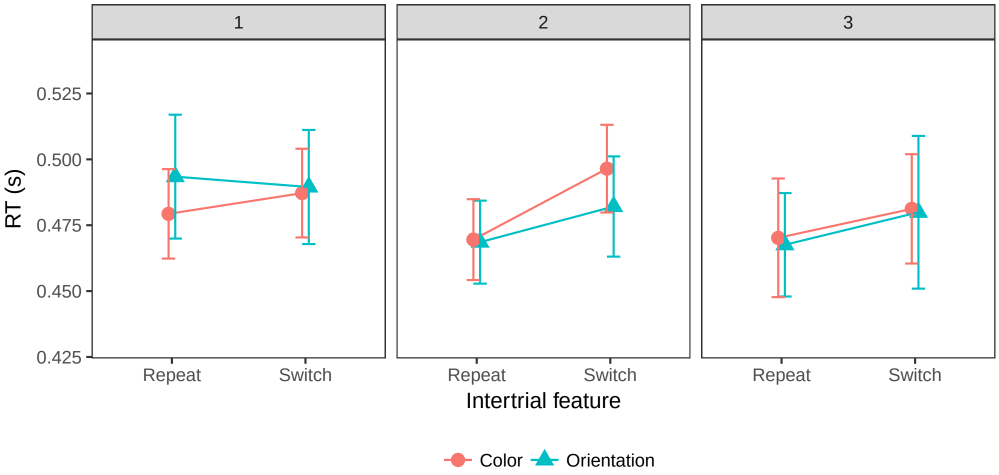

Figure 5: Intertrial effects for feature switch/repetition for all three experiments. Error bars show the standard error of the mean.

the dimension, $F(1,11) = 1.858, p = 0.2, BF = 0.565$.

In this section, we have seen that RTs were faster when target presence/absence or the target dimension was repeated. However, the origin of these inter-trial effects is unclear. The faster RTs after cross-trial repetition could reflect either more efficient stimulus processing (e.g., based on allocating more attention to a repeated stimulus dimension) or response bias (i.e., an inclination to respond based on less evidence). In the next section, we will address this issue by comparing different computational models and determining which parameters are involved in these effects. Because we found feature based inter-trial effects, in only one of the three experiments and they were smaller than inter-trial effects based on either target presence/absence or the target dimension we have chosen to not attempt to model the feature based inter-trial effects.

4 Dynamic Bayesian updating and inter-trial effects

4.1 Factorial comparison of multiple updating models

To identify the origins of the observed inter-trial effects, we systematically compared multiple computational models using the factorial comparison method (Berg, Awh, and Ma 2014). Given that both the DDM and the LATER model provide a good prediction of the RT distributions, we consider the model of RT distributions as one factor (DDM vs. LATER).

Both models have same parameters: the evidence accumulation rate (r), the initial starting point (S_0), and the decision threshold (θ). In addition, the DDM model has one further parameter: non-decision time (T_{er}). Here we also added a non-decision time parameter to the LATER model, and considered the presence vs. absence of a non-decision time as one factor (i.e., non-decision time fixed to zero vs. non-decision time as a free parameter).

One of the main purposes of the model comparison was to investigate through what mechanisms response history and the history of the target dimension influence RTs. To this end, we consider the influence of the history of the response-defining feature (RDF) and the target-defining dimension (TDD) on updating of the parameters of the RT distribution model as two factors. For each factor, we consider six different forms of updating (factor levels).

Level 1 (No update): RDF/TDD repetition/switch does not affect any model parameters.

Level 2 (S_0 with full memory): RDF/TDD repetition/switch updates the initial starting point (S_0) according to the whole prior history. As suggested by Carpenter and Williams (1995) and Gold and Shadlen

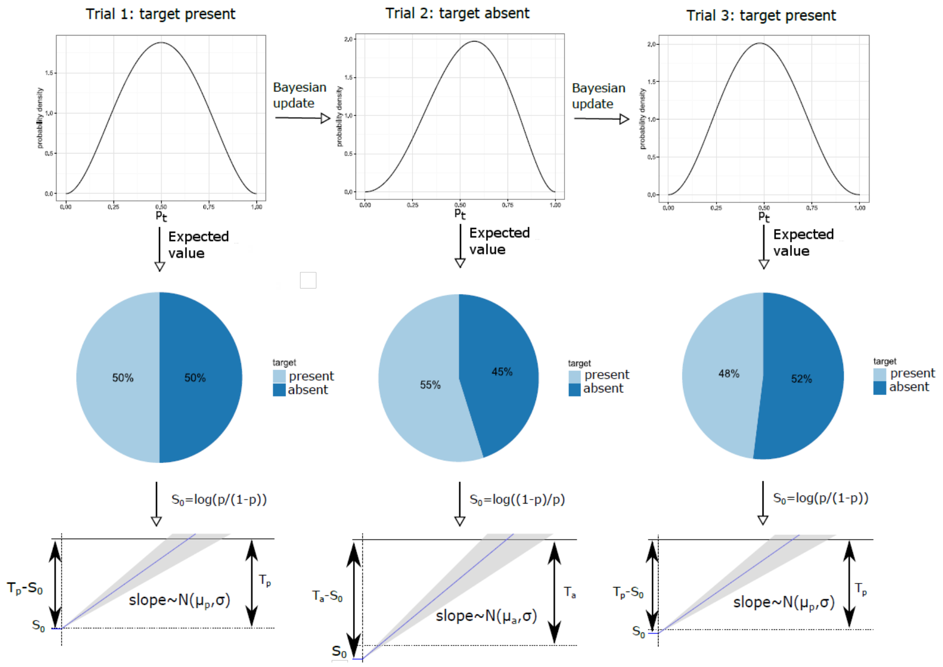

Figure 6: Schematic illustration of prior updating and the resulting changes of the starting point. The top row shows the hyperprior, that is, the probability distribution on the frequency of target present trials (p), and how it changes over three subsequent trials. The middle row shows the current best estimate of the frequency distribution over target-present and -absent trials (i.e., p and $1 - p$). The best estimate of p is defined as the expected value of the hyperprior. The bottom row shows a sketch of the evidence accumulation process where the starting point is set as the log prior odds, for the two response options (target- present vs. -absent), computed based on the current best estimate of p. T_p and T_a are the decision thresholds for the target-present, respectively, the target-absent responses, and mu_p and mu_a the respective drift rates. The sketch of the evidence accumulation process is based on the LATER model rather than the DDM, and is therefore shown with a single boundary (that associated with the correct response). Note that this is not the same boundary on trial 2 (target absent) as on the target-present trials No. 1 and No. 3. In the equivalent figure based on the DDM, there would have been two boundaries, the drift rate on the second trial would have been negative and the starting point on the second trial would have been closer to the upper boundary than on the first trial.

