Mutual interference in working memory updating: a hierarchical Bayesian model

Yiyang Chen[a] Mario Peruggia[b] Trisha Van Zandt [a,*]

[a]Department of Psychology
[b]Department of Statistics
The Ohio State University, Columbus, OH, US

*Corresponding author.
Email addresses: chen.6647@osu.edu (Yiyang Chen); peruggia@stat.osu.edu (Mario Peruggia); van-zandt.2@osu.edu (Trisha Van Zandt).

Highlights

- A joint theory-based framework to account for responses and reaction times in working memory updating.
- A Markov chain structure to characterize probabilities of responses during and after memory updating.
- A Weibull race structure to account for reaction times.
- Application to two data sets shows the mechanisms underlying group differences in memory updating performance.

Abstract

We propose a hierarchical Bayesian model for working memory updating. This model accounts for both the accuracies of responses and reaction times (RT) in the memory updating paradigm, which is a commonly used paradigm to measure working memory capacity. We adapt a mutual interference model from Oberauer & Kliegl (2006) to explain responses. Oberauer & Kliegl (2006) used a Boltzmann equation framework based on the activation levels of items stored in working memory to quantify the probability of correct response at the final recall step after memory updating. We expand the original framework with a Markov chain structure, so that it accounts for the probabilities of all possible responses, correct or incorrect, at both the intermediate steps during memory updating and the final recall step after memory updating. We use a Weibull race framework to characterize RT, where the Weibull parameters are associated with the activation levels of items in working memory. This model allows us to investigate the mechanisms underlying choices and RTs in the memory updating paradigm under a joint theoretical framework. A simulation study shows the effectiveness of this model, and posterior predictive distributions and out-of-sample validations show that this model gives a good account of empirical working memory updating findings. We apply the model to two published data sets, where the estimated model parameters can characterize the cognitive properties of each individual, and reflect the cognitive mechanisms

underlying the group differences in experimental groups to be compared. We discuss the theoretical implication of the modeling results.

Keywords

Working memory; Bayesian hierarchical modeling; Interference theory; Memory updating; Reaction time

1 Introduction

Working memory is a complex process composed of both passive maintenance and active manipulation of information (Vecchi & Cornoldi, 1999; Vecchi et al., 2005; Camos & Barrouillet, 2011; Veltman et al., 2003; Masse et al., 2019). Passive maintenance processes, such as storage and recall, do not change the nature of memorized information, whereas active manipulation processes change the information by transformation and manipulation (Vecchi et al., 2005). Both passive maintenance and active manipulation processes have been studied using a memory updating task designed by Salthouse et al. (1991). This task requires the ability to switch attentional focus (Oberauer, 2006) and remove outdated information from working memory (Schmiedek et al., 2007; Ecker et al., 2014). It is often used to test working memory capacity and efficiency, and sometimes it is used as a training task for working memory abilities (e.g. De Simoni & von Bastian, 2018; Waris et al., 2015).

Salthouse et al. (1991)'s memory updating paradigm requires participants to memorize a sequence of stimuli, then perform specified operations one at a time on each of the stimuli for several steps, and then recall the final outcomes for each stimulus. Varied types of stimuli have been used in the task, including digits, alphabetic letters, arrows and location of items (e.g. De Simoni & von Bastian, 2018). Each stimulus type can isolate either the verbal-numerical or visuo-spatial factors of working memory (Oberauer et al., 2000; Kane et al., 2004). We focus our modeling and analysis on the digit version of the task, which emphasizes the verbal-numerical factor, but the model may

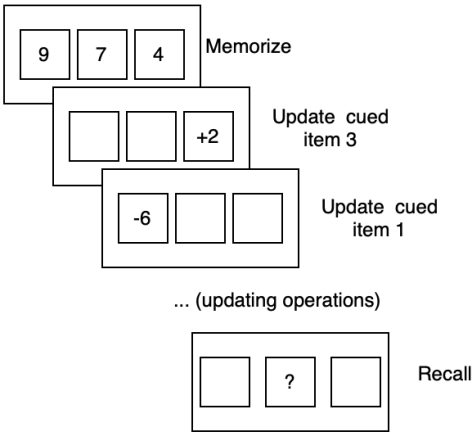


Figure 1: An example of a memory updating task trial. The trial is composed of a memorizing period, multiple updating steps and a final recall period, where participants recall all the items in the order determined by the cue.

be applied to other task versions after minor modifications.

The digit version of the memory updating task usually features a sequence of adjacent boxes, each containing a single digit chosen from 1-9 (see Figure 1). The memory demands of the task increase as the number of boxes increases. After memorizing the digits and their locations in the boxes, the participant is asked to perform a sequence of updating steps by applying a series of arithmetic operations on the digit in the box. During the updating step for each stimulus, the participant must recall the correct digit from working memory and conduct the operation accurately. After the sequence of updating steps, the participant recalls the digits in each cued box one at a time. There are several versions of the memory updating task, including a paradigm that requires intermediate responses after each updating step (e.g. De Simoni & von Bastian, 2018), and a paradigm that does not require such responses (e.g. Oberauer & Kliegl, 2001).

The most commonly used model for the memory updating task data is built on the speed-accuracy trade-off (SAT) function which characterizes the inverse relation between response time and accuracy (Wickelgren, 1977; Heitz, 2014). However, due to its simplicity, the SAT function has only a limited ability to incorporate theory about cognitive mechanisms and corresponding working memory processes. A variety of theory-based models have been proposed to compensate for this shortcoming. The most notable of these are resource theories (e.g. Anderson et

al., 1996; Cowan, 2010), time-based decay theories (e.g. Schweickert & Boruff, 1986; Barrouillet & Camos, 2001; Camos, 2017), and mutual interference theories (e.g. Nairne, 1990; Oberauer & Kliegl, 2006; Oberauer & Lin, 2017). Each theory has a number of corresponding statistical models, such as the time-based resource-sharing model for decay theory (Barrouillet et al., 2004; Oberauer & Lewandowsky, 2011), and activation-framed models for interference theory (Oberauer & Kliegl, 2001, 2006; Oberauer & Lin, 2017). In this paper, we focus on the mutual interference theory and its related modeling.

Interference theory suggests that working memory capacity limitations arise from interference between the items held in working memory. Because the stored items unavoidably share common features with one another, the overwriting of such features impairs the representation of each item. More items in working memory lead to more overwriting, hence the capacity limit. The interference-based mathematical models formalized by Oberauer & Kliegl (2001, 2006) primarily focus on response accuracy but make limited use of RT data. We describe Oberauer and Kliegl's 2001 results in Section 3.1.

In this paper, we propose a hierarchical Bayesian model for the memory updating task. Our model builds on the working memory interference model of Oberauer & Kliegl (2006), and expand it with a Markov chain structure so that the model can account for a wider range of responses at each step of memory updating period, providing a more thorough framework for memory updating performance compared with the original model. The Markov chain structure also allows us to characterize the RTs at each memory updating step under the mutual interference framework. We use a Weibull race component to account for RTs (Logan, 1992; Rouder et al., 2003), and associate the race with the interference component characterized by the Markov chain state at each updating step. Therefore, this model can provide a framework that incorporates both the accuracies of responses and RTs under the interference theory of working memory.

We use a hierarchical structure that allows the parameters from each individual to be informed by group-level hyper-parameters, thus helping to avoid estimation bias caused by potential small sample sizes and outliers (Busemeyer & Diederich, 2010). The model is flexible and can be applied to both the no-intermediate-response paradigm and the intermediate-response paradigm with some

slight modifications. This flexibility allows it to fit data from the majority of memory updating studies.

In what follows, we first describe interference theory and its related model from Oberauer & Kliegl (2006). To develop the hierarchical Bayesian model, we retain the activation-based framework from the original model and characterize it with a Markov chain structure. We link the interference parameters to the RT parameters to formalize the RT race model. We conduct a simulation study to demonstrate the model's parameter recovery ability. We then fit the model to two data sets. The first is from Oberauer & Kliegl (2001); they examined differences in memory performance due to age and did not ask participants to report intermediate results. The second is from De Simoni & von Bastian (2018); they examined working memory training and transfer effects, and they asked participants to report intermediate results. We show that estimated parameters from this model can characterize the group differences shown in these data sets, and provide a theoretical account for the mechanisms underlying the group differences of both responses and RTs.

2 Hierarchical Bayesian model and parameter recovery

In this section, we first describe the mutual interference model proposed by Oberauer & Kliegl (2006). We modify and extend the model to a hierarchical Bayesian framework which incorporates information from both responses and RTs. We then conduct a simulation study to show that the model has reasonable recovery of parameters.

2.1 Interference theory and original model framework

Interference theory assumes that limitations on working memory capacity are a result of mutual interference caused by the shared features of memorized items. Oberauer & Kliegl (2006) assumes that each item is stored as a large number of features in memory. If a proportion A ($A \leq 1$) of features are activated during recall, this item has an activation level A in working memory. Each pair of items is assumed to share a proportion C of features ($0 < C < 1$), but each feature in working memory can only be allocated to one item, and its activation only activates that item.

During memorization, half of the features shared between two items are allocated to each item. Thus, if there is one interfering item present, the target is left with a proportion of $1 - C/2$ features dedicated to it, and it has a maximum activation level of $1 - C/2$ when all these features are fully activated. Suppose that for each pair of items, the features that they share with each other are independent of the features that they share with other items. If there is another interfering item, it shares a proportion C of features among the $1 - C/2$ remaining in the target, and the target is left with a $(1 - C/2) - (1 - C/2)(C/2) = (1 - C/2)^2$ proportion of features after interference with this second interfering item. So, with $n \geq 2$ items present in working memory, one of them being the target and the others distracting competitors, the upper limit of the target's activation level is[1]

$$A_{\text{targ}} = (1 - C/2)^{n-1}.$$

Each competitor item shares a proportion $C/2$ of features with the target. However, it also has interference with the other $n-2$ competitor items, thus the $C/2$ proportion of features are not fully allocated to it. Oberauer & Kliegl (2006) assume that a competitor can maintain $(1 - C/2)^{n-2}$ proportion of features because of its interference with the other competitors, thus the upper limit of a competitor is

$$A_{\text{comp}} = (C/2)(1 - C/2)^{n-2}.$$

Extralist items not present in working memory have an activation level of 0 because no features are allocated to them during memorization.

Oberauer and Kliegl's (2006) model is formulated for the paradigm without intermediate responses and with a time limit imposed for each updating step. It assumes that during the updating steps, the items in working memory gradually activate until their activation levels reach the upper bounds, and this activation process follows a negatively accelerated function (McClelland, 1979; Oberauer & Kliegl, 2001, 2006). Thus, with activation rate r and time limit T, the maximum activation level for the target is

$$a_{\text{targ}} = A_{\text{targ}}(1 - \exp(-rT)),$$

[1] This formula does not exactly partition all features. Some features can be lost during memorization if they are shared by too many items.

and for a competitor is

$$a_{\text{comp}} = A_{\text{comp}}(1 - \exp(-rT)).$$

We add in our model the case when intermediate responses are required and no time limit T is imposed, as shown in Section 2.2.

When applied to the numerical updating task with single digits and arithmetic operations, and when the memory demand is n, the potential recall outcomes are from the digits 1-9, where one of them is the target, $n-1$ are competitors and the remaining $9-n$ are extralist items[2]. It is assumed that participants always choose the item with the highest activation level as the response. Considering the activation process to have a degree σ of noise, the probability of the target having the highest activation among all items is characterized by the Boltzmann equation (Oberauer & Kliegl, 2006)

$$P_{\text{targ}} = \frac{\exp(a_{\text{targ}}/\sigma)}{\exp(a_{\text{targ}}/\sigma) + (n-1)\exp(a_{\text{comp}}/\sigma) + (9-n)\exp(0/\sigma)},$$

where the $9-n$ extralist items have an activation level of 0. Oberauer & Kliegl (2006) give the accuracy of recalling each item correctly as

$$p_{\text{targ}} = 1/9 + (1 - 1/9)P_{\text{targ}}^{m} Q_{\text{targ}},$$

where $1/9$ adjusts for random guessing and $m \geq 0$ is the number of updating steps performed on the current target. In the final recall step, no time limit T is imposed and the activation level can reach the upper bound where $a_{\text{targ}} = A_{\text{targ}}$ and $a_{\text{comp}} = A_{\text{comp}}$. The quantity Q_{targ} is used to characterize the accuracy in the recall step without a time limit imposed.

We base the hierarchical Bayesian model on this scheme, but add some adjustments to formulate the probabilities of choosing competitors and extralist items and to incorporate the information from RTs.

[2]We do not specifically characterize the case when a stimulus appears in more than one box because of its limited occurence and influence.

2.2 Hierarchical Bayesian model

In this section, we expand the framework from Oberauer & Kliegl (2006) with a Markov chain structure to account for the probabilities of all responses at both the updating and recall steps. We also incorporate an RT model into the mutual interference framework so the interference mechanism also explains RTs. We construct the model in the hierarchical Bayesian structure, so that it can fit data sets composed of groups of individuals representing the experimental groups to be compared.

