# Introduction

In recent years, a growing number of psychologists argue that the “replication crisis” is not simply due to problematic practice in methods, but stem from the poor quality of psychological theories (e.g., Borsboom, 2013; Oberauer & Lewandowsky, 2019; Smaldino, 2019). Psychological theories are often formulated in verbal and vague. This vagueness leaves room for researchers to flexibility to change theories, making the theories almost immune from testing (Oberauer & Lewandowsky, 2019). Computational modeling is believed to be a candidate solution of this “theory crisis” (Oberauer & Lewandowsky, 2019; Robinaugh et al., 2021). By instantiating theories into computational models, precision predictions can be made for experiments. If the predictions fit the observations, the theory is believed to be supported by the observations.

However, a piece of important information is missed here: 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. A precise prediction indeed rules out many possible outcomes of an experiment, but how likely these ruled-out outcomes will be observed from this experiment is unknown. If a model does not rule out any plausible outcomes of the experiment, this model will have a negligible risk of providing a prediction consistent with observation.

Vanpaemal (2020) proposed a more complete approach for theory testing in the Bayesian framework. This approach consists of two ingredients: the core predictions of a model and the data priors.

The core predictions contain the outcomes that the model predicts likely to be observed in the experiment. The construction of the core predictions relies on the prior predictive distribution of the model (Equation 1).

where is data, the model, the parameters, the likelihood of being observed predicted by at , and the parameter prior of . The outcomes are considered to be ruled out the model if the model assigns small prior mass to them.

The data priors contain the plausible outcomes of the experiment. The plausibility of outcomes can be decided based on theoretical considerations, previously observed empirical data, and expert knowledge.

Note that the construction of both ingredients should not involve the observed outcomes of the experiment. Otherwise, whether or not the prediction fits the observation is unconvincing (Vanpaemel, 2020). When the core predictions do not fully cover the data prior, the fit is persuasive. In such case, the model will be strongly supported if observations fall into the core predictions, and will be rejected otherwise.

In this study, I applied Vanpaemel’s strong theory testing approach (2020) to two published research in which new models were proposed and tested by experiments. It was expected to see that the models did not rule out any plausible outcome of the experiments in the original research; thus, their conclusions are not that persuasive.

In the next section, I describe the details of the specification of core predictions and data priors. Then the two applications are described. The paper is ended by a discussion of the results.

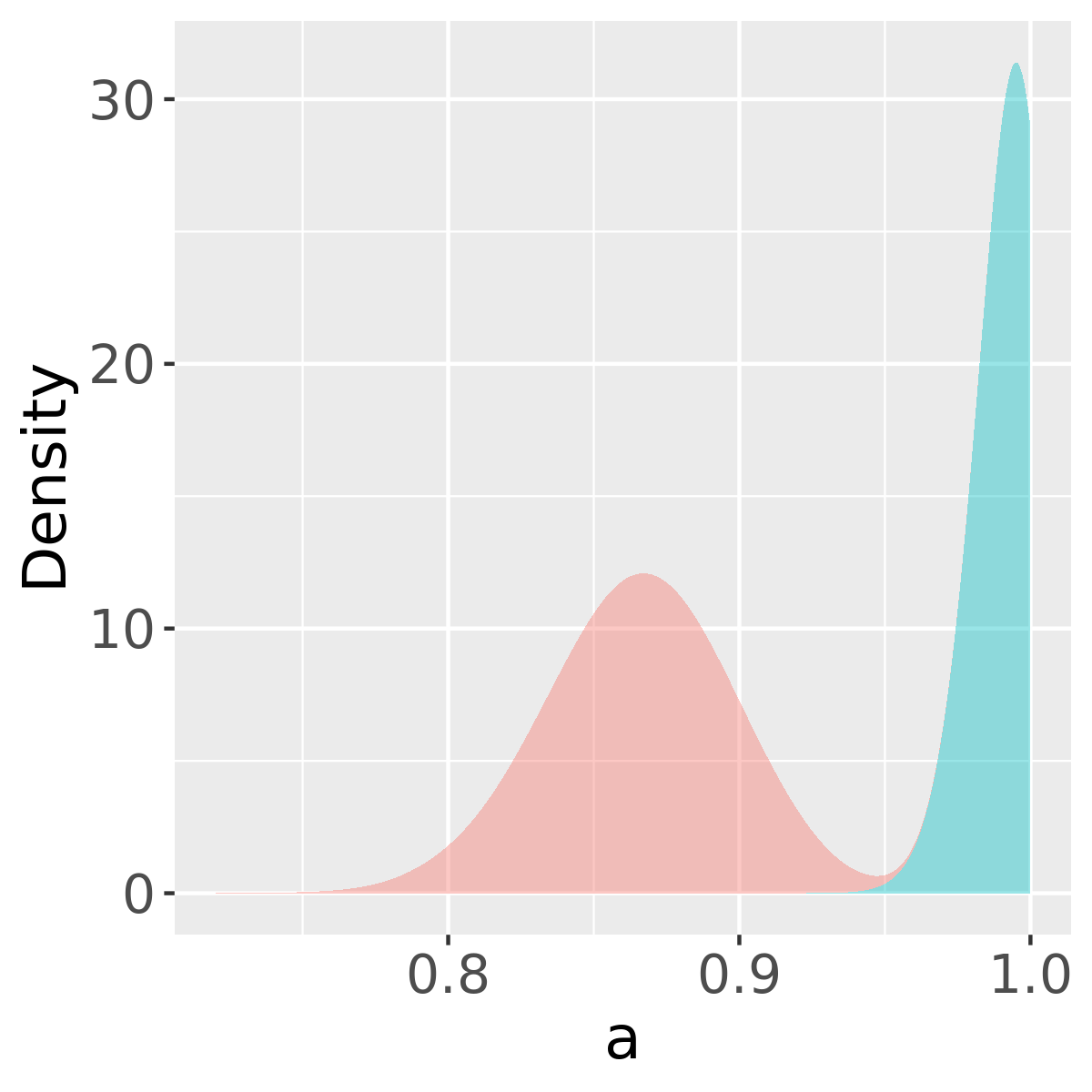
# Method

## 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 may not cause a problem when fitting data if the parameters’ priors are not specified sensibly, as data will shift them to the proper ones. But it is not the case for the prior prediction. 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 , where . Suppose that the value of is below 1 for most people and we know it from previous findings, while a vague prior is assigned to . 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, which is incompatible with our prior knowledge (Figure 1). This will make the model’s predictions inconsistent to observations even if the model is correct. Therefore, it is important to specify the parameter priors to incorporate prior knowledge when constructing prior predictive distributions.

**Figure 1**

*The prior predictive distributions of the proportion of participants choosing the first option.*



*Note.* The green distribution was generated by setting and the red one by setting 50 participants’ , and the rest 50 participants’ . The red distribution is the rightmost prior predictive distribution according to the prior knowledge, but the prior predictive distribution induced by the vague prior has an even higher range.

Lee and Vanpaemel (2018) suggested several sources and methods to develop informative priors. Here I describe the general procedure that used in the following two applications. The study-specific details are described along with the applications.

First, the boundary and order of parameters were decided based on the theoretical assumptions, logic constraints and results of previous studies. Then, if datasets from pilot studies or previous research with similar experimental designs are retrievable, the model was fitted to these datasets. Here, the model was assumed no individual differences since the retrieved datasets contain relatively few participants. The posterior distributions of the parameters were obtained by taking 4000 samples using 4 chains of 1000 iterations, after 1000 warmup iterations. For parameters that posterior distributions can be obtain from the last step, the truncated normal distribution was used. was the posterior mean. and was the lower and upper bounds specified in the first step. The s were set to make the standard deviation of the prior distribution equal to 0.05, 0.1, 0.5, 1 multiplying the absolute values of corresponding posterior means. For example, if the posterior mean of is 10, then the standard deviations were set to be 0.5, 1, 5, 10. This was intended to restrict the variation of the parameters within a proper scale. When the standard deviation of the parameter cannot reach the set standard deviation (e.g., the variance of a unimodal distribution on cannot exceed , the variance of , the parameter was assumed to be uniformly distributed. For parameters whose posterior mean cannot obtained, the prior was set to be , where the lower and upper bounds were the same to the first step. When the lower or upper bound was unknow, different values were tried.

