An interference model for visual working memory: Applications to the change detection task

Lin, Hsuan-Yu ; Oberauer, Klaus

DOI: https://doi.org/10.1016/j.cogpsych.2022.101463

An Interference Model for Visual Working Memory: Applications to the Change Detection Task

Hsuan-Yu Lin

University of Bremen and University of Zurich

Klaus Oberauer

University of Zurich

Author Note

Hsuan-Yu Lin and Klaus Oberauer, Department of Psychology, University of Zurich. This research was supported by grants from the Swiss National Science Foundation (project n° 100014_135002 and 100014_179002) to Klaus Oberauer. We are grateful to Wei-Ji Ma for help with adapting the Bayesian decision rule to the single-probe change-detection task. We thank Jasmin Stöckli for collecting the data.

Correspondence should be addressed to Hsuan-Yu Lin, Department of Psychology, General Psychology Unit, University of Bremen, Bibliothekstraße 1, 28359 Bremen, Germany. E-mail: hslin@uni-bremen.de

The data, the analysis, and the modeling code are available on Open Science Framework at https://osf.io/qahtv /

Abstract

Most studies of visual-working memory employ one of two experimental paradigms: change-detection or continuous-stimulus reproduction. In this study, we extended the Interference Model (IM; Oberauer & Lin, 2017), which was designed for continuous reproduction, to the single-probe change-detection task. In continuous reproduction, participants occasionally report the non-target items instead of the target. The presence of non-target response is predicted by the Interference Model, which relies in part on the interference of non-target items to explain the set-size effect. By presenting a probe matching a non-target item, we can investigate the amount of interference from non-target items in change detection. As predicted by the Interference Model, we observed poorer performance in rejecting a probe matching a non-target item compared to a new probe (i.e., a cost due to intrusions from non-targets). We fitted the IM along with the Variable Precision, the Slot-Averaging, and the Neural-Population model to the data from two change-detection experiments. The models were equipped with a Bayesian decision rule based on the one used in Keshvari, van den Berg, & Ma (2013). The Interference Model and the Neural-Population model successfully predicted the set-size effect and the non-target intrusion cost, whereas the Variable Precision (VP) and Slot-Averaging (SA) models failed to predict the intrusion cost at all. Even with additional assumptions enabling VP and SA to produce intrusion costs, the IM still performed better than the competing models quantitatively.

Keywords: Visual-working memory, Interference Model, Change-detection task, Modeling

An Interference Model for Visual Working Memory: Applications to the Change Detection Task

Visual working memory has been studied primarily with two tasks: Change detection (Awh, Barton, & Vogel, 2007; Cowan, Blume, & Saults, 2013; Luck & Vogel, 1997; Rouder et al., 2008) and continuous reproduction (Bays, Catalao, & Husain, 2009; Blake, Cepeda, & Hiris, 1997; van den Berg, Awh, & Ma, 2014; Wilken & Ma, 2004; Zhang & Luck, 2008). In both tasks, a few items (e.g., color patches) are displayed at the beginning of the trial, and participants are asked to remember the items. After a brief delay during which the items disappear from the screen, one of the item locations is probed. In the continuous reproduction task, participants are asked to reproduce the color at the probed location by selecting it on a continuous color circle. For the change-detection task -- specifically, the single-probe change-detection task -- a color is presented in the probed location, and participants are asked to judge whether the color of the probe is the same as the original color presented at the same location.

Whereas there are are several explanatory computational models built for the continuous reproduction task (Bays, 2014; Oberauer & Lin, 2017; Schneegans & Bays, 2017; van den Berg & Ma, 2018; van den Berg, Shin, Chou, George, & Ma, 2012; Zhang & Luck, 2008), the set of explanatory models for the change-detection task is more limited (Donkin, Nosofsky, Gold, & Shiffrin, 2013; Keshvari, van den Berg, & Ma, 2013; Rouder et al., 2008). Several further models for change detection are measurement models that are fit separately to each experimental condition to investigate how model parameters change between conditions (Cowan et al., 2013; Robinson, Benjamin, & Irwin, 2020), but do not aim to explain the experimental effects with a single set of parameter values (see Oberauer & Lin, 2017; for a detailed discussion of the difference between measurement models and explanatory models).

In this study, instead of creating new models specifically for the change-detection task, we adapted some of the existing models designed for the continuous-reproduction task to explain the data from the change-detection task. We chose four models which jointly representing the main theoretical contenders for explaining visual working memory (VWM): (1) The Slot-Averaging (SA; Zhang & Luck, 2008) model represents the class of models assuming a discrete, slot-like capacity limit on the number of objects or chunks that people can hold in VWM. (2) The Variable-Precision (VP; van den Berg et al., 2012) represents the class of models in which a limited resource is assumed, conceptualized as a continuously varying quantity that can be shared among any number of memory objects. The smaller the resource share assigned to a representation of an object, the lower the precision with which it is represented. (3) We used the Interference Model (IM) that we have developed for explaining findings from continuous-reproduction tasks (Oberauer & Lin, 2017) to represent theories according to which the capacity limit emerges from interference between representations maintained in working memory at the same time, rather than being imposed by a limited commodity such as slots or continuous resources. (4) Finally, the Neural Population model (Schneegans & Bays, 2017) combines the assumptions of a limited resource and of interference.

We chose these models because they describe VWM on a relatively abstract level, as a set of interlocking equations, from which the likelihood can be computed in a straightforward manner with little or no simulation, so that it was feasible to fit them with a likelihood-maximization method. This would not be the case for models on a more detailed level of description, such as the binding-pool model (Swan & Wyble, 2014), the neural-network model of

Manohar, Zokaei, Fallon, Vogels, and Husain (2019), and neural attractor models (Panichello, DePasquale, Pillow, & Buschman, 2019; Wei, Wang, & Wang, 2012).

Our investigation builds on one previous study that adapted models initially designed for the continuous-reproduction task to change detection. Keshvari et al. (2013) adapted SA and the VP model to the change-detection task. They compared the SA and VP models with regard to the full-array change-detection task, in which memory for an array of items is compared to a full test array. The full-array change-detection task requires as many comparisons between memory items and test items as there are in the memory set, which entails two disadvantages: The comparison process is complicated and therefore difficult to model, and the memory set size (i.e., the number of items to be remembered) is confounded with the number of comparisons, making it difficult to separate the effects of the two variables. Therefore, we concentrate our efforts on the single-probe change-detection task, in which a single item selected at random from the array is presented at test.

1.1 The Models

The four models included in the present comparison have originally been designed to account for behavior in the continuous-reproduction task. Each model represents a different theory about the nature of VWM capacity. Here we describe each model verbally; the model equations are given in Appendix A.

1.1.1. The Slot-Averaging Model

The Slot-Averaging (SA) model (Zhang & Luck, 2008) builds on the assumption that VWM has a limited number of discrete slots, and a slot can only store one item or chunk. Once all the slots are used, the remaining items in the array will not be remembered at all. In the continuous

reproduction task, if the target is remembered in a slot, the target will be recalled with the precision afforded by one slot. In that case, the distribution of responses is described as a von-Mises distribution with mean = x_i, the true target feature, and precision κ. If the target is not remembered, however, the participant has to guess. In that case, the response distribution is a uniform distribution over all possible responses. Hence, the SA model is a mixture model: It describes the response distribution as a weighted mixture of a von-Mises distribution and a uniform distribution. The mixture weight of the von-Mises component, P_m, is the probability that the target is represented in (at least) one slot.

One additional assumption in the SA model is that an item can be stored in multiple slots if there are free slots available. When recalling the items stored in multiple slots, the participant will retrieve the item from all the slots and respond with the average of the retrieved memory, which increases the precision of recall.

1.1.2. The Variable-Precision Model

The Variable-Precision (VP) model (van den Berg et al., 2012) incorporates the assumption that a continuous resource limits VWM. The resource can be continuously divided and distributed among all array items, and the precision of the memory representation increases when more resource is allocated to an item. The VP model also includes the assumption that the resource is not evenly distributed among all the items, and the amount of resource allocated also varies from trial to trial. Hence, the precision of memory for the target varies from trial to trial, and from item to item. The response distribution is described by a mixture of von-Mises distributions with the target feature as their common mean, and variable precision parameters that are distributed according to a Gamma distribution.

1.1.3. The Interference Model

The Interference Model (IM, Oberauer & Lin, 2017) is based on the assumption that the limit of VWM capacity is caused by interference between remembered items. Although participants can represent an unlimited number of items in working memory, these representations interfere with each other, and the precision of recall decreases as a result. In the IM, each item's content (e.g., color) is bound to its context (e.g., the color's location in the array). The recall process is driven by a retrieval cue that indicates the target context, and re-activates the contents bound to it – this is primarily the target content, but also to some extent non-target contents. The interaction of the context cue with the memory state generates a distribution of activation over all response candidates (e.g., all the colors on the color wheel), and response candidates with higher activation are more likely to be recalled. The activation of the response candidates comes from three sources. The first source is the context-based activation, which arises from using the context information by which the target is identified (e.g., its location given at test) as a cue to retrieve the content bound to it in memory (e.g., the target color). The second source is the context-independent activation, which arises from persistent activation of all items encoded in the current trial. The last source of activation is the background noise, which arises during the encoding and retrieval process, and activates all response candidates uniformly.

In the IM, several mechanisms are responsible for the capacity limit; that is, the decline of the precision of recall as memory set size increases: First, with larger set size, more response candidates receive activation from the context-independent activation of all array items. Second, because the context representations (e.g., location in the array) have limited precision, a contextual retrieval cue activates not only the target item but also other items in nearby contexts

(e.g., spatial neighbors). With increasing set size, it is more likely that several other items are near neighbors of the target on the context dimension. Third, because each item representation includes a constant amount of background noise, the amount of background noise encoded into VWM increases linearly with set size. Finally, an additional assumption in the IM is that one item can be stored in the focus of attention, and that item has higher precision and is resistant to interference from the other items. As set size increases, the chance that the target is the item held in the focus of attention decreases.

1.1.4 The Neural Population Model

In the Neural Population (NP) model (Bays, 2014) working memory is described as a population of neurons, each of which is tuned to one feature (e.g., a color). At test, each neuron is assumed to send spikes at a rate determined by its gain, and a decoding process takes the spikes from all neurons as input to estimate the most likely feature of the target item. The NP model implements the assumption of a limited resource through divisive normalization: The total gain of all neurons in the population is held constant, so that with more items to be represented, each item receives a smaller share of the total gain, resulting in fewer spikes, and therefore noisier neural information for decoding each item's feature. Recent work by Schneegans, Taylor, and Bays (2020) has shown that, when the number of neurons in the population is very large, the neural-population model of Bays (2014) is mathematically equivalent to the VP model. This is helpful because the VP model is mathematically more tractable, and therefore much easier to fit to data, than the original NP model.

A more recent extension of the NP model (Schneegans & Bays, 2017) includes memory for bindings between features and locations: Each neuron is assumed to be tuned simultaneously

to one feature and one location. The decoder uses the spike train to estimate the location of each item, chooses the item with the location closest to the target location as the target, and reads out spikes from that item to estimate its feature. Building on the equivalence between the original NP model of Bays (2014) and the VP model, we propose that the extended NP model (Schneegans & Bays, 2017) can be represented by an extension of the VP model to include content-context bindings. We have introduced such a model – the VP-Binding model – as part of our investigation of the continuous-response paradigm (Oberauer & Lin, 2017). Like in the IM, in the VP-Binding model arrays are represented as color-location bindings with limited precision on two dimensions, the feature dimension (color) and the context dimension (location). The precision of both the content and the context are limited by the resource allocated to each item. At test, the location of the probe is used as a retrieval cue to retrieve the array item bound to it. At this point, a non-target item could be retrieved instead of the target item due to the limited precision of the context (i.e., location). After retrieving an array item, its feature is recalled with the precision of the content. Here we use our earlier VP-Binding model to represent the NP model with content-context bindings (Schneegans & Bays, 2017).

1.1.5. Memory for Features and for Bindings

Whereas the SA and the VP model characterize the limited precision of features as a function of memory set size, the IM and the NP models also incorporate assumptions about the limited precision of bindings between features (i.e., the to-be-remembered contents) and their contexts. As a consequence, information from non-targets in the array is predicted to be retrieved with an above-chance probability. This enables the IM and the NP models to correctly predict the prevalence of non-target related responses, sometimes referred to as *binding errors* or *swap*

errors, because they appear like mis-bindings, or swaps of the target with a non-target. In the continuous-reproduction task, binding errors appear as the above-chance prevalence of responses close to the feature of a non-target (Bays et al., 2009; van den Berg et al., 2014). In the change-detection task such errors appear as false alarms to probes that match one of the non-targets. These so-called *intrusion probes* are harder to identify as change probes than *new probes* not matching any array item (Donkin, Tran, & Le Pelley, 2015; Rerko, Oberauer, & Lin, 2014).

Binding errors are of theoretical interest because they represent a highly general phenomenon in working memory and attention. A common observation is that such confusions increase as a function of the temporal or spatial proximity between the target and the non-target it is confused with. For instance, transposition errors in serial recall tend to be more frequent the closer the transposed items are in the list (Henson, Norris, Page, & Baddeley, 1996; Lee & Estes, 1977; Murdock & vom Saal, 1967), and transpositions between items in working-memory tests of spatial arrays are more likely to occur between spatially close items (Bays, 2016; Rerko et al., 2014). In the visual-attention literature, illusory conjunctions of features from different, simultaneously presented objects occur more often the closer the objects are in space (Ashby, Prinzmetal, Ivry, & Maddox, 1996). Models explaining these kinds of errors in memory (Brown, Preece, & Hulme, 2000; Burgess & Hitch, 1999; Lewandowsky & Farrell, 2008) and attention (Ashby et al., 1996; Logan, 1996, 2020) all build on the same idea: Items are accessed through their temporal or spatial context, and the context is represented with limited precision, resulting in partial overlap of the contexts of different items; that overlap increases with the items' proximity in context. This idea has also been incorporated into the IM and the NP model.

To gauge the importance of accounting for binding errors for the relative fit of competing models, in one previous comparison of models for continuous-reproduction data van den Berg et al. (2014) considered versions of the SA and the VP model in which the possibility of swap errors was added ad hoc: In these model versions, which we refer to here as SA-swap and VP-swap, respectively, van den Berg and colleagues assumed that swap errors occur with a certain probability, which increases linearly with set size. We characterize this addition as ad hoc because – different from the IM and the NP model – these two models do not offer a causal explanation for why swap errors occur. In addition, the assumption that swap errors increase linearly with set size is implausible because it implies that with a sufficiently large set size the swap probability exceeds one. We nevertheless fit these models to the data because they have been introduced by proponents of these models as versions to account for swap errors, and because the model comparison of van den Berg et al. (2014) has shown that models predicting swap errors fit continuous-reproduction data much better than models not predicting them.

Finally, we also considered a version of the SA model in which swap errors reflect informed guesses, as Pratte (2019) has suggested. We built this assumption into a version of the SA model referred to as SA-NTguess: On a certain proportion of guessing trials (i.e., trials in which the target is not in a slot), the color of one of the non-targets is retrieved, rather than guessing any color at random.

1.2 The Decision Rule for Change Detection

The main difference between the continuous reproduction task and the change-detection task is that in the latter, the information retrieved from memory has to be compared to the probe, and the person has to make a decision as to whether to respond "same" or "change." To adapt

the models designed for the continuous reproduction task to the single-probe change detection task, we therefore need to model this comparison and decision step. To this end, we implemented the Bayesian decision rule as pioneered by Keshvari et al. (2013). We chose this decision for three reasons. First, it is derived in a principled manner from explicit and reasonable assumptions about the knowledge that the decision process brings to bear on making a decision. Second, whereas other decision models for change detection – notably, signal-detection and threshold models – are tied to particular theories about memory (Robinson et al., 2020) – the Bayesian decision rule is neutral with respect to the memory theory it is combined with. Third, the Bayesian decision rule requires few – and in its unbiased version, no – additional parameters, so that it provides only minimal additional flexibility to the models. As such, we use it as a neutral common ground for our model comparison.

