# 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. For example, if the standard deviation of the posterior distribution is 0.02, then the values of are 0.01, 0.05, 0.1, and 0.5. 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 if no other consideration is involved. For parameters that do not have any prior information, the distribution is set to be . The bounds are varied at different levels.

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

**Data priors**

In this study, I use two ways 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. Compare to the core predictions of the newly proposed model,

**Code and data availability**

All codes are provided in . The requirements of systems and R packages, the explanation of the usage of code files, and the links to retrieve all datasets used in this thesis are given in the README.md.

First describe the model and experiment in the original study, and then make prior prediction, then data prior and finally give result. For the example experiments, only details that will be reflected in the predictions of models are included.

The two studies were chosen because they proposed new models and tested it with novel experiments.

# Example 1: Interference model of visual working memory

The first 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). Different theories have been proposed to explain this phenomenon. One of these theories is that the representations of stimuli in WM mutually interfere.

Oberauer and Lin introduced an interference model for visual WM and conducted 4 experiments using the continuous-reproduction paradigm to evaluate this model (Oberauer & Lin, 2017). Here we take the fourth experiment as an example. In the original study, four experiments were conducted. 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. 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.

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.

**The interference model of visual WM**

The interference model (IM) incorporates the following assumptions: (a) 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. (b) 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. (c) Memory performance is limited by interference, which can arise from multiple sources.

In the IM we represent context and content as continuous dimensions, which we refer to as context dimension and feature dimension, respectively.

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.

The probability of choosing each response x out of N response options

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

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

+

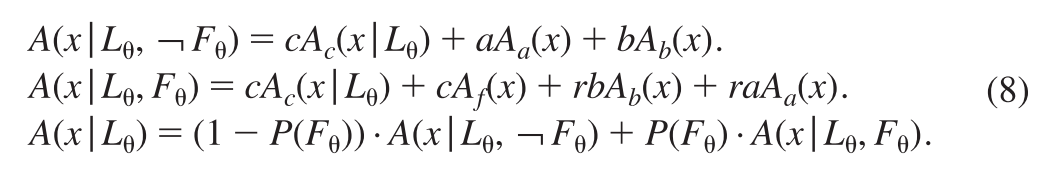
Incorporating attention:

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

When is the target of this trial, activation becomes

The activation of non-cue reduced by parameter .

The full model is



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

**Experimental design**

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. Experiment 4 is the key experiment to test this assumption.

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 viewing distance was 50 cm. 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 with diameter . 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. There were three conditions to indicate 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 and participants did not know which will come until the cue is given. A total of 300 trials for each participant, 100 for each condition. The sequence of three conditions were random.

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.

**Specifying prior predictions**

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

The three experiments in the current paper conducted prior to Experiment 4 can provide information for parameters. Table 2 in (Oberauer & Lin, 2017) provided the mean, standard deviation and median of the estimates across participants. 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.

Bays et al., 2011: Here, we tested participants’ ability to reproduce from memory both the color and orientation of an object indicated by a location cue.

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.

