# Introduction

Computational psychological models, instantiating psychological theories into mathematical equations, enables psychologists to explore the implications of their theories beyond human thinking (McClelland, 2009). Before we ever apply a model to interpret data of a phenomenon that it explains, it is necessary to assess whether this model reflects the nature of this phenomenon.

Often, the evaluation of a model takes the form of assessing the consistency of the model and data by a measure of goodness-of-fit. Suppose a model explaining a phenomenon is wait for assessment, and an experiment about this phenomenon has been conducted. A good fit of the model to the observed outcomes is taken as the support for the model. According to Popper (1959), a scientific theory must be falsifiable, that is, there are some possible outcomes inconsistent with the theory, and the consistent outcomes will (temporarily) support a falsifiable theory. With the possibility that a model may not fit the data well, the test of whether or not a model will provide a good fit to data forms a Popperian test, which makes it a seemingly reasonable choice to use a good fit to support a model.

However, this approach misses a piece of important information: the strength of the support. The strength of the support that a theory gains from an outcome is related to the risk of obtaining that outcome absent the theory (Meehl, 1990). The riskier of the observation is, the stronger support the observation provides. As Roberts and Pashler (2000) pointed out, the test of a good fit neglects the plausibility of the outcomes ruled out by a model. If a model does not rule out any plausible outcomes, i.e., the model can fit all outcomes likely to be observed in the experiment, this model will have a negligible risk of providing a good fit. In such a case, a good fit provides negligible support to a theory. Therefore, to gain strong support from a good fit, it is necessary to show that there are some plausible outcomes ruled out by a model, which implies there is some risk to obtain a good fit.

The dominant strategy, which uses model complexity as additional information to goodness-of-fit, fails to meet this requirement. This strategy claims that the persuasiveness of a good fit is high if the measurement of model complexity is low, and is low if the measurement of model complexity is high (Veksler et al., 2015). As Vanpaemal (2020) pointed out, on the one hand, although controlling the model complexity to be low does restrict the range of the possible outcomes of a model, all plausible outcomes may still be contained in this small range of outcomes; on the other hand, a model with high complexity may still rule out plausible outcomes. As a result, model complexity is not a proper criterion to gauge a good fit's persuasiveness.

Vanpaemal (2020) proposed a more complete approach for the assessment of the persuasiveness of a good fit in the Bayesian framework. This approach consists of two ingredients: the core predictions of a model and the data prior.

The core predictions contain the outcomes of the experiment that can be fit well with the model. The construction of the core predictions relies on the prior predictive distribution. The prior predictive distribution gives the distribution of future outcomes based on the model alone without considering the observed data. The outcomes that are considered to be fit poorly by a model in the sense that the model assigns small prior mass is assigned to them. By defining a bad fit in this way, there are possible outcomes inconsistent with the models.

The data prior contains the plausible outcomes. The plausibility of outcomes can be assessed based on theoretical considerations, previously observed empirical data, and expert knowledge. Two points need to note with the construction of the data prior. First, the plausibility of outcomes depends on the details of the research method, so the data prior should reflect the specialty of the experiment of interest. Second, the data prior should be sensitive to the theory under consideration.

Note that the construction of two ingredients does not involve the observed outcomes of the experiment. It is important since otherwise the fit, whether is good or bad, is unconvincing (Vanpaemel, 2020). When the core predictions do not fully cover the data prior, the fit is persuasive. The model will be supported if all observations fall into the core predictions, and will be rejected if at least one observation falls out the core predictions.

In this study, two psychological models were reexamined by Vanpaemel’s strong testing approach. We expect to see that the models did not rule out any plausible outcomes of the experiments in original research; thus, their conclusions are not that persuasive.

# Methods

## Core predictions

The core predictions are defined as the smallest range of outcomes that cover a predetermined proportion of the prior predictive distribution. The prior predictive distribution of a model is sensitive to the prior distributions of parameters . While the likelihood is well-defined to represent the assumptions in a theory, the prior distributions are either absent in the frequentist framework or often set to be vague in the Bayesian framework. It will cause a problem for the prior predictive distribution if the parameters’ priors are not specified sensibly. Let’s imagine an illustrative example. Assume that 100 people are asked to choose between two options, and the probability that they choose the first option is . The real prior of the parameter is , while a vague prior is assigned to it. When sampling from the vague prior, most participants will be assigned an improper large , and their will also be very large. As a results, the prior predictive distribution of the proportion of participants choosing the first option will be restricted to a small range closed to 1.

Lee and Vanpaemel (2018) provided several sources and methods to develop informative priors. Here I describe the general procedure that used in the following two examples to construct informative priors. As the specification of informative prior is often study-specific, the detailed considerations will be explained with the examples.