1.2.1 A Bayesian Decision Rule

The Bayesian decision rule starts from the assumption that participants first try to retrieve the target feature (e.g., color) at the probed context (e.g., location) in the same way as in the continuous-reproduction task. Because in continuous-reproduction tasks, the to-be-remembered features are usually drawn from a circular dimension (e.g., orientations, or colors on a color circle), and the response options are on a circular scale, we develop the decision rule for features on a circular dimension (as in Keshvari et al., 2013). Whereas models of the continuous-reproduction task can take the retrieved feature directly as the prediction of the chosen response (bar some motor noise), a model of the change detection task needs to describe how the person compares the retrieved feature to the probe feature. According to the Bayesian decision rule, the person evaluates the probability that the retrieved color and the probe came from the *same* condition,

and the probability that both came from the *change* condition. The response is given according to the option with the higher posterior probability.

In the models, we compute a decision variable *d*, as the log posterior odds ratio:

$$d_{\varphi}(x) = \log\left[\frac{P(\text{change}|x,\varphi)}{P(\text{same}|x,\varphi)}\right], \tag{1}$$

where the x is the retrieved color and φ is the probe color. If the log posterior odds of the two conditions, $d_{\varphi}(x)$, is larger than zero, the response will be *change*; otherwise, the response will be *same*. With the help of Bayes' theorem, Equation 1 can be rewritten as:

$$\begin{aligned} d_{\varphi}(x) &= \log\left[\frac{P(x,\varphi|\text{change})P(\text{change})}{P(x,\varphi|\text{same})P(\text{same})}\right] \\ &= \log\left[\frac{P(x,\varphi|\text{change})}{P(x,\varphi|\text{same})}\right] + \log\left[\frac{P(\text{change})}{P(\text{same})}\right] \end{aligned} \tag{2}$$

The values of $P(\text{change})$ and $P(\text{same})$ represent the person's prior probabilities of change trials and of same trials, respectively. When they are identical, they cancel out in the equation, and the decision variable is the log-likelihood ratio of the available data (i.e., the retrieved feature and the probe feature) on the hypothesis of a *change* trial relative to the hypothesis of a *same* trial.

Next, we need to determine $P(x,\varphi|\text{change})$ and $P(x,\varphi|\text{same})$. A first step is to decompose the joint conditional probabilities of the retrieved feature *x* and the probe φ:

$$\begin{aligned} P(x,\varphi|\text{change}) &= P(x|\text{change},\varphi)P(\varphi|\text{change}). \\ P(x,\varphi|\text{same}) &= P(x|\text{same},\varphi)P(\varphi|\text{same}). \end{aligned} \tag{3}$$

For $P(x,\varphi|\text{change})$, the probe is chosen at random from a uniform distribution on the circle, so that every possible probe has probability density $P(\varphi|\text{change}) = \frac{1}{2\pi}$. The retrieved feature *x* comes from a distribution centered around the target feature, and the target feature is

also selected at random from a uniform distribution independently of the probe. Therefore, the probability distribution of the retrieved feature is again a uniform distribution over all possible feature values on the circle. The probability of retrieving x in the *change* condition is

$$P(x|\text{change},\varphi) = \frac{1}{2\pi}. \quad (4)$$

Taking Equation 3 and Equation 4 together, we obtain

$$P(x,\varphi|\text{change}) = \left(\frac{1}{2\pi}\right)^2.$$

As for the $P(x|\text{same},\varphi)$, because the probe condition is *same*, the probe feature is identical to the target feature. As the target feature is drawn at random from a uniform distribution, $P(\varphi|\text{same}) = \frac{1}{2\pi}$. $P(x|\text{same},\varphi)$ is the probability of retrieving feature x, given φ is the target feature. This probability depends on the model in question: Every model of continuous reproduction predicts the probability distribution of retrieving feature x, given memory array **X**, based on two sets of information: the model parameters, and some additional information about the current trial, as we discuss in the next section.

Regardless of the model, the retrieval probability of feature *x* in the *same* condition can be generally expressed as a weighted mixture of a von-Mises distribution centered on the target, φ, and a uniform distribution.[1] The models differ in how they determine the mixture weight P_s.

$$P(x|\text{same},\varphi) = P_s \cdot vonMises(x|\varphi,\kappa) + (1-P_s)\cdot\frac{1}{2\pi}. \quad (5)$$

Taking Equation 5 together with Equation 3, we obtain

[1] The one exception to this expression is the VP model and the NP (=VP-Binding) model, combined with the assumption that the decision process does not know the precision of the target on each trial. As explained in Appendix A, for these models the von-Mises distribution is replaced by a mixture of von-Mises distributions with different precision parameters.

$$P(x, \varphi|\text{same}) = \left[P_s \cdot vonMises(x|\varphi, \kappa) + (1 - P_s) \cdot \frac{1}{2\pi}\right]\frac{1}{2\pi}. \quad (6)$$

Here, P_s is the probability of retrieving the target item, and κ is the precision of retrieval. The precision of retrieval, κ, is a parameter in all four models we compare.

We can now write the decision variable from Equation (2) as the log odds ratio of the two likelihoods, plus the log odds ratio of the priors:

$$d_\varphi(x) = \log\left[\frac{\frac{1}{2\pi}}{P_s \cdot vonMises(x|\varphi, \kappa) + (1 - P_s) \cdot \frac{1}{2\pi}}\right] + \log\left[\frac{p(change)}{p(same)}\right] \quad (7)$$

We can reasonably assume that participants learn the distribution of probe types (i.e., the proportion of *same* and of *change* probes) quickly during an experiment, and adapt their priors to them. Most change-detection experiments have 50% *same* trials. If we assume that people's priors approximately reflect that proportion – either because they learn it, or because it is their prior already before experiencing the experiment – then the priors cancel out in Equations 2 and 7. In the following we will consider two versions of the Bayesian decision rule. One is the simplified version assuming equal priors, and the other is the full version including the prior odds as a free parameter, which could reflect people's experience with unequal proportions of probe types, or their response biases in favor of one or the other response.

To clarify the role of the prior ratio as a criterion or bias parameter – akin to the criterion parameter in signal-detection theory – it is convenient to define the log-likelihood ratio *LRR* as the signal from memory that is compared to a criterion β:

$$LLR_\varphi(x) = \log\left[\frac{\frac{1}{2\pi}}{P_s \cdot vonMises(x|\varphi, \kappa) + (1 - P_s) \cdot \frac{1}{2\pi}}\right] \quad (8)$$

$$\beta = -\log\left(\frac{p(change)}{p(same)}\right) \tag{9}$$

To determine the probability of a "change" response, we compute the probability that $LLR_{\varphi}(x)$ exceeds the criterion β. We do this for all possible values of x, weighted with the probability that x is actually retrieved, $P_{retrieve}(x)$, as given by the memory model, and integrate over all possible retrieved features x:

$$P(\text{change}|\varphi, \boldsymbol{X}) = \int P[LLR_{\varphi}(x) > \beta] \cdot P_{\text{retrieve}}(x)dx, \tag{10}$$

where the $P(\text{change}|\varphi, \boldsymbol{X})$ is the probability of making a "change" response when given the array of memory items $\boldsymbol{X}$ and the probe φ. Note that in Equation (10) the memory model plays two roles: One is as a model of the operation of memory itself, determining $P_{retrieve}(x)$, and the other is as a model of the meta-knowledge about memory used by the decision process, determining $LLR_{\varphi}(x)$. Because we assume the meta-knowledge about the mechanisms and parameters of memory to be approximately accurate, we use the same model in both roles. In the following section we consider different variants of the decision rule that use different levels of knowledge about the parameters of memory.

In the simplified version of the decision rule, assuming equal priors (i.e., β = 0), the decision variable $d_{\varphi}(x)$ is equal to the $LLR_{\varphi}(x)$. It turns out that in this simplified version, P_s can be eliminated from the decision equation, as explained in Appendix B. One advantage of using the Bayesian decision rule is that, as long as a model provides $P_{\text{retrieve}}(x)$ and κ, the model can be adapted to the single-probe change-detection task without an additional parameter (with the simple decision rule assuming equal priors), or with a single parameter for the prior odds. All the

models under consideration here predict $P_{\text{retrieve}}(x)$ because this is the likelihood of responding *x* in the continuous-reproduction paradigm.

1.2.2 Knowledge-Rich and Knowledge-Limited Versions of the Decision Rule

The Bayesian decision rule is derived from knowledge about the task and the person's memory that we assume the decision-making process to have access to. This involves knowledge of (1) the true memory model, (2) the model's parameter values, and (3) knowledge of additional information about each individual item in each individual trial, depending on the model: In the case of the SA model, the additional information is whether on the current trial the target is in memory (so that the probability distribution is a von-Mises distribution centered on the target feature) or is not in memory (so that the person has to guess, and the probability distribution is uniform). In the case of the VP model, the additional information is the precision κ of the target representation on the given trial, which varies randomly from trial to trial. In the NP model, it is knowledge about the precision of the feature and the context representation of every array item in the given trial. In the case of the IM, the additional information consists of whether or not the target is in the focus of attention.

It is unrealistic to assume that participants always have, and use, complete and accurate knowledge of the memory process that generated the retrieved feature *x*. When using the Bayesian decision rule as a psychological hypothesis about how people decide, we need to consider which knowledge they use to estimate the *subjective* probability of retrieving feature x, given array **X**. In other words, we assume that people's decisions are well described by a Bayesian rule that uses the available knowledge and the data (i.e., the probe feature and the retrieved

target feature) in a rational way, but this knowledge is not necessarily complete and accurate in every regard. Therefore, we need to consider what knowledge to attribute to the decision process.

We can reasonably assume that the decision process knows the correct memory model, because that is a constant in the person's life. Likewise, we can reasonably assume a good estimate of the model parameters because these are person parameters that are relatively constant over time, and certainly constant across all trials in an experiment, so the person has ample opportunity to learn about them. It is not clear, however, whether the decision-making process has access to the additional information about each item in each specific trial. For instance, for the SA model, we must ask whether the decision-making process should be assumed to know whether, on a given trial, the target feature is represented in a slot, or whether it should only have knowledge about the probability of having the target in a slot for a given experimental condition. For the VP model and the NP model, we need to ask whether the decision-making process should be assumed to know the precisions with which array items are represented on the current trial, or only know the average precision across trials for a given memory set size. For the IM, we need to ask whether the decision process should be assumed to know whether the target is in the focus of attention.

Nosofsky and Donkin (2016) discuss these questions, distinguishing between *knowledge-limited* and *knowledge-rich* versions of recognition models, and they argue that knowledge-rich model versions, which are imbued with knowledge about each individual trial, strain psychological plausibility. Here we do not want to make a priori decisions on how much knowledge should be attributed to the decision process, and therefore we compared all the models with different levels of knowledge in the decision rule.

To summarize, we used the Bayesian decision rule to make predictions for change-detection behavior from the four core models (SA, VP, IM, and NP) with different levels of knowledge in the decision rule.[2] The details of the models and how the decision rule was applied to the models are explained in Appendix A. We also compared models with a free bias parameter β to models with the bias parameter fixed to 0. We tested the models' predictions with two experiments described next.

2. Experiments A and B

The goal of Experiment A and B was to acquire the data for model comparison using the single-probe change-detection task, varying memory set size from 1 to 6. The simplified decision rule (with equal priors) is based on the assumption that 50% of the probes are exact matches, and the other 50% are drawn at random from a uniform distribution on the circle. In Experiment A, we implemented exactly that sampling scheme, that is, 50% of the probes were identical to the target color, and the remaining 50% of the probes were selected randomly from any other color.

This probe distribution has two potential drawbacks. One is that some of the *change* trials will have probes very similar to the target so that participants are very likely to judge them as "same." As a consequence, they might make more than 50% "same" responses, which could lead them to expect that there are more *same* than *change* trials. Hence, although the objective prior probability of both trial types is identical, the subjective one would not be. The second drawback

[2] For the extended models (SA-Swap, SA-NTguess, and VP-Swap) we only used the level of knowledge that enabled the corresponding core models (SA and VP, respectively) to fit the data best.

is that sampling *change* probes from a uniform distribution results in very few probes that match a non-target item in the array.

To mitigate these two shortcomings of the probe sampling scheme of Experiment A, we ran Experiment B, in which the probes from the *same* condition (50%) were sampled from a region of feature space around the target color for which we expect participants to perceive them as indistinguishable from the target. Also, the probes from the *change* condition were subdivided into two categories: *Intrusion probes* (25%) were sampled from regions around the non-target colors, whereas *new probes* (25%) were sampled from the colors that are not in any region around the target or the non-target colors.

2.1 Method

2.1.1 Participants

Experiment A and Experiment B both included 20 participants recruited from the student population of the University of Zurich. Participants had a normal or corrected-to-normal vision and no color blindness. Participants were rewarded with course credits or 60 Swiss Francs after completing the Experiment A, and 45 Swiss Francs after completing Experiment B.

2.1.2 Material

Items in the memory arrays were color patches selected from a color wheel, which was created in CIE L*a*b* color space with a radius of 60 and centered at luminance set to 70, *a* set to 20, and *b* set to 38. The minimum distance between selected colors was 1 degree, so there were 360 distinct colors. The stimuli were displayed in RGB values with Gamma correction for IEC 61966-2-1 standard on BenQ screens (Model BNQ7F3F, size 24.0 inches, resolution 1920 x 1080, refresh frequency 60 Hz).

2.1.3 Procedure

Experiment A consisted of four identical sessions, and Experiment B consisted of three identical sessions. All the sessions were conducted on different days, and each session took about one hour to complete. Experiment A consisted of 2640 trials, and Experiment B consisted of 1260 trials. The procedure was identical for both experiments.

Participants were instructed to decide whether the probe was identical to the target, and not to compare it to the non-targets (i.e., array items in other locations). They were informed that the color difference between the probe and the target could sometimes be very small. They were not informed further about the probe distributions in Experiments A and B.

At the beginning of each trial, between 1 and 6 color patches were displayed simultaneously on the screen for 500 ms (Experiment A) or 100 ms (Experiment B). This was followed by a blank screen for 500 ms (Experiment A) or 1000 ms (Experiment B). The locations of the color patches were randomly selected from 13 possible locations equidistantly distributed on an invisible circle which was centered on the center of the screen with 5 degrees visual angle when viewed from 70cm away from the screen. After the blank screen, a probe was displayed in one of the stimulus locations, and empty frames were displayed in the remaining locations that had been occupied by a stimulus in the current array. Participants were asked to judge if the probe was the same color as the color patch presented previously at the same location by pressing the left mouse button for “same”, or the right mouse button for “change.” After participants made their response, a blank screen appeared for 500 ms and was followed by the beginning of the next trial.

In Experiment A, the probe matched the target color exactly in 50% of the trials. For the remaining 50% of the trial, the probe was selected randomly with equal probability from all possible colors other than the target color. In Experiment B, the probe color was selected within a boundary around the target color for 50% of the trials, which were defined as *same* trials. In 25% of the trials, the probe color was selected from within regions defined by boundaries around the non-target colors (*intrusion* trials). For the remainder of the trials, the probe color was selected from the remaining colors that were not within the boundaries of the target color and non-target colors (*new* trials). The only exception was in set size one because there was no non-target item, so all 50% of change trials were new trials. The boundaries for selecting the probe color around the target and non-target colors were chosen on the basis of error distributions from continuous-reproduction tasks using the same materials (Oberauer & Lin, 2017) so that they covered the range of colors that participants were likely to reproduce when trying to reproduce the selected color. Because the error distributions usually become broader with larger set sizes, these boundaries varied according to the set size of the trial, which was 10, 14, 16, 17, 18, and 18 degrees for set sizes one to six, respectively. The probe sampling schemes for Experiment A and Experiment B are illustrated in Figure 1.