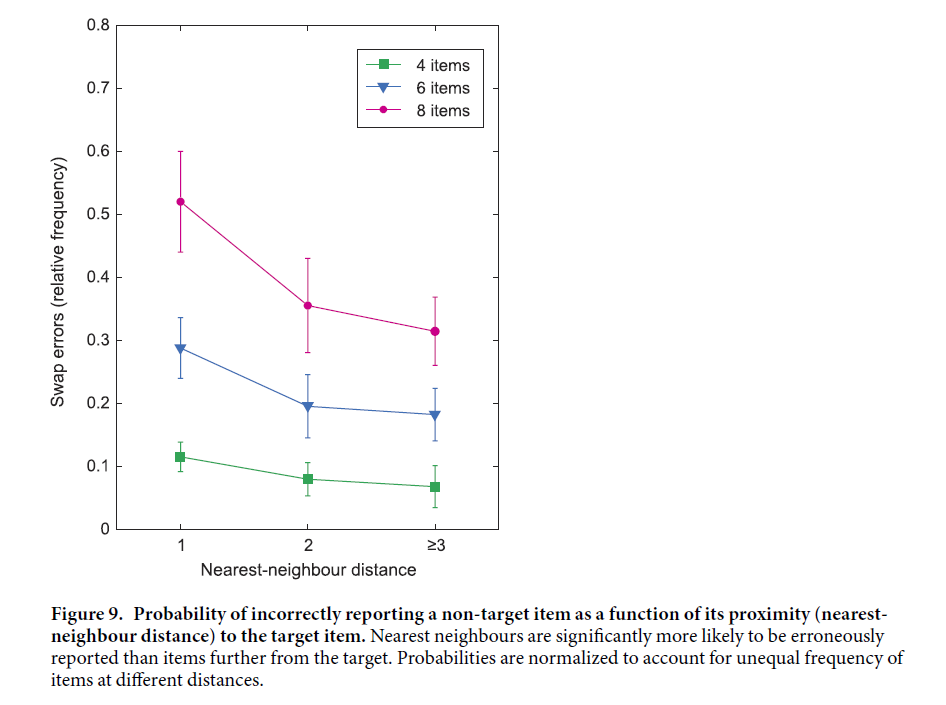
‘binding errors’ in a change detection task: errors caused by incorrectly combining in

memory features that belong to different objects

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

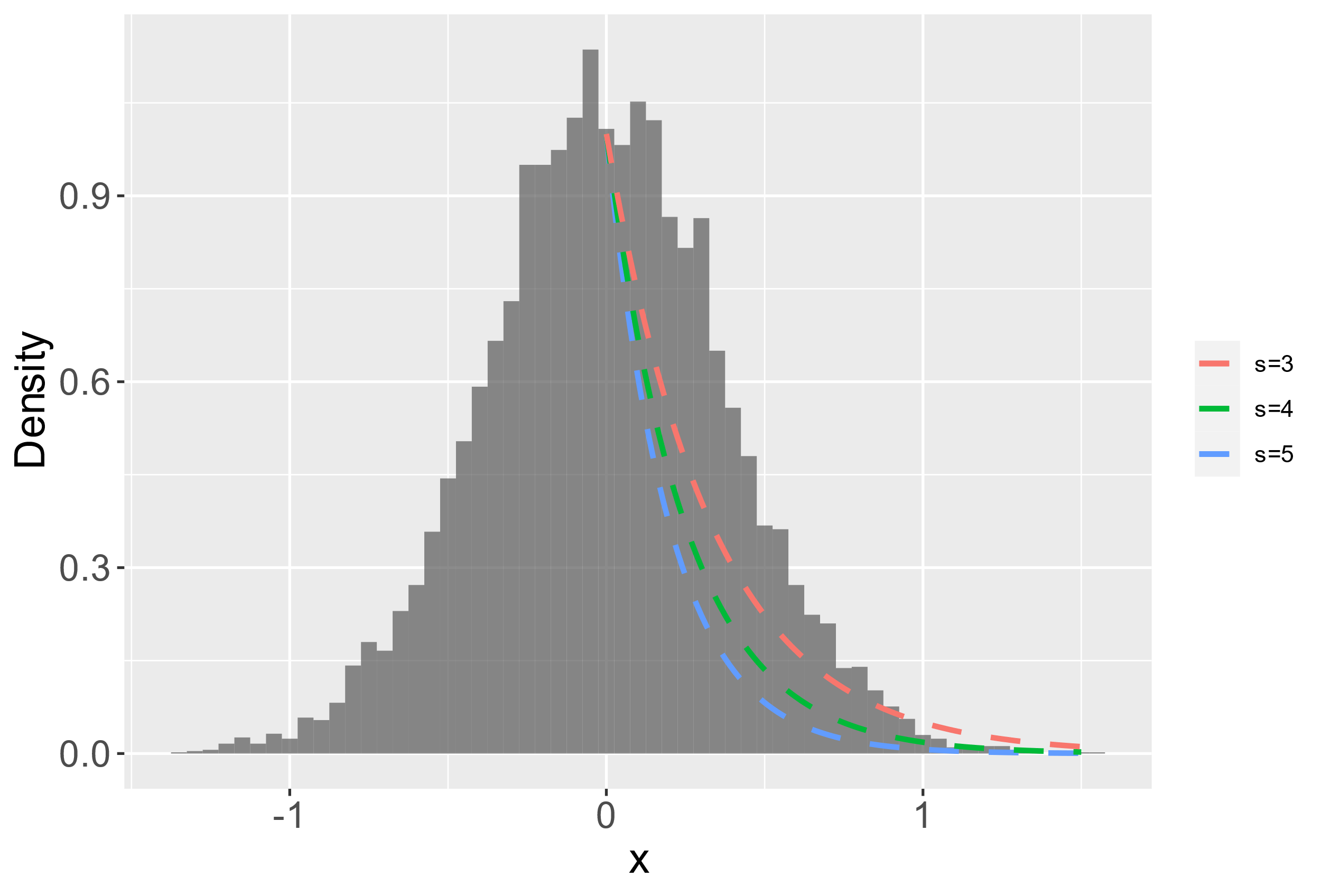
color and orientation with considerable precision. The precision of recall did not significantly differ between feature dimensions. 🡪 assume the kappa and delta similar to exp1.