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 *rstan* package (Stan Development Team, n.d.) was used to fit the models to previous dataset and generate prior predictive distributions. The proportion of the prior predictive distribution that the core prediction should cover was set to be 99.99%. The range of the core prediction was computed by R package *HDInterval* (Meredith & Kruschke, 2020)*.*

## Data priors

Two points need to note with the construction of the data prior (Vanpaemel, 2020). First, the plausibility of outcomes depends on the details of the experimental design, 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. It is often the case in psychology that a model is constructed explicitly incorporating a phenomenon observed before and then tested with a similar experimental design (see Application 2). In such case, even without this model, we still expect to see the reoccurrence of the phenomenon.

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 phenomenon 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 reflect 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 compare the core predictions to the observed data.. All analyses and visualizations were conducted in R programming language version 4.1.2 (R Core Team, 2021). The codes can be retrieved from <https://github.com/sombrelux/Prior-prediction.git>.

# Application 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.

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 manipulating all attributes to compare several utility-based and attribute-based models. Among all models, the RITCH model had the highest Bayes factor. I assessed 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 the reward of option , respectively.

The RITCH model assumes that the absolute and proportional differences of each attribute between two options are considered when making choice (Equation 2-4),

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 hypothetical 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 1). 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 the manipulations on three attributes. 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 1 | | | |
| *Six types of RIC.* | | | |
| Type | Amount | Delay | Probability |
| R vs A |  |  |  |
| D vs A |  |  |  |
| R vs D |  |  |  |
| R vs AD |  |  |  |
| D vs AR |  |  |  |
| DR vs A |  |  |  |

*Note.* R refers risk, A refers amount and D refers delay. is the amount of the reward of option . is the delay of receiving the reward of option . is the probability to receive the reward of option .

## Core predictions

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

Previous to the original 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 original 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.

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 2 gave the posterior means and four sets of s making the standard deviations of the informative priors equal to 0.05, 0.1, 0.5, and 1 multiplying the posterior means.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 2** | | | | | | | | |
| *The posterior means and four sets of s for RITCH.* | | | | | | | | |
| Parameters |  |  |  |  |  |  |  |  |
|  | -0.73 | 0.47 | 0.001 | 0.48 | 0.48 | 0.54 | 0.03 | 0.13 |
|  | 0.036 | 0.024 | 0.000055 | 0.024 | 0.024 | 0.027 | 0.0014 | 0.0065 |
|  | 0.073 | 0.047 | 0.00011 | 0.048 | 0.048 | 0.054 | 0.0028 | 0.013 |
|  | 0.36 | 0.24 | 0.0006 | 0.26 | 0.26 | 0.3 | 0.015 | 0.07 |
|  | 0.73 | 0.47 | 0.0015 | 0.65 | 0.65 | 0.73 | 0.038 | 0.18 |

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

12 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 original 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 the probability of receiving the reward . 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 for the parameters. 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 to the pilot dataset.

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

All parameters are positive according to the logic and theory.

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 original 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 .

The pilot dataset was fitted with the priors specified above. Table 3 shows the posterior means and four sets of s.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table 3** | | | | |
| *The posterior means and four sets of s for HD.* | | | | |
| Parameters |  |  |  |  |
|  | 0.61 | -3.54 | 48.87 | 0.24 |
|  | 0.03 | 0.18 | 2.4 | 0.012 |
|  | 0.061 | 0.35 | 4.9 | 0.024 |
|  | 0.33 | 2 | 25.8 | 0.13 |
|  | 1000 | 4.7 | 65.9 | 0.33 |

*Note.*  were restricted in , so its standard deviations cannot reach . The uniform distribution was approximated by .

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

All parameters are positive according to the logic and theory.

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. Assuming these two parameters are smaller 1 represents 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), while in Myerson et al. (2011), was very closed to 0. Thus, I assumed that . Table 4 shows the posterior means obtained by fitting the pilot dataset and the four sets of s.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 4** | | | | | | | |
| *The posterior means and four sets of s for MHD.* | | | | | | | |
| Parameters |  |  |  |  |  |  |  |
|  | 0.51 | 0.37 | -2.00 | 0.67 | 0.25 | 0.23 | 0.31 |
|  | 0.025 | 0.018 | 0.1 | 0.034 | 0.012 | 0.012 | 0.016 |
|  | 0.05 | 0.037 | 0.2 | 0.067 | 0.025 | 0.023 | 0.031 |
|  | 0.28 | 0.2 | 1.1 | 0.37 | 0.13 | 0.13 | 0.18 |
|  | 0.86 | 1000 | 2.83 | 0.85 | 0.4 | 0.34 | 0.42 |

*Note.* was restricted in , so its standard deviation cannot reach . The uniform distribution was approximated by .

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

The bounds of the PTT’s parameters according to the logic and theory are and .

Baucells and Heukamp (2010)fitted the PTT to an experiment done in EURO and the range of amounts was much narrower than the current study. The values of its estimates was not applicable since amount presented in the current experiment was in dollar, but I made assumptions based the qualitative information they provided.

and control sensitivity to amount. The sensitivity to amount diminishes as amount increases when while the sensitivity increases when . In Baucells and Heukamp (2010)**,** implied slight diminishing sensitivity to amount, 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 and Heukamp (2010)**,** suggested the subjective value was almost inelastic, so I set . PTT transfers delay into risk. The transfer rate decreases as the absolute value of amount increases. Since its value is sensitive to amount, I set without considering the estimate in Baucells and Heukamp (2010). measures subproportionality, the degree of overweighting small probabilities and underweighting larger probabilities. In Baucells and Heukamp (2010), , so I set . Table 5 shows the posterior means obtained by fitting the pilot dataset and the four sets of s.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table 5 | | | | | |
| *The posterior means and four sets of s for PTT.* | | | | | |
| Parameters |  |  |  |  |  |
|  | 0.08 | 0.63 | 0.88 | 1.38 | 0.93 |
|  | 0.0038 | 0.032 | 0.045 | 0.07 | 0.047 |
|  | 0.0077 | 0.063 | 0.11 | 0.14 | 0.093 |
|  | 0.041 | 1000 | 1000 | 0.76 | 0.5 |
|  | 0.11 | 1000 | 1000 | 1.9 | 1.25 |

*Note.*  and were restricted in , so their standard deviations cannot reach 0.5 and . The uniform distribution was approximated by .