(2007), S_0 is determined by the log prior odds of two decision outcomes (H vs. $\sim H$):

$$\beta = log\frac{P(H)}{1 - P(H)} \tag{1}$$

Here we assume the prior probability $P(H)$, rather than being fixed, is updated trial-wise according to Bayesian inference. Thus, the posterior of the prior is:

$$P(H_t|X_t) \propto P(X_t|H_t)P(H_t) \tag{2}$$

This updating can be modeled by using a Beta distribution as the starting distribution on the prior (a hyperprior) and updating after each trial using the Bernoulli likelihood. We assume that participants were unbiased at the beginning of the experiment (i.e., the two parameters of the Beta distribution initially had the same value β) and gradually updated their prior based on the trial history. Figure 6 illustrates the updating.

For updating based on the RDF, a single prior p is being learned, representing the probability of target-present trials (with the probability of a target-absent trial being $1 - p$). For updating based on the history of the TDD, we assume a separate prior is being learned for each dimension.

This factor level contributes one parameter β to the model.

Level 3 (S_0 with decay): Like Level 2, S_0 was updated based on the history of the RDF/TDD through Bayesian updating of the prior. In addition, we incorporated a forgetting mechanism based on the Dynamic Belief Model (DBM) (Yu and Cohen 2008). That is, in addition to Bayesian updating of the probability distribution on the prior H_t, there was, on each trial, a probability α with which the prior is redrawn from the starting distribution H_0. This forgetting mechanism was implemented through the following equation:

$$P(H_t|X_{t-1}) = \alpha P(H_{t-1}|X_{t-1}) + (1 - \alpha)P(H_0) \tag{3}$$

This model is identical to the fixed no-updating model (level 1) when α equals 0, and is identical to the model specified in Level 2 when α equals 1. For intermediate values of α, the prior is partially reset to the initial prior on each trial. This factor level contributes two parameters α and β to the model.

For factor levels 4-6, it is the evidence accumulation rate (r) rather than the starting point (S_0), that was updated from trial to trial. This updating could be based on either the RDF or TDD (in Experiment 2, these were the same), which we will refer to as the update variable (UV). In each case, UV could have two possible values: u_1 and u_2, which would be either color and orientation or target-present and target-absent, depending on which experiment is being modelled.

Level 4 (Binary rate): The RDF/TDD repetition/switch updates the information accumulation rate r in a step-wise manner. The rate depended only on one-trial-back changes of UV: the rate was scaled by a parameter κ, whose value was either κ_0 ($0<\kappa_0<1$) when the UV changed between trials, or 1 when the UV repeated:

$$r_n = \kappa_0^{UV_n \neq UV_{n-1}} \cdot r \tag{4}$$

When updating was performed based on the dimension, it only affected the rate on target-present trials that were immediately preceded by another target-present trial with a target defined in a different dimension. This factor level contributes one parameter κ to the model.

Levels 5-6 were both designed to reduce the evidence accumulation rate after a UV switch, just like the factor level 4, but allowing for an influence from more than one trial back.

Level 5 (Rate with decay): The RDF/TDD repetition/switch updates the rate r with a memory decay, which was accomplished by reducing the rate whenever the UV switched between trials, but increasing it when the same value of the UV was repeated. Specifically, the rate was scaled by κ on each trial if updating was based on the RDF or on each target-present trial if it was based on the target-defining dimension. The starting value of κ was 1, but it was increased by δ after each UV repetition, and decreased by δ after each UV switch. There was also a forgetting mechanism, like the one used at level 3, such that trials further back had less influence:

$$r_{n+1} = \kappa_n \cdot r_0 \tag{5}$$

$$\kappa_n^u = \kappa_{n-1} + (-1)^{UV_n \neq UV_{n-1}} \cdot \delta \tag{6}$$

$$\kappa_{n+1} = \alpha \cdot \kappa_n^u + (1 - \alpha), \tag{7}$$

where κ_{n+1} determines the amount of scaling of the rate on trial n+1 while κ_n^u is the value of κ after being updated based on the stimulus on trial n. When the updating was based on the target-defining dimension, no increase or decrease by δ was performed on target-absent trials, but the forgetting step was still performed. This factor level contributes two parameters δ and α to the model.

Level 6 (Dimension-weighted rate): The RDF/TDD repetition/switch updates the rate r with a shared weighted resource. Level 6, like level 5, allowed for an influence on the rate from more than one trial back. Like at level 4 and level 5 there was a separate rate used for each value of the UV ($r^{(i)}$ for $UV = u_i, i = \{1, 2\}$). Just like at level 4 and level 5, these rates were scaled based on trial history. However, unlike at level 4 or level 5 the factors by which the two rates were scaled summed to a constant value, as if there was a shared resource. After a trial where either value of the UV had occured some weight was moved to the scaling factor associated with that value of the UV (i.e. the target dimension or the target present/absent status depending on whether the rule was used for TDD or RDF based updating). This updating rule was inspired by the dimension-weighting account (Found and Müller 1996). Specifically, the rate ($r^{(i)}$) was scaled by $\kappa^{(i)}$, where the summation of the scaling factor was kept constant to 2, that is,

$$r_{n+1}^{(i)} = \kappa_n^{(i)} \cdot r_n^{(i)}, i = \{1, 2\} \tag{8}$$

$$\kappa_n^{(1)} + \kappa_n^{(2)} = 2 \tag{9}$$

where the scaling factor $\kappa_n^{(i)}$ updates with the following rules,

$$\kappa_1^{(i)} = 1 \tag{10}$$

$$\kappa_n^{(i)u} = \kappa_{n-1} + (-1)^{UV_n = u_i} \cdot \delta \tag{11}$$

$$\kappa_{n+1}^{(i)} = \alpha \cdot \kappa_n^{(i)u} + (1 - \alpha) \tag{12}$$

That is, after each trial some amount of the limited resource determining the scaling of the rate was moved to the scaling factor associated with the value of the UV that had occurred on that trial. In addition, the same forgetting rule was used as that implemented at Level 5. When the updating was based on the target dimension, no scaling of the rate or updating of κ was performed on target absent trials but the forgetting rule was still applied, just like at level 5.