To construct the hierarchical Bayesian model, we denote the group identifier as c and the participant identifier as i. For Trial j with a memory demand of n_j, Participant i has interference model parameters C_{ic}, σ_{ic} and r_{ic}. With no intermediate-response requirement and letting the updating time limit be T_j, denote the choice of target, competitors and extralist items as 1, 2, and 3, respectively. Then the activation levels at the end of each updating step are

$$\begin{aligned} a_{1,ic,j} &= (1 - C_{ic}/2)^{n_j - 1}(1 - \exp(-r_{ic}T_j)), \quad \text{and} \\ a_{2,ic,j} &= (C_{ic}/2)(1 - C_{ic}/2)^{n_j - 2}(1 - \exp(-r_{ic}T_j)). \end{aligned} \tag{1}$$

The corresponding probabilities of choosing the target, competitors and extralist items are then

$$P_{1,ic,j} = \frac{\exp(a_{1,ic,j}/\sigma_{ic})}{\exp(a_{1,ic,j}/\sigma_{ic}) + (n_j - 1)\exp(a_{2,ic,j}/\sigma_{ic}) + (9 - n_j)\exp(0/\sigma_{ic})},$$

$$P_{2,ic,j} = \frac{(n_j - 1)\exp(a_{2,ic,j}/\sigma_{ic})}{\exp(a_{1,ic,j}/\sigma_{ic}) + (n_j - 1)\exp(a_{2,ic,j}/\sigma_{ic}) + (9 - n_j)\exp(0/\sigma_{ic})},$$

and

$$P_{3,ic,j} = \frac{(9 - n_j)\exp(0/\sigma_{ic})}{\exp(a_{1,ic,j}/\sigma_{ic}) + (n_j - 1)\exp(a_{2,ic,j}/\sigma_{ic}) + (9 - n_j)\exp(0/\sigma_{ic})},$$

respectively. To obtain the probabilities of choosing each type of item at each updating step, we

use a Markov chain structure with transition matrix

$$M_{ic,j} = \begin{array}{c} \\ \text{Step x} \end{array} \begin{array}{c} \begin{array}{ccc} \text{Step x+1} & & \\ 1 & 2 & 3 \end{array} \\ \begin{pmatrix} P_{1,ic,j} & P_{2,ic,j} & P_{3,ic,j} \\ \frac{1}{9} & \frac{n_j-1}{9} & \frac{9-n_j}{9} \\ \frac{1}{9} & \frac{n_j-1}{9} & \frac{9-n_j}{9} \end{pmatrix} \end{array} \begin{array}{l} 1 \\ 2 \\ 3 \end{array}.$$

This matrix shows the probability of choosing items of types 1, 2, and 3 given the choice 1, 2, or 3 at the previous step. The first row of $M_{ic,j}$ corresponds to the condition when a participant correctly recalls, updates and memorizes the target at Step x. In this case, when this participant recalls or updates the item in the same location at Step $x+1$, the decision is based on the correct item memorized in Step x. Thus the probabilities of choosing types 1, 2, and 3 are $P_{1,ic,j}$, $P_{2,ic,j}$, and $P_{3,ic,j}$ respectively, shown by the first row of $M_{ic,j}$. The second and third rows correspond to the conditions when a participant incorrectly recalls, updates and memorizes a non-target item at Step x. In this case, recall or update at Step $x+1$ is based on the wrong item from Step x, and we assume that this participant can still obtain the correct result by chance, shown by the second and third rows of $M_{ic,j}$. However, because the participants do not know their mistakes when they make them, they could keep on performing this task regardless of whether their previous updates are correct or wrong. Thus, if a participant successfully gets the correct intermediate result by chance, this participant's subsequent decisions will be based on the correct result, and the state returns to the one corresponding to the first row in $M_{ic,j}$.

We assume that no mistakes are made during the initial memorization period of the task, resulting in a starting state vector $(p_{1,ic,j}, p_{2,ic,j}, p_{3,ic,j}) = (1, 0, 0)$. The final probabilities after updating at the recall step are

$$(p^*_{1,ic,j}, p^*_{2,ic,j}, p^*_{3,ic,j}) = (p_{1,ic,j}, p_{2,ic,j}, p_{3,ic,j}) M^{m_j}_{ic,j} M^*_{ic,j}, \tag{2}$$

where m_j is the number of updates conducted on the item. The matrix $M^*_{ic,j}$ is the matrix for the

recall step without a time limit, computed from

$$a_{1,ic,j} = (1 - C_{ic}/2)^{n_j - 1}, \quad \text{and} \quad a_{2,ic,j} = (C_{ic}/2)(1 - C_{ic}/2)^{n_j - 2}.$$

We model the RTs according to a Weibull race model, where the RT is the minimum of multiple Weibull-distributed racers (Logan, 1992; Rouder et al., 2003; Colonius, 1995). Under the assumptions of the mutual interference theory, each item is stored as a number of features in working memory, where each feature is associated with one specific item. Therefore, we assume that when the participants process a stimulus in this task, they process each of these features simultaneously. For one specific feature, the participants perceive it, activate the item associated with it, (possibly) update the item, make a response, and (possibly) store this item back in working memory. As a result, the processing of each feature is composed of multiple stages. We assume that the time needed for each stage is independently and exponentially distributed with a common rate. Thereby, the processing time for each feature follows a Gamma distribution (McGill & Gibbon, 1965).

Because the participants process multiple features simultaneously, we consider that the response and RT are determined by the first process to finish among them. In this case, when the numbers of features are sufficiently large, the RT can be approximated by the Weibull distribution (Craigmile et al., 2012)[3]. The Weibull density is given by

$$f_w(x|\kappa, \lambda) = (\kappa/\lambda)(x/\lambda)^{\kappa - 1} \exp(-(x/\lambda)^{\kappa}), \quad x \geq 0,$$

where $\kappa > 0$ and $\lambda > 0$ are the shape and scale parameters respectively. The shape parameter κ is associated with the number of stored features: when there are more features stored in working memory, κ decreases; when fewer processing stages are needed to process each feature, κ decreases (Craigmile et al., 2012). As a result, when a participant has a smaller κ, they may be able to store more features of the same item during memorization, and may need fewer stages to process each feature. The scale parameter λ mainly affects the mean and variance of the RT distribution: a larger scale parameter corresponds to a larger mean and variance, which corresponds to the case

[3]See Craigmile et al. (2012) for the conditions under which the Weibull distribution can be used as a limiting distribution.

that a participant needs more time in each processing stage for each feature.

For independent variables $X_k,\ k = 1, 2, \ldots, n$, if

$$X_k \sim f_w(x|\kappa, \lambda_k),$$

the minimum of the X_ks has the distribution

$$\min(X_1, X_2, \ldots, X_n) \sim f_w\left(x\middle|\kappa, \left(\left(\frac{1}{\sum_{k=1}^{n}(1/\lambda_k)^\kappa}\right)^{1/\kappa}\right)\right),$$

and the probability that X_k is the minimum is

$$P(\min(X_1, X_2, \ldots, X_n) = X_k) = \frac{(1/\lambda_k)^\kappa}{\sum_{j=1}^{n}(1/\lambda_j)^\kappa}.$$

To associate the responses to RT data, we use three Weibull racers corresponding to the target[4], competitors, and extralist items respectively. An individual has a limited cognitive capacity α_{ic}. This capacity is divided among the processing of the target, competitors and extralist items, which take up a proportion $p^*_{1,ic,j}$, $p^*_{2,ic,j}$, and $p^*_{3,ic,j}$ of the capacity respectively. Thus we construct the scale parameters as

$$\frac{1}{\lambda_{k,ic,j}} = (\alpha_{ic} p^*_{k,ic,j})^{1/\kappa_{ic}}, \quad k = 1, 2, 3. \tag{3}$$

We assume that when a participant has a larger α_{ic}, they either have a larger capacity, or can make use of that capacity more efficiently. As a result, this participant will have smaller $\lambda_{k,ic,j}$ and faster RTs.

For Participant i in Group c and Trial j, denote the response as $R_{ic,j} \in \{1, 2, 3\}$, where 1, 2, and 3 correspond to the target, competitors and extralist items respectively. Then Equation (3) ensures that

$$\mathrm{P}_n(R_{ic,j} = k) = p^*_{k,ic,j}, \quad k = 1, 2, 3. \tag{4}$$

[4]It is also feasible to construct 9 racers for each digit in a similar format, but due to the characteristics of the Weibull distribution, the 9-racer construct would be statistically equivalent to the 3-racer one and takes more time to run, thus we select the 3-racer construct as a more economical approach.

When $R_{ic,j} = k$, the RT $t_{ic,j}$ has a distribution with density[5]

$$f_n(t_{ic,j}|\kappa_{ic}, \lambda_{ic,j}, R_{ic,j} = k) = f_w(t_{ic,j}|\kappa_{k,ic}, \lambda_{k,ic,j}) \cdot \left(\prod_{m \neq k} (1 - F_w(t_{ic,j}|\kappa_{m,ic}, \lambda_{m,ic,j})) \right) \Big/ \mathrm{P}_n(R_{ic,j} = k). \tag{253}$$

Because the RT data, especially those from relatively complicated tasks, contains some observations from sub-cognitive processes and supra-cognitive processes characterized by very short and long RTs (Kim et al., 2017), we use mixtures to account for such observations. In addition, because the memory updating task often features multiple items, a participant might recall multiple items as a "batch" before items are cued during the recall period. At the start of the recall period, a participant with a sufficient working memory capacity might choose to pre-activate all the items, and keep that information active in working memory. This strategy results in the ability to select a response from this batch of pre-activated items at a faster speed during each step in the recall period (Soto et al., 2008). As a result, some fast RTs from the recall period might be a result of pre-activation, where the participant reads out the items in the pre-activated batch. RTs from pre-activation are likely to be larger than sub-cognitive RTs, but shorter than RTs generated from the algorithmic cognitive process operating on the stored items (see Figure 2). RTs from these pre-activation processes are difficult to distinguish from sub-cognitive processes, thus we integrate them into the same mixture component, and use an additional parameter q_{ic} to estimate response accuracy in the sub-cognitive/pre-activation and supra-cognitive processes. Due to the complex RT structure, the non-decision time will be difficult to identify and estimate, thus we do not include a non-decision time component in the model. We present a general scheme for the model, and adjust the scheme according to the characteristics of each data set.

Denote the mixture proportions for algorithmic, supra-cognitive and sub-cognitive/pre-activation processes by $\phi_{1,ic}$, $\phi_{2,ic}$, and $\phi_{3,ic}$ respectively. For the sub-cognitive/pre-activation processes, because of the inclusion of potential pre-activation process RTs and the difficulty in estimating the mixture component via the model, we use a Weibull distribution to account for such processes for

[5] We replaced $1 - F_w$ with $1.00001 - F_w$ in numerical use to avoid overflow issues.

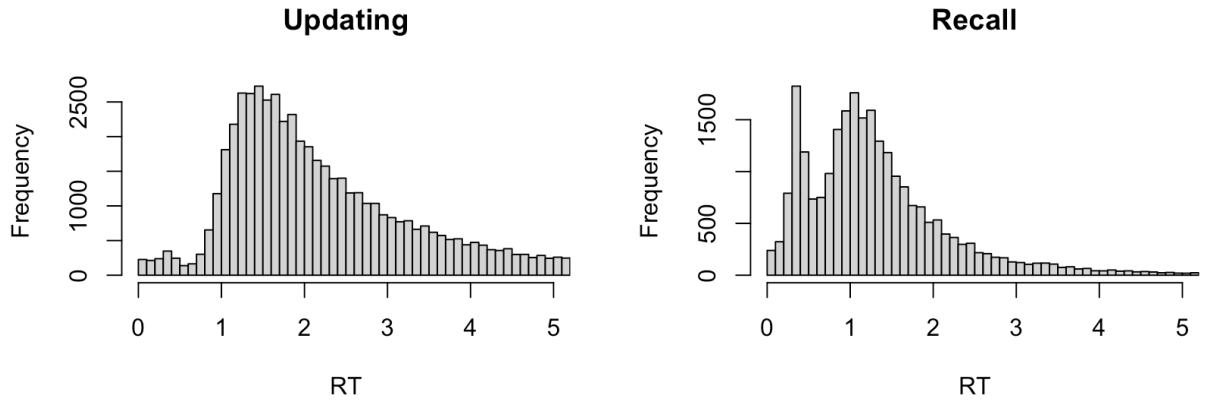


Figure 2: The histogram of the RTs of all 197 participants from De Simoni & von Bastian (2018), from the updating process (left) and the recall process (right). The bin width is taken as 0.1 seconds. The recall process clearly consists of both short sub-cognitive process RTs (close to 0) and supra-cognitive RTs in the tail. Responses from algorithmic cognitive processes are featured by the main peak around 1-1.5 seconds. A sub-peak from pre-activation processes is also present around 0.4 seconds, which is distinct from the main responses but longer than the usual sub-cognitive processes. In comparison, the updating processes contains relatively few sub-cognitive and pre-activation processes.

all participants,

$$t_{ic,j}^{(s)} \sim f_w\left(t \middle| \kappa_{ic}^{(s)}, \lambda_{ic}^{(s)}\right), \tag{5}$$

where $t_{ic,j}^{(s)}$ is the sub-cognitive/pre-activation RT. We use a stepwise procedure: first we estimate the parameters $\kappa_{ic}^{(s)}$ and $\lambda_{ic}^{(s)}$ independently via the short RTs by fitting Equation (5) to the RTs less than a chosen threshold value using a Bayesian model, where the priors are shown in the Appendix. We then plug the estimated posterior means into the full model to avoid identifiability issues in mixture estimation. We model supra-cognitive RTs greater than a boundary b_{ic} as a shifted exponential distribution with rate parameters ψ_{ic},

$$t_{\text{sup},ic,j} - b_{ic} \sim f_w(1, \psi_{ic}), \quad t_{\text{sup},ic,j} \geq b_{ic},$$

where b_{ic} is computed from a quantile of RTs from each data set[6].