First, the boundary and order of parameters are decided based on the theoretical assumptions and logic constraints.

Second, if datasets from pilot studies or previous research with similar experimental designs are retrievable, the model is fitted to these datasets.

Third, the truncated normal distribution is used to represent the distributions of all parameters that have been fitted in the last step, where is the posterior mean. The standard deviation varies at different values according to the scale of the posterior standard deviation. This restricts the values of the parameters would not be too extreme over its proper scale. For example, if the standard deviation of the posterior distribution is 0.02, then the values of are 0.01, 0.05, and 0.1. When the standard deviation cannot reach that high, the parameter is assumed to be uniformly distributed. The lower and upper bounds and of the priors are consistent to that in the first step unless other information has been found. For parameters that do not have any prior information, the distribution is set to be . The bounds are varied at different levels.

The sources of the prior information were restricted to the literatures published at least one year before the original papers.

The prior predictive distribution consisted of 100000 samples for each participant’s each response. The proportion of the prior predictive distribution that the core prediction should cover was set to be 99.99%.

## Data priors

Two ways were used to construct the data priors in this study. Here I briefly describe them and the detailed construction will be given with the applications.

First, for novel experimental designs, I used the practical way suggested by Vanpaemel (2020) to construct the data prior. The joint core predictions of a set of alternative established models explaining the same phenomena were used as the data prior. Compared to the core predictions of the newly proposed model, the estimates in previous studies can provide additional information for the priors.

Second, when previous datasets with similar experimental designs exist, the data priors were constructed by bootstrapping (Efron, 1992) using R package *boot* (Canty & Ripley, 2021; Davison & Hinkley, 1997). The resampling time was 10000. The observed mean and the standard deviations of the bootstrap replicates were computed for all statistics of interest. The confidence intervals were taken as the data priors, where was chosen depending on the extent to which the previous datasets can present the current dataset.

The next two sections are the two applications. For each application, the model and experiment in the original study are described first. Then, I describe the construction of core predictions and data priors. Finally, I assess the experiment and discuss the results in the original study. All the analyses and visualizations were conducted in R programming language version (R Core Team, 2021). The codes can be retrieved from .

# Example 1: Risky intertemporal choice heuristic model

The risky intertemporal choice heuristic (RITCH) model is a model of decisions for risky intertemporal choice (RIC), in which options can differ in three attributes: amount, risk and delay (Luckman et al., 2020). Traditionally, it is assumed that people make choice by computing the utility for each option and choose the one with the highest utility. However, many anomalies found in risky choice and intertemporal choices are not well handled by utility-based models (e.g., Birnbaum et al., 1999; Read et al., 2005). To account these anomalies, an alternative hypothesis has been proposed that people compare each attribute of options and combine them later when making choice (González-Vallejo, 2002; Marzilli Ericson et al., 2015). The RITCH model is based on this hypothesis.

It has been observed in risky choice and intertemporal choice that people tend to wait longer when amount or delay is increased (); when amount or probability decreased, people tend to take risk (). Both utility-based and attribute-based models can accommodate these phenomena but assume different mechanisms. To test whether people make choice by comparing utilities or attributes, Luckman and his colleagues (2020) conducted an experiment containing all kinds of RIC and manipulations to compare several utility-based and attribute-based models. Among all models, the RITCH model had the highest Bayes factors. Here I assess whether this experiment is a strong test for the RITCH model.

## The risky intertemporal choice heuristic model

Suppose the two options of a RIC are and , where , , and represent the amount, delay and probability of each option, respectively.

The RITCH model assumes that the absolute and proportional differences of each attribute between two options are considered when making choice (Equation 1-3),

where , and . s and s are the relative weight for absolute and proportional differences respectively. s are biases toward larger, sooner and safer option. Since people tend to prefer larger over smaller, sooner over later and safer over riskier options, all parameters are constrained to be non-negative.

The choice rule is

## Experimental design

100 participants were asked to complete 390 RICs without time limit. Six of the choices were check questions, each with one dominated option. The rest 384 choices included six types of RIC, which involved all combinations of trade-off among amount, risk and delay (Table x). 16 instances were created for each type and were referred as the baseline set collectively. Three additional choice sets were created by modifying the baseline set according to three key manipulations. The magnitude set was obtained by multiplying the amounts in the baseline set by 10, the immediacy set by adding 12 months to the delays in the baseline set, and the certainty set by dividing the probabilities in the baseline set by five.

10 participants chose dominated options in the check questions more than once and were excluded from the analysis. The raw data can be retrieved from <https://osf.io/bakqj/>.

