# 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 core predictions are defined as the smallest range of outcomes that cover a predetermined proportion of the prior predictive distribution. 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, we will apply Vanpaemel’s approach to assess the pervasiveness of a good fit for psychological models. Specifically, two previous studies are reexamined. We expect to see that some models did not rule out any plausible outcome; thus, their conclusions are not that persuasive.

The structure of the rest of this thesis is as follows. …

# Methods

**Core predictions**

The core prediction is based on the prior predictive distribution of a model which is sensitive to the prior distributions of the model parameters. While the likelihood is well-defined to represent the assumptions in a theory, the prior distributions are either absent in the frequentist framework (the interference model case) or often set to be vague in the Bayesian framework (the RITCH case). 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 references of the original papers to mimic the real prior knowledge of the authors.

In this study, the proportion of the prior prediction that the core prediction should cover was set to be 99.99%.

**Data priors**

In this study, two ways are used to construct the data priors.

First, the practical way suggested by Vanpaemel (2020) to construct the data prior, that is, taking the joint core predictions of a set of alternative established models of the same behavior as the data prior. Compared to the core predictions of the newly proposed model, the previous estimates provide an additional source for informative priors.

Second, when the datasets with similar experimental designs exist, the data priors can be constructed by bootstrapping. Sample size is set to the same to the original sample size, and resampling 100000 times. The 99.99% quantiles are used as the data priors.

All codes are provided in . The requirements of systems and the version of R packages are given in the README.md.

The next two sections are the applications of Vanpaemel’s strong theory testing approach on two previous studies. The model and experiment in the original study are described first. Then, the informative priors are specified and core predictions are made. Then, the data priors are constructed and compared to core predictions. Finally, the core predictions are compared to the observed data. The two studies were chosen because novel models were proposed and the original datasets are retrievable.

# 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 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. Four sets of s were created by multiplying the posterior standard deviation by 1, 5, 10 and 20 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 16 sets of informative priors with different uncertainty were created. The prior predictions contained 20000 samples of 100 participants’ responses. 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

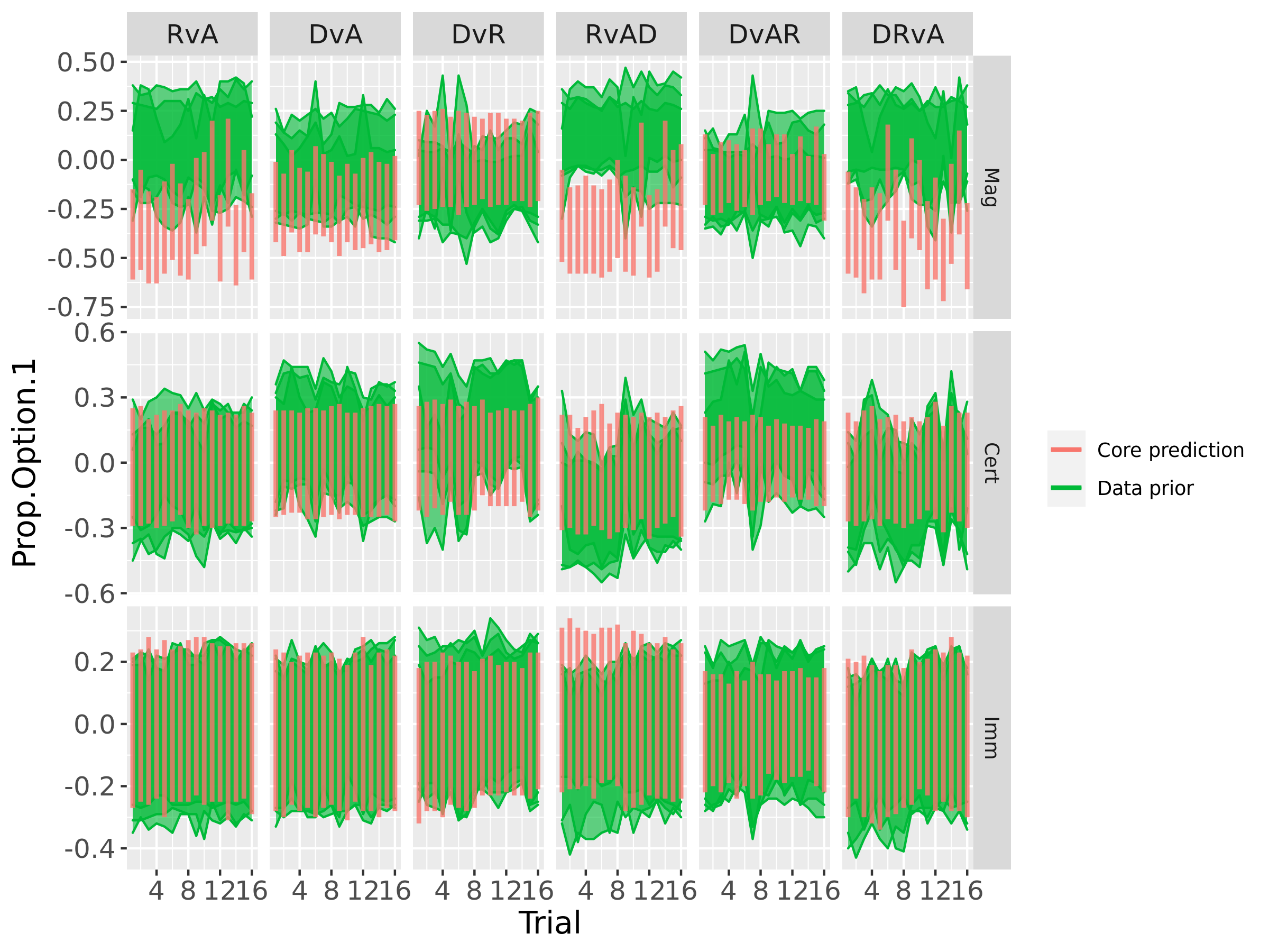
|  |  |  |
| --- | --- | --- |
| Parameter | Mean | Standard deviation |
|  | 0.078879 | 0.043773 |
|  | 0.631985 | 0.098565 |
|  | 0.889788 | 0.089373 |
|  | 1.383843 | 0.431017 |
|  | 0.942771 | 0.412217 |