In the sub-cognitive/pre-activation and supra-cognitive components, because some participants'

[6]We use different quantiles for different data sets because of RT distribution differences. We state the selection in the data analysis section.

responses are from effective cognitive processing, their response accuracies are higher than chance. We use the parameter q_{ic} to model the probability of accurately choosing the target in the sub-cognitive/pre-activation and supra-cognitive components, where

$$\mathrm{P}(R_{ic,j} = 1) = q_{ic}.$$

We assume that a participant has equal probabilities of choosing each non-target item, thus

$$\mathrm{P}(R_{ic,j} = 2) = (1 - q_{ic})(n_j - 1)/8, \quad \text{and} \quad \mathrm{P}(R_{ic,j} = 3) = (1 - q_{ic})(9 - n_j)/8.$$

In sum, the probability of responses $R_{ic,j}$ in the sub-cognitive/pre-activation and supra-cognitive processes are

$$\mathrm{P}_s(R_{ic,j}) = q_{ic} I(R_{ic,j} = 1) + (1 - q_{ic})(n_j - 1) I(R_{ic,j} = 2)/8 + (1 - q_{ic})(9 - n_j) I(R_{ic,j} = 3)/8, \quad (6)$$

where we assume that the RT and choices are independent in the sub-cognitive/pre-activation and supra-cognitive processes. The indicator function $I(S)$ is 1 when S is true and is 0 otherwise.

Therefore, the RT $t_{ic,j}$ and response $R_{ic,j}$ have a joint distribution with density

$$\begin{aligned} f(t_{ic,j}, R_{ic,j} | \phi_{ic}, \kappa_{ic}, \lambda_{ic,j}, \psi_{ic}) = {} & \phi_{1,ic} f_n(t_{ic,j} | \kappa_{ic}, \lambda_{ic,j}, R_{ic,j}) \mathrm{P}_n(R_{ic,j}) \\ & + \phi_{2,ic} f_w(t_{ic,j} - b_{ic} | 1, \psi_{ic}) \mathrm{P}_s(R_{ic,j}) \\ & + \phi_{3,ic} f_w(t_{ic,j} | \kappa_{ic}^{(s)}, \lambda_{ic}^{(s)}) \mathrm{P}_s(R_{ic,j}), \end{aligned}$$

where $\mathrm{P}_n(R_{ic,j})$ and $\mathrm{P}_s(R_{ic,j})$ are determined by Equations (4) and (6).

In the hierarchical Bayesian framework, we select the priors and hyperpriors as shown in Table 1. These priors make use of link functions to constrain the range of some parameters. The standard deviation δ is fixed[7].

For the case where the intermediate results from the updating steps are required, the activation

[7] We use different values of δ for different data sets to avoid potential identifiability problems in the mixture structure during the sampling process.

Priors		
$\text{logit}(C_{ic}) \sim N(C_c, \delta)$	$\log(\sigma_{ic}) \sim N(\log(\sigma_c), \delta)$	$\log(\psi_{ic}) \sim N(\psi_c, \delta)$
$\log(r_{ic}) \sim N(r_c, \delta)$	$\log(\alpha_{ic}) \sim N(\alpha_c, \delta)$	$\log(\kappa_{ic}) \sim N(\kappa_c, \delta)$
$\text{logit}(q_{ic}) \sim N(q_c, \delta)$	$(\phi_{1,ic}, \phi_{2,ic}, \phi_{3_ic}) \sim \text{Dirichlet}(1, 1, 1)$	
Hyperpriors		
$C_c \sim N(C_0, \delta)$	$\log(\sigma_c) \sim N(\log(\sigma_0), \delta)$	$\psi_c \sim N(\psi_0, \delta)$
$r_c \sim N(r_0, \delta)$	$\alpha_c \sim N(\alpha_0, \delta)$	$\kappa_c \sim N(\kappa_0, \delta)$
$q_c \sim N(q_0, \delta)$	$C_0 \sim N(0, \delta)$	$\log(\sigma_0) \sim N(0, \delta)$
$\psi_0 \sim N(0, \delta)$	$r_0 \sim N(0, \delta)$	$\alpha_0 \sim N(0, \delta)$
$\kappa_0 \sim N(0, \delta)$	$q_0 \sim N(0, \delta)$	

Table 1: The model's priors and hyperpriors.

level of items can reach the upper bound, and the gradual activation part of Equation (1) is removed, so that

$$a_{1,ic,j} = (1 - C_{ic}/2)^{n_j - 1}, \quad \text{and} \quad a_{2,ic,j} = (C_{ic}/2)(1 - C_{ic}/2)^{n_j - 2}.$$

The intermediate results are reflected by Equation (2) with the corresponding steps of updates,

$$(p^*_{1,ic,j}, p^*_{2,ic,j}, p^*_{3,ic,j}) = (p_{1,ic,j}, p_{2,ic,j}, p_{3,ic,j})(M^*_{ic,j})^{m_j + 1}.$$

Because the updating steps always include an arithmetic operation, there is limited impact of pre-activation processes, thus we use mixture component parameters that are different from those of the recall processes.

Denoting the entirety of the model parameters by θ, the likelihood of the model is

$$\mathcal{L}(\theta | R, t) = \prod_{i,c,j} f(t_{ic,j}, R_{ic,j} | \phi_{ic}, \kappa_{ic}, \lambda_{ic,j}, \psi_{ic}).$$

Figure 3 shows the model in graphical form.

2.3 Simulation and parameter recovery

We conducted a simulation study to test the model's parameter recovery ability. We investigated parameter recovery under two conditions: when responses and RTs are only recorded during the

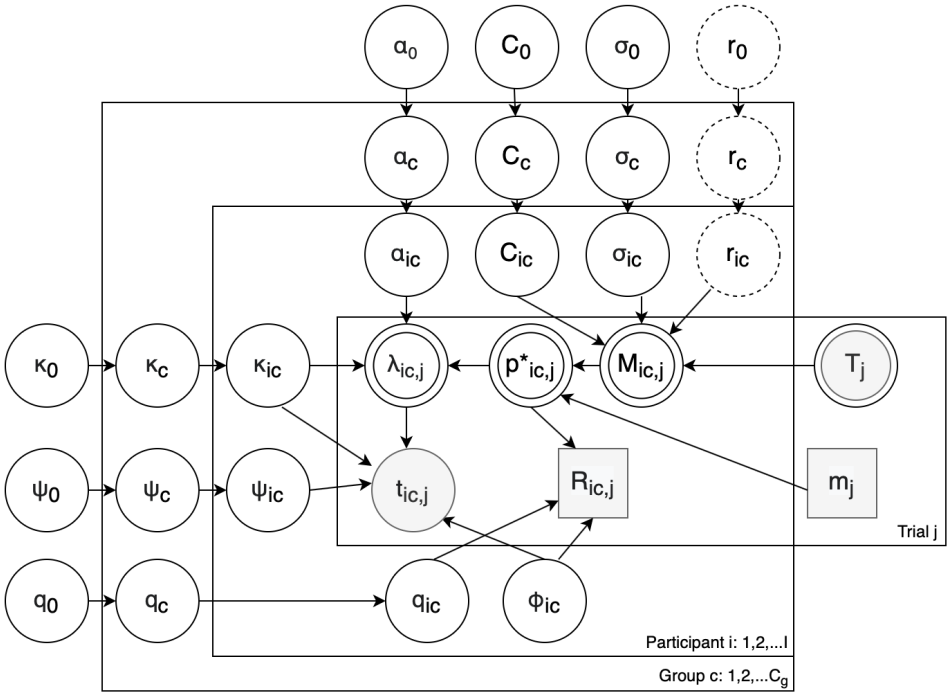


Figure 3: The diagram of the hierarchical Bayesian model. The rectangular boxes contain integers, the round boxes contain real values, and the double-edged boxes contain computed values. The variables from data have gray backgrounds. Parameters to be estimated are unshaded. The arrows indicate dependence. The dashed outlines indicate that the parameters are only used in the no-intermediate-result case. The parameters are embedded in plates representing the hierarchical structure of the model over trials, participants and groups.

recall period for the final results, and when they are recorded both in the updating and recall periods for intermediate and final results. These conditions correspond to the characteristics of the two paradigms and empirical data sets presented by Oberauer & Kliegl (2001) and De Simoni & von Bastian (2018).

For the intermediate condition, we simulated data from 30 artificial participants using the experimental scheme from De Simoni & von Bastian (2018). The parameters of the simulated participants were selected from parameters obtained from Bayesian fits of the model to empirical data from the data analysis section, so they took on values within a reasonable range for an empirical scenario.

Because pre-activation has limited help to the arithmetic updating steps, pre-activation processes rarely contribute during the updating period (see Figure 2). We ignored such mixture

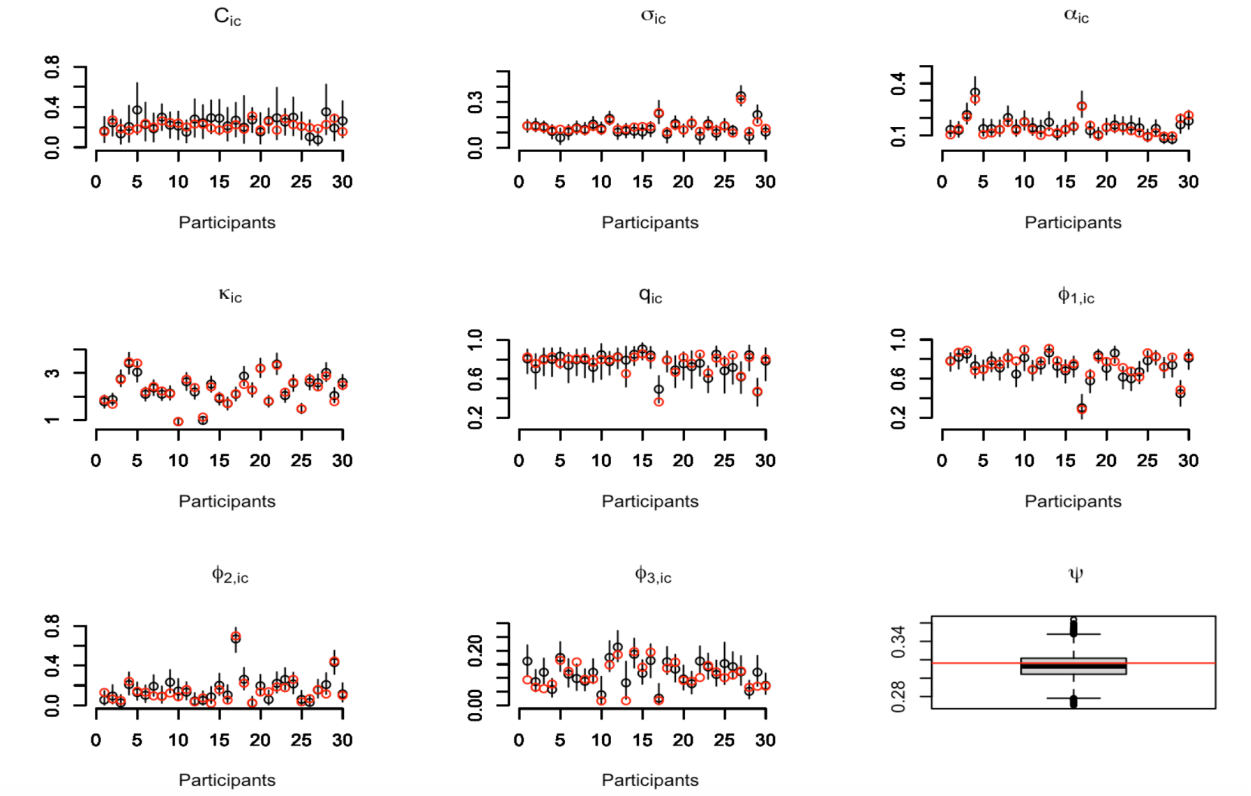


Figure 4: The whisker plots for the parameter recovery results of the simulation. The red points are true parameter values, and the black bars and points are 95% equal-tailed posterior credible intervals and medians.

components and considered $\phi_{1,ic} + \phi_{3,ic}$ to encompass all of the algorithmic cognitive processes in the updating period. We also used a common ψ for the supra-cognitive processes due to the small sample size. According to the scheme, each participant generates 208 responses and RTs over 16 trials. We fit the hierarchical Bayesian model to these data using the software Stan (Stan Development Team, 2018). We obtained 2 chains of parameters, each containing 500 warm-up samples and 4000 iterations, resulting in 8000 posterior samples overall. The effective sample size (Berger et al., 2014) and the Gelman-Rubin $\hat{R}$ statistic (Gelman et al., 1992) ($\hat{R} < 1.01$ for all chains) suggested reliable posterior estimates and satisfactory convergence.

Figure 4 shows the result of the parameter recovery study. Most parameters fall within the 95% equal-tailed credible interval, which indicates a reasonable parameter recovery ability of the model.