Table x.

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Amount | Delay | Probability |
| R vs A |  |  |  |
| D vs A |  |  |  |
| R vs D |  |  |  |
| R vs AD |  |  |  |
| D vs AR |  |  |  |
| DR vs A |  |  |  |

## Core predictions

The logic and theory only restrict the parameters to be positive.

Previous to the current study, the authors collected a dataset involving R vs A, D vs A and R vs D choices from 72 participants (Luckman et al., 2018). These choices were directly used in the current study wherever was possible. According to the design of the previous experiment, 19 choices were filtered from the current datasets. I randomly chose 30 participants from the 90 participants that passed the check and took their responses data to develop informative priors, which is referred as the pilot dataset later.

For this pilot dataset, and are unidentifiable. The combinations of the bias parameters are for D vs A choice, for R vs A choice, and for R vs D choice. Adding any constant to these three parameters would not change the probability of choice. Thus, I defined and , and the RITCH model was degenerated as follows.

The prior distributions of and were set to the standard normal distribution and other parameters to the truncated standard normal distribution . The pilot dataset was fitted by the degenerated RITCH model at group-level and 4000 samples were taken. Table x gave the posterior means and standard deviations of all parameters. The informative priors took the form of . was set to be the posterior mean. Three sets of s were created by multiplying the posterior standard deviation by 1, 5, and 10 to represent different level of uncertainty.

Table x.

|  |  |  |
| --- | --- | --- |
| Parameter | Mean | Standard deviation |
|  | -0.73 | 0.24 |
|  | 0.47 | 0.30 |
|  | 0.001 | 0.0009 |
|  | 0.48 | 0.30 |
|  | 0.48 | 0.38 |
|  | 0.54 | 0.34 |
|  | 0.03 | 0.01 |
|  | 0.13 | 0.10 |

According to the posterior estimates of and , the probability of is high. Therefore, I specified the prior distribution for and let and . Since no information for was found, the prior of was set to be , where and

Totally 9 sets of informative priors with different uncertainty were created. The core predictions of how each manipulation affects the proportion of participants choosing option 1 in each trial were made.

## Date priors

The current study involves types of RIC (R vs AD and D vs AR) that had not been studied before, and the previous findings for other types are not always consistent. Therefore, I used three established models of RIC that were included in the original paper to simulate the data priors.

The three established models, hyperbolic discounting (HD), multiplicative hyperboloid discounting (MHD) and probability and time trade-off (PTT), are all utility-based models. The general form of them is , where is the subject value of the amount and is the discounting function of delay and probability . The probability function to choose between two options is , where and are the utilities of option 1 and 2 respectively.

In the following, I briefly introduce the forms of and in the established models, and construct four sets of informative priors of the parameters as what was done to the RITCH model. denotes the utility of an option and the odd against . Since most previous studies are deterministic, no information was found for the value of . The prior of was set to before fitting.

### *HD* *model* (Yi et al., 2006).

, .

controls how the subject value changes with the amount . In (Luckman et al., 2015), was significant below 1 on average for both risky choice and intertemporal choice, which implied diminishing sensitivity to amount as the amount increases. While in (Abdellaoui et al., 2013), the value function for intertemporal choice is likely to be linear or convex. Thus, I relaxed the bound of and set the prior distribution as . HD transforms risk into delay. In (Yi et al., 2006), was set to 35.3. This value was obtained by asking subjects to choose between a risky $1000 and a delayed $1000 (Rachlin et al., 1991). Since this paradigm is different from the current study, I assigned higher standard deviation to and the prior distribution was . Discounting rate parameters are often found right-skewed (Myerson et al., 2001), thus I assumed distributed lognormally. Also, for intertemporal choice, usually below 1 when delay was present in month (e.g., Takahashi et al., 2007; Vanderveldt et al., 2015; Yi et al., 2006), i.e., is likely to be negative. The prior of was set to . Table x shows the posterior means and standard deviation of all parameters.

Table. x.

|  |  |  |
| --- | --- | --- |
| Parameter | Mean | Standard deviation |
|  | 0.61 | 0.10 |
|  | -3.54 | 0.28 |
|  | 48.87 | 6.70 |
|  | 0.24 | 0.12 |

### *MHD model* (Vanderveldt et al., 2015)

, , .