Bays, 2016: location as cue, recall color, orientation or direction; sensitive to distance when 1<d<2. When s>10, exp(-10)<<1, unlikely to be even larger.



The estimated reflect the distance sensitivity of location-cue for color-reproduction. It may be applicable for orientation-reproduction . Similarly,

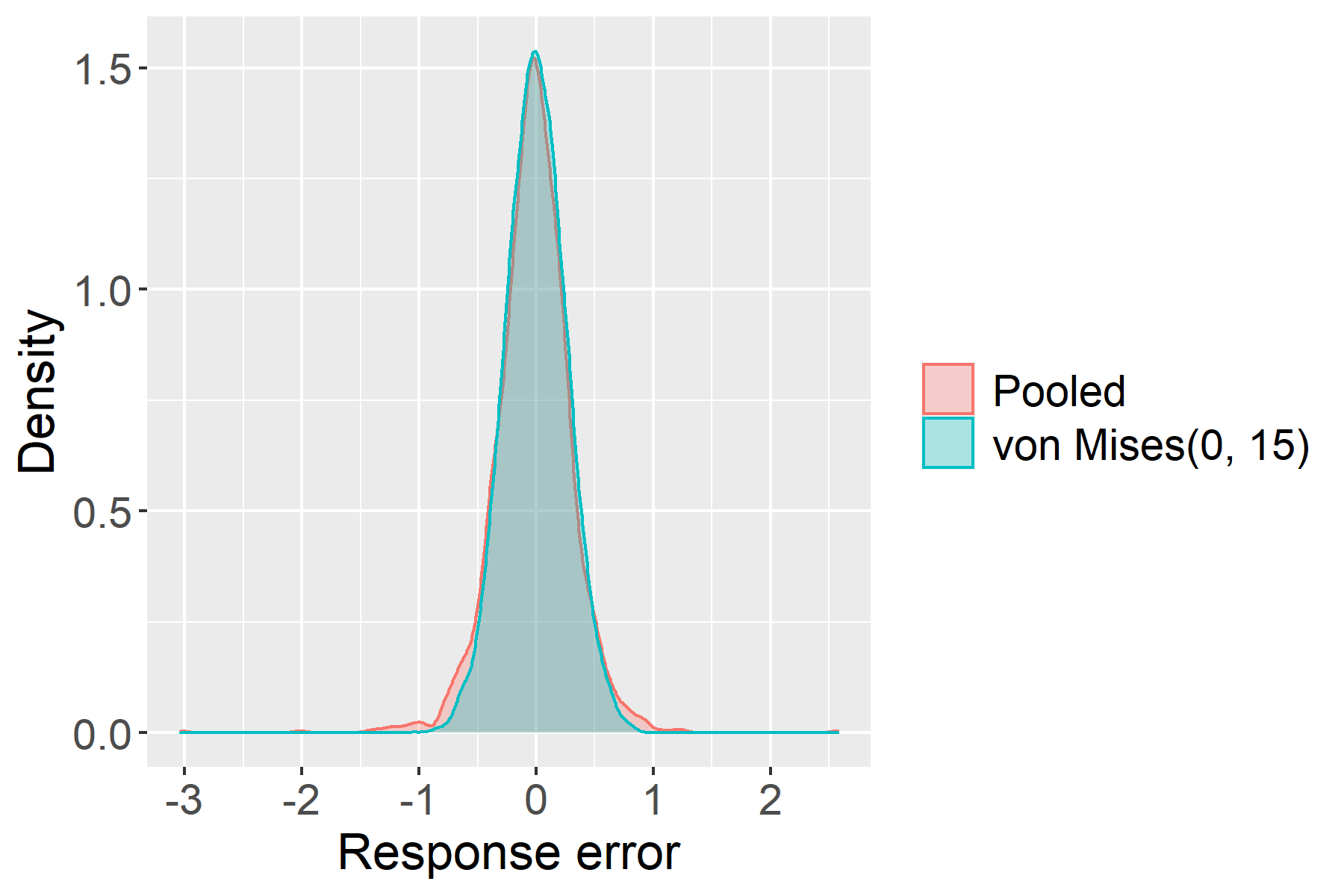
The dispersion of for color-reproduction may provide information for color-cue as it is still need to remember (Pertzov & Husain, 2014). I set in degree, and the SD in radian is The relation between and is approximate by 7.3. The activation of color feature of single item follows the shape . I used to approximate half of the activation to make two functions diminish at the close point. The approximate to 4. When , it seems diminished to quick, while seems to slow. I set