This level contributes two parameters δ and α to the model.

4.1.1 Models comparison

With the full combination of the four factors, there were 144 (2 x 2 x 6 x 6) models altogether for comparison: non-decision time (with/without), evidence accumulation models (DDM vs. LATER), RDF-based updating (6 factor levels), and TDD-based updating (6 factor levels). We fitted all models to individual participant data across the three experiments, with 12 participants per experiment, yielding 5184 fitted models. Several data sets could not be fitted with the full memory version of the starting point updating level (i.e., level 2) of the dimension-based updating factor, due to the parameter updating to an extreme. Thus, we excluded this level from further comparison.

Experiment 1: Target detection task with variable ratio of target-present vs. -absent trials

Figure 7 shows the mean Akaike Information Criteria (AICs), averaged across all participants, for all models with a non-decision time component in Experiment 1 (in Experiment 1 the task was to detect whether a target was present, the ratio of target present/absent trials was varied between blocks, and the target dimension, color or orientation, changed only between shorter mini-blocks). The AIC is a measure of the quality of a model, which considers both goodness of fit (as measured by the likelihood) and penalizing model complexity. Lower AIC vaulues indicate better model performance. In this figure, as well as in Figures 8 and Figure 9, only models with a non-decision time component have been included since these generally performed better, in terms of AIC. This was particularly the case when the DDM was used for RT distribution modelling, but also, to a lesser extent, with the LATER model. The model with the lowest AIC for each experiment incorporated a non-decision time component, regardless of whether LATER or DDM was used for modelling

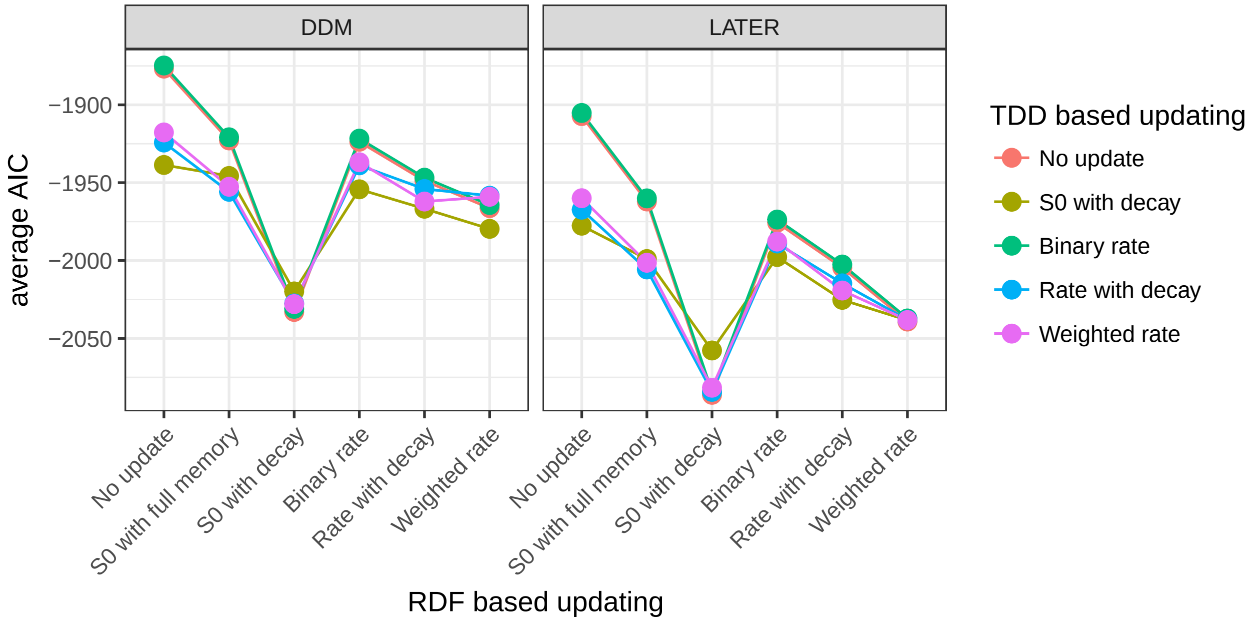

Figure 7: Mean AICs as a function of the tested models in Experiment 1. The response based updating rules are mapped on the x-axis while the dimension based updating rules are mapped to different colors. The left panel contains results for the DDM while the right panel contains results for the LATER model. Only models with a non-decision time component are included in the figure. Models without a non-decision time component generally performed worse, and the best fitting model had a non-decision time component (see also Table 1)

RT distributions. In general, models using LATER for the RT distribution performed better than those using DDM, but the pattern across other factors was very similar. For example, the (other) factor levels of the model with the lowest AIC turned out to be the same whether the DDM or the LATER model was used (see also Supplementary text S1 for figures of the AIC for the models without a non-decision time component). Importantly, for the target/response switch/repetition, the best-fitting model was revealed to be that which updates the initial starting point with partial forgetting. For the dimension switch/repetition, by contrast, the various updating rules yielded comparable results, but no other rule was better than the no-update rule. The latter is unsurprising given that, in Experiment 1, the dimensions were separated in different blocks, that is, effectively there was no dimension switch condition (except for the infrequent changes between blocks).

Experiment 2: Dimension discrimination with variable ratio of color vs. orientation

Figure 8 depicts the mean AICs, averaged across all participants, for all models with a non-decision time component in Experiment 2, in which there was a target present on each trial and the task was to report the dimension of the target, color or orientation, which changed randomly from trial to trial, and the ratio of color target to orientation target trials was varied between blocks. Similar to Experiment 1, models using LATER did overall better than those using DDM, and the best factor level for the target/response-based updating involves updating of the initial starting point with partial forgetting. The best factor level for the updating based on the dimension is updating of the accumulation rate with partial forgetting (i.e., the “rate with decay” level of the dimension-based updating factor).