The prior of was set the same to that of HD. and measures discounting as a function of delay and odds against receiving the outcome respectively and were assumed distributed lognormally. In (Vanderveldt et al., 2015), and , so I assumed that and . The parameters and measure sensitivity to delays and risks, respectively. Assume these two parameters are smaller 1 representing diminishing sensitivity, which is expected (Gonzalez & Wu, 1999; Zauberman et al., 2009). The priors of and were set to . measures how sensitivity to risk changes with amount. The peanuts effect that people tend to take risk when the amount is small implies (Weber & Chapman, 2005). In (Myerson et al., 2011), was very closed to 0. Thus, I assumed that . Table x shows the posterior means and standard deviation of all parameters.

Table x.

|  |  |  |
| --- | --- | --- |
| Parameter | Mean | Standard deviation |
|  | 0.51 | 0.13 |
|  | 0.37 | 0.18 |
|  | -2.00 | 0.89 |
|  | 0.67 | 0.49 |
|  | 0.25 | 0.19 |
|  | 0.23 | 0.24 |
|  | 0.31 | 0.26 |

### *PTT model* (Baucells & Heukamp, 2010).

, .

(Baucells & Heukamp, 2010)was done in EURO and the range of amounts was much narrower than the current study. To what extent its estimates can applied to the current study is unknow, therefore only qualitative information was taken here.

and control sensitivity to amount. The sensitivity to amount diminishes as amount increases when while the sensitivity increases when . In (Baucells & Heukamp, 2010)**,** implies sensitivity to amount diminishes slightly, so I set . controls how the subjective value is proportionally affected by the change in amount, i.e., the elasticity of the subjective value. In (Baucells & Heukamp, 2010)**,** suggests the subjective value is almost inelastic, so I set . PTT transfers delay into risk. The transfer rate decreases as the absolute value of amount increases. I set since its value does not have qualitative influence on the discounting function. measures subproportionality, the degree of overweighting small probabilities and underweighting larger probabilities. In (Baucells & Heukamp, 2010), , so I set . Table x shows the posterior means and standard deviation of all parameters.

Table x

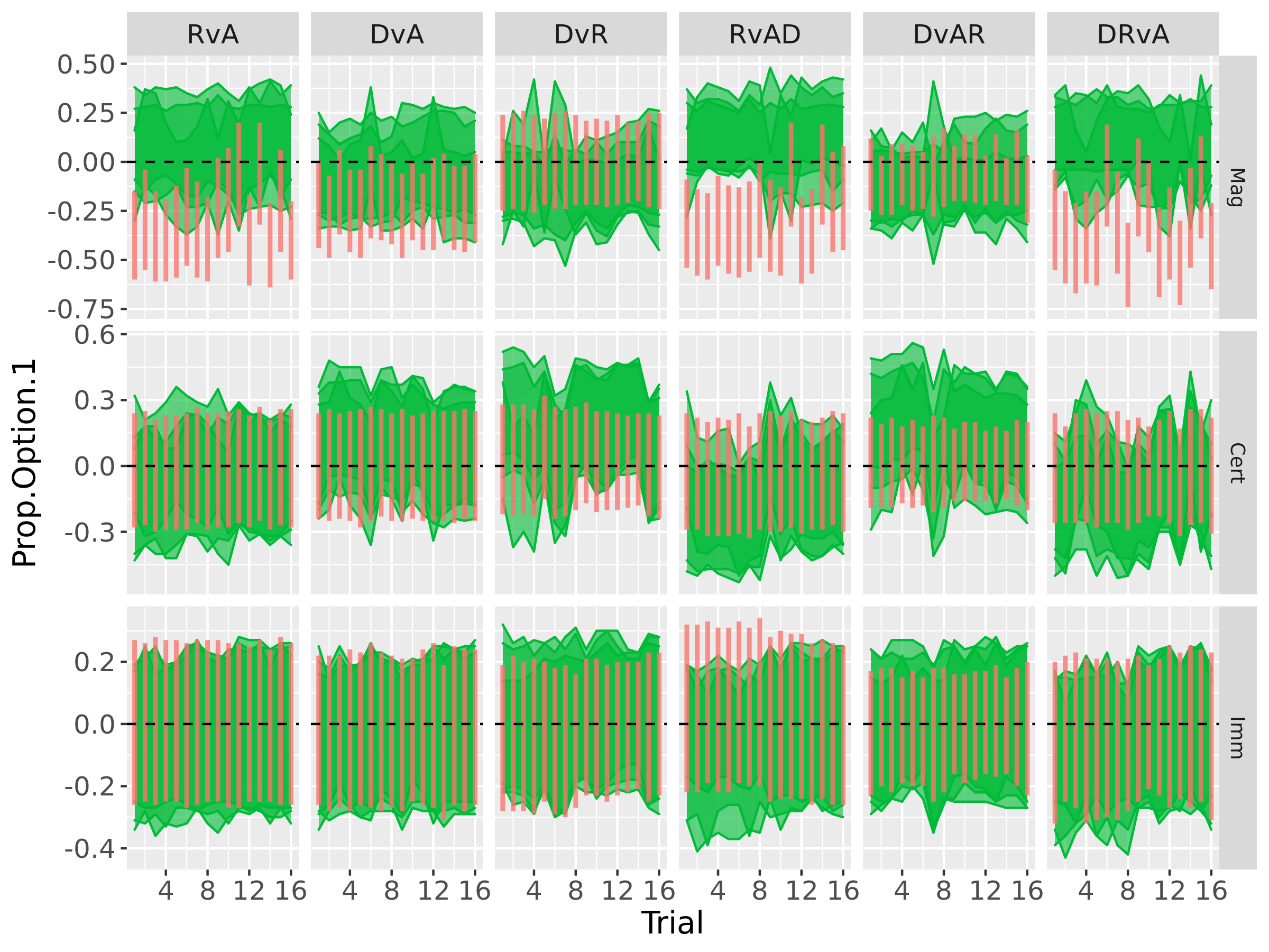
|  |  |  |
| --- | --- | --- |
| Parameter | Mean | Standard deviation |
|  | 0.08 | 0.04 |
|  | 0.63 | 0.10 |
|  | 0.88 | 0.10 |
|  | 1.38 | 0.43 |
|  | 0.93 | 0.41 |