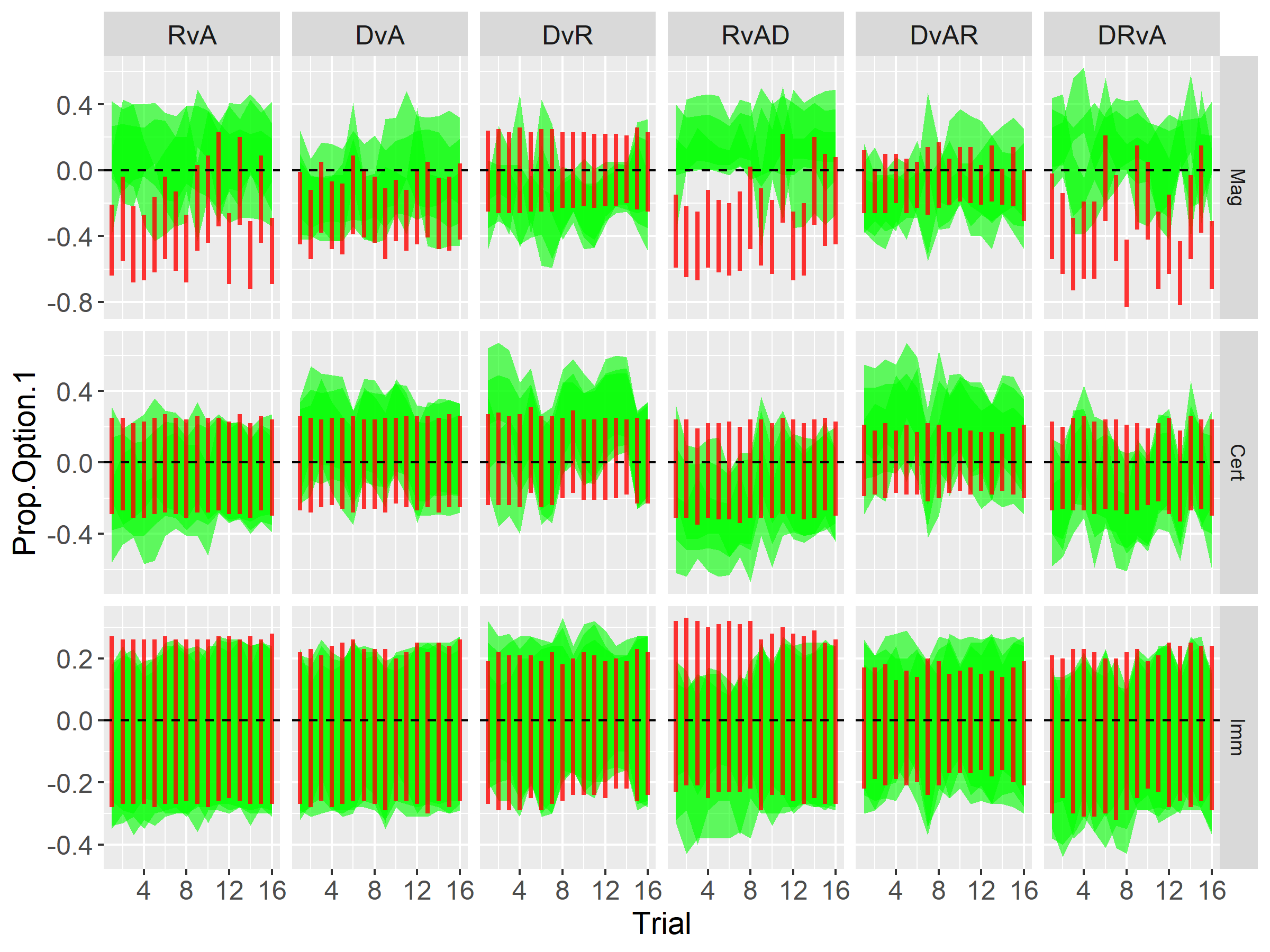
Since all parameters here can be identified with the pilot dataset, only four sets of informative priors were obtained for each model. The joint of the core predictions of HD, MHD and PTT was taken as the data priors.

## Assessing the experiment and model

Figure 2 shows the data priors and the core predictions when s were 0.05 multiplying the posterior means and . The core predictions do not cover all the data priors, especially for magnitude manipulation. According to the data priors, for R vs A, R vs AD and DR vs A choices, increasing the amounts of two options by 10 times will make more people choose the option with larger amount and higher risk, while the RIC model’s predictions are contrary to it. Thus, the experiment was a strong test for the RITCH model.

**Figure 2**

*Assessment of the experiment.*

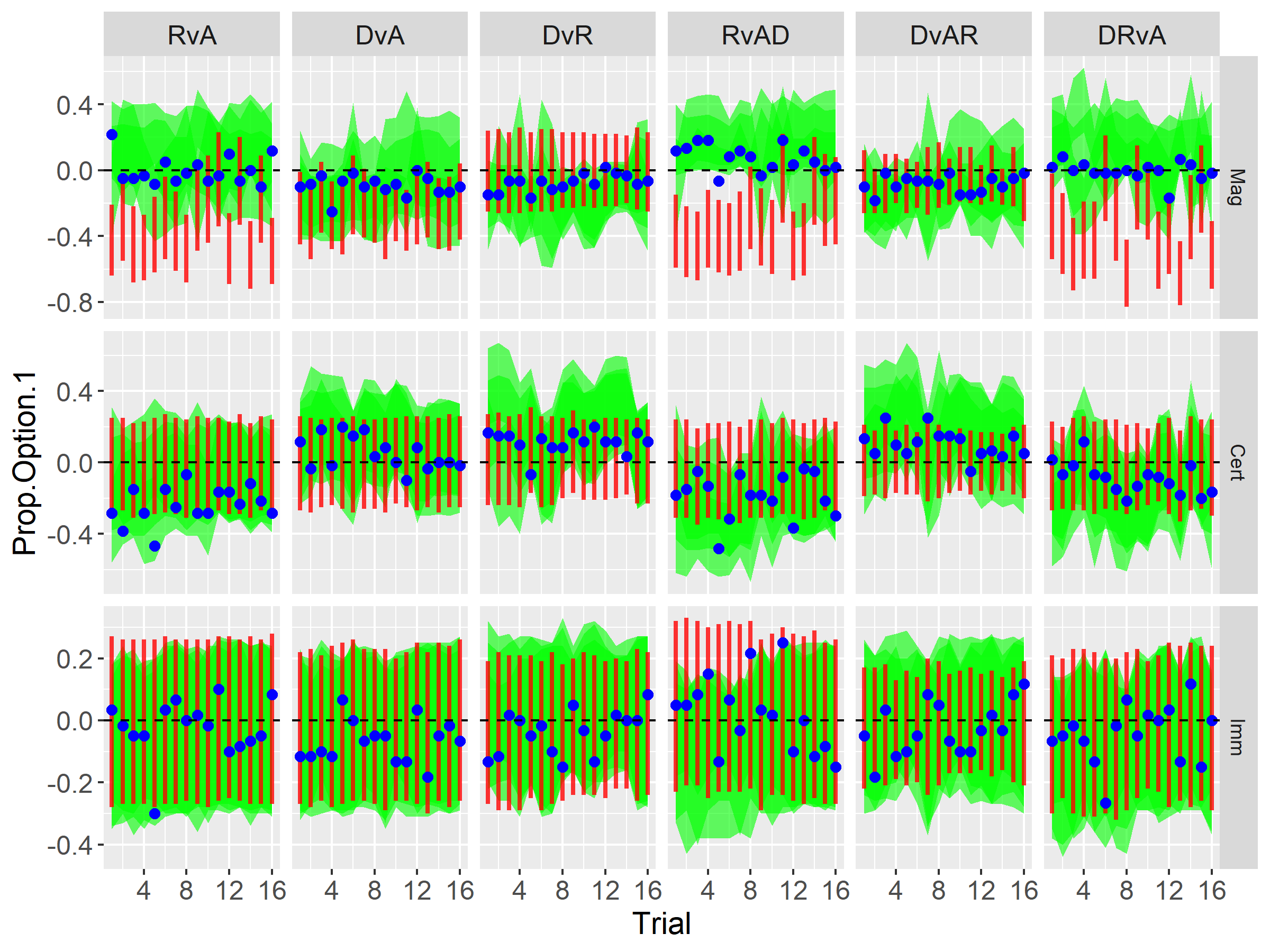


*Note.* The red segments are core predictions of the difference between the proportion of participants choosing option 1 in the manipulated choice set and that of the baseline choice set of the RITCH model. The green ribbons are the data priors.

However, when comparing to the behavioral data of the rest 60 participants, the RITCH model did not consistent with it (Figure 3). When the amounts increased, participants preference of options did not significantly change in R vs A and DR vs A choices, and participants tended to run higher risk in most of R vs AD choices. Other sets of informative priors gave qualitatively same results (Supplementary materials S1). Thus, although the RITCH model superior to other models in term of Bayes factors, it was still refuted by the observation.

**Figure 3**

*Comparison of the observation, the core predictions and the data priors.*



*Note.* The blue dots are the observed difference between the proportion of participants choosing option 1 in the manipulated choice set and that of the baseline choice set of each choice.The red segments are core predictions of the RITCH model. The green ribbons are the data priors.

# Application 2: Interference model

The second example comes from the domain of visual working memory (WM). One of the most robust and general phenomenon 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). 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.

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). Experiment 1 used location as retrieval cue and color as memory contents, which were same to previous studies. Experiment 2 and 3 were intended to test other assumption. Here I focus Experiment 4 which involved novel retrieval cues and memory contents. I assessed whether it was a strong test for the IM.