Experiment 3: Standard pop-out search task with equal target-present vs. -absent trials

Experiment 3 used a standard pop-out search detection task (target-present vs. –absent response), with color and orientation targets (on target-present trials) randomly mixed within blocks. Like Experiments 1 and 2, the LATER model and the response-based updating of the initial starting point outperformed the other model alternatives (see Figure 8). For the dimension switch/repetition, again a form of accumulation rate updating won over the other factor levels. The top two models both involved rate updating, with a slightly

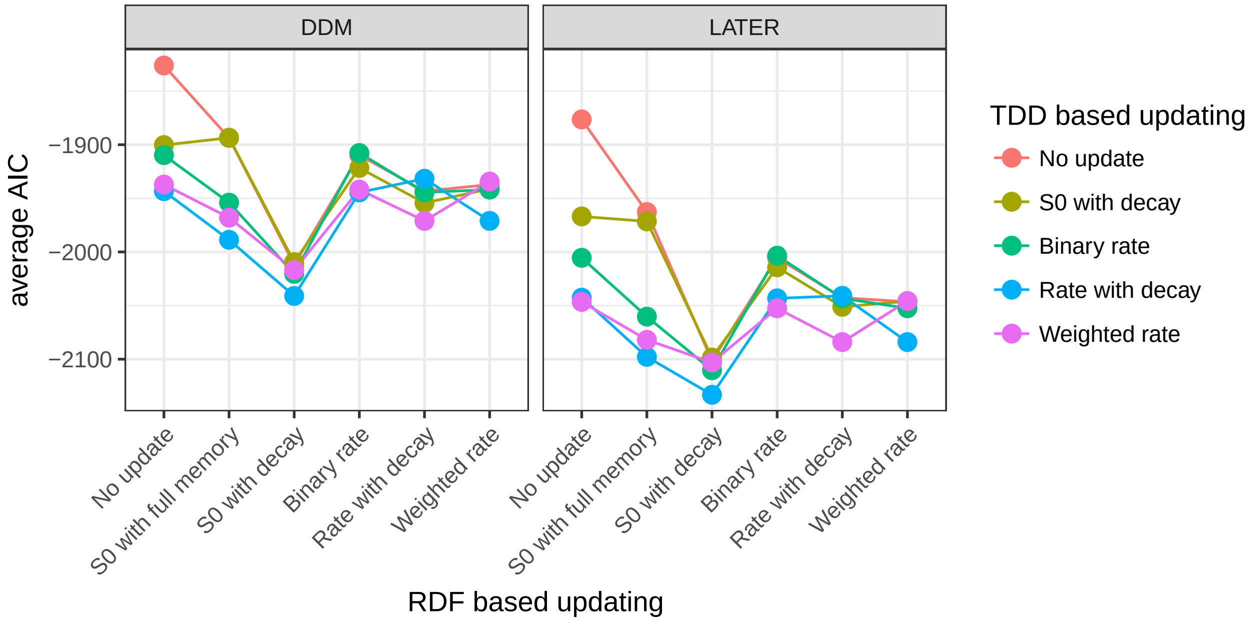

Figure 8: Mean AICs of the tested models for Experiment 2. The response based updating rules are mapped on the x-axis while the dimension based updating rules are mapped to different colors. The left panel contains results for the DDM while the right panel contains results for the LATER model. Only models with a non-decision time component are included in the figure. Models without a non-decision time component generally performed worse, and the best fitting model had a non-decision time component (see also Table 1).

superior AIC score for the model implementing a weighting mechanism with a memory of more than one trial back ('dimension-weight rate') compared to the model in which the rate updating was based only on whether the dimension was repeated/switched compared to the previous trial ('binary rate').

To summarize, for all three experiments the best models, in terms of AIC, were based on LATER rather than the DDM and used updating of the starting point with partial forgetting based on the response. For the two experiments in which color and orientation targets were randomly interleaved within each block, that is, in which dimension switching occured, the best model was that updating of the evidence accumulation rate based on the dimension.

Another way of comparing the models is by picking the best model, in terms of AIC, for each participant and counting how often each level of each factor appears in the resulting list. Table 1 summarizes the results of such an analysis. The table has four sections corresponding to the different factors of our factorial model comparison: RT distribution model, non-decision time, response-based updating and dimension-based updating. For the response-based updating factor, some of the levels are not included in the table as these did not appear in any of the best-fitting models. Almost all the best-fitting models were based on the LATER model, rather than the DDM - consistent with the analyses based on average AIC values above. Also, almost all best-fitting models included a non-decision time parameter. This would be unsurprising if the best-fitting models were based on the DDM; however, almost invariably, the LATER model has previously been used without including a non-decision time component. Our results suggest that adding such a component improves the fit to the data sufficiently to motivate the extra parameter.

For the response-based updating, almost all the best-fitting models were based on updating of the starting point rather than the rate. In addition, most of them were based on updating with forgetting, although for the data from Experiment 2, models that use updating with full memory almost equally often. It is worth noting that in Experiment 2, response repetition/switch exactly coincided with dimension repetition/switch; that is, there is a possibility of "cross-talk" between the two. The better relative performance of the full memory version of the updating rule may have been a consequence of such cross-talk or, alternatively, something