For the established models, informative priors were created by taking the posterior mean as. Since all parameters here can be identified with the pilot dataset, only three sets of informative priors were obtained. The joint of the core predictions of HD, MHD and PTT is taken as the data priors.

## Strong testing

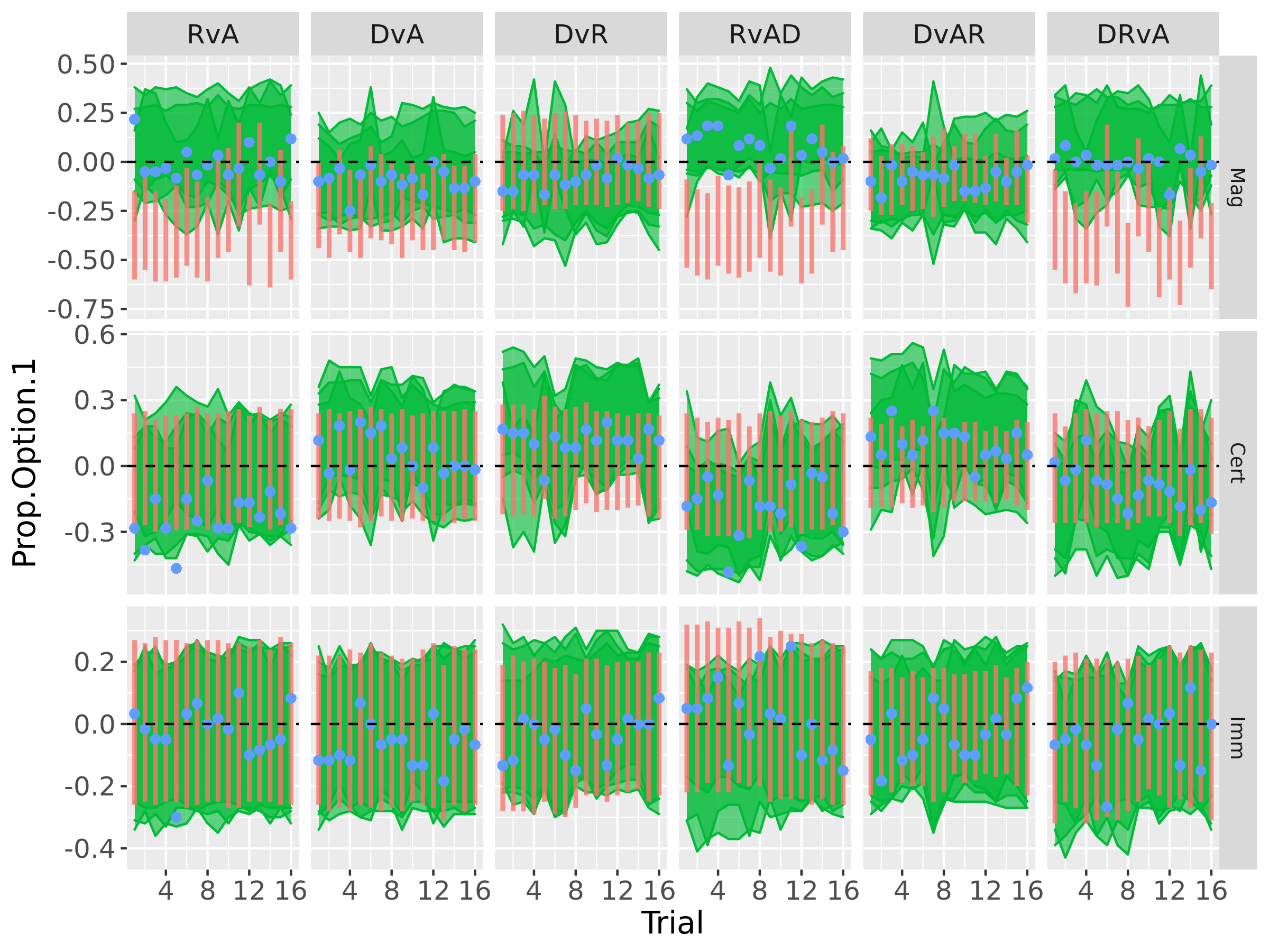
Figure x shows the data priors and the core predictions when s are the posterior standard deviations and . The core predictions do not cover all the data priors, especially for magnitude manipulation. Thus, the experiment was a strong test for the RITCH model.