3. Results

3.1. Experimental Effects

To bring the result of Experiment A in line with the result of Experiment B, we relabeled the probes in Experiment A in a similar fashion as Experiment B. The *same* probes stayed the

same[3]. The *change* probes were relabeled into *intrusion probe* or *new probe*, where the former was any probe that was within the boundaries around a non-target color, and the latter was any probe not within the boundaries of any non-target color. The boundaries were the same as the boundaries used to sample probes in Experiment B.

After relabeling the probes, we compared mean accuracy for the three different probe types (same, intrusion, and new) and the six set sizes between experiments with a Bayesian linear model using the BayesFactor package for R (Morey & Rouder, 2015; R_Core_Team, 2020) with default priors and a random intercept. The Bayes factor strongly supports the assumption that there was no difference between Experiments (BF_{01} = 1111.11). Therefore, the following analysis will be conducted with both experiments collapsed together. The Bayes factor showed evidence supporting both the main effects of set size and of probe type (BF_{10} = 5.92e+41 and BF_{10} = 3.23e+18, respectively). Evidence also supported the interaction between set size and probe type (BF_{10} = 1.59e+54). The statistics are summarized in Table 1, and these effects are visualized in the right panel of Figure 2.

3.2 Model Comparison

We fitted the Slot-Averaging Model, the Variable Precision Model, the Interference Model and the Neural Population model to the data of each experiment after embellishing each model

[3] Even when a *change* probe was very similar to the target, the *change* probe was not relabeled into a *same* probe. We also ran the analyses after relabeling the *change* probes within the similarity-region around the target into *same* probes, and the conclusion drawn from Bayesian model comparison did not differ.

with the Bayesian decision rule. The models were implemented in Python 3.6, and the parameters were estimated with the differential evolution algorithm in SciPy (Jones, Oliphant, & Peterson, 2001). Differential evolution is a global optimization algorithm. The starting population of parameters was automatically chosen by the differential-evolution algorithm (Storn & Price, 1997). The model was fit to each participant's data separately. The goodness-of-fit of the models was estimated by the deviance, which is –2*log(Likelihood), averaged over participants. To compensate for the different number of parameters in the models, we relied primarily on the Bayesian Information Criterion (BIC) to compare between models. The BIC adds a penalty to the deviance for the model's free parameters, which is heavier than in the equally popular AIC. Because our IM has more free parameters than the competing models, using the BIC is the more conservative choice as it assigns the competing models a relatively lighter penalty. Nevertheless, we also report model comparisons based on AIC for completeness.

We applied the same Bayesian decision rule to both experiments. One assumption in the Bayesian decision rule is when a *change* occurs, the probe is sampled uniformly from all the possible response candidate. This assumption does not match the sampling scheme of Experiment B. However, the behavior data didn't show a significant difference between the results from Experiments A and B, and participants were not informed about the sampling scheme of Experiment B. Therefore, for both experiments we used the Bayesian rule with the simple assumption that change probes are sampled uniformly[4].

[4] We also tested the Bayesian decision rule with the assumption that the *change* distribution matches the probe sampling scheme of Experiment B. The goodness-of-fit and the estimated parameters did not vary much comparing to the decision rule with uniform sampling assumption.

We did model comparisons on two levels. On a first level, we identified the best-fitting version of each of the four competing models (SA, VP, IM, and NP). On a second level, we compared the best-fitting versions of each model to each other. We next introduce the different model versions of each competing model.

3.2.1 Model Versions with Decision Rules using Different Degrees of Knowledge

As mentioned above, there is some room for interpretation in implementing the Bayesian decision rule (Nosofsky & Donkin, 2016), depending on which knowledge we assume the decision rule to have available about individual trials. We next consider for each model the different levels of knowledge that we could assume the decision process to use.

For the SA model, we initially tested four versions of the decision rule with different levels of knowledge. The first had knowledge of how many slots hold a representation of the target in a given trial. If the target is held in at least one slot, the retrieved color always comes from the target color, that is, $P_s = 1$. The precision of the von-Mises distribution assumed in the decision process depends on the number of slots that hold a representation of the target: Averaging across multiple slots yields higher precision. If the target is not in memory, however, $P(x|\text{same}, \varphi)$ is a uniform distribution, and as such identical to $P(x|\text{change}, \varphi)$. As a consequence, $LLR_{\varphi}(x)$ is 0, and the decision process can only guess. The probability of guessing "change" is the person's prior probability of a change trial, which we can express as a function of the bias parameter by solving Equation (9):

$$p(change) = \frac{\exp(-\beta)}{1 + \exp(-\beta)} \tag{11}$$

Hence, in this model version the bias parameter enters twice, once as the criterion to compare the *LLR* signal to when the target is in memory, and again as determining the guessing probability when it is not.

The second version of the decision rule is like the first, except that it assumes a constant precision κ_1, the precision of a single slot, for all trials regardless of set size. The third and fourth versions of the decision rule do not use knowledge of whether the target is in the memory or not on a given trial, and therefore rely only upon the probability of retrieving the target for computing P_s. The third version uses knowledge of how the precision of available memory information improves by averaging across multiple slots at lower set sizes, and therefore assumes higher κ for smaller set sizes. The fourth version always assumes precision κ_1. The fitting result showed that the two model versions without knowledge of whether the target is in a slot fit better; the third version, assuming a precision parameter adjusted to the trial's set size, performing best (see Table 2).

Upon inspecting the predictions of the four model versions, it became clear that the versions with knowledge of the target's memory state on every trial and the versions without such trial-by-trial knowledge mis-predicted the data in complementary ways (see Figure 3, first vs. second row). We therefore also tried a mixture of the second and the fourth version, assuming that decisions are made with knowledge of the target's memory state with probability p_k, and without knowledge of the target's memory state with probability 1-p_k. This partial-knowledge model reflects the assumption that on a proportion p_k of all trials the person knows whether or not the target is in a slot, and uses that knowledge to decide, whereas on the remaining trials that knowledge is not available (or not used), and the decision instead is based on the probability

that the target is in a slot, given the trial's set size. That fifth version of the SA model turned out to fit better than the first four. Therefore, for the following comparison between competing models, we used the decision rule that has a mixture of knowledge states about whether the target is remembered or not on a given trial, and assumes precision κ_1 for all set sizes.

For the VP we tested two versions of the decision rule. The first one has knowledge of the precision of the target item in the current trial, and the second version only has knowledge of the average precision at the current set size. The fitting result showed that the version which only has knowledge of the average precision performed better, as shown in Figure 5, a result which was in contradiction to the previous study (Keshvari et al., 2013). We selected the VP model with the decision rule which only has knowledge of the mean precision of the current set size.

For the IM, we were unsure if the Bayesian decision rule has knowledge of whether or not the target is in the focus of attention on a given trial. If the target is in the focus of attention, the target would have higher precision, and higher resistance to non-target interference. Having knowledge of whether the target item is in the focus of attention or not affects the assumed precision of retrieval κ in the decision rule. We therefore compared two versions of the Bayesian rule in the context of the IM. The version with a decision rule equipped with knowledge of whether the target is in the focus of attention or not fit the data slightly better (see Table 2). The comparison of the IM with competing models was based on that version.

For the NP model we considered two knowledge versions analogous to the VP model: In one version, the decision rule used knowledge of the precision of every array item's feature and context. In the other version, it only used knowledge of the average precision at a given set size. As for the VP, the NP version with knowledge of only the average precision fit the data better.

3.2.2. Model Versions with Decision Rules Using Equal and Unequal Priors

A second matter we considered pertains to the priors for *same* and *change* trials in the decision rule. If we assume that people use approximately equal priors for these two trial types, the prior ratio drops out of the decision rule. As this is the simplest version of the decision rule, requiring no additional free parameter, we should prefer it for its parsimony if the data warrant that simplification. To test whether that is the case, we compared model versions with equal priors to model versions in which the ratio of the priors was a free parameter.[5] When treated as a free parameter, the prior odds ratio has the effect of a freely estimated decision criterion by which a person's bias in favor of *same* or *change* responses can be implemented, akin to the criterion parameter in signal-detection theory. We therefore refer to this free parameter as the bias parameter. In Appendix A we detail how we included the bias parameter in each model.

We found that for the SA model and the IM, model versions with and without a bias parameter performed about equally well, and the BIC, which penalizes the additional bias parameter more, favored the version without that parameter in both cases. In contrast, the VP model and the NP model fit better with a free bias parameter. For the subsequent comparison of the four models, we considered the bias-free versions of the SA model and the IM, and VP and NP models with bias.

3.2.3. Comparison of the Best Versions of Each Model

Figures 3 to 6 show the fits of the SA, the VP, the IM, and the NP model, respectively. Overall, all four models were able to fit the similarity gradient effect for *change* probes (left

[5] We did this for most combinations of knowledge attributed to the decision process, except for a few versions that were performing most poorly among the bias-free models.

panels), and the set-size effect for the *same* probes (right panels). However, only the IM and the NP model were able to predict the worse performance of the *intrusion* probes compared to the *new* probes, as shown in the right panels of each figure. Quantitatively, IM was the best fitting model for both experiments according to both AIC and BIC. On average, the IM won over the SA, the VP, and the NP model by 38.77, 20.53, and 9.85 BIC points per participant, respectively; see Table 2 for the summary of the goodness of fit statistics. The SA and the VP fitted the data poorly mostly because they failed to accommodate the intrusion cost, and the fitting algorithm had to compromise between reproducing the performance on intrusion probes and new probes.

One reason for why the NP fit worse than the IM is that in the NP model, swap errors arise only from the imprecision of representations along the spatial dimension. As a consequence, the model predicts false alarms to intrusion probes to be highest for probes coming from non-targets spatially close to the target. In the present data, no such spatial gradient on intrusion cost was observed (see Figure 7). The NP model cannot predict intrusion costs to occur independent of the spatial distance between the target and the non-target that the probe is taken from. In contrast, the IM has two routes through which non-target intrusions arise: One is their re-activation by the location cue, which generates a spatial gradient of non-target activation; the other is the context-independent activation of all array features, which produces the non-target activation regardless of the spatial proximity between the target and the non-targets. The IM can accommodate the lack of a spatial gradient on intrusion costs by attributing them largely to the context-independent activation of array features.

Tables 3 to 6 present the best-fitting parameters for the SA, the VP, the NP, and the IM, respectively.

3.2.4. Extended Versions of Slot-Averaging and Variable-Precision Models with Swap Errors

The failure of predicting the intrusion cost is due to the fact that the SA and VP models have no mechanism for generating an elevated chance of retrieving colors similar to the non-target colors. In both models, the non-target features play no role in predicting the distribution of retrieved colors. This results in the same prediction for both intrusion probes and new probes. The IM, in contrast, predicts that non-target features are retrieved more often than new features because non-targets receive activation at retrieval. As a consequence, features matching an intrusion probe are more likely to be retrieved than features matching a new probe. This enables the IM to explain the intrusion cost. Similarly, the NP model assumes that on some trials, a non-target item is retrieved instead of the target item due to the limited precision of context representations. When that happens, the feature of the non-target is retrieved, and if the intrusion probe happens to match that non-target's feature, the model predicts a false alarm, explaining the intrusion cost. To compensate for their lack of ability to predict the intrusion cost, we also fitted variants of the SA and the VP which were able to predict elevated retrieval probabilities for non-targets, namely: SA-Swap, SA-NTguess, and VP-Swap. For these extensions, we made the assumptions concerning knowledge of the decision process that fit best for each model class before the extension (i.e., partial-knowledge with constant precision for SA; knowing only the average precision per set size for VP).

All three extended models were able to capture the general trend in the data, including the intrusion cost, as shown in Figure 8. However, AIC and BIC values – listed in Table 2 – indicated that SA-Swap, SA-NTguess, and VP-Swap still fit the data more poorly than the IM. The SA-Swap model underestimated the memory precision, which lowered the predicted performance on the

new and *intrusion* probes in order to compensate for the slightly decreased performance of the *intrusion* probes. Lower memory precision, however, led to misfit on the *same* and *new* probes in the smaller set sizes. The SA-NTguess model was able to account for the presence of the intrusion cost, but failed to account for its invariance across set sizes: In this model, the probability of guessing any one of the non-targets increases with set size n once $n > K$; it is proportional to $1-(n/K)$. The probability that the guessed non-target happens to be the one matching an intrusion probe, however, decreases with set size according to $1/(n-1)$. The two effects of set size follow different functions, and therefore don't compensate each other. As a result, the SA-NTguess model predicts a decrease of the intrusion cost with set size for set sizes > 3 (see Figure 8). In the VP-Swap model, non-target intrusions arise independently of a non-target's spatial proximity to the target. This model did better than the NP model, in which non-target intrusions are a function of proximity. Still, this model fit slightly worse than the IM.

Within each model class we selected the version that provided the best fit. These best-fitting versions entered a competition for fitting the data of each participant. The last column of Table 2 shows that the IM won the competition for 18 of the 40 participants; the SA-Swap and the VP-Swap model fit best for 6 and 8 participants, respectively; the SA model won the competition for 5, and the VP model for the remaining 3.

4. General Discussion

In this study, we tested four models representing four different theories about visual working memory capacity on the single-probe change detection task. The Interference Model (IM) performed best, closely followed by the Neural Population (NP) model; the Slot-Averaging

(SA) and Variable-Precision (VP) models fared worse, mainly because they failed to account for the intrusion cost observed in the two experiments.

4.1. Non-Target Intrusions in Models of Change Detection

The intrusion cost – an elevated false-alarm rate to probes matching a non-target compared to new probes not matching any array item – is consistently observed in the change-detection task (Donkin et al., 2015; Rerko et al., 2014). The intrusion cost was also observed in verbal materials, for instance, in the local-recognition task (Oberauer, 2008). The intrusion cost in the change-detection task is conceptually related to the above-chance tendency to report features of non-targets in the continuous-reproduction task (Bays et al., 2009): Both can be explained by a tendency to erroneously retrieve a non-target of the current memory set instead of the target. This tendency is also observed in serial-order tests of WM, where transposition errors – reports of list items in the wrong order – are common (Henson et al., 1996; Murdock & vom Saal, 1967).

The SA and the VP models both failed to explain the intrusion cost unless the assumption of swap errors was added ad hoc, as in the SA-Swap and the VP-Swap models (van den Berg et al., 2014). The extension of these two models by a swap-error probability that increases linearly with set size is ad hoc because it does not represent a mechanism that causes swap errors to occur within the assumed memory architecture. One possible mechanism could be that swap errors occur with a certain probability at encoding (Donkin et al., 2015). For instance, in the SA model, each slot could have a constant probability of encoding the feature of one array item paired with the location of another. At test, the probe location is used to find the slot in which that location is stored, and if that location has been paired with another item's feature, then the feature of a

non-target is erroneously retrieved as the target feature. One difficulty for such a model is that the probability of swap errors has to increase with set size to accommodate the observed intrusion cost across set sizes. Donkin et al. (2015) have shown that a slot model with constant swap error probability would predict a decrease of the intrusion cost with increasing set size. This is because not every swap error leads to an intrusion error: For an intrusion error to happen, the target must be swapped with the one non-target that happens to match the intrusion probe. With a constant probability of making any swap error, the probability of swapping the target with one particular non-target decreases with set size. The prediction of decreasing intrusion costs with larger set sizes contradicts the findings in Donkin et al. (2015) and the present study.