For the no-intermediate condition, we generated data from 33 artificial participants using a

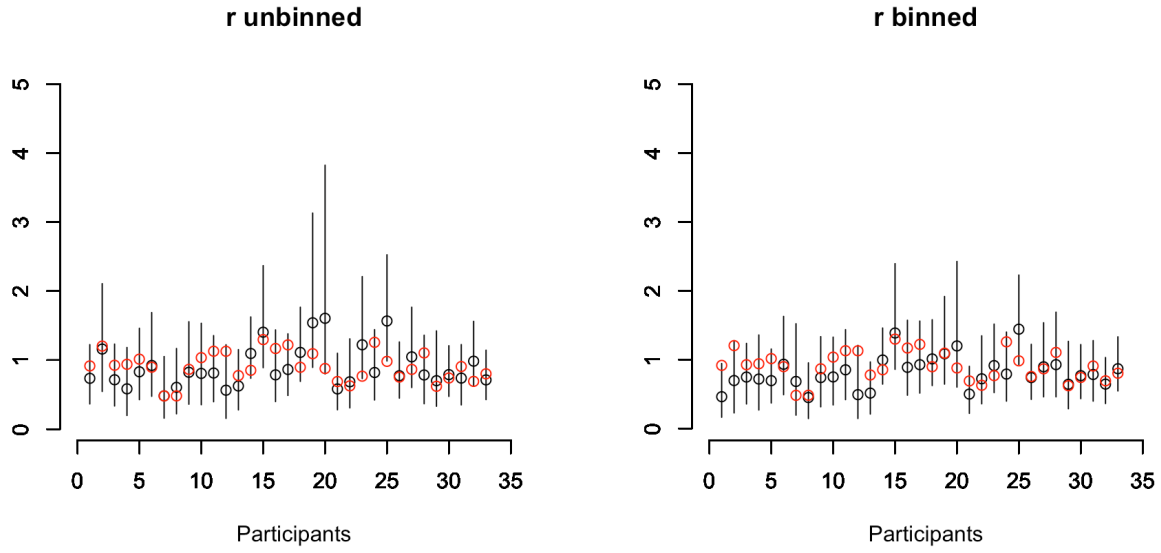


Figure 5: The whisker plots for the parameter recovery results of the parameter r. The red points are true parameter values, and the black bars and points are 95% equal-tailed posterior intervals and medians.

scheme similar to that of Oberauer & Kliegl (2001). Because the full design of the experiment generates large-scale data sets, we reduced the number of trials by randomly selecting each trial in the simulated data set with probability 0.08. Because the updating time limit T_j is variable in the design, and direct application to Equation (1) can be time consuming for large data sets, we binned T_j to reduce the amount of calculation. We binned T_j according to the 3 quartiles, and used the mean in each of the bins as the T_j value for Equation (1).

Using Stan, we generated 2 chains consisting of 500 warm-up samples and 4000 iterations, one for the unbinned T_j case and the other for the binned case. Most parameter recovery results are similar to those from the intermediate simulation. For the activation rate r_{ic}, Figure 5 shows the contrast between the unbinned and binned case. The model has reasonable recovery for parameter r_{ic} for both cases, thus we considered the binning approach to be an acceptable simplification.

3 Empirical data analysis and results

In this section, we report the result of fitting the hierarchical Bayesian model to two empirical data sets, one evaluating age differences in working memory (Oberauer & Kliegl, 2001) and the other evaluating the transfer effect of working memory training (De Simoni & von Bastian, 2018). We first introduce the original studies and the associated constructs of each data set. We made small adjustments to the hierarchical Bayesian model of Section 2 to accommodate the data structure of each study. We show that the model accounts for these data by examining the posterior predictive distributions and using out-of-sample validations. We present the modeling results and discuss the theoretical implications of these results, with a focus on the potential cognitive mechanisms underlying group differences.

3.1 Age differences in working memory

The influence of age on the capacity and efficiency of working memory has been a common topic for investigation (e.g. Wingfield et al., 1988; Salthouse & Babcock, 1991; Oberauer, 2005; Cragg et al., 2017). Older adults are found to exhibit poorer performance in multiple aspects of working memory, such as decreased capacity (Wingfield et al., 1988), a decreased ability to actively manipulate working memory items (Dobbs & Rule, 1989), less accurate recall (Salthouse & Babcock, 1991), and the need to use more resources for the same task (Reuter-Lorenz & Sylvester, 2005), which indicates a potential decline in working memory.

Multiple theories have been proposed to explain such a decline, among which the theory of inhibition (Hasher & Zacks, 1988) has played an important role. According to the inhibition theory, older adults may be more likely to have deficits in their abilities to remove outdated information from memory and replace it with new, updated information, thus wasting usable memory capacity. This deficit in turn increases the limitations on working memory capacity, which leads to comparatively poorer performance for the older adults. The mutual interference theory is compatible with the inhibition theory (Oberauer & Kliegl, 2001), where the increased interference between items in older adults leads to a loss in inhibition. Oberauer and Kliegl evaluated the interference theory

using a statistical model and estimated the interference parameter to be significantly higher for the older group.

However, in addition to accuracy, RTs may also serve as an important, meaningful source of information reflecting the capabilities of working memory. Figure 6 shows that the older group displays overall longer RTs compared with the younger group. Even when an older adult (like Participant 24) has a similar response accuracy as a younger adult (like Participants 3, shown in Figure 6), the older adult can still have longer RTs. The original interference models (Oberauer & Kliegl, 2001, 2006) did not include a mechanism to explain such a difference in RTs.

There can be different potential underlying reasons for the RT difference between older and younger adults. It is possible that the older adults can retain a similar number of features for each item, and use a similar number of stages to process each feature compared to younger adults. However, due to a reduced capacity or efficiency, their processing speed is slower than that of younger adults, and they have larger RT means and variances. This mechanism can be reflected by similar shape parameters κ_{ic} and smaller capacity parameters α_{ic} for older adults. Another possible mechanism could be that older adults store fewer features and require more stages to process each feature. This could be due to reduced inhibition and a reduced ability to identify features of the items. This mechanism could be reflected in a larger shape parameter κ_{ic} for older adults.

We fit the hierarchical Bayesian model to Oberauer & Kliegl (2001)'s data, aiming to characterize the interference mechanism and other potential RT differences between older and younger individuals. In the next section, we first describe the data set, model fit, and some modifications that we made to the model to accommodate the data's characteristics. To demonstrate goodness of fit, we examine the posterior predictive distributions and out-of-sample validation results. We then present the results from the Bayesian model fit and discuss their implications.

3.1.1 Data and model fit

The data set from Oberauer & Kliegl (2001) consisted of 18 younger participants (average age 19.1, sd 0.68) and 18 older participants (average age 68.8, sd 3.55). The original study only analyzed the

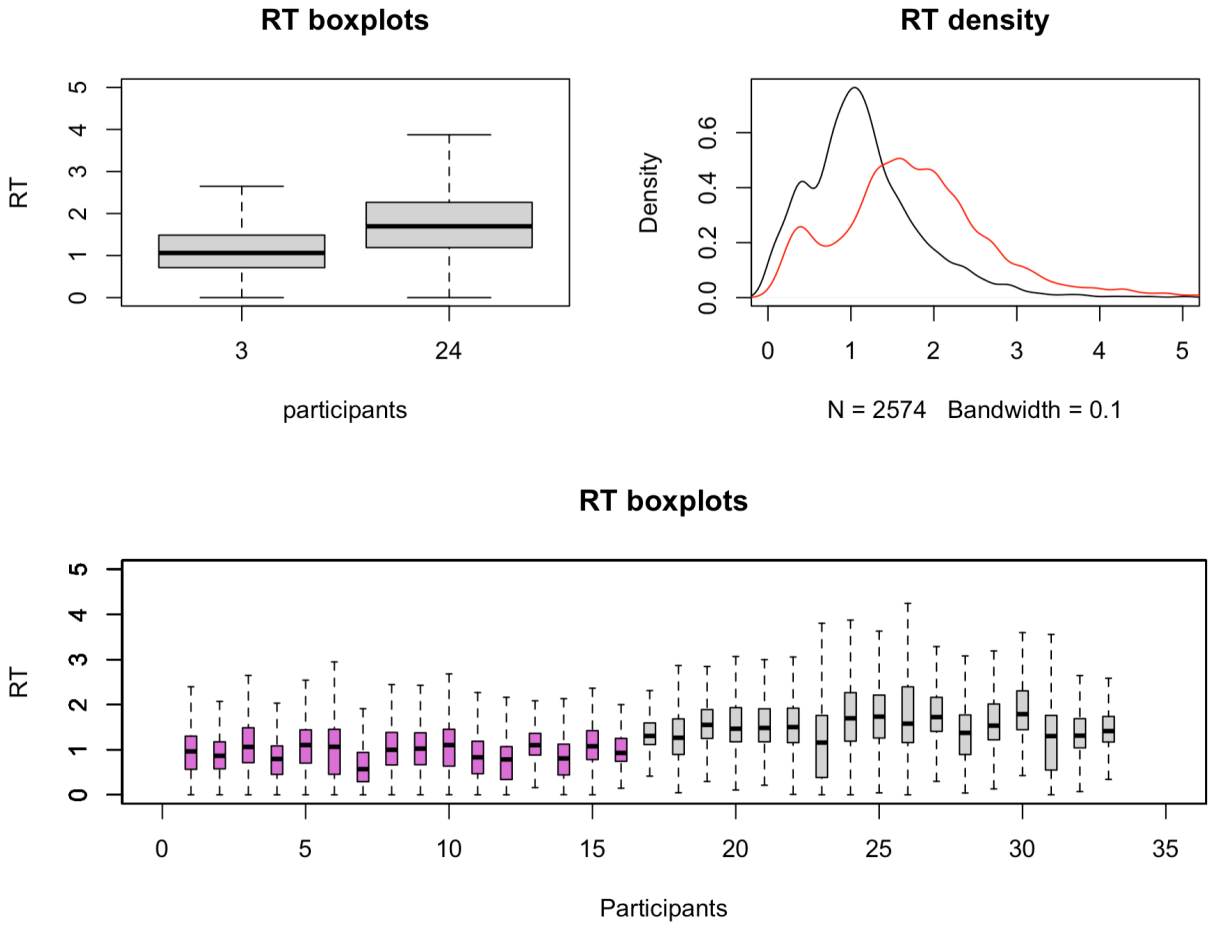


Figure 6: Contrast of RTs from the younger and older groups. The upper figures display the contrast of Participant 3 (Younger) and Participant 24 (Older). The black and red density lines each correspond to Participants 3 and 24 respectively. Despite having similar response accuracy (0.656 for Participant 3 and 0.657 for Participant 24), the RTs are overall much higher in Participant 24. The lower figure shows the box-plots of all participants (purple for younger and gray for older). The older participants generally have slower RTs than younger participants, where the fastest individual median RT from the older participants (1.16s) is larger than the slowest individual median RT from the younger participants (1.10s).

data from 16 younger participants and 17 older participants who completed the entire experiment, thus we also restricted our analyses to these participants. The experiment was composed of two parts: the first part included trials with a memory demand of 1-4, and the second with a memory demand of 4-6. Clear evidence of a learning effect was present between the two parts. For simplicity, we applied the model only to data from the first low-demand part of the experiment.

We fit the hierarchical Bayesian model to the data set using Stan (Stan Development Team, 2018). Due to the mixture characteristics of these data, we used a stepwise fitting approach. We first fit the Weibull distribution to all RTs less than 0.6 seconds using a Bayesian model (see Appendix for details), obtaining the Weibull parameters $\kappa_{ic}^{(s)}$ and $\lambda_{ic}^{(s)}$ for each participant. We then selected the supra-cognitive process boundary b_{ic} to be the maximum of the 95% RT quantile and 2 seconds. We set the parameters' prior standard deviation δ to 1, which is weakly informative and can avoid potential identifiability problems while fitting the mixture model. We obtained 2 chains, each containing 500 warm-up samples and 4000 iterations. The Gelman-Rubin statistic $\hat{R}$ ($\hat{R} < 1.01$ for all chains) and the effective sample size were both satisfactory, indicating good convergence and reliability of the posterior estimates.

To determine goodness of fit, we examined the posterior predictive distributions, and conducted an out-of-sample validation analysis to evaluate how well the model can generalize to new data. Figures 7 and 8 show posterior predictive summaries plotted together with the observed accuracies and RTs[8]. The observed accuracies are consistent with the estimated posterior predictive accuracies. There is more discrepancy between the empirical and the posterior predictive RTs, but the shapes of the densities and the mixture components are close for most participants.

For out-of-sample validation, we extracted a subset of the data set by selecting 50% of the observations at random. We fit the model to this subset and simulated accuracies from the posterior predictive distributions. Figure 9 shows the simulated box-plots of accuracies compared to the true accuracies of the other 50% of data. Despite some differences, most of the simulated accuracies are close to the true values, which indicates that the estimation results from a small subset can be generalized to the remainder of the data set. Thus, we consider the model to have a satisfactory

[8] All densities are computed via the function "density" in R with the shown bandwidth.