Figure x.



However, when comparing to the behavioral data of the rest 60 participants, the RITCH model did not conform to observation (Figure x). This means the RITCH model was rejected by the observations. In the original paper, similar manipulation effects were predicted by the RITCH model (see Figure 3-5 in Luckman et al., 2020), though the posterior estimates are very different from here (see Table 7 Luckman et al., 2020).

Figure x



# Example 2: Interference model

The second example comes from the domain of visual working memory (WM). One of the most robust and general phenomena for WM is its limited capacity, the accuracy of response decreases as the set size of memory items increases (Oberauer et al., 2018). The interference model (IM) for visual WM assumes that the limited capacity is due to the mutual interference among representations of stimuli (Oberauer & Lin, 2017). A key assumption of the IM is that the strength of interference from a non-target item depends on the difference between the cue feature of it and that of the target item. This relationship between intrusion of non-target items and the differences of cues have been observed in previous studies about visual WM (Bays, 2016; Rerko et al., 2014). Previous models such as the slots model (Zhang & Luck, 2008) and the resource model (Bays & Husain, 2008) cannot accommodate this phenomena.

Oberauer and Lin (2017) conducted four continuous-reproduction experiments to compare the IM to slots and resource models. The IM provided the lowest AIC in most cases and only slightly worse than a version of resource model in one experiment.

Experiment 1-3 used location as retrieval cue and color as memory contents, which was commonly used in previous visual WM studies. Here I focus Experiment 4 which involved novel retrieval cues and memory contents. I assess whether Experiment 4 was a strong test for the IM.

## The interference model

In the following, I briefly describe the assumptions of the IM. More details can be found in (Oberauer & Lin, 2017).