Since previous models such as the slots model (Zhang & Luck, 2008) and the resource model (Bays & Husain, 2008) do not accommodate this phenomenon, I used the second way of constructing the data priors.

## 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*

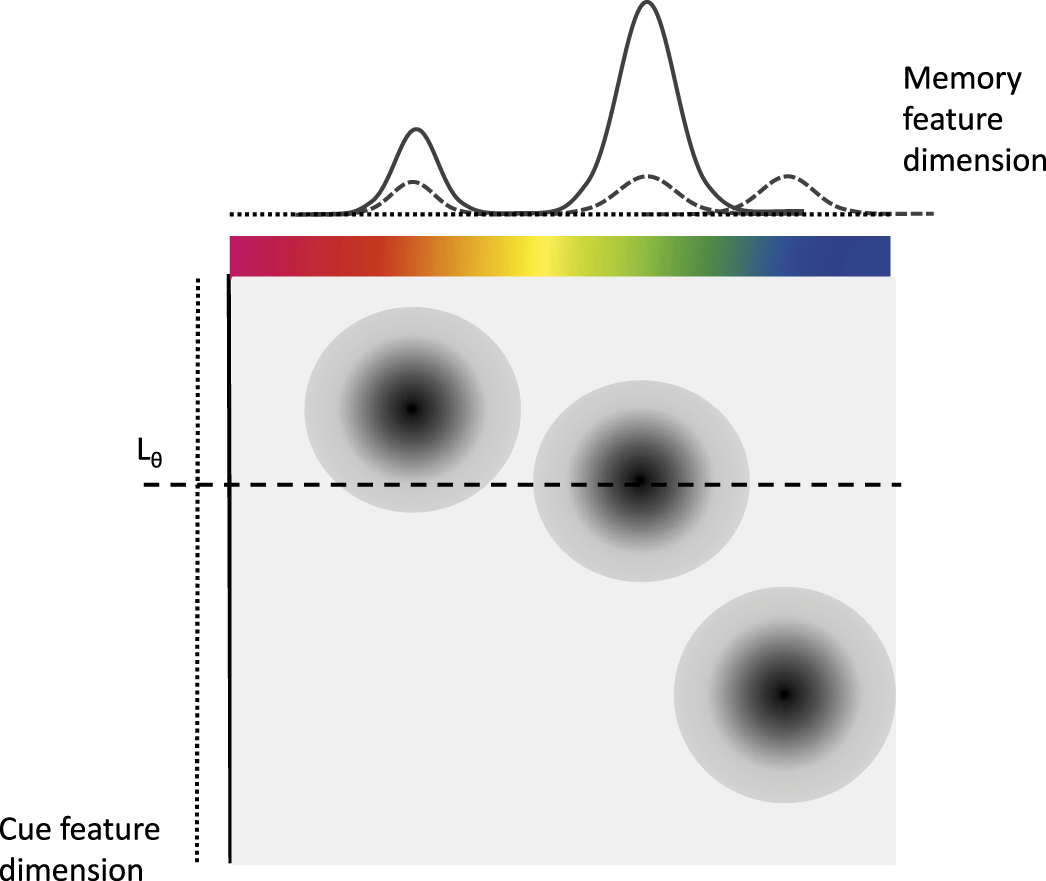
In visual WM tasks, memory contents are often features varying on a continuous dimension (e.g., colors or orientations) and the retrieval cue is often spatial location. Therefore, the IM represents the memory contents and retrieval cues as continuous dimensions, which are referred as memory feature dimension 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 4). 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 4**

*Schema of the bindings between the memory feature dimension and the cue feature dimension in the binding space.*



*Note.* is the retrieval cue. The darkness level is the binding strength. The continuous line represents the cue-based activation, the dashed line the cue-independent activation, and the dotted line the background noise. Reprinted from “Hierarchical Bayesian measurement models for continuous reproduction of visual features from working memory,” by Oberauer, K., Stoneking, C., Wabersich, D., & Lin, H. Y., 2017, *Journal of Vision*, *17*(5), p. 11-11. Reprinted under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

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 original 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 16)

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 17).

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 18), 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 19.

### *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 located 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 would be in the present trial 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. Since and are not closed related to the stimuli, I assumed the priors of them followed Experiment 1’s estimates. The distribution of response errors when using location to retrieve orientation was similar to that when retrieving color (Bays et al., 2011). It was also found that when participants did not know whether color or location would be used as cue to reproduce orientation before cue was presented, the performance using color or location as cue were similar (Pertzov and Husain, 2014). Therefore, I also assumed the precision parameters and followed the same prior distributions as Experiment 1. I fitted the Bayesian version of the IM to the dataset of Experiment 1 to develop the informative priors for these parameters.

The prior of and were set to and to . 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 between the cues of the target and non-target items when it was between 1 and 2. When , , which makes the influence of distance negligible when distance is larger than 1. Therefore, I assumed . The prior of was set to . The probability of response in the IM (Equation 25) is proportional to (Equation 23). When set size is one, . I approximated by , where it was expected that . It has the same form as in Bays et al. (2011), where the standard deviation of the von Mises distribution for response errors was when orientation was memory content and spatial location was retrieval cue. I assumed that the range of the standard deviation to be . Transforming this range into precision, the range of precision was . I further relaxed the upper bound to 30, thus . No extra information was found for , thus its upper bound was set to . Table 6 gave the posterior means and four sets of s.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Table 6** | | | | | |
| *The posterior means and four sets of s for IM.* | | | | | |
| Parameter |  |  |  |  |  |
|  | 0.16 | 0.15 | 0.13 | 8.32 | 9.97 |
|  | 0.008 | 0.0075 | 0.0065 | 0.42 | 0.5 |
|  | 0.016 | 0.015 | 0.013 | 0.83 | 1.0 |
|  | 0.087 | 0.081 | 0.07 | 4.55 | 5.4 |
|  | 0.22 | 0.20 | 0.18 | 20 | 13.44 |

For , and , no available dataset informs their distributions. and were set to follow was set to follow .

Four sets of informative priors with different uncertainty were created. The core predictions were made for the mean absolute error (MAE) of target-centered responses, average nontarget-centered responses and nontarget-centered responses at each distance along cue feature dimensions for each cue condition.

## 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 MAE 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.

For the confidence intervals , was set to 5 for the MAE of target-centered responses, average nontarget-centered responses and nontarget-centered responses at each distance on the location dimension. For the MAE of nontarget-centered responses at each distance on the color dimension, was set to 10 since more coarse approximation was used. The obtained data priors were applied to all three cue conditions.