The core predictions of HD, MHD and PTT were made as for the RITCH model. The joint of them is taken as the data priors.

## Strong testing

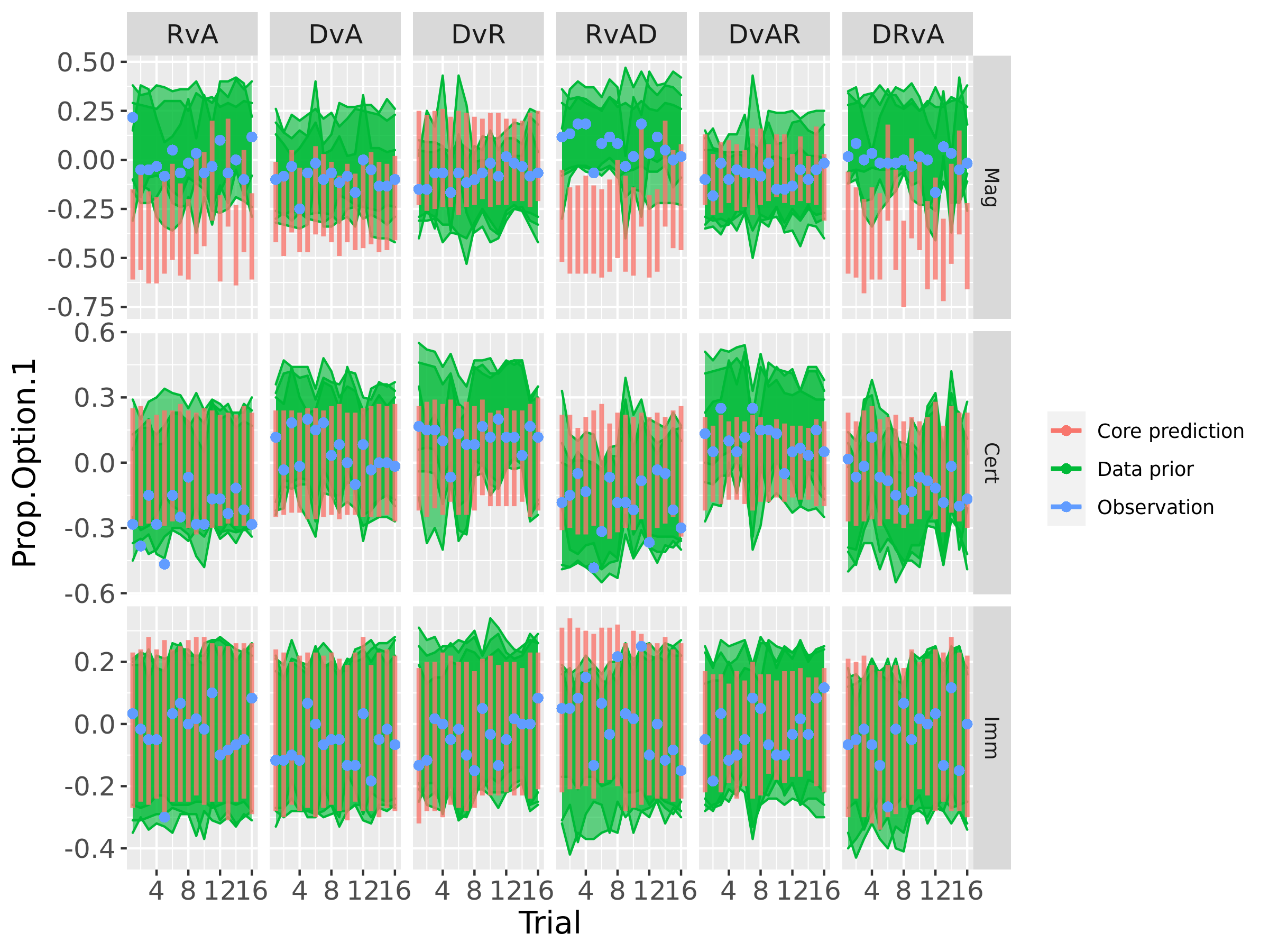
Figure x shows the data priors and the core predictions when s are the posterior standard deviations and . The core predictions do not cover all the data priors, especially for magnitude manipulation. Thus, the experiment was a strong test for the RITCH model.

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 of visual working memory

The second example comes from the domain of working memory (WM). One of the most robust and general phenomena for working memory (WM) is its limited capacity (Oberauer et al., 2018). Explanations include slots and resource. One of explanations of this phenomena is that the representations of stimuli in WM mutually interfere. Oberauer and Lin (2017) proposed an interference model (IM) for visual WM and conducted four experiments to compare IM with slot and resource models. The AIC supported IM.

## The interference model of visual WM

The IM is built on the following assumptions.

*Memory contents and retrieval cues*

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

*Bindings*

Binding between context and feature: a continuous 2-dimensional binding space