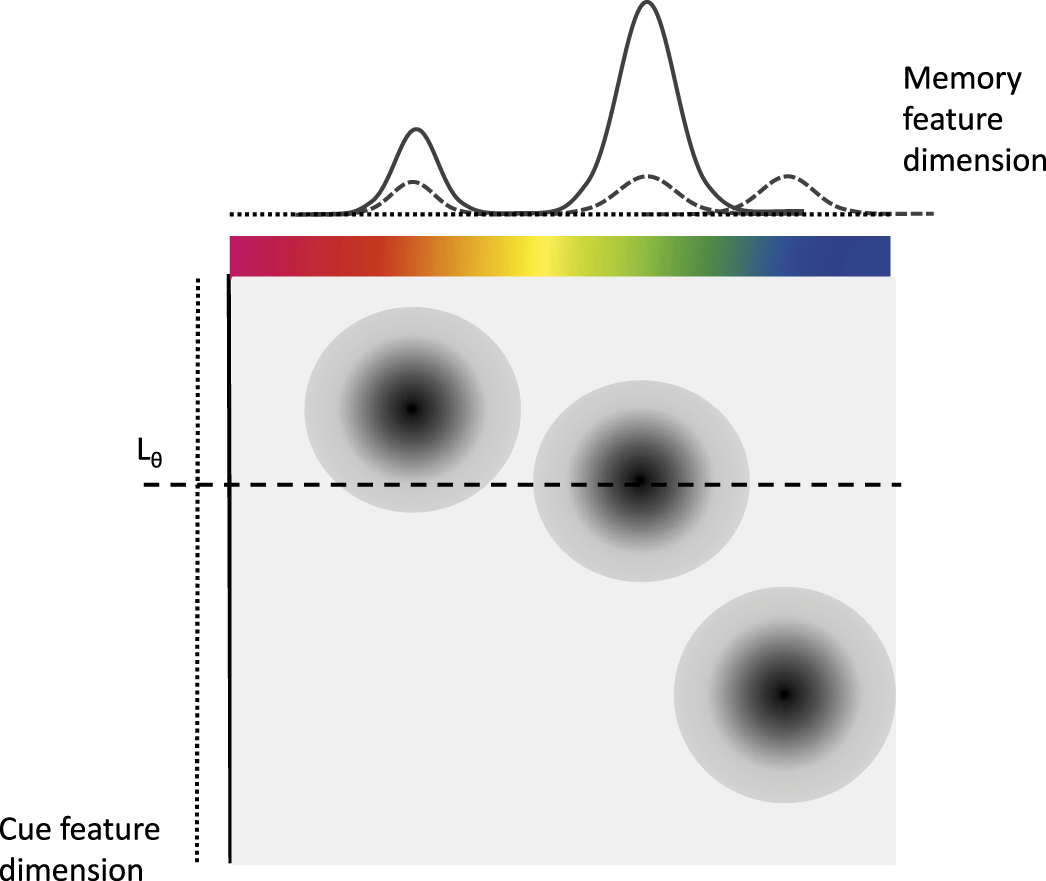
### *Memory contents and cues*

In visual WM tasks, memory contents are often features varying on a continuous dimension (e.g., colors or orientations) and the retrieval cue is often spatial location. Therefore, the IM represents the memory contents and retrieval cues as continuous dimensions, which are referred as memory feature dimension and cue feature dimension respectively. Suppose the number of the memory items in each trial is . The values of memory contents are represented by , and the values of the cues by . The representations of memory contents and cues are both unimodal distributions on the memory and cue feature dimensions respectively.

### *Bindings*

Encoding involves binding the representations of memory contents to the representations of corresponding cues. The binding strengths are conceptualized as bivariate distributions in a 2-dimensional space which referred as the binding space (Figure x). When a cue is given, an activation distribution on the memory feature dimension is generated by the representation of the cue on the cue feature dimension through the bindings. The precision of the bindings is limited, so the distributions of bindings may overlap along the memory and cue feature dimensions. At retrieval, the representations of memory contents of non-target items which are closed to the target item along the cue feature dimension may also be reactivated by the cue. The amplitude of the activation of a memory content’s representation depends on the marginal distribution of the binding-strength distribution at the value of the present cue.

Figure x O et al., 2017.



The representation of a memory content is modeled by a von Mises distribution, a Gaussian distribution on the circle, since memory contents involved in the current study are circular. The density function of the von Mises is

where is the memory content of the target item and is the order 0 modified Bessel function. controls the precision of the representation.

The representation of a cue and the binding space are not explicitly modeled in the IM. The activation of cue-based retrieval is created by weighted sums of each memory content’s representation at retrieval (see next section).

### *Activations*

The IM assumes that the probability of a candidate response being chosen depends on its activation at retrieval. Given the cue of the target item, the activation distribution over response candidates generated at retrieval is a weighted sum of three components (Equation x)

The first component is the cue-based activation. It is the weighted sum of the representations of these reactivated items, where the weight exponentially decreases with the distance between the cue and item ’s cue on the cue feature dimension, (Equation x).

The spatial gradient parameter controls the speed of the decrease of the weight according to distance.

The second component is the sum of representations of all memory contents in the present memory sets (Equation x), representing activation from cue-independent information of memory contents.

The third component is the background noise. It represents activation irrelevant to the present memory items and is modeled by a uniform distribution. The IM assumes each item is accompanied with the same amount of background noise. The third term has the form of Equation x.

### *Attention*

An additional assumption is that the WM system has a focus of attention which can hold one memory content and its cues. The representation of the focused item has higher precision. Formally, this assumption can be expressed by an additional component with higher precision:

where is the attended item and .

When the target item is attended (represented by ), the activation distribution of response becomes

The contributions of the cue-independent component and background noise are reduced by .

The complete form of the response activation is

where is Equation x. When no specific item is expected to receive attention, is assumed to be .

### *Multiple cues*

Experiment 4 used color and location of items as retrieval cue, so a further assumption is needed to accommodate multiple cues in the cue-based activation.

Suppose and is the location and color of the target item. When both color and location are probed, the cue-based component is

and are the spatial gradient parameter for the color dimension and location dimension respectively. controls the weight of the activation of the location cue.

### *Response rule*

Suppose the number of the response options is and the options are . Given the retrieval cue (or ), the probability of responding is

Since this probability would not change by multiplying a constant to and simultaneously, is set to 1.