With the assumption that the probability of swap error increases linearly with set sizes, the SA-swap model and the VP-swap model can explain the intrusion cost we observed. However, the assumption that swap probability increases with set size is difficult to justify in the context of a slot model, especially when the set size exceeds the capacity limit. For instance, if the capacity limit is three slots, then three array items are stored regardless of whether the set size is three or anything larger than three. If swap errors arise because the target is confused with one of the stored non-targets (either at encoding or at retrieval), the probability of such a confusion should not increase beyond set size three.

Alternatively, in the context of a slot model, swap errors could be explained as strategic guesses. Pratte (2019) has proposed that explanation for the continuous reproduction task: A person not knowing the target feature could intentionally choose the feature of a non-target as response on the assumption that this might be a better guess than choosing any feature at random. Could the increased false-alarm rate to intrusion probes in change detection also be

explained as a form of sophisticated guessing? If the participant did not remember the target color but remembered a non-target color, and that non-target color appeared as a probe in the target location, a rational participant should reject the probe because they remembered that its color had been in a different location in the array. As a consequence, we should observe an intrusion benefit. A more plausible explanation could start from the assumption that colors are occasionally remembered after losing their location information, so that people remember having seen the probe color but have no information about where in the array it was (Cowan et al., 2013). In that case, selecting for retrieval a color that is represented in a slot without any information about its location could be a reasonable guessing strategy, as there is a 1/*n* chance that it is the color of the target. We implemented the sophisticated-guessing explanation in the SA-NTguess model. Including the possibility of guessing a non-target improved the fit of the SA model, but it still gave a worse account of the data than the IM and the NP models.

The IM and the NP model predict swap errors, and their increased prevalence across set sizes, through their core assumptions. In the NP model, swap errors occur because of the limited precision of representations on the context dimension (here: spatial location) to which items are bound. Swap error probability is predicted to increase with set size for two reasons: There are more non-targets in the spatial vicinity of the target; and the precision of context representations declines, rendering the target location less discriminable from non-target locations, so that non-targets become more likely to be retrieved in response to the target location as a cue.

The IM also predicts that the probability of swap errors increases with set sizes. Swap errors arise from two sources. One is the generalization of the spatial context cue to neighboring items in the array according to a spatial gradient that reflects the imprecision of context

representations, similar to the NP model. The other source is the context-independent activation of all array items. The overall amount of the context-independent activation is the same for each non-target regardless of set size. As a result, the probability of each individual non-target feature to be erroneously retrieved remains constant, implying that the overall probability of some swap error increases with set size. The contribution of spatial generalization of the target location to each individual non-target is also independent of set size, because the average distance between the target and any non-target did not change with set size in our experiments. Thereby, the probability of erroneously retrieving the one non-target feature that happens to match the current intrusion probe is predicted to be constant across set sizes. There is, however, one element of the IM that leads to a slight increase of intrusion costs with set size: With larger set size, the probability that the target item is in the focus of attention – and thereby, partially protected from interference from non-targets – decreases.

The two sources of the intrusion cost in the IM can be mapped onto the two sources of memory information – familiarity and recollection – in dual-source models of recognition (for dual-source models of working memory: Göthe & Oberauer, 2008; Oberauer, 2008; for episodic long-term memory: Wixted & Mickes, 2010; Yonelinas, 2002). Familiarity reflects the summed similarity between a recognition probe and all items in memory regardless of its context. The context-independent activation of the features of all array items in the IM generates such a familiarity signal: It increases the chance that any feature in the array, regardless of its location, is retrieved, so that a probe matching any array feature has an increased chance of matching the retrieved feature. This leads to an intrusion cost that is independent of the spatial distance between the target and the non-target matching the intrusion probe. Recollection generally refers

to the retrieval of memory contents together with their contexts, which reflects the strength of content-context bindings in memory. In tasks such as change detection, where the probe is given in a specific context (here: a location in the array), recollection refers to the retrieval of contents that are bound to the context of the probe. In the IM, the context-dependent activation of array features as a function of their spatial proximity to the probe is responsible for the recollection signal. The assumption that the context representation has limited precision implies that recollection can fail, and when it does, it leads primarily to retrieval of spatially close non-targets.

In the present experiments, it turned out that the contribution of context-dependent activation (i.e., recollection) to intrusion errors was minimal: We observed no spatial gradient on the prevalence of intrusion costs. This is the main reason why the NP model, which necessarily predicts such a spatial gradient, fit the present data slightly less well than the IM. In the context of the IM, the absence of a spatial gradient means that the intrusion cost in the present experiments reflects primarily familiarity. This contrasts with other change-detection experiments, in which intrusion costs from contextually close non-targets were more frequent than from far non-targets (see Logan, Cox, Annis, & Lindsey, in press, for verbal stimuli; Rerko et al., 2014, for visual stimuli). In the continuous-reproduction task, there is a similar variability across experiments, with some observing a spatial gradient on non-target intrusion (Bays, 2016; Experiments 2 and 3 in Oberauer & Lin, 2017), and others showing no evidence for it (e.g., Experiment 1 in Oberauer & Lin, 2017). So far, it is not clear whether that reflects true differences between experiments, or random fluctuation of an effect that is sometimes too small to be statistically identifiable. The IM is well equipped to accommodate this variability because it treats the precision of context representations as a free parameter; when this precision is high, the

generalization gradient of the cue becomes so steep that it causes only a negligible amount of non-target intrusions. Together with previous studies providing evidence for a spatial gradient on intrusion costs, the present finding of substantial intrusion costs that are independent of spatial distance provide strong evidence for the existence of both sources of non-target intrusions.[6]

We considered whether the NP model could also be equipped with two sources of non-target intrusions, corresponding to familiarity and recollection. We could add a component that generates non-target retrievals independent of their proximity to the target. This would require additional assumptions that stretch the model considerably beyond its core assumptions. Specifically, we would have to assume two layers of representations: One of features bound to their context – with precision on both the feature and the context dimension depending on the resource allocated to the item, as in the NP model – and another with representations of features independent of context, generating a familiarity signal. This model would come very close to a version of the IM with the addition of a resource limit on precision. This extension of the NP model would raise the question whether the context-independent feature representations are also resource limited in their precision, and if yes, whether they draw on the same resource as the feature-context bindings, or have their own resource pool. We believe that an extension of the NP model along these lines is promising, but it involves substantive theoretical decisions that we decided to leave to proponents of continuous-resource models.

[6] Further evidence for the existence of content memory dissociated from their context comes from a recent study of Hedayati and Wyble (2020), who showed that on some trials people remembered the identity of a briefly presented letter without any memory for its location in an array.

4.2. The Decision Rule

Most models of change detection assume a decision rule derived either from signal detection theory or from high-threshold models of memory. The choice of decision rule is usually linked to the model's assumption about memory: Theorists assuming that all array items are represented with some continuously varying degree of strength or precision use signal-detection rules (Wilken & Ma, 2004). Theorists assuming two qualitatively different memory states, where some items are represented with fairly good precision, and others not at all, choose a high-threshold model (Nosofsky & Donkin, 2016; Rouder et al., 2008), where the decision depends on a comparison of the probe to the retrieved target feature when the target is in memory, and on (potentially biased) guessing when it is not.

In the present study, we employed a Bayesian decision rule for all models, independent of the assumptions about memory that are incorporated in these models. The Bayesian decision rule implements a signal-detection decision process, or a high-threshold decision process, and this depends not only on the model of memory that it is combined with, but also on the assumptions we make about what knowledge is available to the decision process.

For most model versions we investigated here, the decision rule comes down to a signal-detection model, in which a signal – a random variable drawn from some distribution – is compared to a criterion to determine the decision. The contribution of the Bayesian analysis is to provide a principled rule for computing the signal from the data of each trial – the probe feature and the target feature retrieved from memory – and the available knowledge about memory. There are two equivalent ways of describing the Bayesian rule as a signal-detection model. One is to think of the log-likelihood ratio $LLR_{\varphi}(x)$ as the signal, and the log of the

(negative) prior odds ratio β as the criterion, or bias. A second description can be given on the to-be-remembered feature dimension: The angular deviation of the retrieved feature x_i from the probe feature φ is the signal; when it exceeds a criterion, the probe is judged as "change". The criterion is the angular deviation for which the $LLR_{\varphi}(x)$ is equal to β (see Equations 8 and 9). For the simplified version of the Bayesian decision rule with equal priors for *same* and *change* trials, β is zero, and the criterion is the angular deviation at which the density of the von-Mises distribution, centered on zero and with precision κ, crosses the density of the uniform distribution, $\frac{1}{2\pi}$ (see Figure 9 for an illustration). Unequal priors introduce a bias that shift the criterion towards smaller or larger angular deviations.

Whereas the Bayesian rule translates into a signal-detection decision in most cases, for some models – namely the SA versions with full knowledge of whether the target is in memory on a given trial, including the partial-knowledge version – the decision rule is not a simple signal-detection rule, but rather a mixture of signal-detection and high-threshold models, akin to the decision rule that Nosofsky and Donkin (2016) proposed for mixed-memory state models: When the target is assumed to be in memory, the retrieved feature is compared to the probe to compute a signal that is compared to the criterion, as described above. When the target is assumed not to be in memory, there is no signal, and the decision process just guesses based on the priors. In that case, the priors play the role of the guessing parameter in high-threshold models, which is the source of bias in those models. To conclude, the Bayesian decision rule provides a unifying framework for the decision rules of continuous-strength models and of discrete-state models of memory.

4.3. Knowledge-Rich and Knowledge-Limited Model Versions

The Bayesian decision model makes transparent that memory-based decisions can be informed by more or less meta-cognitive knowledge about one's memory, resulting in substantially different predictions independent of our choice of memory model. We found that equipping the Bayesian decision rule with higher degrees of trial-by-trial knowledge benefits or harm to the model's fit depending on the model. For the IM, attributing to the decision process knowledge of whether the target is in the focus of attention or not slightly improved the model performance. However, assuming knowledge of whether the target is in memory or not reduced the performance of the SA model, and knowledge of the precision of the target on an individual trial reduced the fit of the VP and the NT model. The detriment to the SA model arises because when the decision process knows that the target is not in memory, it guesses in proportion to the priors, which results in poor performance for both *same* and *change* trials. As set size increases, the target is more often not in memory, and thus, the predicted performance for both *same* and *change* probes declines as the set size increases. This is contrary to the data of both experiments, in which set size predominantly affected accuracy on *same* trials. The knowledge-rich SA model could perhaps circumvent this problem by assuming a bias in favor of *change* that increases with set size, but that would detach the bias from any reasonable priors, and thereby render the decision rule irrational. We regard that as a theoretically unattractive alternative.

For the VP and the NP model, the higher degree of knowledge leads to similar predictions as the knowledge-rich SA model. Under lower precision in the current trial, the distribution of retrieved features from VP and NP will be similar to a uniform distribution. If the decision rule is assumed to know that, on a given trial, target precision is very low, it will use a very low precision

parameter for computing the likelihood $P(x|\text{same}, \varphi)$. In that case, the decision rule can barely distinguish between the *change* and the *same* likelihood, and the response will be almost like random guessing, approximating the priors. When the set size gets larger, the decision rule is more likely to encounter trials with low precision, and more guessing-like decisions will be made on both *same* and *change* trials. This forces the models to predict about equally strong set-size effects on accuracy for both kinds of trials, which is contrary to the observed data. To conclude, our finding that set size affects accuracy for *same* probes much more strongly than for *change* probes requires that the SA, VP, and NP models are equipped with a knowledge-limited decision rule, or at least a mixture of knowledge-rich and knowledge-lean (as in the case of the SA).

This is good news for these models, because assuming a knowledge-rich decision process would require a rather rich set of additional assumptions about how that meta-knowledge is acquired and processed on every trial. In addition to control processes for extracting the required information from the current memory set, an additional form of working memory would be required to temporarily maintain the meta-knowledge, such as representing for each array item whether or not it is in a slot, or with what precision it is represented in WM. When these assumptions are spelled out (see van den Berg, Yoo, & Ma, 2017, for an example of an explicit model of obtaining trial-by-trial meta-knowledge of precision in the VP model), they would add considerably to the complexity of the models. The knowledge-limited versions, by contrast, require only the assumption that the average number of items one can keep in WM (for the SA model), or the average precision of a visual-feature representation in WM, given a memory set size (for the VP and NP models). This could be achieved by a simple prediction-error based learning mechanism. The value learned in this way is used to tune the decision process. For the

VP and the NT models, for which the decision rule is a form of signal-detection process, an even simpler approach would be to learn the criterion directly by a gradual learning process that monitors the long-run proportion of "same" and "change" responses and adjusts the criterion until that proportion matches the prior odds ratio of *same* and *change* trials.

For the IM, the knowledge-rich version fit slightly better than the knowledge-limited version. The knowledge-rich version requires a process of determining whether the target is in the focus of attention, and feeding that knowledge into the decision process to adjust the assumed memory precision accordingly. So far, we have no formally explicit model of that meta-cognitive process. If it turns out to substantially add to the model's complexity, then taking that added complexity into account in the model comparison would give the knowledge-limited version of the IM an advantage over the knowledge-rich version, because their difference in fit was only very slight (about 3 BIC units per subject).[7] The knowledge-limited version only needs to learn the average memory precision and set the decision criterion accordingly, as discussed above for the VP and the NP model.

4. 4. Conclusions

We can draw three main conclusions from the present results. First, the IM is a viable model of change detection, as it is for continuous reproduction (Oberauer & Lin, 2017). It accounts well for the effects of two experimental variations that have been instrumental for motivating the slot model: Set size (Luck & Vogel, 1997) and the degree of change in change

[7] Analogously, the partial-knowledge version of the SA fit only slightly better than the knowledge-limited version. Taking into account the complexity of processes required for the trial-to-trial meta-knowledge in the partial version is likely to tilt the balance of fit and parsimony in favor of the knowledge-limited version of the SA.

probes (Nosofsky & Donkin, 2016). The IM does not include the assumption of a limited resource for memory maintenance, be it in the form of discrete slots or a continuously varying quantity. Therefore, we conclude that such an assumption is unnecessary to explain the limited capacity of visual WM.

Second, a successful model of visual WM requires a principled, mechanistic explanation for non-target intrusions. The need to account for non-target intrusions has already been demonstrated for the continuous-distractor paradigm (Oberauer & Lin, 2017; van den Berg et al., 2014); the present model comparison shows that for change detection, too, the main feature that distinguishes between the more and the less successful models is whether they account for intrusion errors. Intrusion errors reinforce an insight about access to information in WM that has been generally accepted among theorists in the list-memory tradition (Lewandowsky & Farrell, 2008), but sometimes neglected by theorists of visual WM: The contents of WM are not directly available for processing or report – they need to be accessed through a selection process that is fallible, sometimes leading to confusions. From a memory perspective this process can be described as cue-based retrieval (Oberauer & Lin, 2017); from an attention perspective it can be described as selective attention to contents of WM (Chun, Golomb, & Turk-Browne, 2011).

The third conclusion pertains to the decision process in change detection. We found that a Bayesian decision rule (Keshvari et al., 2013) can be combined with all extant models of visual WM, incorporating the decision models preferred by theorists of different persuasions as special cases. The Bayesian decision model provides a framework for exploring different degrees of meta-knowledge assumed to be used in the decision process. We found that the decision model works well in combination with all memory models when it is equipped with knowledge about constant

characteristics of memory, but little or no knowledge about the state of memory representations in a particular trial. Although people, when asked, show fairly accurate meta-knowledge about the quality of their current memory representations, change-detection decisions appear to make little use of that knowledge. This insight facilitates modelling of change detection – and more generally, recognition – because it enables us to make simple assumptions about the decision process.

Tables

Table 1. Summary of statistics

Effect Tested	Model With Effect	Model Without Effect	Bayes Factor
Experiment	ProbeType * Setsize * Experiment	ProbeType * Setsize	0.0009
Set Size x Probe Type Interaction	ProbeType * Setsize	ProbeType + Setsize	1.59 * 10^54
Set Size	Setsize	Intercept only	5.92 * 10^41
Probe Type	ProbeType	Intercept only	3.23 * 10^18

Note: Model expressions denote combinations of main effects and interactions of predictor variables through an asterisk (“*”); models including only the main effects combine the variables with a plus (“+”). All models also included a random intercept effect of subject.