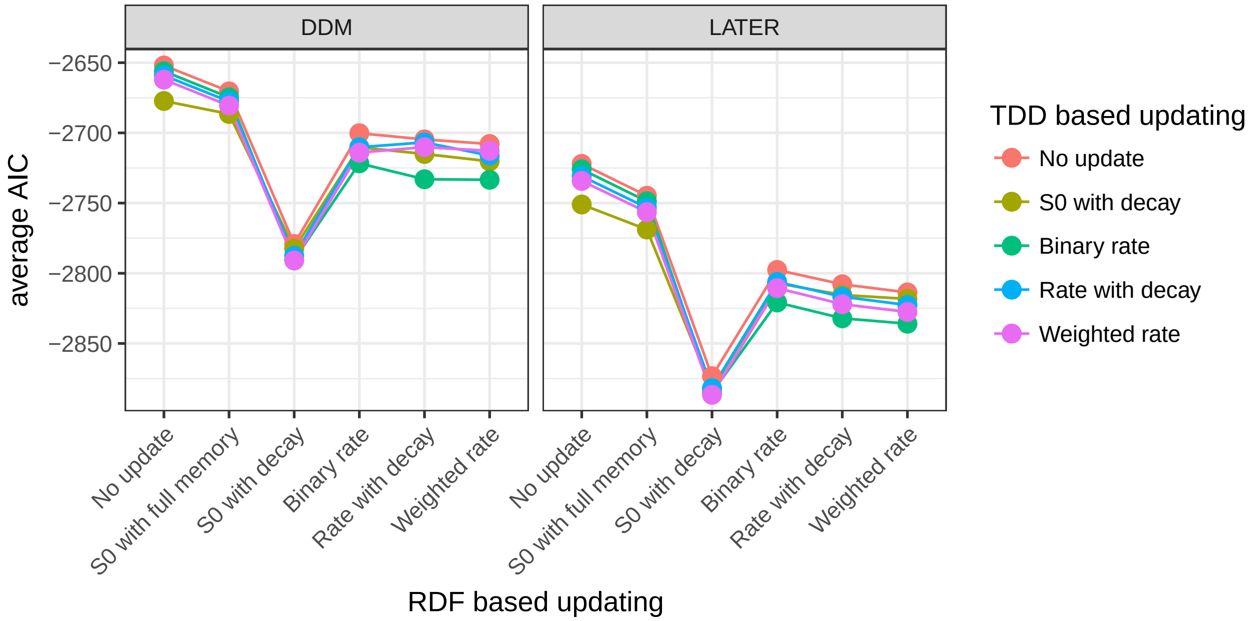

Figure 9: Mean AICs of the tested models for Experiment 3. The response based updating rules are mapped on the x-axis while the dimension based updating rules are mapped to different colors. The left panel contains results for the DDM while the right panel contains results for the LATER model. Only models with a non-decision time component are included in the figure. Models without a non-decision time component generally performed worse, and the best fitting model had a non-decision time component (see also Table 1).

about the explicit dimension (color vs. orientation) discrimination task used in Experiment 2, which may have caused the memory of the response on older trials to decay more slowly, compared to the simple detection (target-present vs. -absent) task used in Experiments 1 and 3.

Finally, for the dimension-based updating, the best-fitting models differed among experiments. In Experiment 1, it was most frequently the case that no form of dimension-based updating improved fits sufficiently to motivate the extra parameter(s). In Experiments 2 and 3, by contrast, the various rate-based updating models most frequently provided the best explanation of the data. Given that the dimension only varied between blocks in Experiment 1, as compared to varying randomly within each block in Experiments 2 and 3, it is little surprising that dimension-based updating played less of a role in modeling the data from Experiment 1. It is less clear why the "Rate with decay" model consistently outperformed the other two rate-based updating models in Experiment 2 while the "Binary rate" and "Dimension-weighted rate" models performed better in Experiment 3. There were two important differences between Experiments 2 and 3. First, the former used a dimension discrimination task, the later a detection task. Also, there were no target-absent trials in Experiment 2, by contrast, 50% of the trials were 'target-absent' in Experiment 3, so that dimension-based updating would affect only half of the trials. This could perhaps explain why it was worth the extra parameter to have a longer memory than a single trial back for the updating in Experiment 2, but not in Experiment 3 (where the "Binary rate" model, with a memory of only a single trial back but one less parameter, most frequently provided the best fit to the data). The better performance of the "Rate with decay", compared to "Dimension-weighted rate"-based version of rate updating, in Experiment 2 is harder to explain with the present design, but is likely related to either the use of a discrimination (rather than a detection) task or the absence of target-absent trials, which requires further investigation.

Overall, the results of this form of model comparison closely matched those of comparing models based on average AIC values. If the factor level that occured most frequently in the list of the best models was selected for each factor, this would result in the same factor levels being selected as in the model with the lowest average AICs, with only one exception: the "Binary rate" level would be selected instead of "Dimension-weighted rate" for the dimension-based updating factor for Experiment 3. Importantly both

Table 1: Model comparison across individual participants

models	exp1	exp2	exp3
RT distribution model: DDM	0	0	0
RT distribution model: LATER	12	12	12
Without non-decision time	1	2	1
With non-decision time	11	10	11
Response: S0 with full memory	1	5	1
Response: S0 with decay	10	7	10
Response: Rate with decay	1	0	1
Dimension: No update	7	0	0
Dimension: S0 with decay	0	0	3
Dimension: Binary rate	0	1	7
Dimension: Rate with decay	2	11	0
Dimension: Dimension-weighted rate	3	0	2

the “Binary rate” and the “Dimension-weighted rate” updating rules are based on updating the evidence accumulation rate, although they differ in that the “Dimension-weighted rate” rule has a memory of more than one trial back.

4.1.2 Prediction of RTs and model parameter changes

To get a better picture of the best model predictions, we plotted predicted vs. measured RTs in Figure 10. Each point represents the average RT over all trials from one ratio condition, one trial condition, and one inter-trial condition in a single participant. There are 144 points each for Experiments 1 and 2 (12 participants x 3 ratios x 2 trial conditions x 2 inter-trial conditions) and 108 for Experiment 3 (12 participants x 3 trial conditions x 3 inter-trial conditions). The predictions were made based on the best model for each experiment, in terms of the average AIC (see Figures 7, 8, and 9). The best linear fit has an r^2 value of 0.85 for the data from Experiment 1, 0.86 for Experiment 2, 0.98 for Experiment 3, and 0.89 for all the data combined.