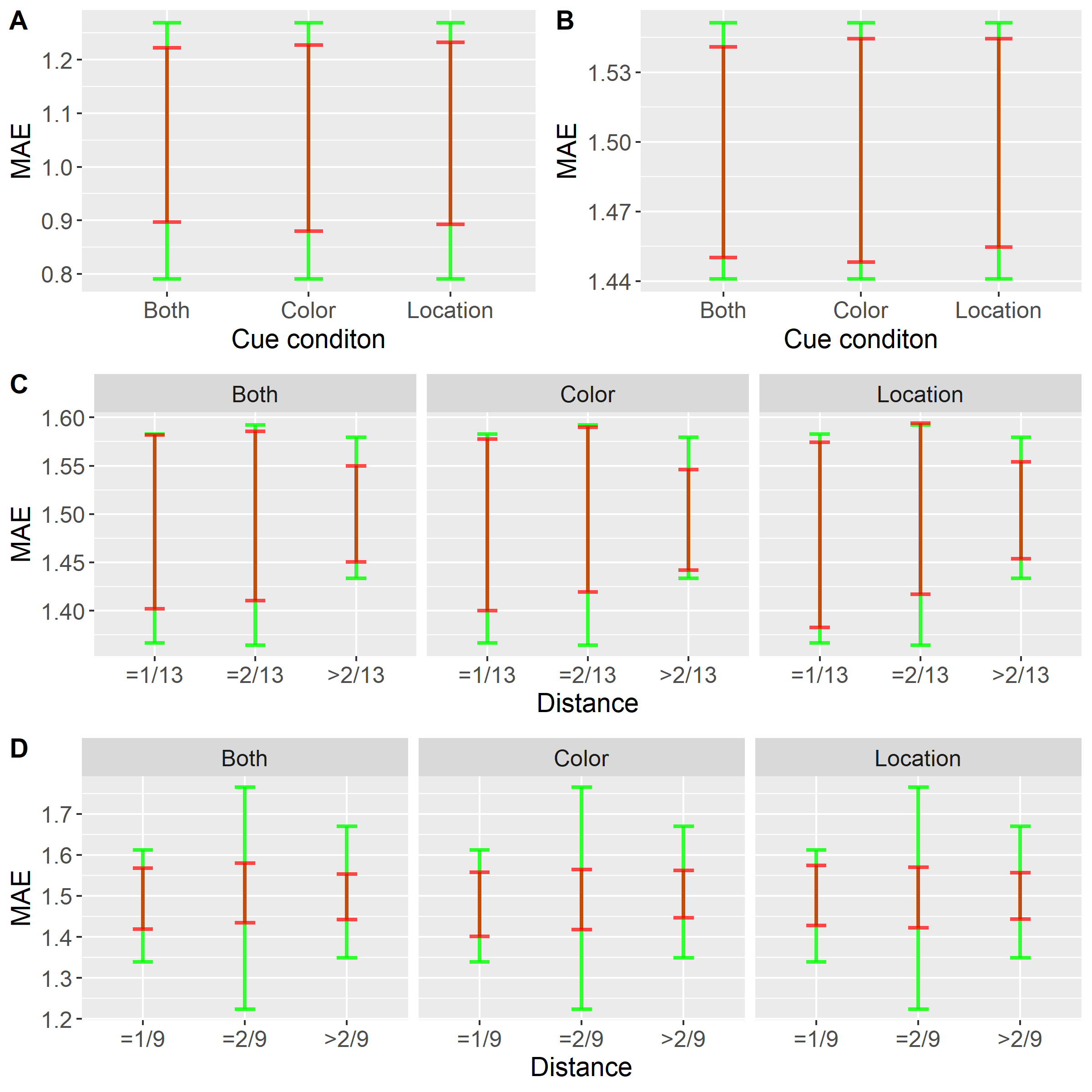
## Assessing the experiment and model

Figure 5 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.

Figure 2 shows the data priors and the core predictions when s were 0.05 multiplying the posterior means and . The core predictions do not cover all the data priors, especially for magnitude manipulation. According to the data priors, for R vs A, R vs AD and DR vs A choices, increasing the amounts of two options by 10 times will make more people choose the option with larger amount and higher risk, while the RIC model’s predictions are contrary to it. Thus, the experiment was a strong test for the RITCH model.

**Figure 5**

*Assessment of the experiment.*

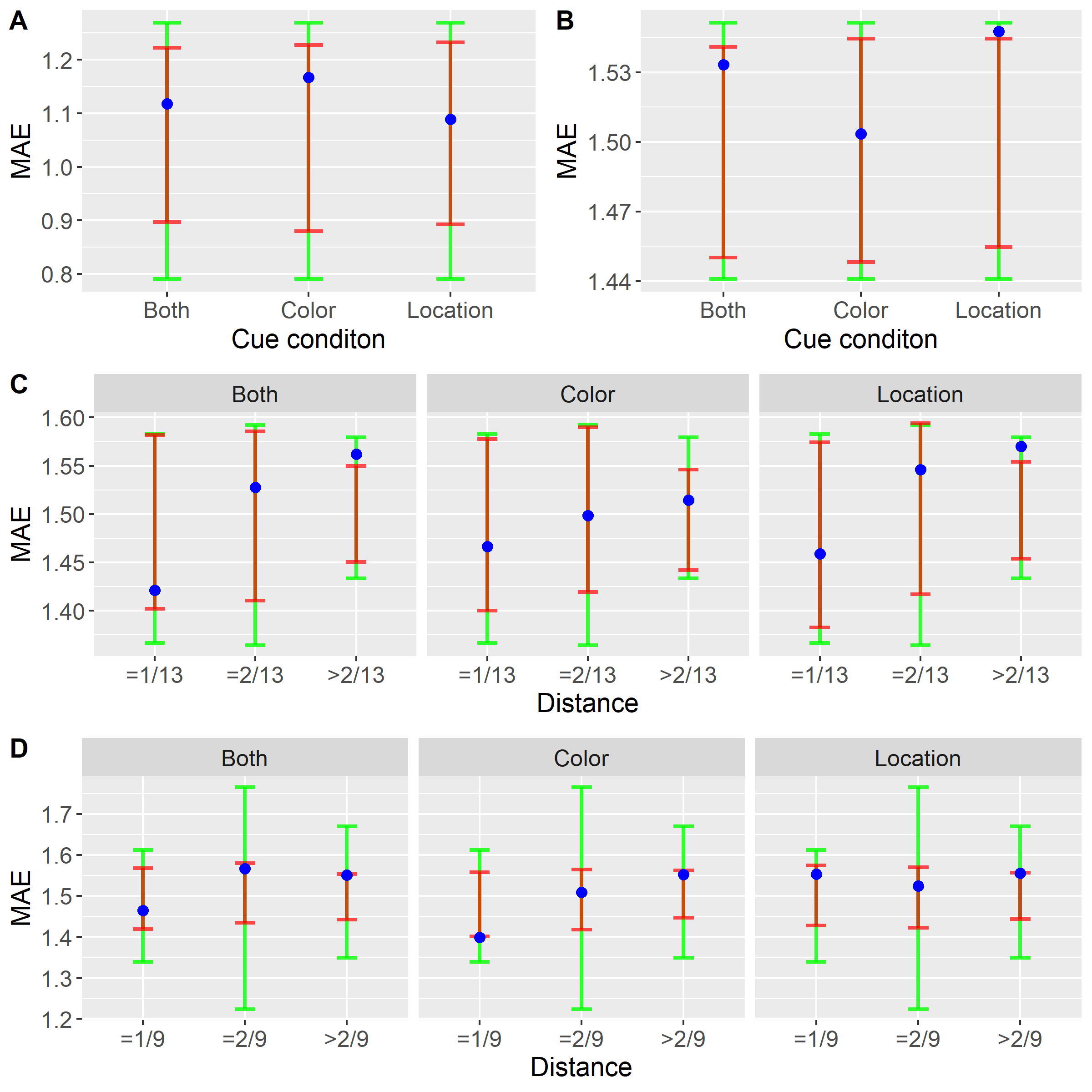


*Note.* The red segments are core predictions of the difference between the proportion of participants choosing option 1 in the manipulated choice set and that of the baseline choice set of the RITCH model. The green ribbons are the data priors.

However, when comparing to the behavioral data of the rest 60 participants, the RITCH model did not consistent with it (Figure 3). When the amounts increased, participants preference of options did not significantly change in R vs A and DR vs A choices, and participants tended to run higher risk in most of R vs AD choices. Other sets of informative priors gave qualitatively same results (Supplementary materials S1). Thus, although the RITCH model superior to other models in term of Bayes factors, it was still refuted by the observation.

**Figure 6**

*Comparison of the observation, the core predictions and the data priors.*



*Note.* The blue dots are the observed difference between the proportion of participants choosing option 1 in the manipulated choice set and that of the baseline choice set of each choice.The red segments are core predictions of the RITCH model. The green ribbons are the data priors.

# Discussion

In this study, I applied the strong theory testing approach (Vanpaemel, 2020) to assess two experiments in previous research. I developed informative priors for the models and made prior predictions. Two methods were used to construct data priors. It was expected to see that the two experiments were not strong tests for the models. However, the results were contrary to the hypothesis. In addition, the predictions of the models were inconsistent to the experimental data, suggesting the models were refuted.

For the failure of seeing the hypothesized result, a direct explanation is publication bias. The original studies must have novel findings to get published. Another reason is that I did not have much prior information for constructing the core predictions and data priors, but this was not the case for the authors. Several experiments were conducted before researchers proposed a new model (Luckman et al., 2015, 2017, 2018 for the RITCH; Rerko et al., 2014, Souza et al., 2016 for the IM)(e.g., The core predictions and data priors they construct would be different from what I constructed. The experiments might not be strong tests based on their prior knowledge.