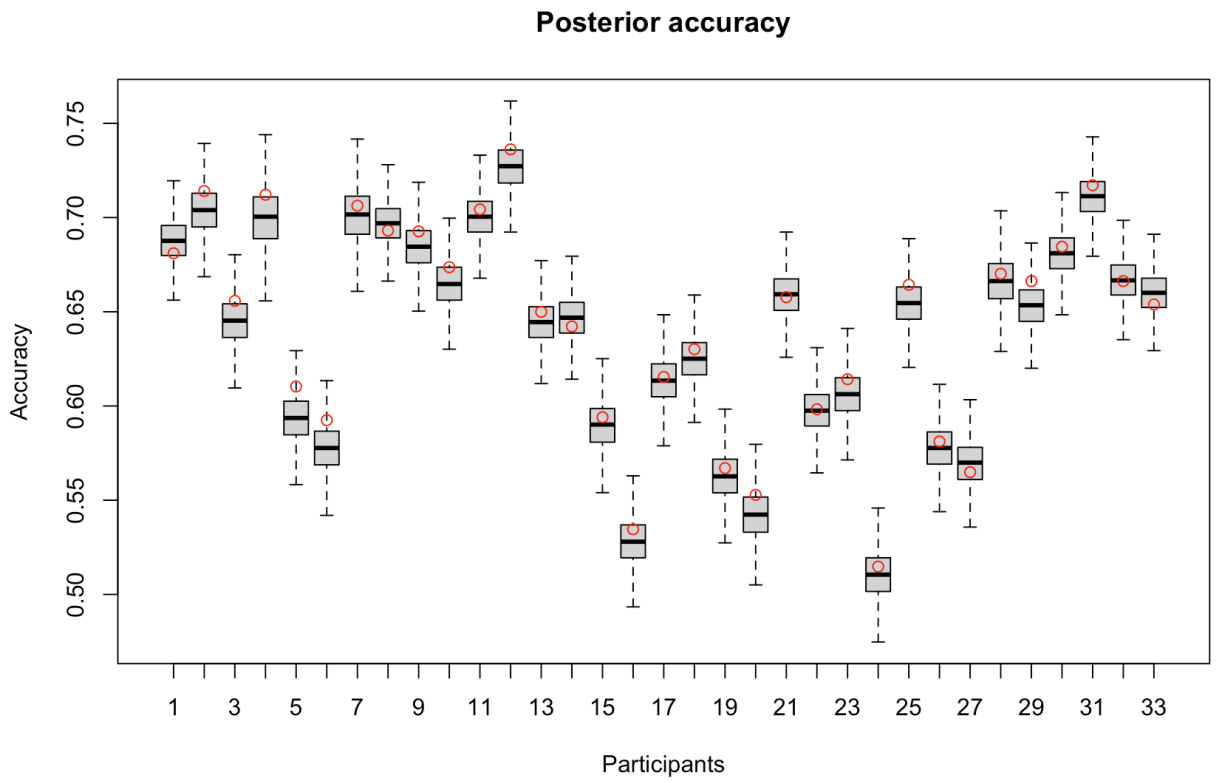


Figure 7: Posterior accuracy for each participant in the data set from Oberauer & Kliegl (2001). The red points are the observed accuracy from the participants, and the corresponding box-plots are the posterior predictive accuracies.

fit to the data and a reasonable ability to generalize.

3.1.2 Results and implications

To analyze the mechanism and implications of the age data, we present the group and individual estimated posterior results. Figure 10 displays the estimated posterior distributions of the group parameters. The interference parameter C and the capacity parameter α have the largest between-group differences: the estimated posterior probability that the younger group has a lower interference C is 0.901, and the estimated posterior probability that the younger group has a larger capacity α is greater than 0.99. The noise parameter σ and the Weibull shape parameter κ have relatively small differences: the estimated posterior probability that the younger group has a lower noise σ is 0.648, and the estimated posterior probability that the younger group has a smaller

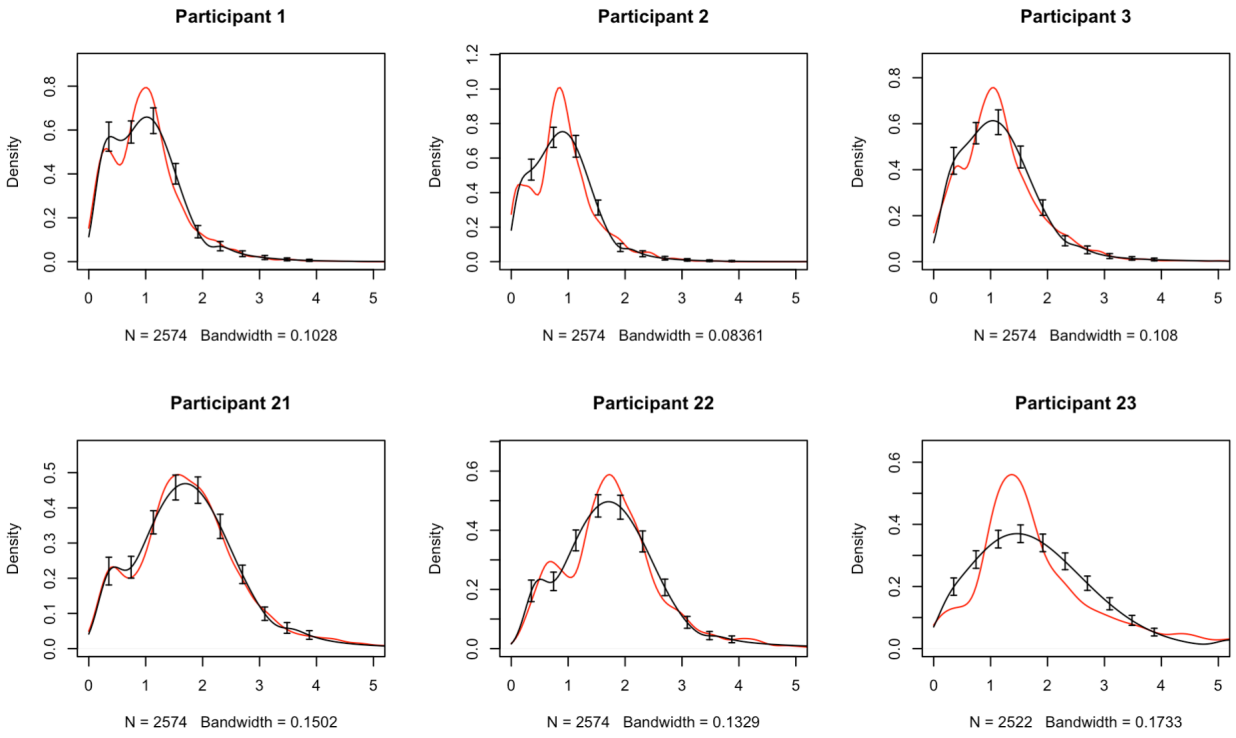


Figure 8: The posterior predictive RT distributions contrasted with the empirical RT distributions for Participants 1-3 (younger) and 21-23 (older). Results of the other participants can be found in the supplemental materials. The red lines are densities from the empirical data. The black lines are the means of the estimated posterior predictive densities, and the black box-and-whisker plots are the pointwise 95% intervals for the estimated posterior densities.

Weibull shape parameter κ is 0.710. The recall rate r has almost no difference: the estimated posterior probability that the younger group has a larger recall rate is 0.496. This implies that the amount of interference C may be a key factor in the performance deficits in the older participants, who may experience a higher level of interference than the younger participants. The difference in the capacity parameter α suggests that another main difference in RT performance may be slower and more variable RTs for the older group.

Figure 11 shows the estimated posteriors for the individual parameters. The parameters C and α display a clear difference in the majority of younger and older participants, further supporting the existence of group-level differences. The parameter κ however, does not show a difference between groups except for a small number of young participants. There are individual differences across all parameters, and some of the participants have clear differences from the majority of the group.

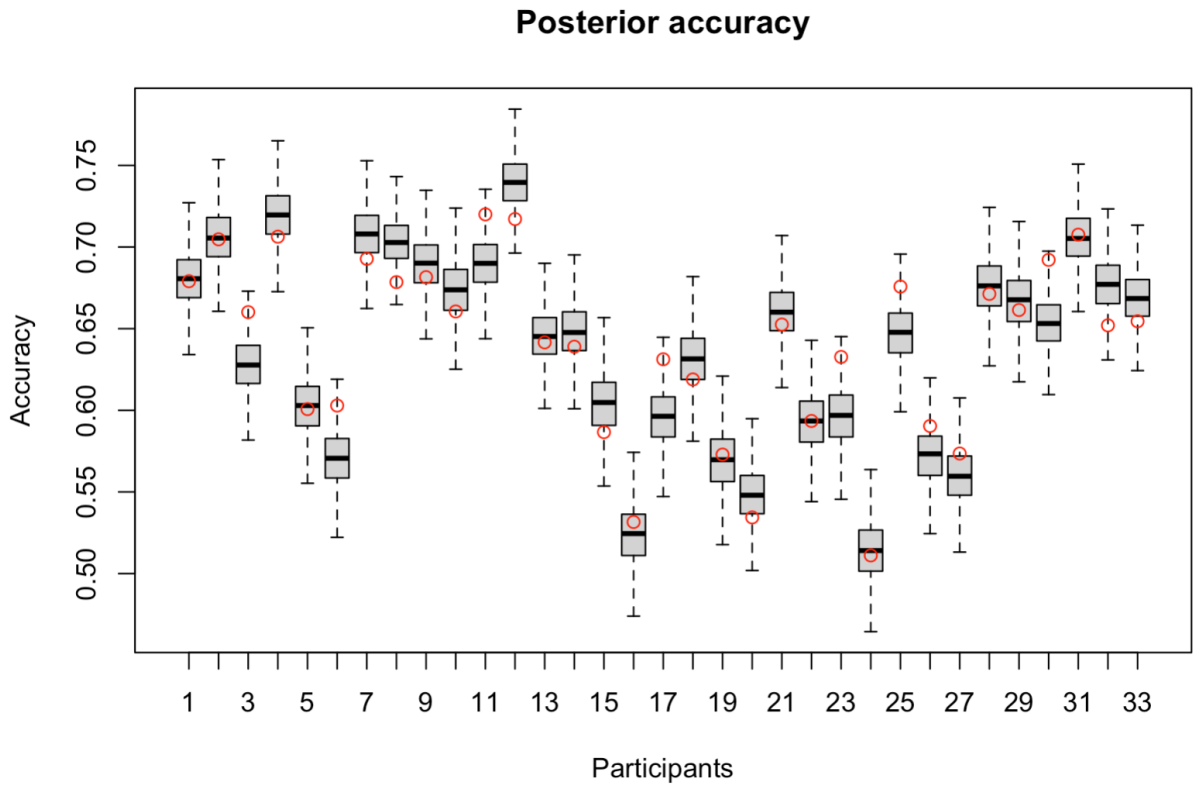


Figure 9: The box-plots of accuracies simulated from the posterior predictive distributions based on the generated out-of-sample validation parameters, compared to the observed accuracies from the data (red points).

The results are consistent with those from the Oberauer & Kliegl (2001) study. Evidence suggests that the interference parameter C is higher for the older group, and the older adults may have a decreased inhibition mechanism with a lowered ability to resist mutual interference between items. The results find no clear evidence to support differences in noise σ, which is also shown in Oberauer & Kliegl (2001). A lack of difference in noise may indicate that the younger and older groups do not generally differ in the variability of their mental representations, nor is the variability a cause of the older groups' working memory attenuation. It is consistent with past findings (Vecchi & Cornoldi, 1999; Vecchi et al., 2005) that aging affects the accuracy of active manipulation processes (here represented by inhibition and interference) more than passive maintenance processes (here represented by noise in storage and recall).