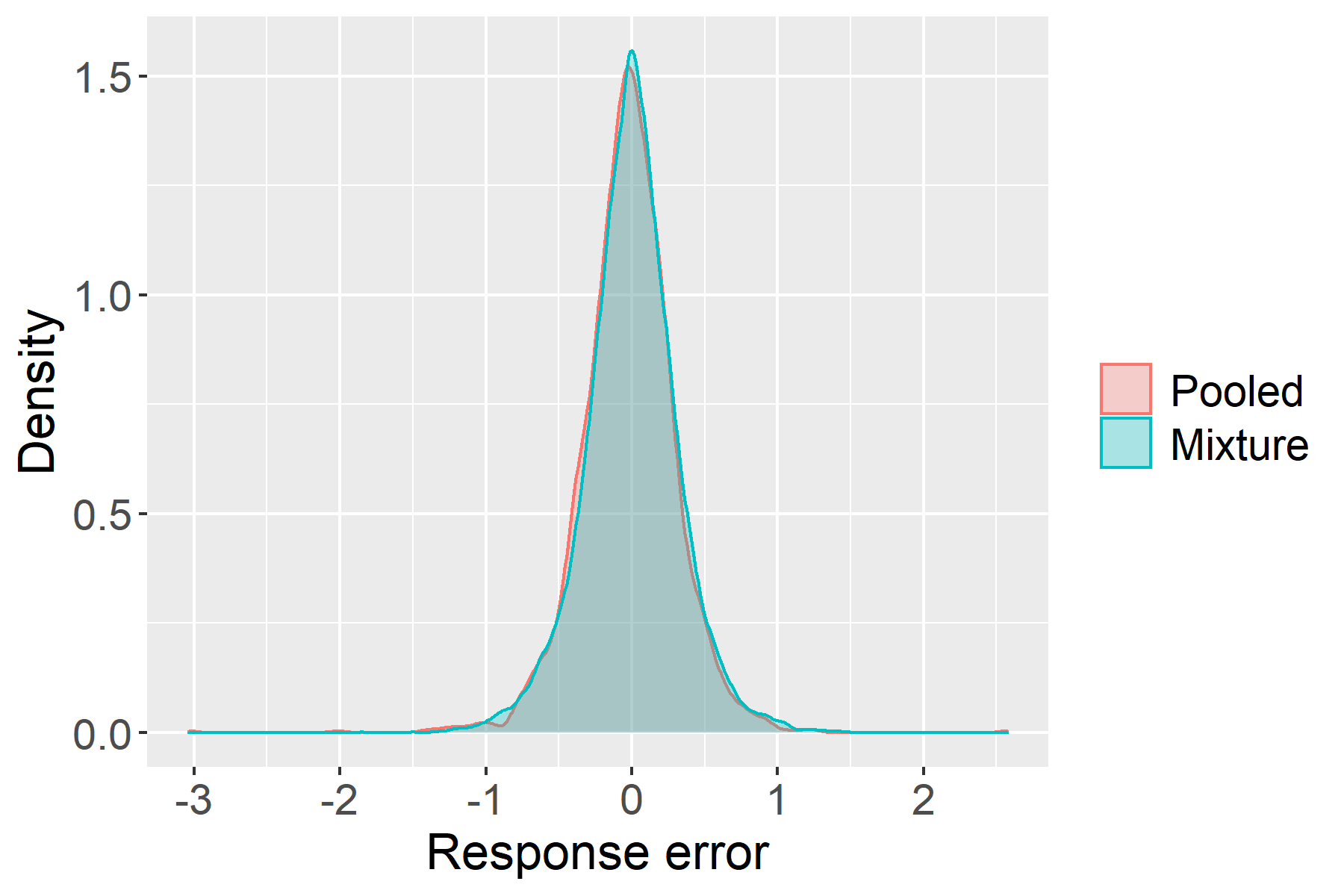
To inform feature-specific parameters and , I found retrievable previous data sets using location-cue reproduce orientation from (van den Berg et al., 2014). Only the set size of each trial, the difference between the response and the target of each trial, and the differences between the response and the non-target items of each trial were provided in these data sets (see README.txt in van den Berg’s codes), so they could not be fitted with the IM.