Table 2

Goodness-of-fit for all modes considered. Each model was tested in conjunction with different levels of knowledge attributed to the decision rule, and with an unbiased prior or a prior with bias as a free parameter. The N indicates the number of free parameters of the models. The ΔAIC and ΔBIC represent the differences between each model's AIC or BIC, and the lowest AIC or BIC value, respectively, of any model, averaged over participants; smaller values reflect better fits. Within each model class, the knowledge version with the best fit has fit values printed in bold, and was included in the by-participant competition; the final column gives the number of participants for which each model won the competition.

Model	Knowledge Attributed to the Decision Rule	Bias	N	ΔAIC	ΔBIC	N(win)
Slot-Averaging (SA)	Knowing whether the target is in a slot, assumed precision varying with set size	No	2	177.58	155.55	
	Knowing whether the target is in a slot, assumed precision constant	No	2	170.29	148.25	
	Not knowing whether target is in a slot; assumed precision varying with set size	No	2	78.68	56.65	
	Not knowing whether the target is in a slot, assumed precision constant	No	2	63.59	41.56	
	Partial knowledge of whether the target is in a slot	No	3	**55.29**	**38.77**	5
	Knowing whether the target is in a slot, assumed precision constant	Yes	3	89.15	72.63	
	Not knowing whether target is in a slot; assumed precision varying with set size	Yes	3	70.64	70.64	
	Not knowing whether the target is in a slot, assumed precision constant	Yes	3	60.96	44.43	
	Partial knowledge of whether the target is in a slot	Yes	4	51.72	40.71	
Variable-Precision (VP)	Knowing the precision of the current trial	No	3	176.69	160.16	
	Only knowing the average precision	No	3	70.79	54.27	
	Knowing the precision of the current trial	Yes	3			
	Only knowing the average precision	Yes	4	**31.55**	**20.53**	3
Interference Model (IM)	Knowing whether the target is in the focus of attention	No	6	**0**	**0**	18
	Not knowing whether the target is in the focus of attention	No	6	9.78	9.78	

	Knowing whether the target is in the focus of attention	Yes	7	-2.46	3.05	
	Not knowing whether the target is in the focus of attention	Yes	7	-2.17	3.34	
Neural Population (NP)	Knowing the precision of the current trial	No	4	145.82	134.80	
	Only knowing the average precision	No	4	23.20	12.18	
	Knowing the precision of the current trial	Yes	5	202.04	191.02	
	Only knowing the average precision	Yes	5	**15.36**	**9.85**	0
SA-NTguess	Partial knowledge	No	4	59.13	48.12	
	Partial knowledge	Yes	5	51.45	45.94	
SA-Swap	Partial knowledge	No	4	**39.80**	**28.78**	6
	Partial knowledge	Yes	5	51.95	46.44	
VP-Swap	Only knowing the average precision	No	4	19.47	8.45	
	Only knowing the average precision	Yes	5	**10.47**	**4.96**	8

Table 3

Summary of the parameters estimates for the Slot-Averaging model. Given are the medians of the estimated parameters, with means in parentheses.

Knowledge Attributed the Decision Rule	Bias	κ	K	b	β	p_k
Slot-Averaging Model (SA)						
Knowing whether target is in a slot; assumed precision varying with set size	No					
Knowing whether the target is in a slot, assumed precision constant	No	1.816 (1.865)	27.535 (30.502)			
Not knowing whether target is in a slot; assumed precision varying with set size	No	1.823 (2.026)	18.256 (19.902)			
Not knowing whether the target is in a slot, assumed precision constant	No	1.803 (2.040)	16.339 (18.029)			
Partial knowledge	No	1.813 (1.927)	23.340 (26.892)			0.080 (0.246)
Knowing whether the target is in a slot, assumed precision constant	Yes	3.882 (2.953)	16.232 (16.922)		0.161 (0.073)	
Not knowing whether target is in a slot; assumed precision varying with set size	Yes	1.837 (2.055)	15.629 (17.625)		-0.062 (0.070)	
Not knowing whether the target is in a slot, assumed precision constant	Yes	2.344 (2.484)	12.523 (23.028)		0.108 (0.119)	
Partial knowledge	Yes	1.866 (2.106)	13.941 (19.851)		0.147 (0.172)	0.177 (0.369)
Slot-Averaging Model with Non-Target Guessing (SA-NTguess)						
Partial knowledge	No	1.803 (1.789)	22.528 (26.285)	0.678 (0.741)		0.114 (0.244)

Partial knowledge	Yes	1.850 (2.280)	10.322 (15.087)	0.532 (0.511)	0.209 (0.185)	0.871 (0.821)
Slot-Averaging Model with Swap Errors (SA-Swap)						
Partial knowledge	No	3.064 (3.525)	17.087 (20.974)	0.077 (0.076)		0.184 (0.222)
Partial knowledge	Yes	2.615 (3.128)	10.75 (19.887)	0.015 (0.023)	0.307 (0.1435)	0.279 (0.369)

Table 4

Summary of the parameters estimates for the Variable-Precision model. Given are the medians of the estimated parameters, with means in parentheses.

Knowledge Attributed to the Decision Rule	Bias	$\overline{J_1}$	τ	α	b	β
Variable-Precision (VP)						
Knowing the precision of the current trial	No	83.041 (77.641)	93.619 (79.372)	0.964 (0.935)		
Only knowing the average precision	No	90.092 (78.153)	71.462 (71.213)	0.966 (0.909)		
Only knowing the average precision	Yes	123.995 (150.875)	147.029 (156.214)	1.226 (1.173)		0.096 (0.122)
VP-Swap						
Only knowing the average precision	No	77.603 (70.292)	60.195 (58.463)	0.541 (0.582)	0.088 (0.094)	
Only knowing the average precision	Yes	123.565 (147.100)	146.797 (173.115)	0.720 (0.723)	0.081 (0.078)	-0.006 (0.009)

Table 5

Summary of the parameters estimates for the Interference model. Given are the medians of the estimated parameters, with means in parentheses.

Knowledge Attributed to the Decision Rule	Bias	b	a	s	κ	κ_{focus}	r	β
Knowing whether the target is in the focus of attention	No	0.016 (0.040)	0.200 (0.271)	8.776 (9.240)	17.991 (21.328)	33.873 (44.870)	0.040 (0.171)	
Not knowing whether the target is in the focus of attention	No	0.017 (0.039)	0.201 (0.270)	9.678 (8.604)	22.301 (23.145)	28.405 (29.782)	0.054 (0.218)	
Knowing whether the target is in the focus of attention	Yes	0.021 (0.024)	0.206 (0.281)	9.546 (8.677)	16.612 (26.428)	37.826 (46.718)	0.039 (0.187)	-0.437 (-0.497)

Table 6

Summary of the parameters estimates for the Neural Population model. Given are the medians of the estimated parameters, with means in parentheses.

Knowledge Attributed to the Decision Rule	Bias	$\overline{J_1}$	τ	α	b	β
Knowing the precision of the current trial	No	90.844 (80.236)	92.377 (78.031)	0.574 (0.599)	0.144 (0.174)	
Only knowing the average precision	No	86.496 (75.891)	84.016 (71.771)	0.967 (0.931)	2.371 (2.413)	
Knowing the precision of the current trial	Yes	51.58 (51.150)	18.554 (21.119)	0.662 (0.632)	2.513 (2.609)	0.0007 (0.001)
Only knowing the average precision	Yes	122.873 (137.261)	144.930 (156.464)	0.923 (0.835)	2.146 (2.500)	0.003 (0.002)

Figures

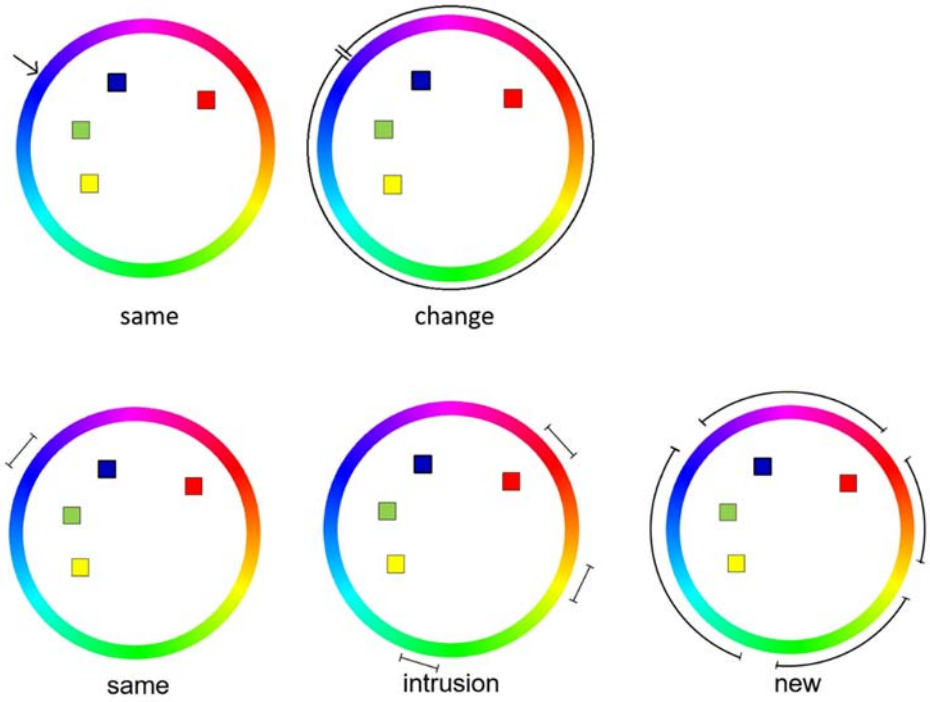

Figure 1. The probe sampling scheme for Experiment A and Experiment B. The four colors used in the trials are blue, red, green, and yellow, and the probe is presented at the blue color location. The top row is the probe sampling scheme for Experiment A, and the bottom row is the probe scheme of Experiment B.

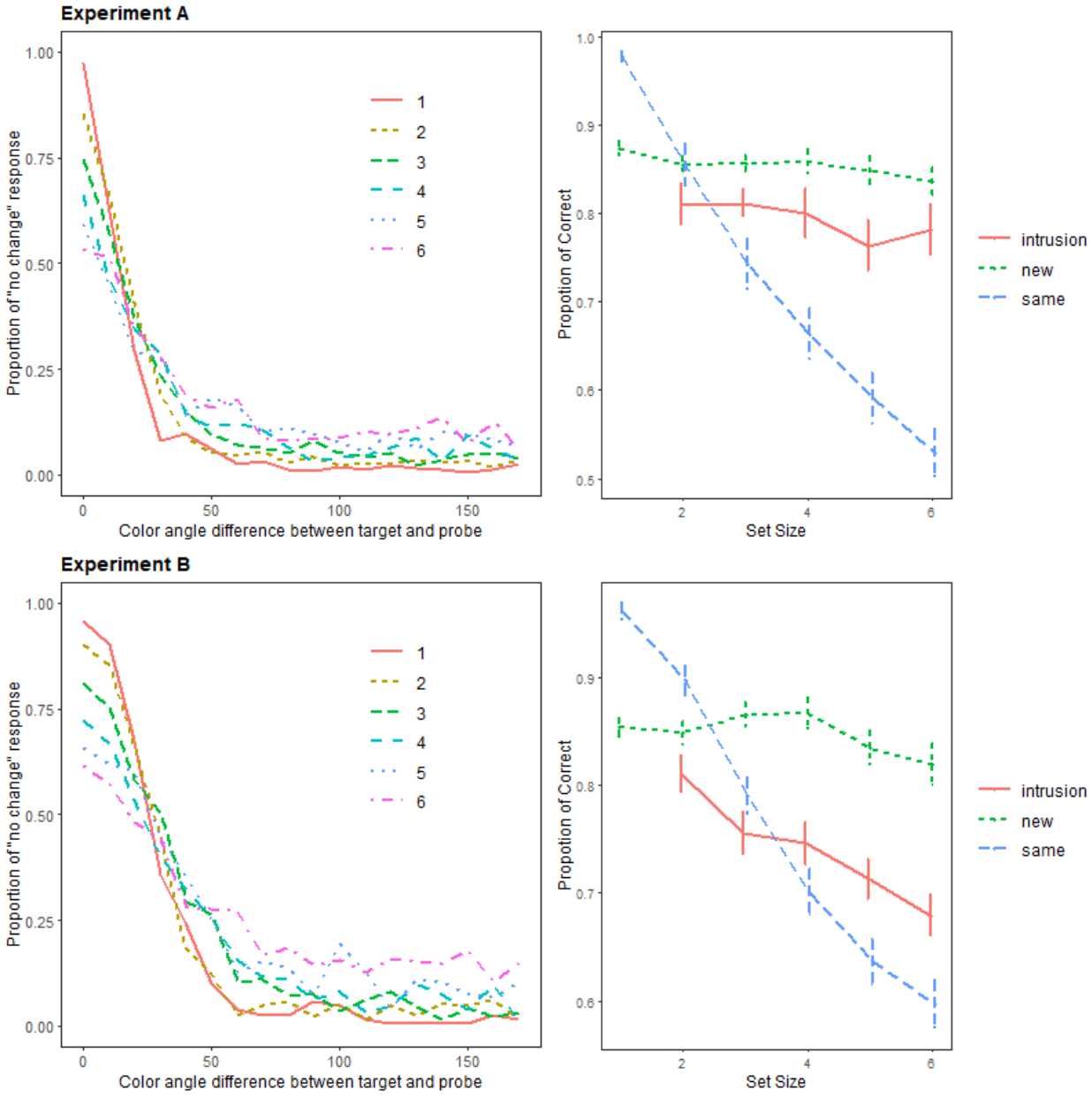

Figure 2. The proportion of "same" responses as a function of the similarity of the probe to the target color, and the proportion of correct responses for the three probe types as a function of set size. The error bars indicate one standard error for within-subjects comparison.