In the RT results, no evidence is found to support any differences in the numbers of stored fea-

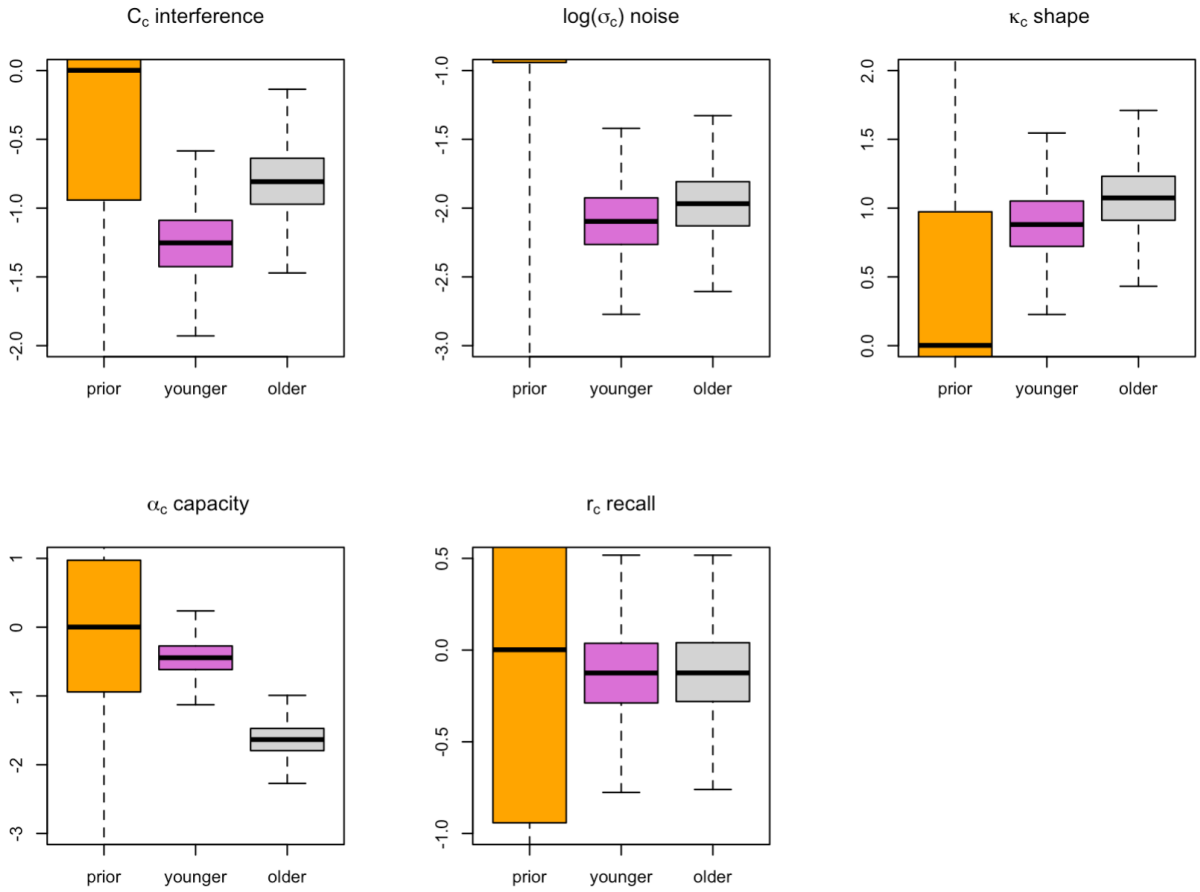


Figure 10: The box-plots for the priors and estimated posteriors of the group-level parameters. Orange box-plots are the priors. The purple and gray box-plots corresponds to the younger and older groups, respectively. These parameters can have negative values because of the transformations shown in Table 1.

tures and processing stages between the younger and older groups, shown by the shape parameters that have little group differences. The older adults may differ from the younger adults only in the speed to process each stage, such as recalling and responding, shown by difference in the parameter α. This may indicate that they have a smaller cognitive capacity compared to younger adults, or they cannot use their capacity as efficiently. Past studies (Salthouse & Babcock, 1991; Caplan et al., 2011) also find older adults to be slower to process and execute elementary working memory operations. Therefore, the older adults might use a similar general cognitive structure to process the tasks, but spend more time executing the same processes than the younger adults.

The individual parameters show some results worthy of notice. The RT shape parameter is

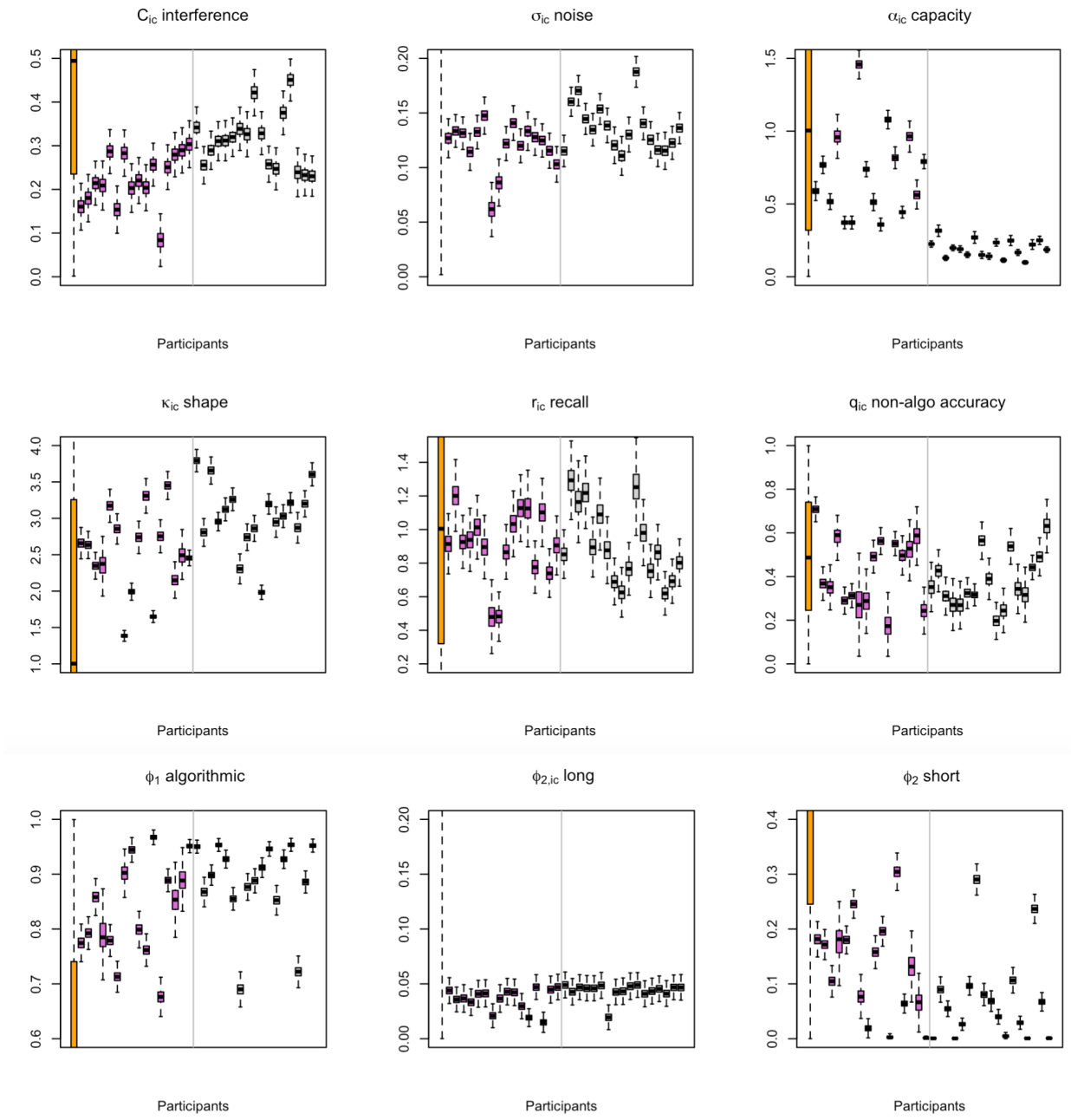


Figure 11: Box-plots for the priors and estimated posteriors of the individual-level parameters. Younger participants are shown in purple (left of the vertical gray line) and older participants are shown in gray (right of the vertical line).

clearly smaller for a number of participants, mostly in the younger group. These participants may tend to store more features and use fewer processing stages for each feature than the others. It might be an indication that these participants potentially adapted strategies to simplify the process. The mixture components for sub-cognitive/pre-activation processes are higher for younger than older participants despite generally better accuracy, which may show a noticeably larger proportion of pre-activation processes for the younger adults. This may imply that the younger group is more capable of using cognitive resources for pre-activation processing than the older group, who relies more on algorithmic processing.

3.2 Transfer effects of working memory training

De Simoni & von Bastian (2018) conducted a study to evaluate the transfer effect of working memory training on the improvement of cognitive abilities. Working memory is related to many cognitive abilities and related human performance (e.g. Oberauer et al., 2008; Cragg et al., 2017), leading to research about the effect of working memory training transferring to other abilities (e.g. Borella et al., 2010; Schwaighofer et al., 2015; von Bastian & Oberauer, 2013). The transfer effect assumes that the improvement of working memory ability, gained via training, can be transferred to improve other related abilities (De Simoni & von Bastian, 2018; Shipstead et al., 2010, 2012). Transfer effects are categorized as near transfer and far transfer effects (e.g. Shipstead et al., 2010). In near transfer, the benefit of training for a specific type of working memory task transfers to performance of other working memory tasks (e.g. Hovik et al., 2013). In far transfer, working memory task training transfers to abilities outside the working memory domain (e.g. Bigorra et al., 2016). In the literature, there is substantial evidence both for (e.g. Minear et al., 2016) and against (e.g. Sala & Gobet, 2017) the general benefit of transfer effects.

De Simoni & von Bastian (2018) designed an experiment to search for both near and far transfer effects. For near transfer, participants were divided into three groups, two of them receiving training in different working memory tasks, namely memory updating tasks and binding tasks, and a control group that received training in visual search tasks. All participants performed all three types of tasks along with a battery of cognitive tests before the training, then received training of their

specific allocated task across five weeks, followed by the same battery of cognitive tasks. Such a design allowed recording of performance and evaluation in both near transfer to different working memory tests and far transfer to other cognitive abilities. De Simoni & von Bastian (2018) used measurement statistics and a latent-variable confirmatory factor analysis to investigate transfer effects. Despite improved performances in the trained tasks, little evidence was found to support the theories of near and far transfer. They concluded that working memory training is more likely to induce the use of stimulus-specific strategies than general transfer effects.

With the intention to investigate the near transfer effect to memory updating performance, we applied the hierarchical Bayesian model to the pre-test and post-test memory updating data. If the training from both memory updating and binding training lead to improvement in memory updating performance, then the estimated parameters of both groups should differ from those of the control group. One plausible result is a decrease in the interference parameter C that corresponds to reduced mutual interference. There could also be changes in the storage and processing of features, and the speed of response processes, shown by RT parameters κ and α respectively.

Apart from near transfer, other mechanisms could also cause non-transfer practice effects. De Simoni & von Bastian (2018) considered reduction of extralist errors to be a large factor in performance improvement. An extralist error is the recall of an item that is not in the current memory array, which we referred to as response type 3 earlier. Their analysis shows that memory updating training can largely reduce the number of extralist errors compared with the binding task training. Such a mechanism may arise as a reduced noise parameter σ, which results in a smaller chance of choosing an extralist digit. If extralist errors happen more often during sub/supra-cognitive processes, a reduction of the mixture parameter ϕ may arise for these processes, so there are less frequent activations of sub/supra-cognitive processes and as a result fewer extralist errors. Another possibility is that some participants could use task-specific strategies to improve their performance. This event would be difficult to identify, but if the parameters of a participant sub-group are different from the others, then they might have used task-specific strategies.

3.2.1 Data and model fit

The data set of De Simoni & von Bastian (2018) includes 216 participants. They excluded 19 participants from the analysis for reasons such as programming errors and abnormal response patterns. Thus we also used the data from the remaining 197 participants for the hierarchical model analysis. The memory updating training group had 59 participants, the binding training group had 66 participants, and the visual search control group had 72 participants. Each participant provided data from pre-test and post-test memory-updating sessions. Each session contained 16 trials, where each trial was composed of 9 updating steps and either 3 or 5 recall steps corresponding to two different memory demands. As such, each participant generated 416 observations overall.

To accommodate the pre/post-test nature of this study, we permitted the parameters C, σ, α and κ to differ for pre-test and post-test conditions. Denoting a parameter as θ, it is modeled as

$$\theta_{ic}^{(k)} \sim N(\theta_c^{(k)}, \delta), \quad \theta_c^{(k)} \sim N(\theta_0^{(k)}, \delta), \quad \theta_0^{(k)} \sim N(0, \delta), \quad k = 1, 2,$$

where $k = 1, 2$ corresponds to the pre-test and post-test conditions respectively. The mixture proportions ϕ_{ic} were also allowed to be different for pre and post-test conditions. Because of the small individual sample sizes, we applied more informative priors by selecting δ to be 0.3 during model fitting. Because the proportion of potential sub-cognitive processes are small and hard to estimate during the updating step (see Figure 2), we eliminated it and let $\phi_{1,ic} + \phi_{3,ic}$ be the mixture proportion for the algorithmic cognitive process during updating. Due to the small sample size from each participant, we used common parameters ψ, $\kappa^{(s)}$ and $\lambda^{(s)}$ for all participants. We fit the Weibull distribution to RTs smaller than 0.65 seconds to obtain the sub-cognitive/pre-activation process parameters $\kappa^{(s)}$ and $\lambda^{(s)}$ (see the Appendix for more detail), and used these values in the full model. We used each individual's 90% RT quantile as the supra-cognitive process boundary b_{ic}.