Figure 11 shows examples of how the starting point (S_0) and rate were updated in the best model (in terms of AIC) for each experiment. For all experiments, the best model used starting point updating based on the response-defining feature (Figure 11A, C, E). In Experiments 1 and 2, the trial samples shown were taken from blocks with an unequal ratio; so the updating results for the starting point are biased towards the (correct) response on the most frequent type of trial (Figure 11A, C). In Experiment 3, the ratio was equal; so, while the starting point has a small bias on most trials (Figure 11E), it is equally often biased towards each response. Since, in a block with unequal ratio the starting point becomes biased towards the most frequent response, the model predicts that the average starting-point boundary separation for each response will be smaller in blocks in which that response is more frequent. This leads to a prediction, namely, that RTs for a stimulus requiring a particular response should decrease with increasing frequency of that stimulus in the block, which is what we observed in our behavioral data. In addition, since after each trial the updating rule moves the starting point towards the boundary associated with the response on that trial, the separation between starting point and boundary will be smaller on trials on which the same response was already required on the previous trial compared to a response switch. This leads to faster predicted RTs when the same response is repeated, matching the pattern in the behavioral data. The forgetting mechanism used in the best models ensures that such intertrial effects will occur even after a long history of previous updates.

In Experiment 1, the best model did not use any updating of the drift rate, but a different rate was used for each dimension and for target absent trials (Figure 11B). In Experiment 2 the best model updated the rate based on the “Rate with decay” rule described above. The rate is increased whenever the same dimension is

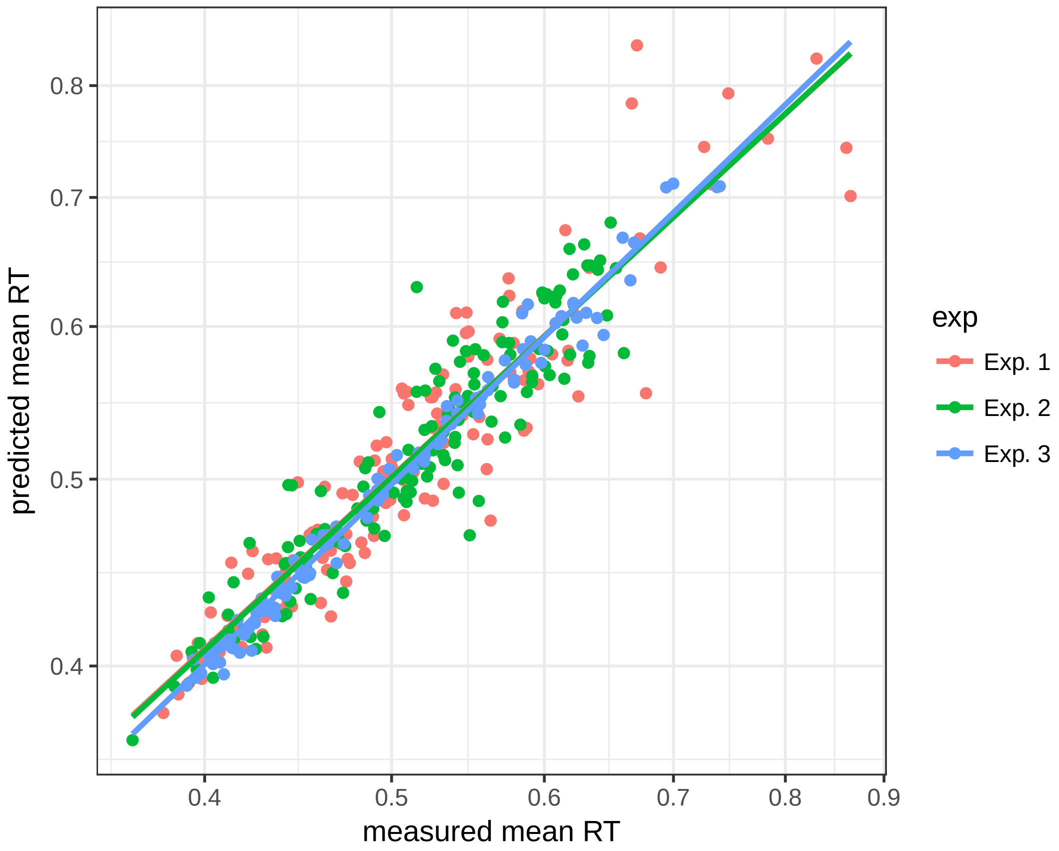

Figure 10: Scatterplot of predicted vs. measured mean RTs for all experiments, participants, ratio conditions and inter-trial conditions. Lines are correspondent linear fitting.