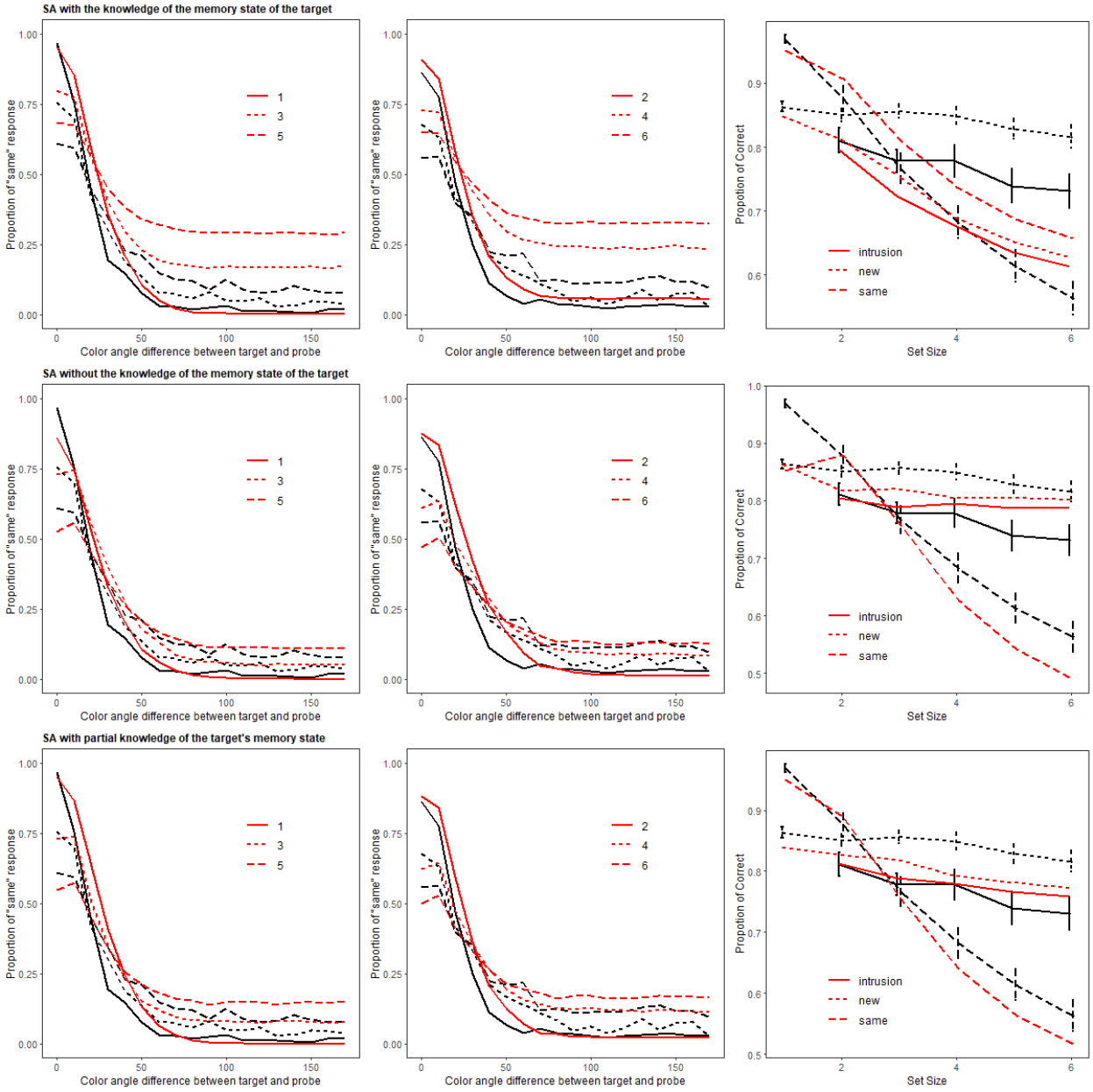

Figure 3. The model fits of SA model versions with different levels of knowledge attributed to the decision process without biased prior. The black lines are the data, and the red lines are the model predictions. The top panels show predictions of the SA with knowledge of whether or not the target is in a slot on the current trial. The middle panels show predictions of the SA without knowledge of the status of the target, and assuming the precision of a single slot for all set sizes. The bottom panels show predictions of the SA with partial knowledge.

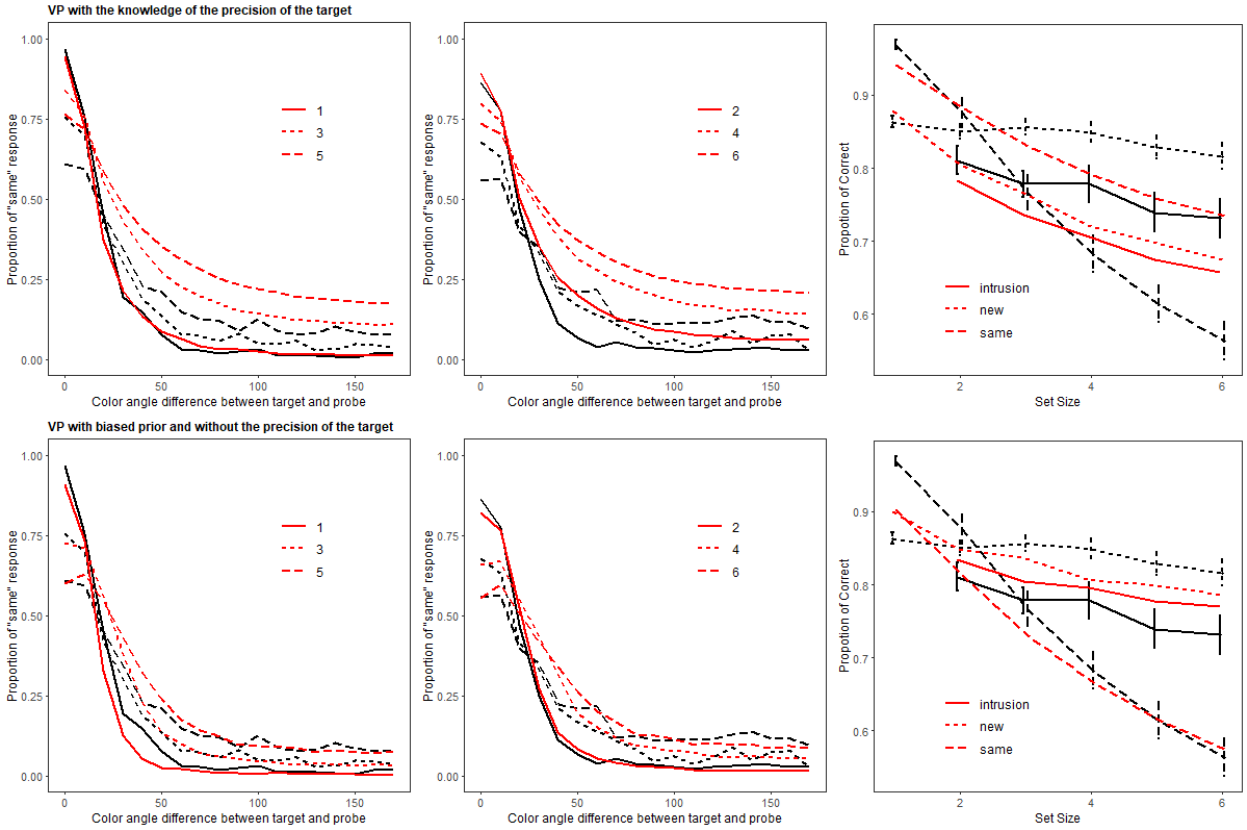

Figure 4. The model fits of the VP model versions with different levels of knowledge attributed to the decision process; both with biased prior. The black lines are the data, and the red lines are the model predictions from VP. The top panels show the model version with knowledge of the precision of the target in the current trial; the bottom panels show the version with knowledge only of the average precision at each set size.

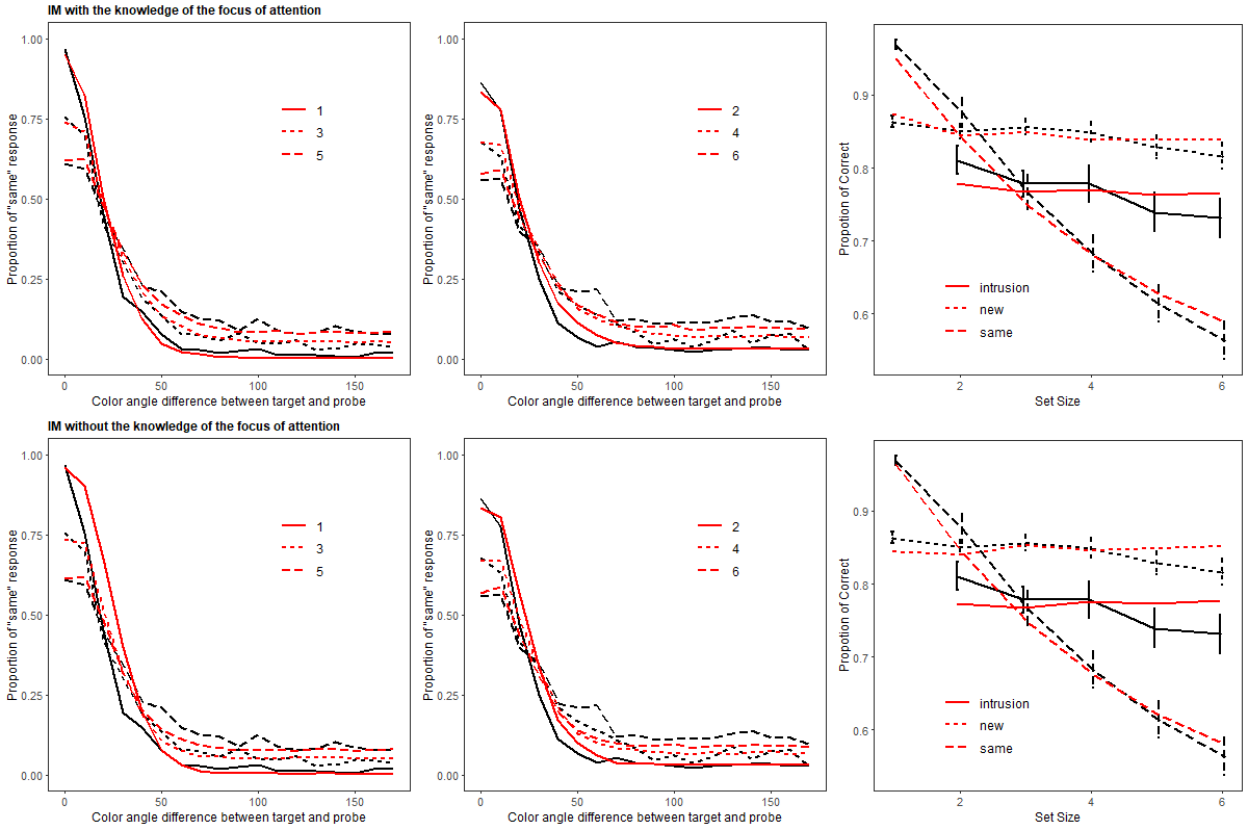

Figure 5. The model fits of the IM with different levels of knowledge attributed to the decision process; both with unbiased prior. The black lines are the data, and the red lines are the model predictions from IM. Model predictions were simulated for each participant with their best-fitting parameter values, then averaged across participants.

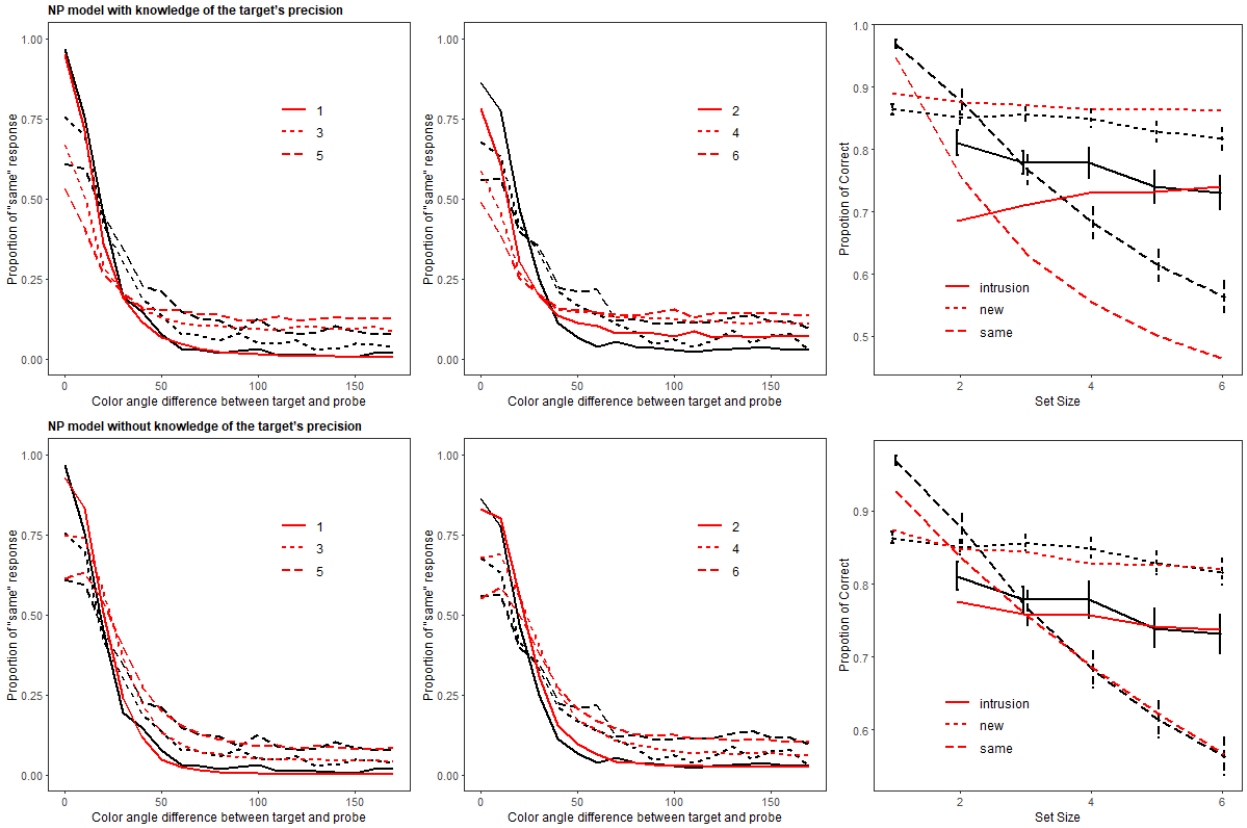

Figure 6 The model fits of the the NP model with different levels of knowledge attributed to the decision process; both with biased prior. Each panel shows the data in black, and the model predictions in red.

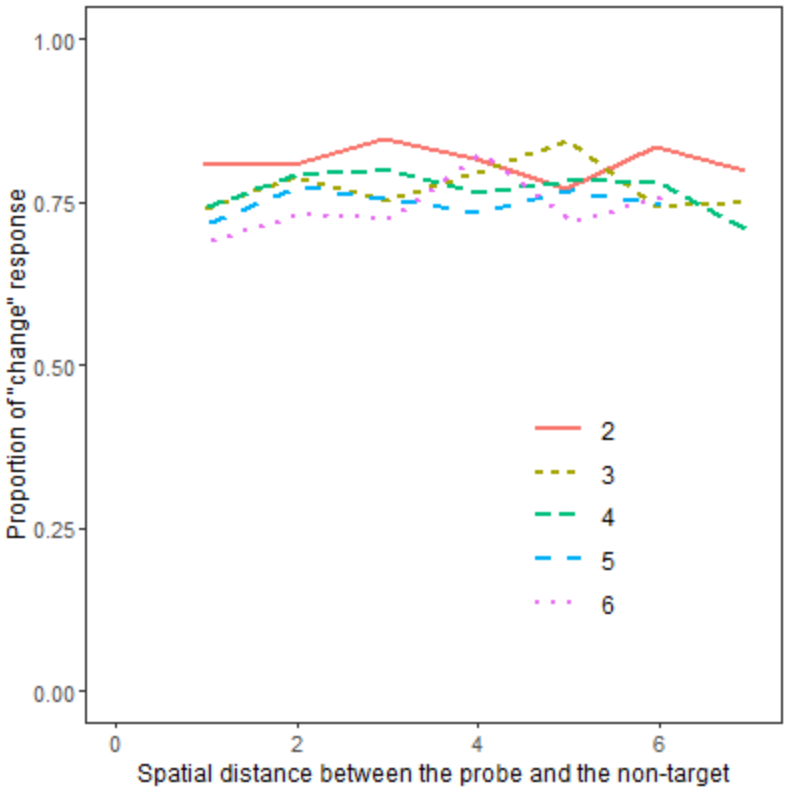

Figure 7. The probability of a *change* response to an intrusion probe as a function of the distance between the probed location and the location of the non-target from which the probe's color is taken. The line color indicates the set size.