I consider the data when set size is one to avoid the influence from cues of non-target items since two studies used different settings would affect the intrusion when set size>1. Only the data set of van den Berg et al (2012) could be used, since the papers of other three data sets with set size 1 were retracted. When set size is 1, the probability to respond at with the target orientation is proportioned to

I first approximated it with a von Mises distribution with trial-and-error. When , the von Mises distribution’s peak is closed to the pooled error distribution (Fig \*). Since should have higher peak than the real mixture distribution, is likely to be higher than 15. Similarly, is likely to be smaller than 15.



Then I decomposed the pooled response errors with set size one to a mixture of two von Mises distributions where and . The maximum likelihood estimates are (sd=0.32) and (sd=2.13) (Fig \*).



I set and . Let to satisfy the order constrain.

There is no information for . I made bold assumptions based on the literature. Color and location were approximately equally good retrieval cues for orientations (Pertzov & Husain, 2014). But this may because the color active location then retrieve orientation. The location has privilege role according to experiment 1 in PH, 2014. When both location and color are cues, I assume that more weight will be assigned to location-cue. Here I set .

One thing needs to be care is that the distance is in radius and will be affected by viewing distance and eccentricity. The absolute distance is different across studies.

Complexity increase: for experiment1-3, only color and location; for experiment 4, color, loc, and orientation

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Article | Set sizes | Eccentricity  (degrees) | Display duration  (ms) | Delay  (ms) | Possible location | Possible response | Viewing distance  (cm) |
| Van den Berg et al., 2012 | 1-8 | 8.2 | 110 | 1000 | 8 | 180 | 50 |
| Oberauer & Lin, 2017 | 6 | 5.1/2 | 1000 | 1000 | 13 | 360 | 50 |

**Prior predictions of the IM**

The informative prior predictions are as followed.

Since the standard deviation may be too small, I add two extra sets of prior distributions with larger standard deviations. SD\*2 and SD\*3

**Data priors**

The IM makes a unique prediction that the response will be closer to a non-target item if the context-cue, i.e., location, of the non-target item is closer to the context-cue of the target item. I reanalyzed the dataset from Bays et al. (2009) where the distance did not show impact on the responses (Fig 4). Since the established explanations (slot, resource) have not incorporate the effect of the distance between target and non-target items, I take the distribution of the deviation between the response and the non-target item, uniform on -pi, pi, (Fig 4.B) as the data prior.

color and location were approximately equally good retrieval cues for orientations (Pertzov & Husain, 2014).