describe representations of individual features as a von-Mises distribution on the feature dimension.

The precision of feature memory is governed by the concentration parameter k

Binding each feature to context generates a bivariate distribution of binding strengths in binding space

limits the precision with which a feature can be reactivated, given its context.

*Activations*

activation comes from 3 sources in memory: cue-based retrieval using context cues, context-independent memory for relevant contents, and noise;

The activation distribution over response candidates generated at retrieval is a weighted sum of three components:

* first term: given context cue, conditioned on location of target . Parameter reflects the sensitivity to context cue, i.e., the distance between the target item and the item.

+

+ a further assumption to accommodate multiple context cue: independent sensitivity and

* second: which feature values are in the current memory set, independent from context

+

* third: background noise

+

*Attention*

1 memory object and its context can be held in the focus of attention, where it is represented with higher precision, and partly shielded against interference.

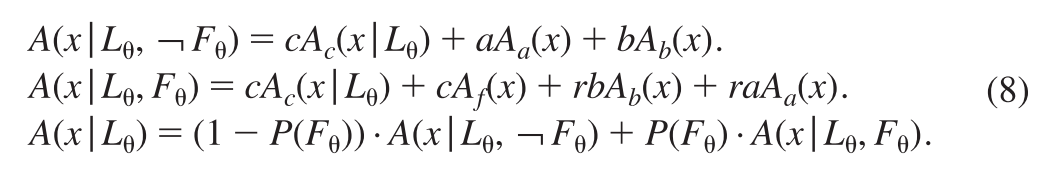
: the relative strength of activation of each candidate *x* at retrieval

The activation of non-cue reduced by parameter .