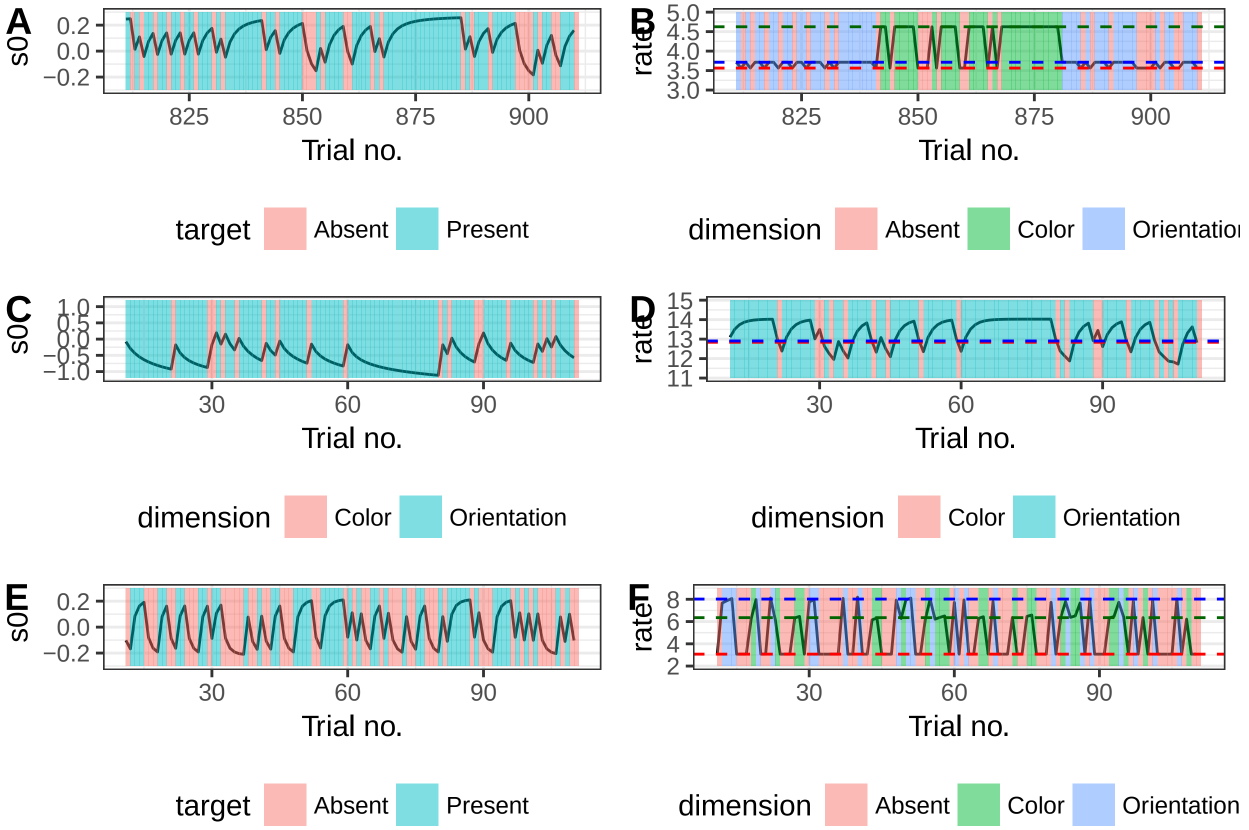

Figure 11: Examples of the updating of the starting point (s0) and rate. Panels A-C show examples of starting point updating for a representative sample of trials from typical participants from Experiments 1-3. Panels E-F show updating of the rate for the same trial samples from the same participants, the dashed lines represent the baseline rates before scaling for target-absent, color target and orientation target trials (i.e., the rate that would be used on every trial of that type if there was no updating). In each case, updating was based on the best model, in terms of average AIC, for that experiment.

repeated and decreased when it switches between trials and these changes could build up across repeated repetitions/switches, but with some memory decay (Figure 11D). Since the target dimension was the response defining feature in Experiment 2, the rate updating would contribute to the "response-based" intertrial effects. In Experiment 3 the best model used the "Dimension weighted rate" rule. Notice that the rate tends to be below the baseline level (dashed lines) after switching from the other dimension but grows larger when the same dimension is repeated (Figure 11F). This leads to a prediction of faster RTs after a repeated dimension compared to a switch, which is what we observed in our behavioral data.

5 Methods

5.1 Experiment 1

5.1.1 Participants

Twelve subjects participated in Experiment 1 (eight females; age range 20 and 33 years). All had normal or corrected-to-normal vision and naive to the purpose of the experiment. All participants gave informed consent prior to the experiment. The study was approved by the LMU Department of Psychology Ethics Committee and conformed to the Helsinki Declaration and Guidelines.

5.1.2 Apparatus and Stimuli

Stimuli were presented on a CRT monitor (screen resolution of 1600 x 1200 pixels; refresh rate 85 Hz; display area of 39x29 cm). Participants were seated at a viewing distance of about 60 cm from the monitor.

Each stimulus display consisted of 39 bars, arranged around three concentric circles (see Figure 12). The distractors were turquoise-colored vertical bars. When a target was present, it was always on the middle circle. Targets were bars that differed from the the distractors in terms of either color or orientation, but never both. Color targets were either green or purple, while orientation targets were tilted 30° clockwise or counterclockwise from the vertical. The search display subtended approximately 7.5° x 7.5° of visual angle and each individual bar had a size of approximately 0.5° x 0.1°.

5.1.3 Procedure

The experiment consisted of 30 blocks of 40 trials, divided into three equally long sections with different proportions of target-present (and, correspondingly, target-absent) trials: 75% [target-absent: 25%], 50% [50%], and 25% [75%]. A text message informed participants about the current proportion of target-present trials at the start of each block. Alternating trial blocks presented exclusively color targets or orientation targets, on target-present trials. The task was to report as quickly and accurately as possible whether a target was present or absent, using the left and right mouse buttons, respectively. Each trial started with the presentation of a fixation dot for 700-900 ms followed by the stimulus display, which was displayed until the participant responded. After the response, there was another 400-600 ms delay before the next trial started with the presentation of the fixation dot, so the total interval from response on one trial to presentation of the search display on the next trial was 1100-1500 ms.

5.2 Experiment 2

5.2.1 Participants

Twelve new participants took part in Experiment 2 (six females; age range 18 and 33 years). All had normal or corrected-to-normal vision and were naive as to the purpose of the experiment. All participants gave

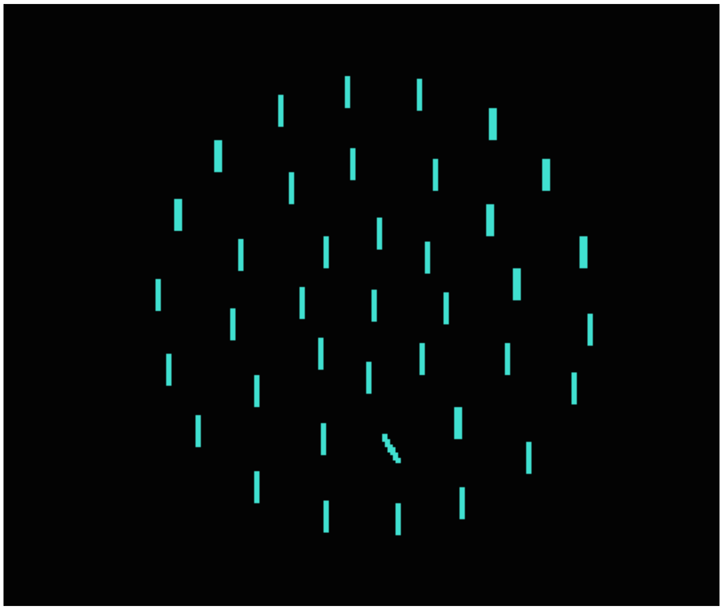

Figure 12: Example of visual search display with an orientation target.

informed consent before the experiment. The study was approved by the LMU Department of Psychology Ethics Committee and conformed to the Helsinki Declaration and Guidelines.