For the inconsistency between the observation and the core predictions, the two applications are likely due to different reasons. For the RITCH, Bayes factor has long been pointed out only reflect relative behavior of models. The superiority does not mean the model is a good explanation (Lee’s book; Blue sea). For the IM, it may due to the misspecification of the parameter priors as the core predictions were closed to the observation. But the posterior distribution could not be obtained from the dataset since the original dataset only contained one set size , making the parameter unidentifiable.

The difficulty of applying Vanpaemel’s approach comes from the scarcity of prior information. This may not as bad as here for researchers as I explained above. This situation will also be alleviated by the encouragement of data sharing in psychology (e.g., Open Science Framework, 2016).In addition, a recent research did parameter systematic review, making the specification of parameter priors easier.

Third, lack of generalizability. The estimates cannot be taken seriously. They change with stimuli.

Fourth, poor theory development.

Generalizability

The structure of models changes across studies. What the instantiation affects the results? In the IM study, the authors discard attention activations due to fitted difficulty. To what extent can assumptions being discard?

Psychologists start to work on theory development.

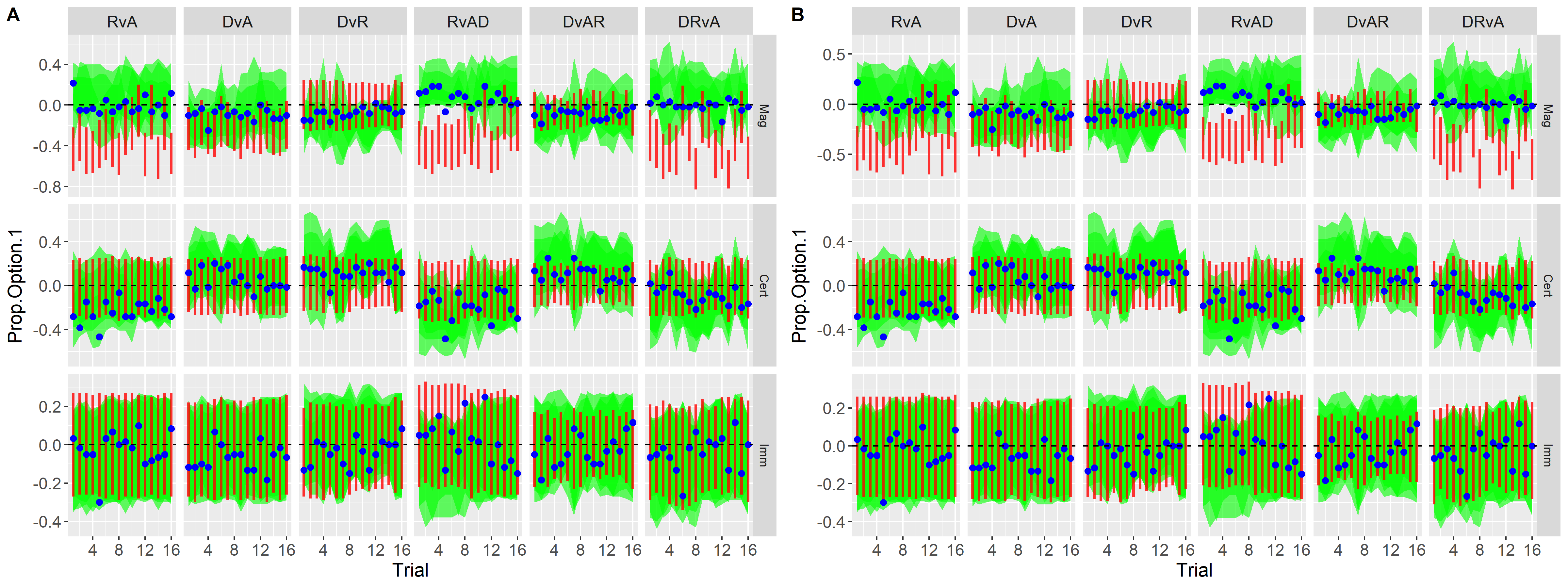
Eronen – validity – not only computational models are need, but more profound.

# References

# Supplementary materials

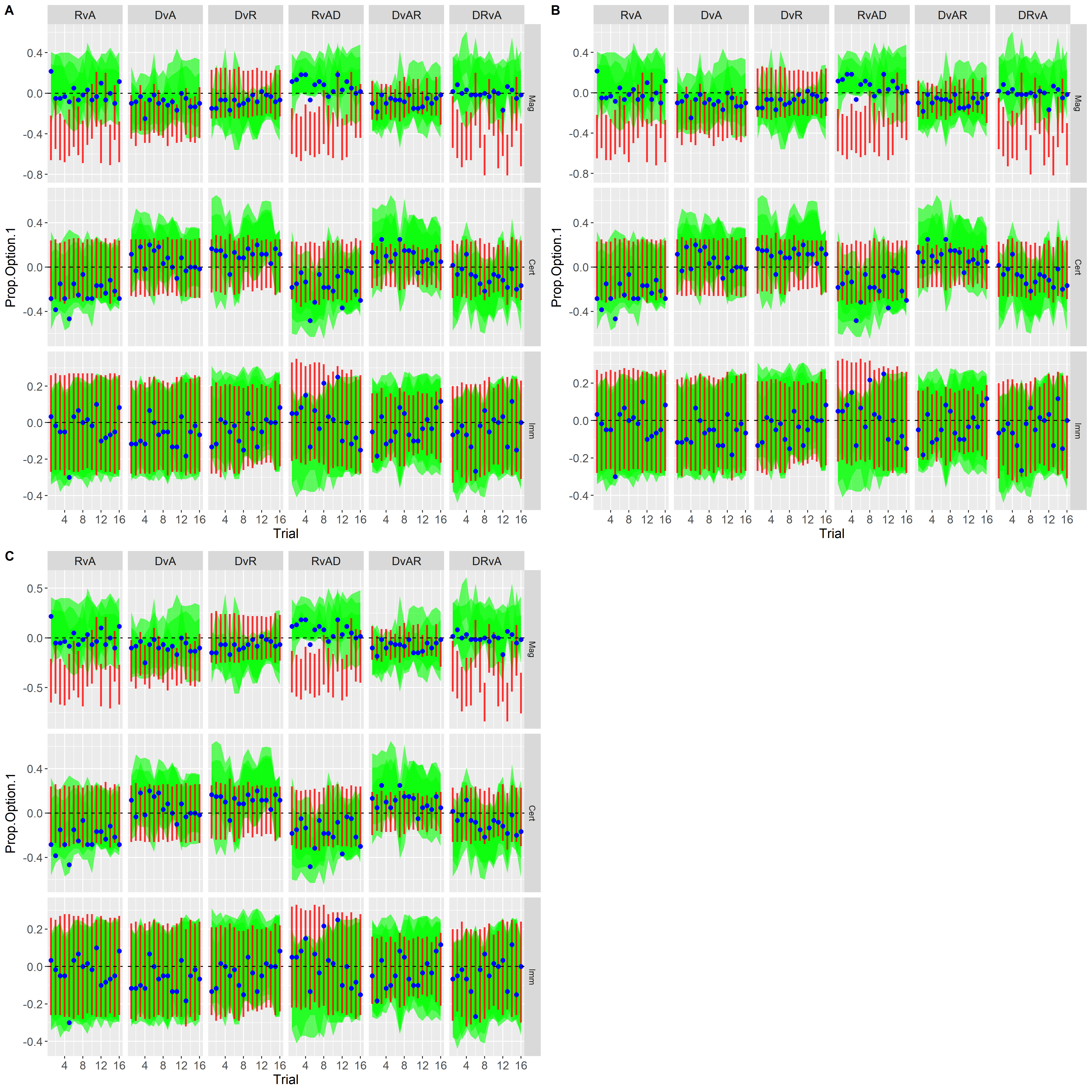
## S1. Assessment of the experiment in Luckman et al. (2020)