# Example 2: Risky intertemporal choice heuristic model

**Original study**

It is observed that people tend to choose sooner safer larger reward. How amount, delay and risk interplay? Two questions are debating: whether the comparison is between utility or attributes? And whether the delay and the risky are translatable? Luckman and his colleagues conducted a large experiment by manipulate three features and six conditions and fitted the existed and newly proposed models of RIC to evaluate them. Their results suggested that attribute-comparison and untranslatable of delay and risk. The risky intertemporal choice heuristic model provided best account of the data. Here I assess whether the experiment put RITCH face a strong test.

***The risky intertemporal choice heuristic model***

The risky intertemporal choice heuristic model (RITCH) assumes that the absolute and relative differences of two gambles at each attribute are considered when making choice between them. Suppose the two gambles are and . , , and represent the amount, delay, and risk of each gamble respectively. The RITCH has the form as eq. x-x.

where , and . Since human beings have tendency to choose larger, sooner and safer gamble, all parameters are constrained to be non-negative.

***Experimental design***

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).

The experiment consisted of six types of RIC and four conditions. The six types of RIC are

|  |  |  |  |
| --- | --- | --- | --- |
| Type:  attribute is preferred vs attribute is preferred) | Amount | Delay | Risk |
| R vs A |  |  |  |
| D vs A |  |  |  |
| R vs D |  |  |  |
| R vs AD |  |  |  |
| D vs AR |  |  |  |
| DR vs A |  |  |  |

Four conditions: baseline, magnitude, immediacy and certainty. Baseline condition contain 96 choices, 16 for each type. Magnitude condition is obtained by multiply the amounts in the baseline condition by 10, immediacy condition by add 12 months to the delays in the baseline condition, and certainty condition by divide the probabilities in the baseline condition by five. Six choices with one dominated option are included as check questions. Total 386 trials. Sample size was 100, which was decided before data collection.

**Expected results of the experiment**

Unlike example 1, the experimental design of RIC varies largely from study to study and the observed results are unstable. The ways to analyze data also changed largely. No proper dataset could be used as reference. I will use the established models of RIC: PTT, HD, MHD to simulate the expected results.

I use the information of the effects related to three manipulations than was reviewed in the original study to tuning the prior parameters. In addition, one dataset for risky choice and one dataset for intertemporal choice have been found. I also use these two datasets to refine the priors. Other literatures results will also be used.

***Established models***

The three established models are all utility-based model. The general form of the is , where is the subject value of the amount and is the discounting function of delay and uncertainty . In the following, I use to denote the utility of an option and to denote the odd against .

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

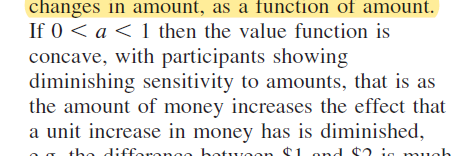
As in MHD, is assumed between 0 and 2.

Risk is transformed into a delay, which is then added to the existing level of delay before a discount function is applied. This transformation assumes that risky choices are made by calculating the amount of time it would take, on average, to achieve the desired amount, if the gamble was repeated multiple times. Larger values of *i*, therefore, lead to a more negative view of risk as the inferred delay is longer. In (Yi et al., 2006), was set to 35.3 (Rachlin et al., 1991, Subjects were asked to state their preference between a card that represented a risky $1,000 versus a delayed $1,000.).

the estimate of log(h) at x=10 is -6.82, at x=1000 is -7.48. Takahashi: k~0.015, h~1.02, i~70

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

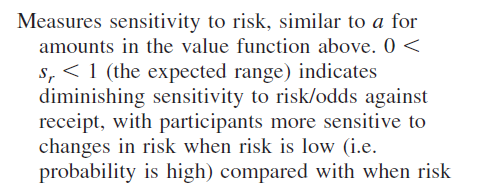
Decreasing *a* increases the concavity of the value function, or the extent to which participants show diminishing sensitivity to increasing amounts. the power function has constant elasticity. Diminished sensitivity to amount (Loewenstein & Prelec, 1992; Tversky & Kahneman, 1992): （assume .）



But, (Luckman et al., 2015 fig3, grey dots) .