When pay attention to an item , higher precision is given. Formulize as .

When is the target of this trial, activation becomes

The full model is



The possibility an item receive attention was used to incorporate the manipulation of attention. Without cueing, P(F)=1/n.

*Response rule*

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

## Experimental design

Four experiments using the continuous-reproduction paradigm were conducted in the original research to evaluate this model. The raw dataset is available on https://osf.io/wgqd5/.

The first experiment was similar to what (Rerko et al., 2014) observed and not surprise to the authors but in fact the motivation to incorporate distance into model. The second and third were about attention assumption. The fourth experiment tested the generalization to another context-cue and may be surprise and worth to be tested. Experiment 4 is the key experiment to test this assumption. Here I focus the fourth experiment.

21 young adults were included. Participants were asked to do a continuous-reproduction task. In each trial, six colored discs were presented, each with a rectangular 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 were equidistant on the 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 . After a short retention interval, one disc was randomly chosen and participants were asked to reproduce its orientation by moving the mouse cursor; the orientation always pointing toward the mouse cursor. Three conditions indicated the chosen disc: color-cue, location-cue, and both-cue. In color-cue condition, a chosen disc was presented and the target with the same color to the probe-cue. In location-cue condition, the target was the one at the same location. In both-cue condition, the probe-cue had the same color and location with the targe. 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.

If Experiment 4 was designed effectively, then the IM, with plausible prior distributions, would predict the response error would be more disperse as the distance between target and non-target along the cue dimension increasing and these should exclude some plausible observation.

## 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 paradigm as Experiment 4, except that the memory contents are colors and only spatial locations as retrieval cues. In (Bays et al., 2011), the reproduction of colors and orientations were similar in terms of precision, dispersion 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 assumed that and followed the same prior distributions and fitted the dataset of Experiment 1 with the Bayesian version of the IM 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).

I made further assumptions in regard of the upper bounds of the parameters. Since the original estimates of and are much smaller than 1 across first three experiments, I assumed that and have smaller proportion than . Thus, the upper bounds of and were set to 1. 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, the upper bound of and were set to 20. 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 . I further relaxed the upper bound to 30, thus . No extra information was found for , thus the upper bound was set to .

For each parameter, the informative prior took the posterior means as , 0 as L and the assumed upper bound as U. Three sets of were specified to make the standard deviations of these truncated normal distributions equal to 5, 10, 50, 100 multiplying the scales of the parameters.

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

Table x.

The dataset of Experiment 1 did not provide information for , but it was found that location has special role for visual WM. Therefore, was set to to assign more weight for location cues in the both-cue condition.

Now 6 sets of informative priors with different uncertainty were created. 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. Fig. x shows the core predictions when the 10\*scale informative priors were used.

## 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 response errors, average deviation from non-target items and deviations for each location distance were calculated for each participants, and the population statistics were calculated by bootstrapping with 10000 resampling times. Mean+\_5\*std

The deviations for each color distance were approximated by the location distance, therefore I increased the range of CI into mean+\_10\*std.

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.

## Comparing with observations

# 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

Abdellaoui, M., Bleichrodt, H., l’Haridon, O., & Paraschiv, C. (2013). Is There One Unifying Concept of Utility?An Experimental Comparison of Utility Under Risk and Utility Over Time. *Management Science*, *59*(9), 2153–2169. https://doi.org/10.1287/mnsc.1120.1690

Baucells, M., & Heukamp, F. H. (2010). Common ratio using delay. *Theory and Decision*, *68*(1–2), 149–158. https://doi.org/10.1007/s11238-008-9130-2

Birnbaum, null, Patton, null, & Lott, null. (1999). Evidence against Rank-Dependent Utility Theories: Tests of Cumulative Independence, Interval Independence, Stochastic Dominance, and Transitivity. *Organizational Behavior and Human Decision Processes*, *77*(1), 44–83. https://doi.org/10.1006/obhd.1998.2816

Gonzalez, R., & Wu, G. (1999). On the Shape of the Probability Weighting Function. *Cognitive Psychology*, *38*(1), 129–166. https://doi.org/10.1006/cogp.1998.0710