**Figure S1.1**



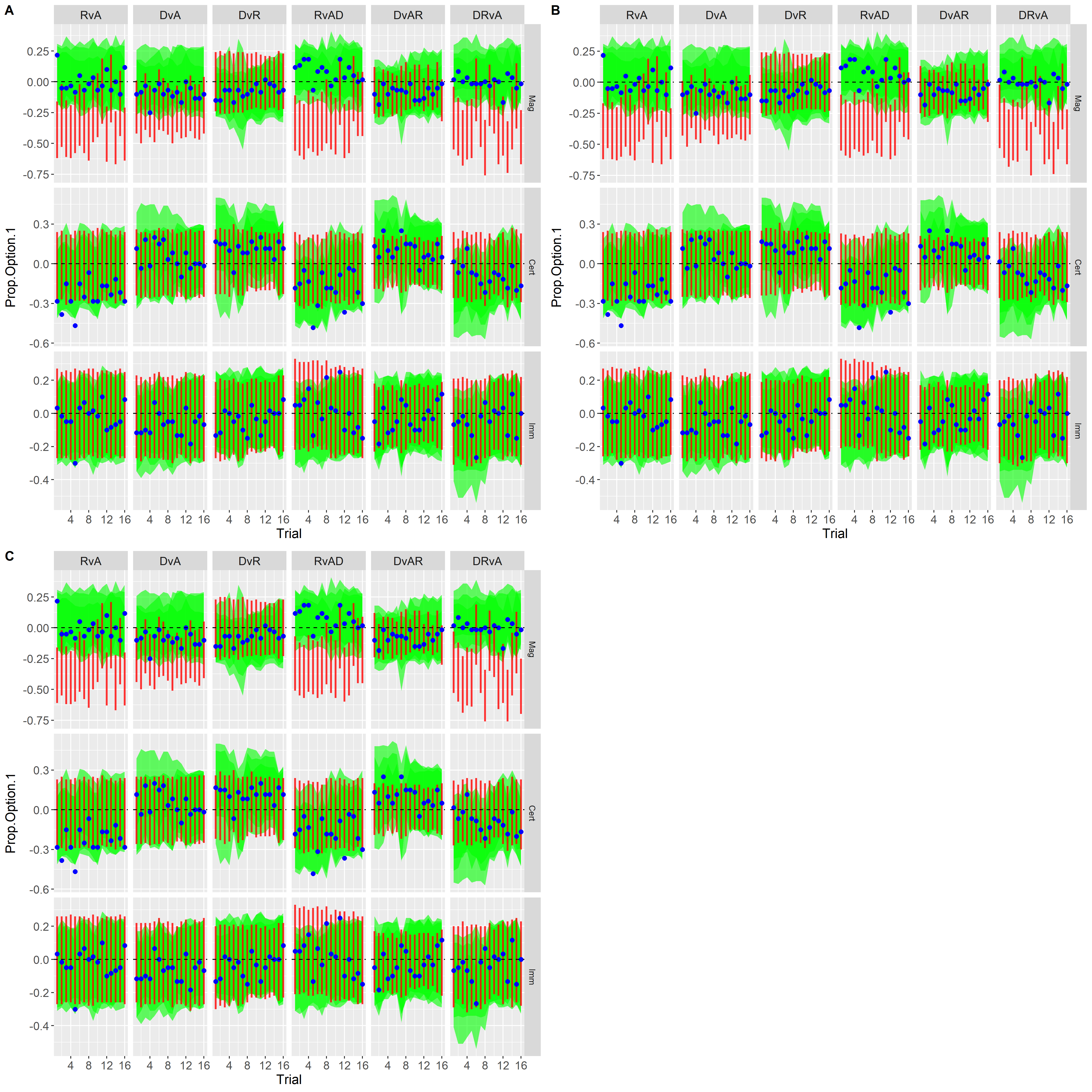
*.*The observation (blue), the core predictions of the RITCH model (red) and the data priors (green) of the difference between the proportion of participants choosing option 1 in the manipulated choice set and that of the baseline choice set when s were used. (A) . (B) .

**Figure S1.2**



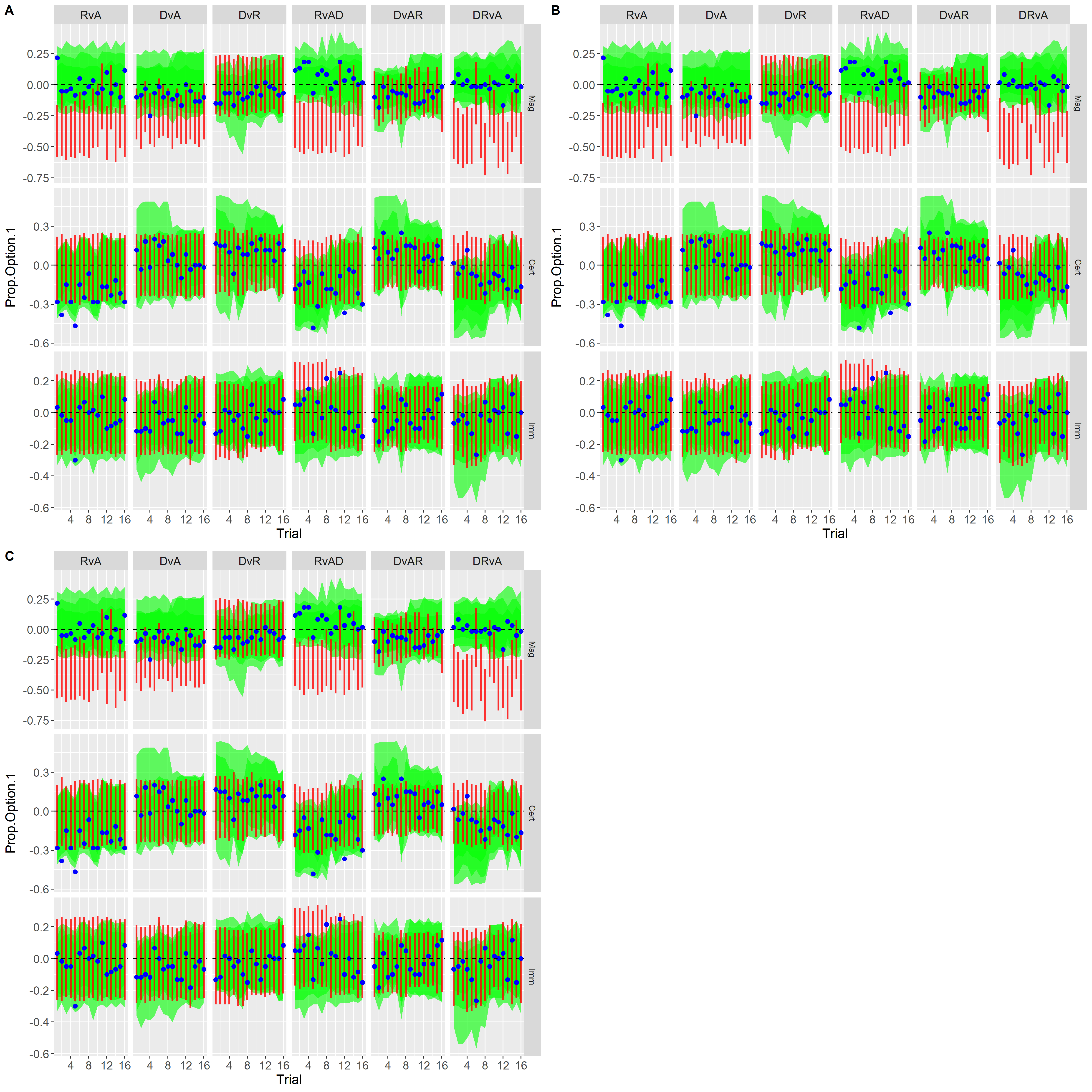
The observation (blue), the core predictions of the RITCH model (red) and the data priors (green) of the difference between the proportion of participants choosing option 1 in the manipulated choice set and that of the baseline choice set when s were used. (A) . (B) . (C) .