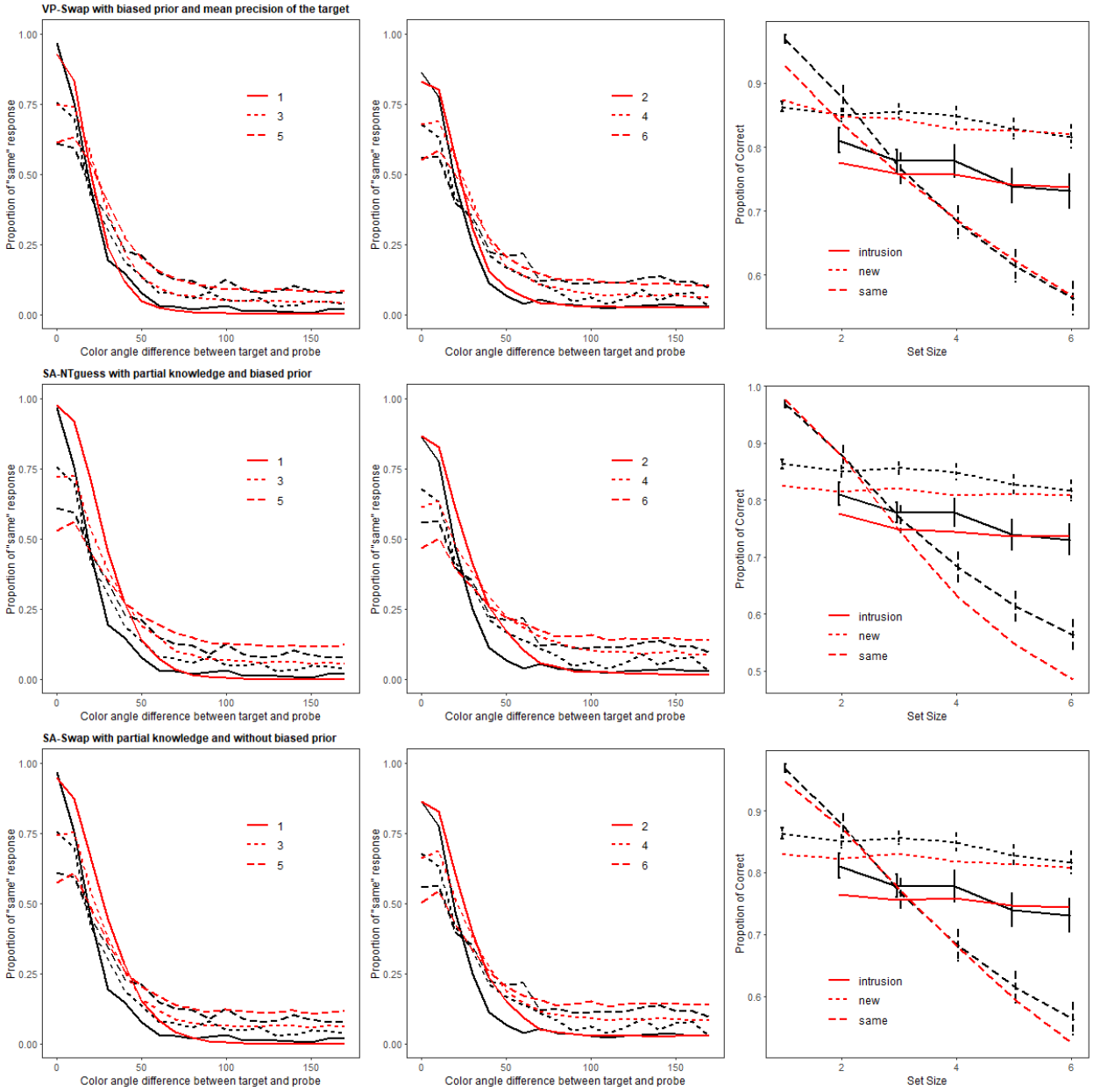

Figure 8: Model fits of the three extended SA and VP models, equipped with additional assumptions to accommodate swap errors.

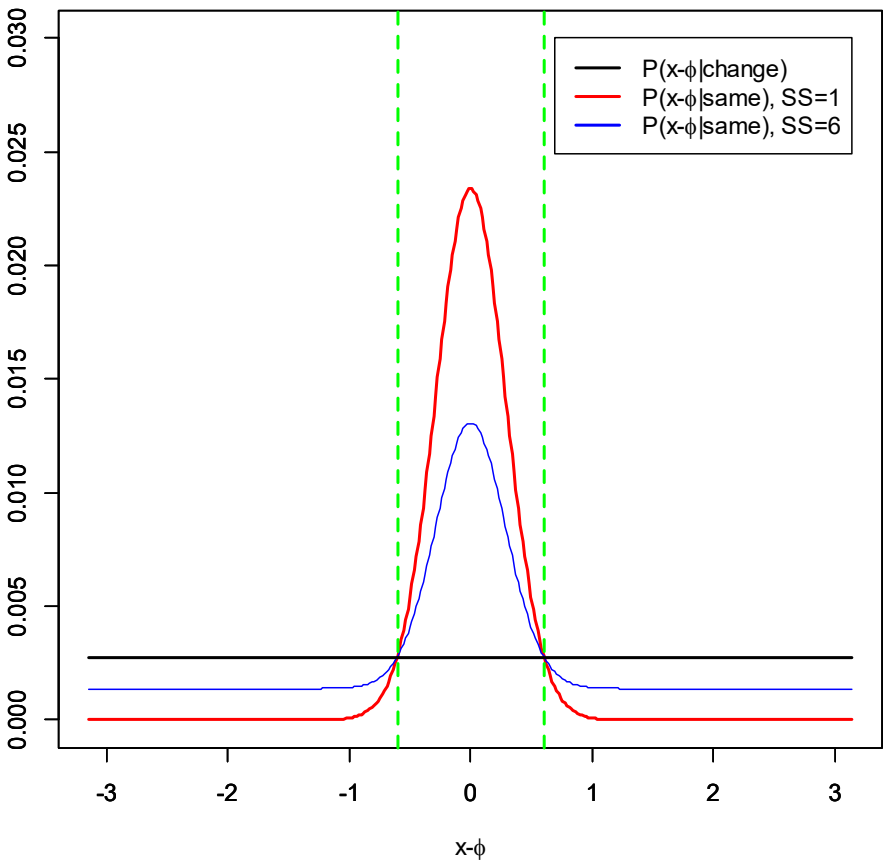

Figure 9. Likelihoods of the deviation of the retrieved feature x from the probe feature φ, given a *change* probe, or given a *same* probe. The likelihood for a *same* probe depends on the probability that the retrieved feature is drawn from a memory-precision distribution centered on the target, P_s, which decreases with set size. Nevertheless, the absolute deviation at which the likelihood for change probes exceeds that of same probes is the same regardless of the value of P_s, as explained in Appendix A. An unbiased criterion on the absolute deviation between probe and retrieved feature, shown here as the green vertical lines, is set to where the likelihoods cross, regardless of set size.

Appendix A

We tested three core models and their variants in the single-probe change-detection task. In this Appendix, we explain these models in detail, and how the Bayesian decision rule was applied with different levels of trial-by-trial knowledge attributed to the decision process.

Slot-Averaging Model

The assumption of the SA model is that participants can only remember the number of items up to the capacity of working memory, K. Thus, the probability of the target being stored in memory is

$$P_m = \begin{cases} 1, & n < K \\ \frac{K}{n}, & n \geq K \end{cases} \quad \text{(B.1)}$$

Here, n stands for set size of the current trial, and P_m is the probability of the target being in memory. If the target is in memory, the retrieved feature is distributed around the target feature. If the target is not in memory, the participant has to guess, and the retrieved feature will be uniformly distributed among all the possible responses.

Assuming the target item is stored in a slot, the precision of retrieval is depended on how many slots are contributing to the retrieval. If each slot holds a unique item in the trial, which happens when $n \geq K$, only one slot contributes to memory retrieval, and the precision of retrieval is the free parameter κ_1. However, when multiple slots hold the same item at once, which happens when $n < K$, multiple slots will be sampled during the retrieval process, which results in higher precision of retrieval. The precision of retrieval is determined by how many slots are contributing during the retrieval process. We follow the equations used in Zhang and Luck

(2008). When an item is stored in m slots, the standard deviation of the retrieval distribution, σ_m, is

$$\sigma_m = \frac{\sigma_1}{m^{0.5}}, \tag{B.2}$$

and the σ_1 is the standard deviation when the item is stored in one slot (which is a function of κ_1). Once the standard deviation of the retrieval is determined, we then convert it to the precision of retrieval κ by using the *sd2k* function provided by Bays et al. (2009).

For the Slot-Averaging model with non-target guessing, we introduced a free parameter *b* for the conditional probability of guessing a non-target, given that one has to guess (because the target is not in a slot). When a non-target is guessed, the non-target that is retrieved is any of the *n*-1 non-targets with equal probability. Therefore, the probability of retrieving the non-target that matches an intrusion probe is

$$P(retrieve\ matching\ NT) = (1 - P_m)\frac{b}{n-1}. \tag{B.3}$$

For the Slot-Averaging Swap model, we assumed that there is a certain probability that a swap error could occur during encoding, and the probability of a swap error increases linearly with set size. The probability that a swap error occurs between the target and any other item is b at set size 2, which is the lowest possible set size for a swap error to occur. Thus, the probability of retrieving the target P_m in SA-Swap is

$$P_m = \begin{cases} 1 \cdot [1 - b(n-1)], & ss < K \\ \frac{K}{n} \cdot [1 - b(n-1)], & ss \geq K \end{cases}, \tag{B.4}$$

where b is the probability of a swap error occurring at set size 2. When a swap error occurred, we assumed that all the non-target items have an equal probability of being remembered instead of the target. Therefore, at retrieval, each non-target item has probability b of being retrieved, regardless of set size.

To implement the decision rule for the SA, SA-NTguess, and SA-Swap models, we had to make two assumptions about the knowledge used in the decision. One is whether the decision process knows the target's precision of the current trial as a function of how many slots store the target. If that is the case, the decision rule uses the κ value computed from σ_m; otherwise it uses κ_1. The other assumption is whether the decision rule has knowledge of whether the target is in memory or not, or does not have knowledge regarding the memory state of the target. If the decision rule has knowledge of the memory state of the target, the probability of responding *change* is

$$P_{change}(\varphi) = (1 - P_m) \times \frac{\exp(-\beta)}{1 + \exp(-\beta)} + P_m \int P[LLR_{\varphi}(x) > \beta] \cdot vonMises(x|x_i, \kappa)dx. \quad \text{(B.5)}$$

Here, in case the target item is known not to be in memory (which happens with probability $1 - P_m$), the decision process is to guess, and the optimal guess is to respond "change" according to the prior probability of change, p(*change*), which is equal to $\frac{\exp(-\beta)}{1+\exp(-\beta)}$. In case the target is known to be in memory, the response is determined by comparing the log-likelihood ratio for each possible retrieved feature *x* to the criterion β, weighted with the probability of retrieving x given by a von Mises distribution centered on the true target feature x_i, and precision = κ. In this case, P_s in Equation (8) equals 1, and the log-likelihood ratio simplifies to:

$$LLR_{\varphi}(x) = \log\left[\frac{\frac{1}{2\pi}}{vonMises(x|\varphi,\kappa)}\right] \tag{B.6}$$

If the decision rule has no knowledge of the memory state, the probability of a "change" response is

$$P_{change}(\varphi) = \int P[LLR_{\varphi}(x) > \beta] \cdot P_{retrieve}(x)dx. \tag{B.7}$$

with

$$LLR_{\varphi}(x) = \log\left[\frac{\frac{1}{2\pi}}{P_m \cdot vonMises(x|\varphi,\kappa) + (1-P_m)\cdot\frac{1}{2\pi}}\right] \tag{B.8}$$

$$P_{retrieve}(x) = (1-P_m) \times \frac{1}{2\pi} + P_m vonMises(x|x_i,\kappa)$$

Variable Precision Model

The assumption in the VP model is that the total amount of resources is fixed, and the amount of resources put into an item varies from trial to trial. The mean precision at set size n is

$$\overline{J_n} = \frac{\overline{J_1}}{n^{\alpha}} \tag{B.9}$$

$\overline{J_1}$ is the mean Fisher information at set size 1, which reflects the total amount of resource available, and the α parameter controls the steepness of decline in Fisher information as the set size increases, and the resource is divided among *n* items.

The precision κ of each item in an array depends on the Fisher information J assigned to it; the relationship between κ and J is $J = \kappa\frac{I_1(\kappa)}{I_0(\kappa)}$. There is no analytic solution to compute κ directly from J, therefore we computed multiple Js from a fine grid of κs and interpolated the κ

at the given J. The Fisher information J_i assigned to the target at each trial is a random variable. It is drawn from a Gamma distribution with the mean of $\overline{J_n}$ and scale of τ. For a computational reason, instead of integrating across the Gamma distribution, we draw the J_i values from the 1000-quantiles of the Gamma distribution to simulate the integration.

To compute the probability of retrieving each possible feature value x, for each set size *n* we simulated the variable-precision distribution $G_{VP}(x|x_i,\overline{J_n})$ by averaging 1000 von-Mises distributions with the true target feature x_i as its mean, and κ sampled from the Gamma distribution on $\overline{J_n}$:

$$P_{retrieve}(x) = G_{VP}(x|x_i,\overline{J_n}) = \frac{1}{m}\sum_{i=1}^{m} vonMises(x|x_i,\kappa_i);$$

$$\kappa_i \sim Gamma(\overline{J_n},\tau) \qquad \text{(B.10)}$$

For implementing the VP with the Bayesian decision rule, we need to again consider the trial-by-trial knowledge assumed to be available. If the decision rule has knowledge of the precision of the target in the current trial, κ is easy to determine in the VP model: The κ drawn from the Gamma distribution on $\overline{J_n}$ is is used as the precision parameter of a von-Mises distribution in the computation of the log-likelihood ratio for the decision rule. If the decision process only has knowledge of the average precision of the current set size, the log-likelihood ratio is instead computed by a variant of Equation (8) in which the von-Mises distribution is replaced by the simulated variable-precision distribution $G_{VP}(x|\varphi,\overline{J_n})$:

$$LLR_{\varphi}(x) = \log\left[\frac{\frac{1}{2\pi}}{G_{VP}(x|\varphi,\overline{J_n})}\right], \qquad \text{(B.11)}$$

The probability of a "change" response is again given by Equation (10):

$$P_{change}(\varphi) = \int P[LLR_{\varphi}(x) > \beta] \cdot P_{retrieve}(x)dx, \tag{B.9}$$

Interference Model

The assumption of the IM is that there are three sources of activation contributing to the retrieval process. Activation A_C is the context-dependent information, which arises from using the location to re-activate the item bound to it,

$$A_c(x|L_\theta) = \sum_i^n \exp[-s \cdot D(L_i, L_\theta)] \cdot vonMises(x|x_i, \kappa). \tag{B.12}$$

The mean of the von-Mises, x_i, is the color of item i, and $D(L_i, L_\theta)$ is the distance between the location of item i and the location of the probe θ. The s parameter is the rate of decline of the spatial generalization gradient of the probe location as a retrieval cue, that is, the precision of spatial-location memory.

Activation A_a is the content-independent information, which arises from recent experience with every array item.

$$A_a(x) = \sum_i^n vonMises(x|x_i, \kappa) \tag{B.13}$$

Activation A_b is the background noise activation. Every response candidate *x* shares the same background-noise activation, and the amount of activation has a linear relationship with set sizes, as

$$A_b = \frac{n}{2\pi}. \tag{B.14}$$

The activation of a response candidate is the sum of the three activations, with weights *a*, *b*, and *c* for each source of activation, as

$$A(x|L_\theta) = cA_c(x|L_\theta) + aA_a(x) + bA_b(x), \quad \text{(B.15)}$$

and we used the Luce's choice rule to determine the probability of retrieving x based on the $A(x|L_\theta)$ through

$$P_{retrieve}(x) = \frac{A(x|L_\theta)}{\sum_i^N A(i|L_\theta)}, \quad \text{(B.16)}$$

where *N* is the number of possible response candidates (i.e., *N*=360 for the 360 colors on the color wheel).