González-Vallejo, C. (2002). Making trade-offs: A probabilistic and context-sensitive model of choice behavior. *Psychological Review*, *109*(1), 137–155. https://doi.org/10.1037/0033-295X.109.1.137

Lee, M. D., & Vanpaemel, W. (2018). Determining informative priors for cognitive models. *Psychonomic Bulletin & Review*, *25*(1), 114–127. https://doi.org/10.3758/s13423-017-1238-3

Luckman, A., Donkin, C., & Newell, B. (2015). *Exploring the Concept of Utility: Are Separate Value Functions required for Risky and Inter-temporal Choice?*

Luckman, A., Donkin, C., & Newell, B. R. (2018). Can a single model account for both risky choices and inter-temporal choices? Testing the assumptions underlying models of risky inter-temporal choice. *Psychonomic Bulletin and Review*, *25*(2), 785–792. https://doi.org/10.3758/s13423-017-1330-8

Luckman, A., Donkin, C., & Newell, B. R. (2020). An evaluation and comparison of models of risky intertemporal choice. *Psychological Review*, *127*(6), 1097–1138. https://doi.org/10.1037/rev0000223

Marzilli Ericson, K. M., White, J. M., Laibson, D., & Cohen, J. D. (2015). Money Earlier or Later? Simple Heuristics Explain Intertemporal Choices Better Than Delay Discounting Does. *Psychological Science*, *26*(6), 826–833. https://doi.org/10.1177/0956797615572232

McClelland, J. L. (2009). The place of modeling in cognitive science. *Topics in Cognitive Science*, *1*(1), 11–38. https://doi.org/10.1111/j.1756-8765.2008.01003.x

Meehl, P. E. (1990). Appraising and Amending Theories: The Strategy of Lakatosian Defense and Two Principles that Warrant It. *Psychological Inquiry*, *1*(2), 108–141. https://doi.org/10.1207/s15327965pli0102\_1

Myerson, J., Green, L., & Morris, J. (2011). Modeling the Effect of Reward Amount on Probability Discounting. *Journal of the Experimental Analysis of Behavior*, *95*(2), 175–187. https://doi.org/10.1901/jeab.2011.95-175

Myerson, J., Green, L., & Warusawitharana, M. (2001). Area Under the Curve as a Measure of Discounting. *Journal of the Experimental Analysis of Behavior*, *76*(2), 235–243. https://doi.org/10.1901/jeab.2001.76-235

Oberauer, K., Lewandowsky, S., Awh, E., Brown, G. D. A., Conway, A., Cowan, N., Donkin, C., Farrell, S., Hitch, G. J., Hurlstone, M. J., Ma, W. J., Morey, C. C., Nee, D. E., Schweppe, J., Vergauwe, E., & Ward, G. (2018). Benchmarks for models of short-term and working memory. *Psychological Bulletin*, *144*(9), 885–958. Scopus. https://doi.org/10.1037/bul0000153

Popper, K. R. (1959). *The logic of scientific discovery* (p. 480). Basic Books.

Rachlin, H., Raineri, A., & Cross, D. (1991). Subjective probability and delay. *Journal of the Experimental Analysis of Behavior*, *55*(2), 233–244. https://doi.org/10.1901/jeab.1991.55-233

Read, D., Frederick, S., Orsel, B., & Rahman, J. (2005). Four Score and Seven Years from Now: The Date/Delay Effect in Temporal Discounting. *Management Science*, *51*(9), 1326–1335. https://doi.org/10.1287/mnsc.1050.0412

Rerko, L., Oberauer, K., & Lin, H.-Y. (2014). Spatial Transposition Gradients in Visual Working Memory. *Quarterly Journal of Experimental Psychology*, *67*(1), 3–15. https://doi.org/10.1080/17470218.2013.789543

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*(2), 358–367. APA PsycArticles®. https://doi.org/10.1037/0033-295X.107.2.358

Takahashi, T., Ikeda, K., & Hasegawa, T. (2007). A hyperbolic decay of subjective probability of obtaining delayed rewards. *Behavioral and Brain Functions*, *3*. https://doi.org/10.1186/1744-9081-3-52