## Experimental design

21 participants were asked to do a continuous-reproduction task. In each trial, six colored discs were presented, each with a rectangular gap, for a short while and then disappear. After a short retention interval, one disc was randomly chosen and participants were asked to reproduce the orientation of its gap. The colors of six discs were randomly selected from nine equidistant colors on a color circle. The color circle contained 360 colors which were created from the CIE L\*a\*b color model and equidistant on this circle. The locations of six discs were randomly selected from 13 locations equidistantly spaced along an invisible circle. The orientations of six gaps were randomly selected from to . The disc could be probed by its color, location or both. A disc with the same color of the probed disc was present at the screen center in the color-cue condition; a black disc appeared at the probed disc’s location in the location-cue condition; a disc with the same color and location of the probed disc was presented in the both-cue condition. The conditions appeared randomly for each trial and participants did not know which will be until the cue was given. Each participant took a total of 300 trials, 100 for each condition. The sequence of three conditions were randomized for each participant. The raw dataset is available on https://osf.io/wgqd5/.

## Core predictions

Seven free parameters are in the IM for Experiment 4. Based on the logic and theory, the boundary and order of parameters are

Let , then . The informative prior was developed for to maintain the order between and .

Experiment 1 in the original paper used the same paradigm as Experiment 4, except that the memory contents are colors and only spatial location was used as retrieval cue. I fitted the Bayesian version of the IM to the dataset of Experiment 1 to develop the informative priors for Experiment 4. The prior distributions of and were set as truncated standard normal distribution , and the prior of was set to be . Table x gave the posterior means and standard deviations of all parameters. These posterior estimates are closed to the estimates in the original research (see Table 1 in Oberauer and Lin, 2017).

Table x.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Parameter | Mean | Standard deviation | set 1 | set 2 | set 3 |
| A |  |  |  |  |  |
| B |  |  |  |  |  |
| R |  |  |  |  |  |
| S |  |  |  |  |  |
| kappa |  |  |  |  |  |
| delta |  |  |  |  |  |

Since and are not closed related to the stimuli, I assumed the priors of them followed Experiment 1’s estimates.

In (Bays et al., 2011), the reproduction of colors and orientations were similar in terms of precision, standard deviation of response errors and the proportion of misreporting errors when using location as retrieval cues. Experiment 3 of (Pertzov and Husain, 2014) used the color and location as cues to reproduce orientations, and the performance when color and location as cues was similar. Therefore, I also assumed the precision parameters and followed the same prior distributions as Experiment 1. I made further assumptions in regard of their upper bounds based on previous studies. When set size is one, the probability of response in the IM is proportional to . Since and were assumed to be small here, . I further approximated by and it was expected that . In (Bays et al., 2011), when only one item was present, the standard deviation of the von Mises distribution for response errors for orientation was . I assumed that the range of the standard deviation to be . Transforming this range into precision, the range of is likely to be . It is closed to of in Experiment 1. I further relaxed the upper bound to 30, thus . No extra information was found for , thus the upper bound was set to .

For and , the informative prior took the posterior means as . Three sets of were specified by multiplying the posterior standard deviations by 1, 5, 10. The upper bounds were specified as assumed.

For and , no available dataset informs their distributions. was set to follow . In (Bays, 2016), where location was retrieval cues and color, orientation and direction were memory contents, the swap errors was sensitive to the radial distance when it was between 1 and 2. When , , which makes the influence of distance negligible. Therefore, and were set to follow .

There sets of informative priors with different uncertainty were obtained. The core predictions of response errors, average deviations between response and non-target items and average deviations between response and non-target item at each distance were made.

## Data priors

The data priors were constructed from the dataset of Experiment 1 using bootstrapping. The number of participants of Experiment 1 is 19. The mean absolute error of target-centered responses, average nontarget-centered responses and nontarget-centered responses at each distance on the location dimension were computed for each participant. Since Experiment 1 did not use color as the retrieval cue, I approximated the nontarget-centered responses at each distance along the color dimension by that along the location dimension. The six location distances were . The four color distances were . Responses centered at were approximated by the mean of responses centered at and , responses centered at by responses centered at , and the mean of responses centered at the rest color distances by the mean of responses centered at the rest location distances.

The obtained bootstrap confidence intervals were taken as data priors for all three cue conditions.

## Strong testing

Fig. x showed the comparison between data prior and core predictions with s set 3. The core predictions excluded plausible results. Therefore, Experiment 4 was a strong test for the IM.

# Discussion

**Two frameworks**

Falsification Verification

Qualification difference of two models = falsifying one model, but verifying another one, but the strength is unknown. 🡪 falsify other model does not increase the credibility of the new one

# References

# Appendix A