5.2.2 Apparatus and Stimuli

The same equipment and stimuli were used as in Experiment 1.

5.2.3 Procedure

The procedure was the same to Experiment 1, except that instead of reporting whether a target was present or absent, participants had to report whether the target differed from distractors in terms of color or orientation. As in Experiment 1 there were three sections, each consisting of 10 blocks of 40 trials. Unlike in Experiment 1, a target was present on every trial and it was the proportion of color (or, respectively, orientation) targets that differed between the three sections, using the same ratios of 75% [orientation: 25%], 50% [50%], and 25% [75%]. Also unlike in Experiment 1, participants were not informed in advance of what that the proportion of color trials would be in any section of the experiment, nor were they informed that this proportion would differ across the different sections of the experiment.

5.3 Experiment 3

5.3.1 Participants

12 participants took part in Experiment 3 (six females; age range 23 and 33 years). All had normal or corrected-to-normal vision and were naive as to the purpose of the experiment. All participants gave informed consent before the experiment. The study was approved by the LMU Department of Psychology Ethics Committee and conformed to the Helsinki Declaration and Guidelines.

5.3.2 Apparatus and Stimuli

The same equipment and stimuli were used as in Experiment 1.

5.3.3 Procedure

As in Experiment 1, participants had to report on each trial whether a target was present or absent. However, the procedure differed from Experiment 1 in two important ways. First, in Experiment 3, the target-present/absent ratio was fixed at 50% throughout the whole experiment. Second, color targets and orientation targets were interleaved within each block. We used a De Bruijn sequence generator (Brimijoin and O'Neill 2010; Bruijn 1946) to obtain a trial sequence where each of the four possible target types (i.e., purple, green, left-tilted, and right-tilted) were equally often followed by each target type (including itself) and were also equally often followed by a target-absent trial as by a target-present trial. Having such a trial sequence within each block requires 65 trials per block instead of 40 as in Experiments 1 and 2.

5.4 Modelling

To find the model that best explained our data, we performed a factorial model comparison. Full descriptions of the four factors and their levels are given in the modelling section. Here we describe the general procedure used for the model fitting, which was the same for all models.

Each model consisted of an evidence accumulation model: either the LATER model or the DDM, and two updating rules, each of which specified how one parameter of the evidence accumulation model should change from trial to trial, based on the stimulus history. There was one such updating rule for the starting point and one for the evidence accumulation rate, and in each case one of the factor levels specified that no updating at all should take place. For the DDM, we used a closed-form approximation (Lee, Fuss, and Navarro 2006), adding a scaling parameter that determined the size of the random component of the drift diffusion model. This was necessary since our rule for updating the starting point made the scale non-arbitrary.

Models were fitted using maximum likelihood, using the R function "constrOptim" to find minimum value of the negative log likelihood. Error trials and outliers were excluded from the calculation of the likelihood, but were included when implementing the updating rules. Outliers were defined as trials with reaction times more than 1.5 interquartile ranges below the mean or longer than 2 seconds.

To make sure we found the best possible fit for each combination of factor levels, we used an inner and an outer optimization process. The inner optimization process was run for each combination of parameters that was tested by the outer optimization process, to find the best possible values of the inner parameters for those values of the outer parameters. The inner parameters were the parameters of the evidence accumulation model itself, except for the non-decision time which was an outer parameter (because one level of one of the factors specified that the non-decision time should be fixed to zero). For the LATER model, the inner parameters were the starting point boundary separation, and the mean and standard deviation of the distribution for the rate. For the DDM, the inner parameters were the starting point boundary separation, the rate, and the scaling parameter. These parameters could differ between target absent trials, as well as between the two different target dimensions, meaning that there were nine inner parameters for Experiments 1 and 3 and six for Experiment 2 (where there were no target absent trials). The outer parameters were the non-decision time (when this wasn't fixed to zero), and 0 to 2 parameters for each updating rule (see the modelling section for details). This means that models could have 0 to 5 outer parameters in total depending on the factor levels.

Berg, Ronald van den, Edward Awh, and Wei Ji Ma. 2014. "Factorial Comparison of Working Memory Models." *Psychological Review* 121 (1): 124–49. doi:10.1037/a0035234.

Brimijoin, W. Owen, and William E. O'Neill. 2010. "Patterned Tone Sequences Reveal Non-Linear Interactions in Auditory Spectrotemporal Receptive Fields in the Inferior Colliculus." *Hearing Research* 267 (0): 96–110.

doi:10.1016/j.heares.2010.04.005.

Bruijn, N. G. de. 1946. "A Combinatorial Problem." *Proc. Akademe van Westeschappen* 49 (2): 758–64. http://ci.nii.ac.jp/naid/10003051015/.

Carpenter, R. H. S., and M. L. L. Williams. 1995. "Neural Computation of Log Likelihood in Control of Saccadic Eye Movements." *Nature* 377 (6544): 59–62. doi:10.1038/377059a0.

Found, Andrew, and Hermann J. Müller. 1996. "Searching for Unknown Feature Targets on More Than One Dimension: Investigating a 'Dimension-Weighting' Account." *Perception & Psychophysics* 58 (1): 88–101. doi:10.3758/BF03205479.

Gold, Joshua I., and Michael N. Shadlen. 2007. "The Neural Basis of Decision Making." *Annual Review of Neuroscience* 30 (1): 535–74. doi:10.1146/annurev.neuro.29.051605.113038.

Lee, Michael D., Ian G. Fuss, and Daniel J. Navarro. 2006. "A Bayesian Approach to Diffusion Models of Decision-Making and Response Time." In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, 809–16. NIPS'06. Cambridge, MA, USA: MIT Press. http://dl.acm.org/citation.cfm?id=2976456.2976558.

Müller, Hermann J., Dieter Heller, and Johannes Ziegler. 1995. "Visual Search for Singleton Feature Targets Within and Across Feature Dimensions." *Perception & Psychophysics* 57 (1): 1–17. doi:10.3758/BF03211845.

Yu, Angela J., and Jonathan D. Cohen. 2008. "Sequential Effects: Superstition or Rational Behavior?" *Advances in Neural Information Processing Systems* 21: 1873–80. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4580342/.