We also assumed that the focus of attention is playing a role in the VWM. The focus of attention can maintain one item with higher precision, and that item is resistant to interference from the other items in the trial. However, with the simultaneous presentation of the memory items, we cannot know which item is in the focus of attention. Therefore, we assumed that all the items have an equal probability of being in the focus of attention. When the target is in the focus of attention, the feature precision κ_f is used instead of the usual κ, thus the activation A_c is

$$A_c(x|L_\theta) = \sum_i^n \exp[-s \cdot D(L_i, L_\theta)] \cdot vonMises\big(x|x_i, \kappa_f\big), \quad \text{(B.17)}$$

The activation of the response candidate becomes

$$A(x|L_\theta) = cA_{c(x|L_\theta)} + r[aA_a(x) + bA_b(x)], \quad \text{(B.18)}$$

where r is a free parameter ranging from 0 to 1, reflecting the proportional reduction of the interference from A_a and A_b activations.

The decision rule for the IM is

$$P_{change}(\varphi) = \int P[LLR_\varphi(x) > \beta] \cdot P_{retrieve}(x)dx. \quad \text{(B.19)}$$

with

$$LLR_{\varphi}(x) = \log\left[\frac{\frac{1}{2\pi}}{P_s \cdot vonMises(x|\varphi,\kappa) + (1-P_m)\cdot\frac{1}{2\pi}}\right] \quad \text{(B.20)}$$

and $P_{retrieve}(x)$ as given in Equation B.14.

For the computation of the LLR we assumed that the κ parameter used as precision of the von-Mises distribution depends on the level of knowledge attributed to the decision process. If the decision process does not have knowledge of whether the target item is in the focus of attention or not, the κ value used to compute the LLR is the same as the κ in the IM without the focus of attention. However, if knowledge of whether the target is in the focus of attention or not is assumed to be available to the decision process, we assumed that κ_f is used when the target item is in the focus of attention, and κ when it is not.

The probability of retrieving a feature from the von-Mises distribution centered on the target, P_s, is computed as follows: As $P_{retrieve}(x)$ is proportional to the distribution of activation at retrieval, which is a weighted mixture of von-Mises distributions and a Uniform, we can also write it as a mixture of probability distributions with an appropriate normalization constant (Oberauer, Stoneking, Wabersich, & Lin, 2017). The probability of retrieving a feature from the von-Mises distribution centered on the target in Equation (8), P_s, is the mixture weight of the target distribution normalized by the sum of all mixture weights:

$$P_s = \frac{c+a}{c\sum_{i=1}^{n} exp[-s\cdot D(L_i, L_\theta)] + na + nb} \quad \text{(B.21)}$$

The distance between the targets and the non-targets, $D(L_i, L_\theta)$, varies from trial to trial, and therefore, we computed P_s separately for each trial, on the assumption that the decision process has knowledge of the distances in each trial. The alternative, not assuming such trial-by-trial noise, would be to take the average distance between items in arrays across trials in the experiment. We ran both model versions and found only a negligible difference between them. As explained in Appendix A, the P_s term plays a role only in the IM version with a free bias parameter. The best-fitting version was the IM without bias parameter, and therefore we did not further investigate different levels of knowledge about the target to non-target distances.

The Neural-Population Model

In the NP model, the precision of each item's feature is computed in the same way as for the VP model (Equation B.7 and the subsequent description). In addition, the NP model incorporated the assumption that participants used the location to retrieve the correct item, and the resource does not only affect the precision of the retrieved feature but also the precision of location information of all array items, and thereby the probability of retrieving the correct item, given the target's location. We assumed that the resources allocated to the binding between an item and its location are tied to the resources allocated to the memory item. The Fisher information of binding is the Fisher information of the memory item, multiplied with a scaling parameter b, and the precision of binding κ_{bn} is derived from $b \times J_n$. At test, the location of the probe is used as a cue to retrieve the array item in that location. Given the probe location L_θ, the probability of retrieving each possible location L_i on the virtual circle on which the array items are placed is $G_{VP}(L_i|L_\theta,\ b \times \overline{J_n})$. The probability of retrieving the ith item is defined as the relative probability of retrieving each item's location:

$$P_r\left(i\middle|L_p\right) = \frac{G_{VP}(L_i|L_\theta,\ b \times \bar{J_n})}{\sum_j^n G_{VP}\left(L_j|L_\theta,\ b \times \bar{J_n}\right)}. \quad \text{(B.22)}$$

The probability of retrieving every possible feature x is now a probabilistic mixture of *n* distributions, one for each of the *n* array items that could be retrieved, centered on the feature value x_i of item *i*:

$$P_{retrieve}(x) = \sum_i^n P_r(i|L_p)\, G_{VP}(x|x_i, \bar{J_n}) \quad \text{(B.23)}$$

The Bayesian decision rule is applied in the same way as for the VP model (Eq. B.8 and B.10).

Appendix B

In the Bayesian decision rule, the likelihood of a *same* trial, $P(x, \varphi|\text{same})$ is determined by both P_s and κ. For the simplified decision rule with equal priors, P_s does not influence the expected outcome of the decision, as expressed in $P(\text{change}|\varphi, \boldsymbol{X})$. In this Appendix, we explain why P_s drops out of the simplified decision rule.

According to Equation 9, the probability of change response, $P_{\text{change}}(\varphi)$, is determined by the integral across x of $P\left[LLR_\varphi(x) > \beta\right] \cdot P_{retrieve}(x)$. We can interpret Equation 9 as the sum of the probabilities of recalling any color which is sufficiently similar to the probe color to decide *same*, and how similar the recalled color has to be for a *same* response is determined by the difference between $LLR_\varphi(x)$ and β. If the recalled color is similar enough to the probe color, which results in $LLR_\varphi(x) < \beta$, a *same* response is given, and otherwise a *change* response is given. When we assume equal priors, β = 0. Thus, the minimum similarity for a recalled color to be given a *same* response is determined by $LLR_\varphi(x) = 0$. Because $LLR_\varphi(x)$ is the log-likelihood ratio pf $P(x, \varphi|\text{change})$ and $P(x, \varphi|\text{same})$, the exponential of $LLR_\varphi(x)$ is the likelihood ratio:

$$\exp\left[LLR_\varphi(x)\right] = \frac{P(x, \varphi|\text{change})}{P(x, \varphi|\text{same})}.$$

Setting $LLR_{\varphi}(x)$ = 0 implies $\exp\left[LLR_{\varphi}(x)\right] = 1$, so we obtain:

$$1 = \frac{\frac{1}{2\pi}}{P_s \cdot vonMises(x|\varphi,\kappa) + (1-P_s)\cdot\frac{1}{2\pi}}$$

$$\frac{1}{2\pi} = P_s \cdot vonMises(x|\varphi,\kappa) + (1-P_s)\cdot\frac{1}{2\pi}$$

$$P_s\frac{1}{2\pi} = P_s \cdot vonMises(x|\varphi,\kappa)$$

$$\frac{1}{2\pi} = vonMises(x|\varphi,\kappa) \quad \text{(A.1)}$$

Thus, when simply calculating the minimum similarity between retrieved feature and probe for giving a *same* responses, P_s is no longer a factor in the equation: The minimum similarity needed for a *same* response is the angular distance at which the von-Mises density (with precision κ) crosses the density of the uniform distribution. Hence, we left out the P_s when applying the simplified decision rule to all the models.

References

Ashby, F. G., Prinzmetal, W., Ivry, R., & Maddox, W. T. (1996). A formal theory of feature binding in object perception. *Psychological Review, 103*(1), 165-192. doi:10.1037/0033-295X.103.1.165

Awh, E., Barton, B., & Vogel, E. K. (2007). Visual working memory represents a fixed number of items regardless of complexity. *Psychological Science, 18*, 622-628.

Bays, P. M. (2014). Noise in neural populations accounts for errors in working memory. *Journal of Neuroscience, 34*, 3632-3645.

Bays, P. M. (2016). Evaluating and excluding swap errors in analogue tests of working memory. *Scientific Reports, 6*. Retrieved from doi:10.1038/srep19203

Bays, P. M., Catalao, R. F. G., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision, 9*, 1-11.

Blake, R., Cepeda, N. J., & Hiris, E. (1997). Memory for visual motion. *Journal of Experimental Psychology: Human Perception and Performance, 23*, 353-369.

Brown, G. D. A., Preece, T., & Hulme, C. (2000). Oscillator-based memory for serial order. *Psychological Review, 107*, 127-181.

Burgess, N., & Hitch, G. J. (1999). Memory for serial order: A network model of the phonological loop and its timing. *Psychological Review, 106*, 551-581.

Chun, M. M., Golomb, J. D., & Turk-Browne, N. B. (2011). A taxonomy of external and internal attention. *Annual Review of Psychology, 62*, 73-101.

Cowan, N., Blume, C. L., & Saults, J. S. (2013). Attention to attributes and objects in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 731-747. doi:10.1037/a0029687

Donkin, C., Nosofsky, R. M., Gold, J. M., & Shiffrin, R. M. (2013). Discrete-slot models of visual working-memory response times. *Psychological Review, 120*, 873-902.

Donkin, C., Tran, S. C., & Le Pelley, M. (2015). Location-based errors in change detection: A challenge for the slots model of visual working memory. *Memory & Cognition, 43*, 421-431. doi:10.3758/s13421-014-0487-x

Göthe, K., & Oberauer, K. (2008). The integration of familiarity and recollection information in short-term recognition: Modeling speed-accuracy trade-off functions. *Psychological Research, 72*, 289-303.

Hedayati, S., & Wyble, B. (2020). Memories of visual events can be formed without specific spatial coordinates. *Journal of Cognition, 3*. Retrieved from doi:10.5334/joc.104

Henson, R. N. A., Norris, D. G., Page, M. P. A., & Baddeley, A. D. (1996). Unchained memory: Error patterns rule out chaining models of immediate serial recall. *Quarterly Journal of Experimental Psychology, 49A*, 80-115.

Jones, E., Oliphant, T., & Peterson, P. (2001). SciPy: Open source scientific tools for python: http://www.scipy.org/.

Keshvari, S., van den Berg, R., & Ma, W. J. (2013). No evidence for an item limit in change detection. *PLoS Computational Biology, 9*. Retrieved from doi:10.1371/journal.pcbi.1002927

Lee, C. L., & Estes, W. K. (1977). Order and position in primary memory for letter strings. *Journal of Verbal Learning & Verbal Behavior, 16*, 395-418.

Lewandowsky, S., & Farrell, S. (2008). Short-term memory: new data and a model. In B. H. Ross (Ed.), *The Psychology of Learning and Motivation* (Vol. 49, pp. 1-48). London, UK: Elsevier.

Logan, G. D. (1996). The CODE theory of visual attention: An integration of space-based and object-based attention. *Psychological Review, 103*(4), 603-649. doi:10.1037/0033-295X.103.4.603

Logan, G. D. (2020). Serial order in perception, memory, and action. *Psychological Review*. doi:10.1037/rev0000253

Logan, G. D., Cox, G. E., Annis, J., & Lindsey, D. R. B. (in press). The episodic flanker effect: Memory retrieval as attention turned inward. *Psychological Review*.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390*, 279-281.

Manohar, S. G., Zokaei, N., Fallon, S. J., Vogels, T. P., & Husain, M. (2019). Neural mechanisms of attending to items in working memory. *Neuroscience & Biobehavioral Reviews, 101*, 1-12. doi:10.1016/j.neubiorev.2019.03.017

Morey, R. D., & Rouder, J. N. (2015). BayesFactor (Version 0.9.12.2). Retrieved from http://cran.at.r-project.org/web/packages/BayesFactor/index.html

Murdock, B. B., & vom Saal, W. (1967). Transpositions in short-term memory. *Journal of Experimental Psychology, 74*, 137-143.

Nosofsky, R. M., & Donkin, C. (2016). Qualitative contrasts between knowledge-limited mixed-state and variable-resource models of visual change detection. *Journal of Experimental Psychology: Learning, Memory and Cognition*.

Oberauer, K. (2008). How to say no: Single- and dual-process theories of short-term recognition tested on negative probes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(3), 439-459.

Oberauer, K., & Lin, H.-Y. (2017). An interference model of visual working memory. *Psychological Review, 124*, 21-59.

Oberauer, K., Stoneking, C., Wabersich, D., & Lin, H.-Y. (2017). Hierarchical Bayesian measurement models for continuous reproduction of visual features from working memory. *Journal of Vision, 17*, 1-27. doi:10.1167/17.5.11

Panichello, M. F., DePasquale, B., Pillow, J., & Buschman, T. (2019). Error-correcting dynamics in visual working memory. *nature Communications, 10*. Retrieved from doi:10.1038/s41467-019-11298-3

Pratte, M. S. (2019). Swap errors in spatial working memory are guesses. *Psychonomic Bulletin & Review, 26*, 958-966. doi:10.3758/s13423-018-1524-8

R_Core_Team. (2020). R: A language and environment for statistical computing (Version 3.6.2). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from URL: http://www.R-project.org

Rerko, L., Oberauer, K., & Lin, H.-Y. (2014). Spatially imprecise representations in working memory. *Quarterly Journal of Experimental Psychology, 67*, 3-15. doi:10.1080/17470218.2013.789543

Robinson, M. M., Benjamin, A. S., & Irwin, D. E. (2020). Is there a K in capacity? Assessing the structure of visual short-term memory. *Cognitive Psychology, 121*. Retrieved from doi:10.1016/j.cogpsych.2020.101305

Rouder, J. N., Morey, R. D., Cowan, N., Zwilling, C. E., Morey, C. C., & Pratte, M. S. (2008). An assessment of fixed-capacity models of visual working memory. *Proceedings of the National Academy of Sciences, 105*, 5975-5979.

Schneegans, S., & Bays, P. M. (2017). Neural architecture for feature binding in visual working memory. *Journal of Neuroscience, 37*, 3913-3925. doi:10.1523/JNEUROSCI.3493-16.2017

Schneegans, S., Taylor, R., & Bays, P. M. (2020). Stochastic sampling provides a unifying account of visual working memory limits. *Proceedings of the National Academy of Sciences, 117*(34), 20959. doi:10.1073/pnas.2004306117

Storn, R., & Price, K. (1997). Differential Evolution – A Simple and Efficient Heuristic for global Optimization over Continuous Spaces. *Journal of Global Optimization, 11*(4), 341-359. doi:10.1023/A:1008202821328

Swan, G., & Wyble, B. (2014). The binding pool: A model of shared neural resources for distinct items in visual working memory. *Attention, Perception & Psychophysics*. doi:10.3758/s13414-014-0633-3

van den Berg, R., Awh, E., & Ma, W. J. (2014). Factorial comparison of working memory models. *Psychological Review, 121*, 124-149. doi:10.1037/a0035234

van den Berg, R., & Ma, W. J. (2018). A resource-rational theory of set size effects in human visual working memory. *eLIFE, 7*, e34963. Retrieved from https://doi.org/10.7554/eLife.34963 doi:10.7554/eLife.34963

van den Berg, R., Shin, H., Chou, W.-C., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences, 109*, 8780-8785.

van den Berg, R., Yoo, A. H., & Ma, W. J. (2017). Fechner's law in metacognition: a quantitative model of visual working memory confidence. *Psychological Review, 124*, 197-214. doi:10.1037/rev0000060

Wei, Z., Wang, X.-J., & Wang, D.-H. (2012). From distributed resources to limited slots in multiple-item working memory: A spiking network model with normalization. *Journal of Neuroscience, 32*, 11228-11240.

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision, 4*, 1120-1135.

Wixted, J. T., & Mickes, L. (2010). A continuous dual-process model of remember/know judgments. *Psychological Review, 117*, 1025-1054. doi:10.1037/a0020874

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46*, 441-517.

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature, 453*, 233-236.