We fit the hierarchical Bayesian model using Stan to generate 2 parameter chains, each consisting of 500 warm-up samples and 4000 iterations. The Gelman-Rubin statistic $\hat{R}$ ($\hat{R} < 1.01$ for all chains) and the effective sample size were both reasonable, indicating satisfactory convergence and

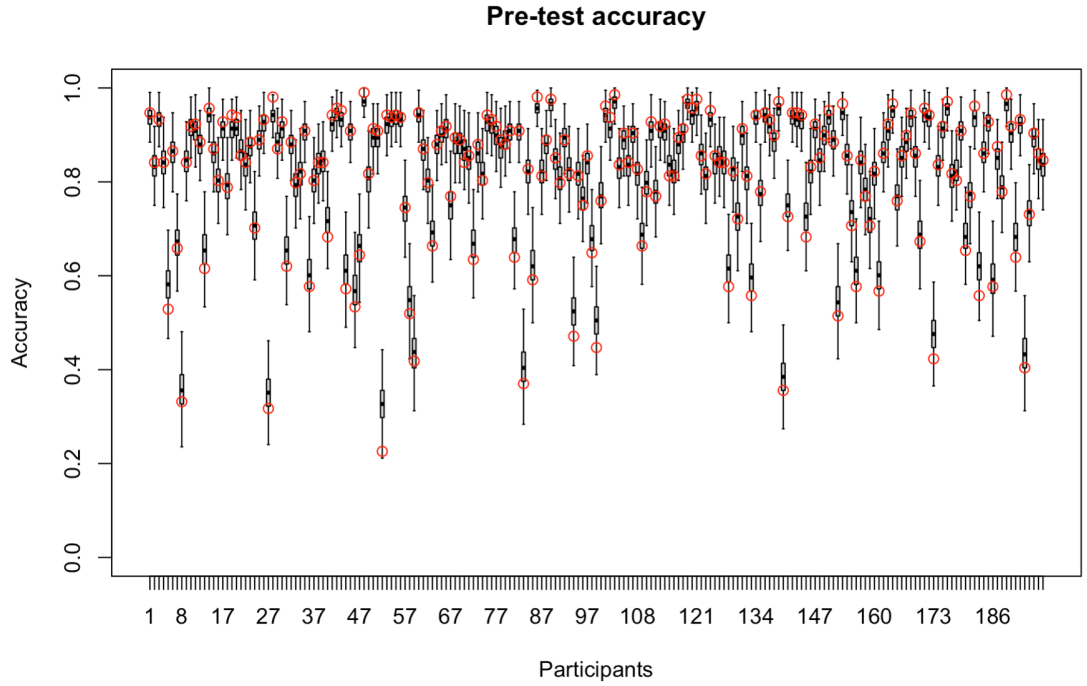
reliable posteriors.

To demonstrate model fit, we present the estimated posterior predictive distributions in Figures 12 and 13. These figures show that the posterior predictive accuracies can recover the observed accuracies reasonably well. The posterior predictive RT densities show some level of divergence from the empirically estimated densities, where the peaks are underestimated, but their shapes and locations are generally consistent.

To perform out-of-sample validation, we drew 50% of the data set at random. We applied the model to the subset, simulated accuracies from the posterior predictive distributions and plotted the mean predictive accuracies against the observed accuracies for the rest of the data in Figure 14. Although small accuracies are overestimated, the divergence from a linear pattern is not major considering the small sample size, which indicates reasonable generalizability of the model.

3.2.2 Results and implications

In this section, we display a subset of estimated group and individual posterior parameters, and discuss their patterns and implications. Figure 15 shows the estimated posterior distributions of the pre/post-test differences of group parameters. Among them, the estimated posteriors of $C_c^{(2)} - C_c^{(1)}$ and $\alpha_c^{(2)} - \alpha_c^{(1)}$ show considerable group differences. Participants who receive memory updating training (purple group) show a clear decrease in the interference parameter C compared with other groups, where the estimated posterior probability that $C_c^{(2)} - C_c^{(1)}$ is less than 0 is 0.877, 0.999, and 0.142 for the control, updating-trained and binding-trained groups, respectively. These results indicate that memory updating training can help to reduce the level of interference, and may also indicate that binding training might not only fail to reduce, but also increase the level of interference. The capacity parameter α increases in all groups, with the memory-updating group showing the largest increase, where the estimated posterior probability that its increase is larger than other groups is 1. The binding-trained group has the second largest increase in α, where the estimated posterior probability that its increase is larger than the control group is 0.952. The control group shows the smallest increase, where the estimated posterior probability that its increase is larger than 0 is 0.998. This may indicate that training, regardless of type, can reduce the



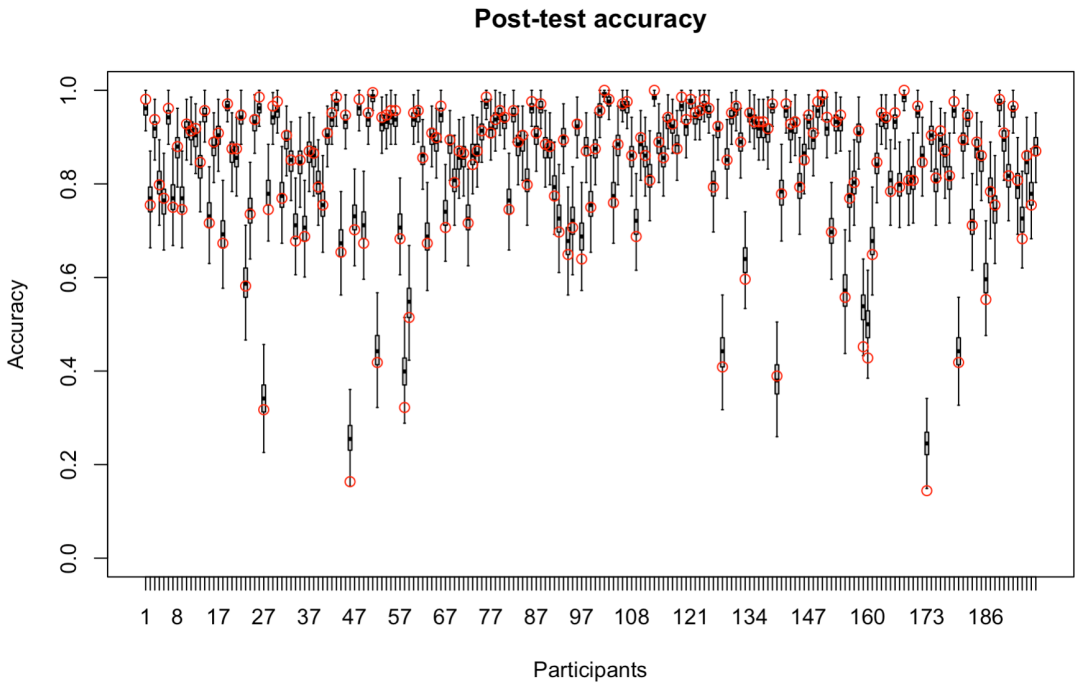


Figure 12: Posterior accuracy for each participant in De Simoni & von Bastian (2018)'s experiment. The red points are the observed accuracy from the participants, and the corresponding box-plots are the posterior predictive accuracies.

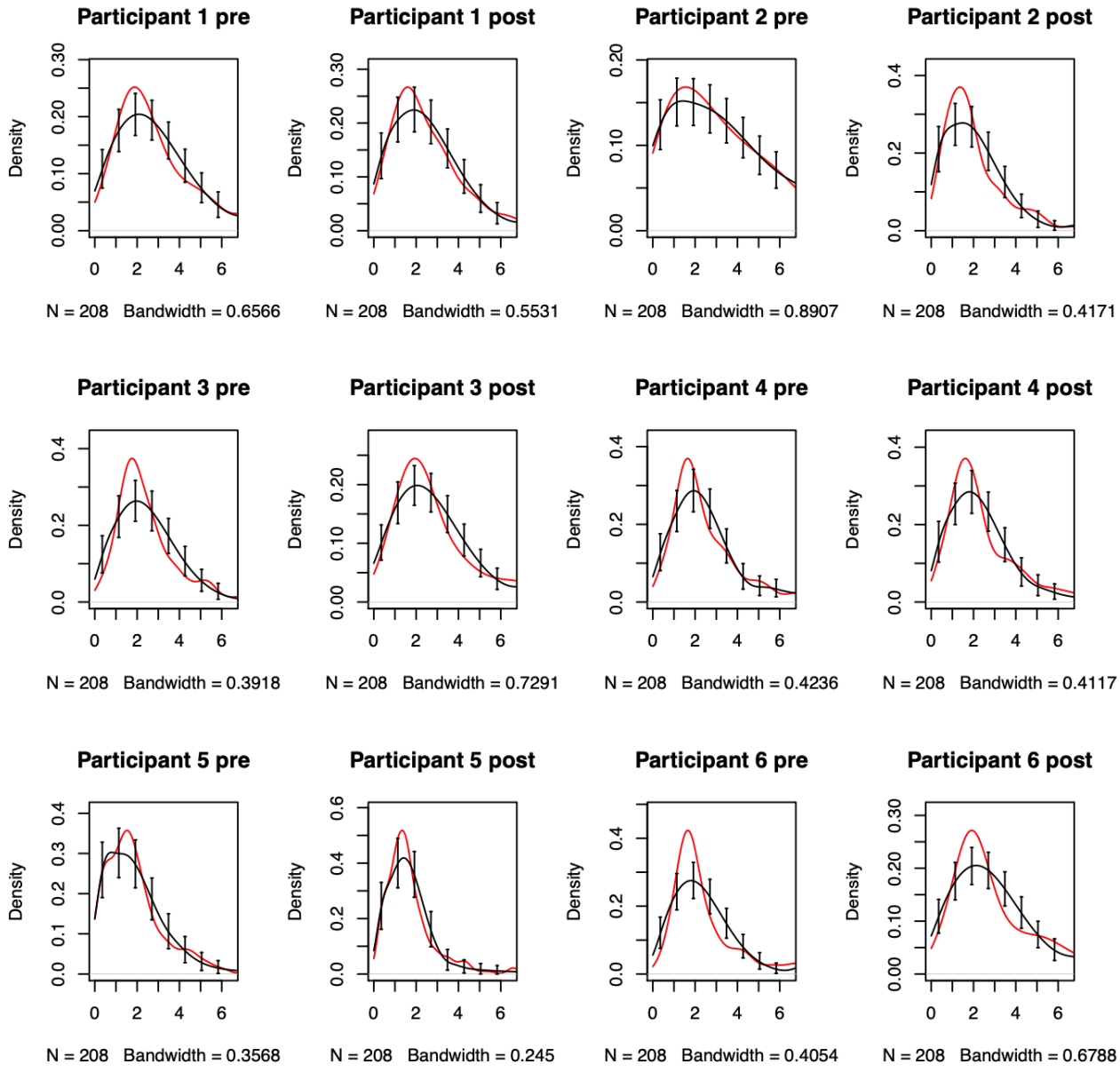


Figure 13: The posterior predictive RT densities (black lines) contrasted with the empirical RT densities (red lines) for Participants 1-6. Results of the other participants can be found in the supplemental materials. The black lines are plotted through the means of the estimated posterior predictive densities, and the box-and-whisker plots are the pointwise 95% intervals for the estimated posterior densities.

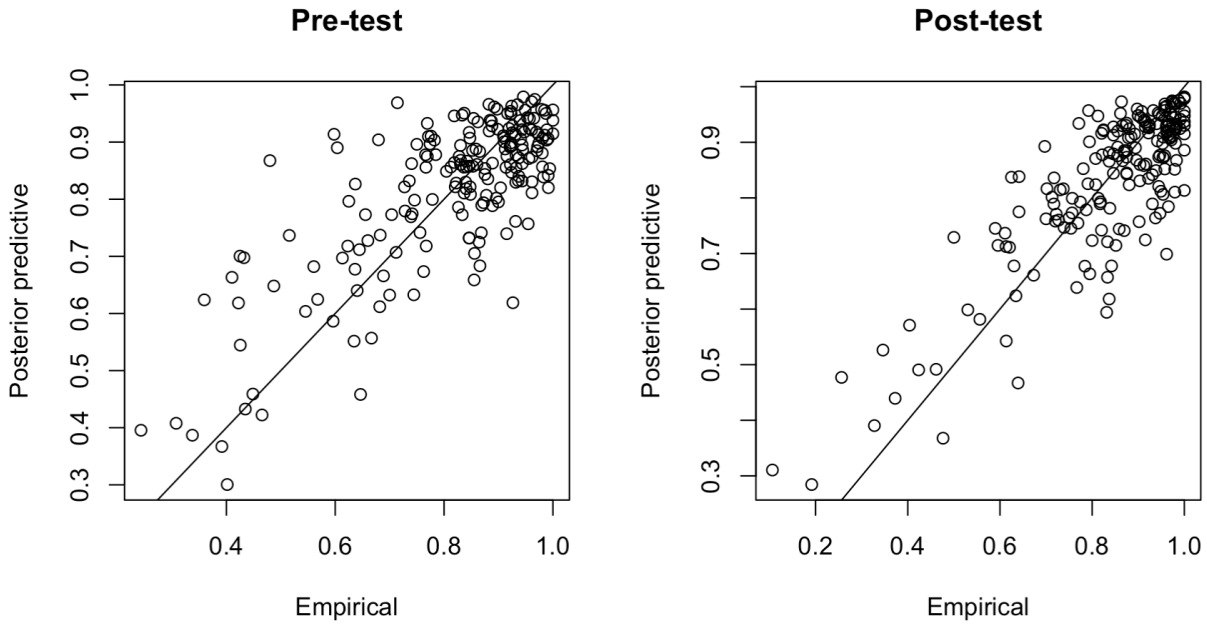


Figure 14: The scatter-plot of mean accuracies simulated from the posterior predictive distributions based on the generated out-of-sample validation parameters (y-axis), compared to the observed empirical accuracies from data (x-axis). The black line is the identity line.

mean and variance of the RT distribution, possibly by improving the speed of executive working memory functions (Covey et al., 2019). The other parameters do not show clear pre/post-test differences, and thus may have not received as much influence from training.

Figures 16, 17 and 18 display individual pre-post test differences of the interference parameters C, capacity α and the mixture weight of the algorithmic component ϕ_1. Figure 16 shows a pattern of effects on the interference parameter C that is consistent with that of the group-level parameters, where most participants in the memory-updating group show a decrease in C. In the memory-updating group, 31 out of 59 participants have an estimated posterior probability larger than 0.8 that $C_{ic}^{(2)} < C_{ic}^{(1)}$; the control group has 19 such participants out of 72, and the binding group has 8 such participants out of 66. Thus, memory-updating training may reduce the degree of mutual interference in working memory processing, while binding training, despite its importance for working memory abilities, cannot transfer its benefit to the memory updating task by influencing active manipulation processes such as interference.