Vanderveldt, A., Green, L., & Myerson, J. (2015). Discounting of Monetary Rewards that are Both Delayed and Probabilistic: Delay and Probability Combine Multiplicatively, not Additively. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *41*(1), 148–162. https://doi.org/10.1037/xlm0000029

Vanpaemel, W. (2020). Strong theory testing using the prior predictive and the data prior. *Psychological Review*, *127*(1), 136–145. https://doi.org/10.1037/rev0000167

Veksler, V. D., Myers, C. W., & Gluck, K. A. (2015). Model flexibility analysis. *Psychological Review*, *122*(4), 755–769. APA PsycArticles®. https://doi.org/10.1037/a0039657

Weber, B. J., & Chapman, G. B. (2005). Playing for peanuts: Why is risk seeking more common for low-stakes gambles? *Organizational Behavior and Human Decision Processes*, *97*(1), 31–46. https://doi.org/10.1016/j.obhdp.2005.03.001

Yi, R., de la Piedad, X., & Bickel, W. K. (2006). The combined effects of delay and probability in discounting. *Behavioural Processes*, *73*(2), 149–155. https://doi.org/10.1016/j.beproc.2006.05.001

Zauberman, G., Kim, B. K., Malkoc, S. A., & Bettman, J. R. (2009). Discounting Time and Time Discounting: Subjective Time Perception and Intertemporal Preferences. *JOURNAL OF MARKETING RESEARCH*, 14.

# Appendix A

Since both the core prediction and data prior rely on subjective choices,

For a psychological model, the parameters should reflect our existing knowledge about the phenomena. It reflects the possible prediction. However, this possible prediction depends the specification of prior distribution.

Denote the prior distribution of parameters as and the likelihood of the model as , where is the vector of parameters and is the vector of observed variables. The prior predictive distribution is . However, when a theory is proposed, its assumptions are made based on the empirical evidence which already provide constrains or information to the parameters. Otherwise, why do psychologists ever include that parameter? Neglecting this information unconsciously or intendedly would increase the flexibility of the model and make the core prediction broader than it should be. Therefore, in order to make proper strong theory testing, the prior should be specified to reflect the prior information in hand.

what is considered plausible depends on the exact research design, instructions, and so on.

one should set a data prior in such a way that it is sensitive to the theory under consideration. 🡪 the data prior should come from the same information as the theory.

As the psychological theories are often lack of generalizability, the experiments that used to test a theory are typically closely related to the empirical evidence motivating the theory.

For experiments that have similar previous studies, the results are combined to develop data priors. For experiments that no similar design, use other established models to predict the results.

For the example experiments, only details that will be reflected in the predictions of models are included.

I started from explaining the model of interest and the experimental design. Then I review previous empirical evidence and data sets to inform prior distributions of the model’s parameters and set the data prior. Finally, I made the strong theory testing and compared the result to the claim in the original study.

1. Access to contents of visual WM relies on cue-based retrieval. The core assumption of the IM, which distinguishes it from most other models of visual WM, is that access to individual memory contents depends on cue-based retrieval, which gives rise to interference.
2. Both memory contents and their potential cues are represented in a distributed fashion, such that their similarity is reflected in the degree of overlap between representations.
3. Memory performance is limited by interference, which can arise from multiple sources.

It involves a standard paradigm and rich empirical evidence, and several previous data sets using the same paradigm are retrievable. These made the specification of the two ingredients more trackable and precise.

The IM makes a unique prediction about people’s tendency to report the feature of a nontarget: That tendency should increase with the similarity between the retrieval-relevant context of the target—that is, the context presented as the retrieval cue—and the context of the nontarget.

The viewing distance was 50 cm.

with diameter

Here, by analysing the frequency of these ‘misreporting’ errors within and across feature dimensions,

we confirm that they are the result of misbinding features held in independent memory

stores, consistent with the storage of visual features in separate sensory representations

Instead these results support the proposal of Wheeler and Treisman (2002) that visual features in different dimensions are maintained in independent memory stores.