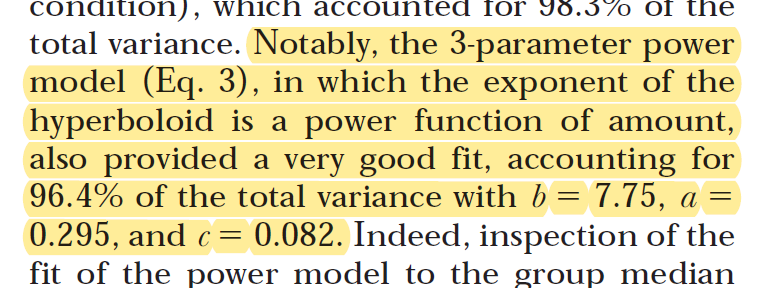
and measures discounting as a function of delay and odds against receiving the outcome respectively. In Vanderveldt et al., 2015, , . 🡪 Assume hd<1,hr>1

The parameters and measure sensitivity to delays and risks, respectively. Assume these two parameters are smaller 1 representing diminishing sensitivity (see (Zauberman et al., 2009) for delay; (Gonzalez & Wu, 1999) for probability).



incorporates peanut effect. Myerson, Green, and Morris

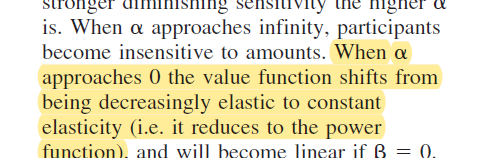
(2011)

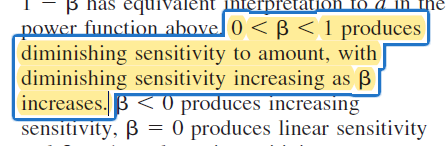


. In Vanderveldt et al., 2015, the indifference point when and 40000$ was measured. Choose between a smaller immediate certain option and a larger later riskier option. In this paradigm, and and are unidentified. The formula can be written as . Then could be approximate by the ratio of the estimate of when x=40000 and 800: => .

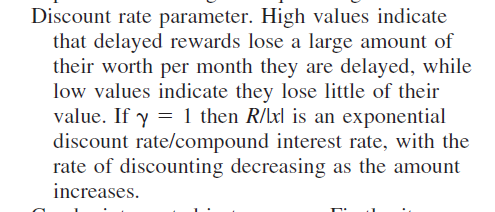
**PTT** (Baucells & Heukamp, 2010)**.** , .

and control sensitivity to amount. When , diminish sensitivity to amount. effects the degree of diminishing sensitivity, and also the elasticity of the function (which is important for explaining the magnitude effect). Decreasing elasticity means that proportional increases in outcome magnitude, *x*, lead to smaller proportional increases in subjective value, *v*(*x*), for larger values of *x* than for smaller values of *x.* capture some “peanuts” type effects. In (Baucells & Heukamp, 2010)**,**  which implied the elastic did not decrease much; , sensitivity did not diminish much.





Delays are combined with probabilities in the weighting function. *R*/|*x* | is the discount rate for time. decreases as a function of the amount, accomodate intertemporal magnitude effects. .



Decreasing increases the subproportionality of probability weighting, or the degree to which small probabilities are overweighted and medium to large probabilities underweighted. the expo-power function the later has decreasing elasticity. In (Baucells & Heukamp, 2010), .

(Baucells & Heukamp, 2010)was done in EURO and the range is much narrower than the current study, to what extent it can applied to dollar is unknow.

The probability function to choose between two options is , where and are the utilities of option 1 and 2 respectively.

***Retrieved data***

Four previous datasets for risky choice (Erev et al.,2002; González-Vallejo, 2002) and intertemporal choice (Ericson et al., 2014; Scholten & Read, 2010) were found. The dataset from (Ericson et al., 2014) is the raw data, while other three sets only provide the proportion to choose one option for each choice.

There are also studies given indifference point for risky or intertemporal choice through titration procedure, but may not be informative to parameters of models for current paradigm.

***Novel phenomena***

**Prior predictions of the RITCH**

***Bounds of parameters***

***Uniform distributions***

***Truncated distributions***

***LKJ distributions***

# 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

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

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

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

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