**Figure S1.3**



The observation (blue), the core predictions of the RITCH model (red) and the data priors (green) of the difference between the proportion of participants choosing option 1 in the manipulated choice set and that of the baseline choice set when s were used. (A) . (B) . (C) .

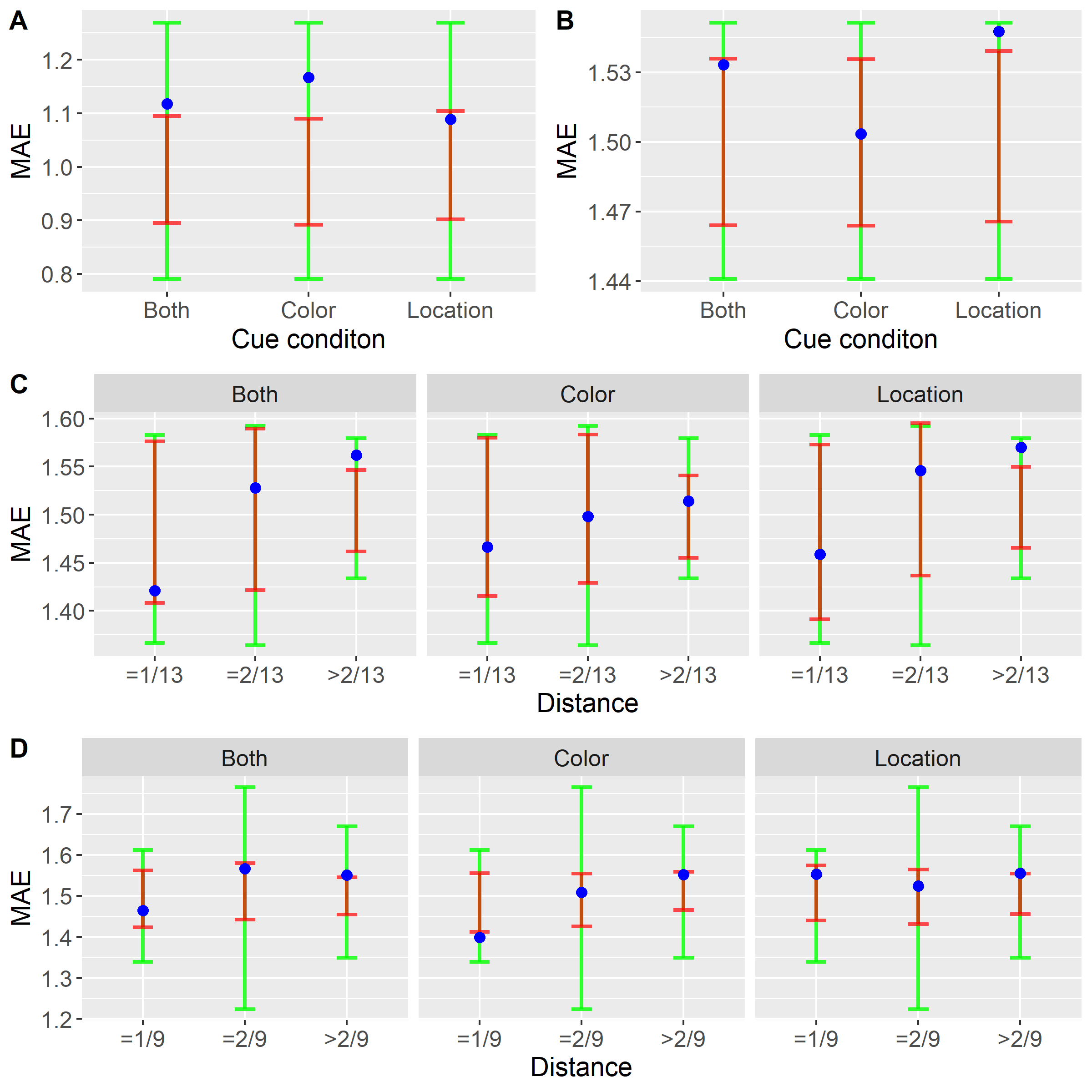
**Figure S1.4**



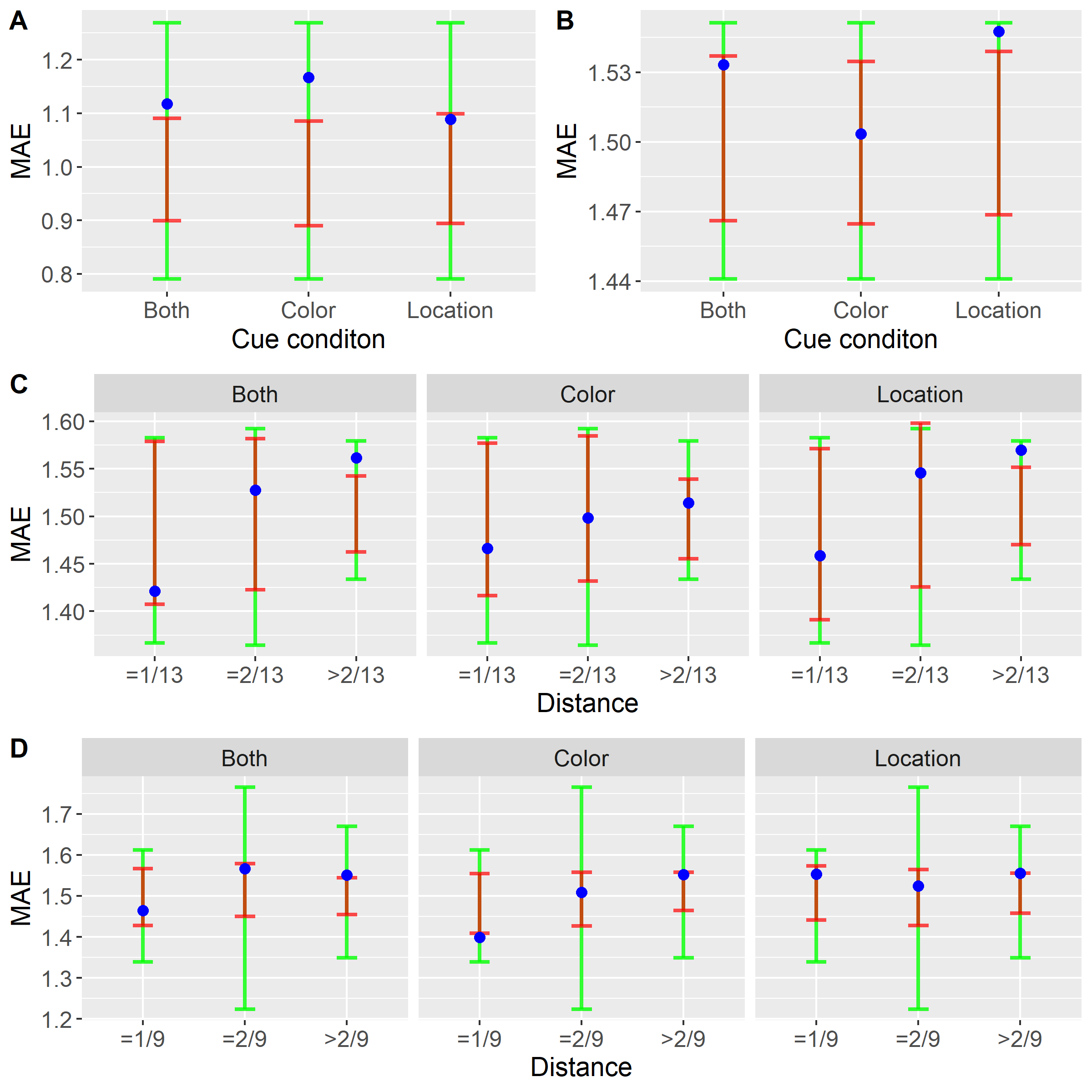
*.* The observation (blue), the core predictions of the RITCH model (red) and the data priors (green) of the difference between the proportion of participants choosing option 1 in the manipulated choice set and that of the baseline choice set when s were used. (A) . (B) . (C) .

## S2. Assessment of Experiment 4 in Oberauer & Lin (2017)

**Figure S2.1**



**Figure S2.2**



**Figure S2.3**