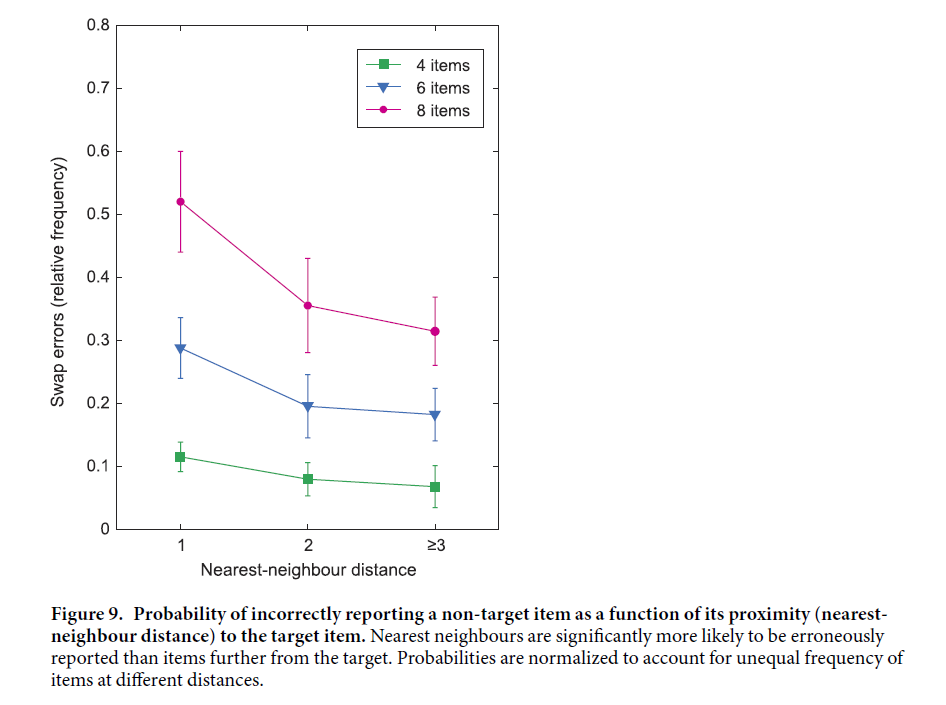
‘binding errors’ in a change detection task: errors caused by incorrectly combining in

memory features that belong to different objects

When only one item was present in the memory array (low-load), subjects recalled both

color and orientation with considerable precision. The precision of recall did not significantly differ between feature dimensions. 🡪 assume the kappa and delta similar to exp1.

Estimates of parameters are stable across experiments except for and at Experiment 3. Since Experiment 2 and 3 were intended to manipulate attention and thus different from Experiment 4, I used the estimates of Experiment 1 as informed priors for and , which are not directly related to cues and features. I used truncated normal distributions as the prior distributions. The mean is set to median since it is more stable, and the standard deviation is set to SD of estimates.



Two questions are debating: whether the comparison is between utility or attributes? And whether the delay and the risky are translatable?

attribute-comparison and untranslatable of delay and risk.

The authors reviewed effects related to three manipulations than have been robustly observed in previous risky and intertemporal choice studies and designed their experiment based on these effects (Table \*; see Luckman et al., 2020 for more details).

|  |  |
| --- | --- |
| Effect | Explanation |
| Magnitude effect. | A standard intertemporal choice (e.g., $50 now vs. $100 in 6 months) if you increase the magnitude of the outcomes of both options by multiplying them by a common multiplier, people become much more likely to choose the larger later option. |
| Peanut effect | a standard risky choice between a smaller but safer reward, and a larger but riskier reward (e.g., $50 for certain or $100 with probability 0.5), multiplying both amounts by a common multiplier (e.g., 10) has generally been shown to make people more likely to choose the safer option. |
| certainty effect (Common ratio effect) | When the probabilities of the outcomes presented in a risky choice are reduced by a common ratio, people are more risk seeking (by the same amount). |
| Immediacy effect (common difference effect) | increasing delays by a common amount, which leads to an increase in preference for the larger more delayed outcome. |

Combined certainty and immediacy. adding risk to an intertemporal choice, had a similar effect to adding a common delay in that participants became more likely to choose the larger later option. adding a common delay to a risky choice has a similar effect to increasing the risk, in that participants became more likely to choose the larger riskier option.

Prefer delay than uncertainty. a recent study found that people prefer delaying an outcome to taking a risk when the two are in direct competition, despite giving the risky and delayed outcome similar values in isolation (Luckman et al., 2017).