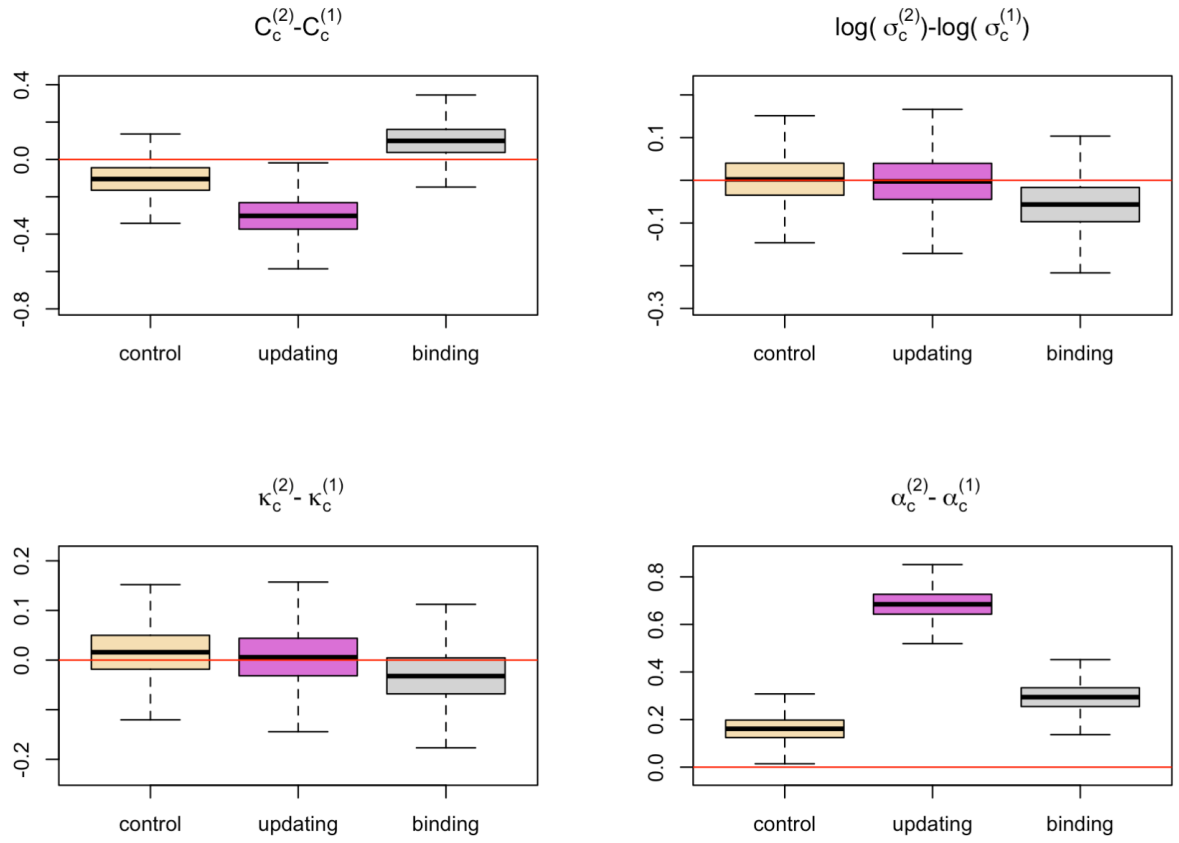


Figure 15: Posterior box-plots for the pre-post test differences of group parameters. Parameters $\theta_c^{(1)}$ ($\theta \in (C, \sigma, \kappa, \alpha)$) corresponds to the pre-test condition, and $\theta_c^{(2)}$ corresponds to the post-test condition. The control group, memory-updating group and binding group are each colored in brown, purple and gray respectively.

The capacity parameter α in Figure 17 is larger for the memory-updating group after training. In the memory-updating group, 55 out of 59 participants have an estimated posterior probability larger than 0.8 that $\alpha_{ic}^{(2)} > \alpha_{ic}^{(1)}$; the control group has 39 such participants out of 72, and the binding group has 42 such participants out of 66. For the binding group, the parameter α is slightly but consistently larger in post-test, shown in Figure 17. The control group shows mixed patterns, where quite a few participants have smaller or unusually larger parameters after training. This variability in the control group may be an indication that visual search training cannot result in a consistent change. Because the shape parameter κ has almost no pre/post-test change, the parameter α's change may indicate a decrease of RT mean and variance that is unrelated to how

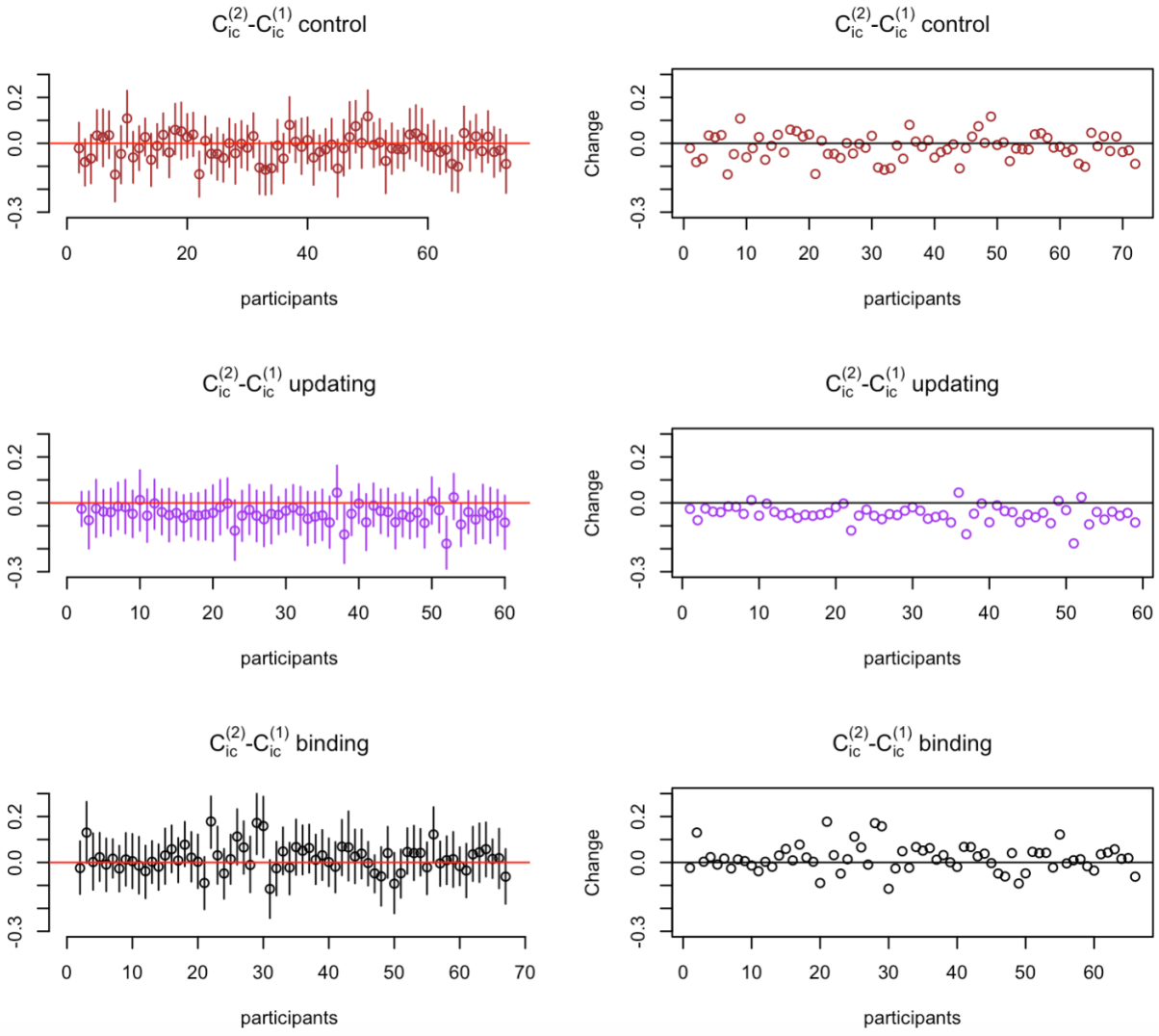


Figure 16: The whiskerplots of pre-post test differences (left) and the change in means (right) of individual posteriors for the interference parameter C. The brown plots in the top panel are for the control group, the purple plots in the center are for the memory-updating group, and the gray plots on the bottom are for the binding group. The left column shows whiskerplots, where the points are placed at the posterior medians, and the whiskers are the 95% equal-tailed credible intervals. The right column shows the means of the estimated posterior distribution of $C_{ic}^{(2)} - C_{ic}^{(1)}$.

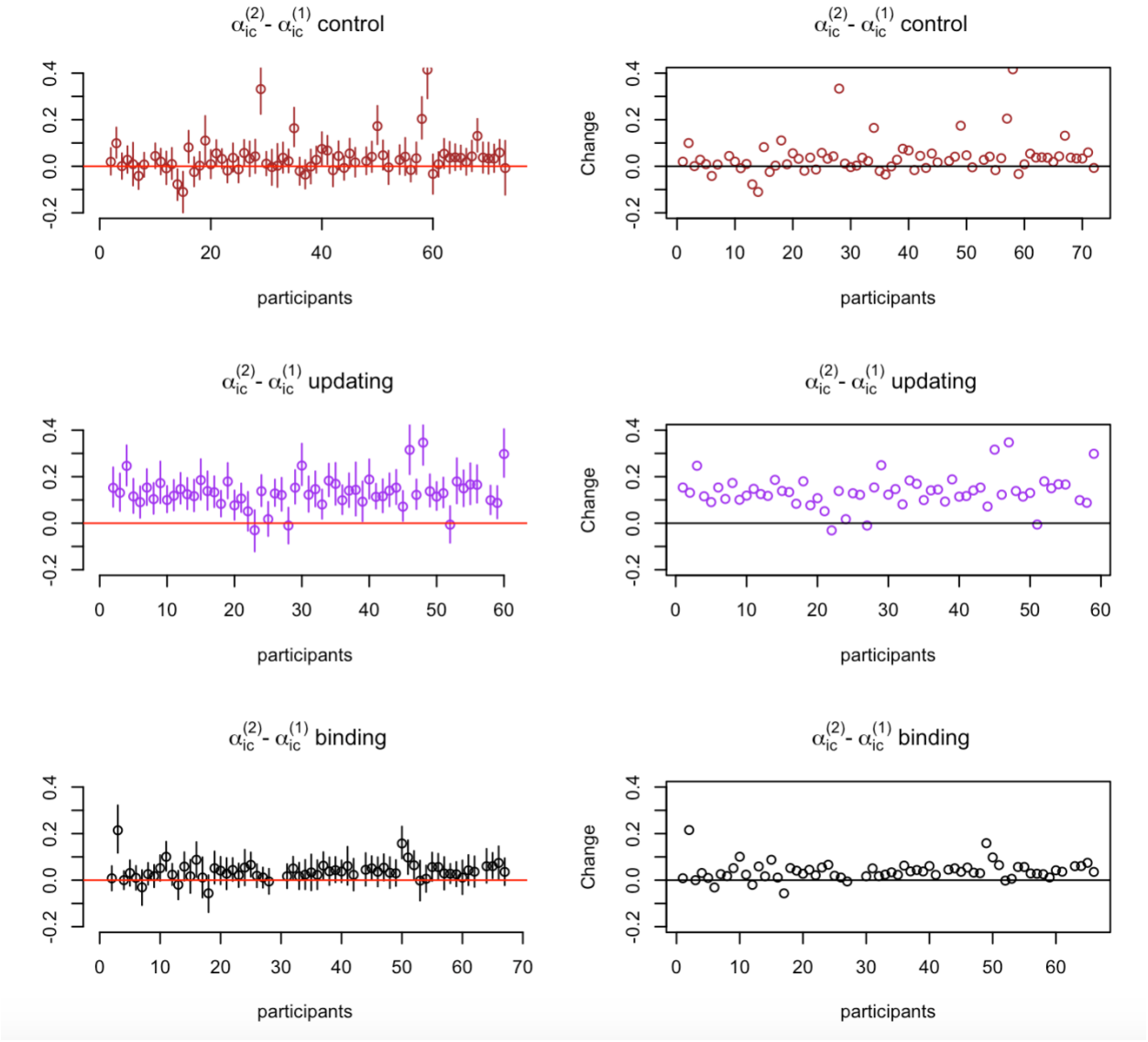


Figure 17: The whiskerplots of pre-post test differences (left) and the change in means (right) of individual posteriors for the capacity parameter α. The brown plots in the top panel are for the control group, the purple plots in the center are for the memory-updating group, and the gray plots on the bottom are for the binding group. The left column shows whiskerplots, where the points are placed at the posterior medians, and the whiskers are the 95% equal-tailed credible intervals. The right column shows the means of the estimated posterior distribution of $\alpha_{ic}^{(2)} - \alpha_{ic}^{(1)}$.

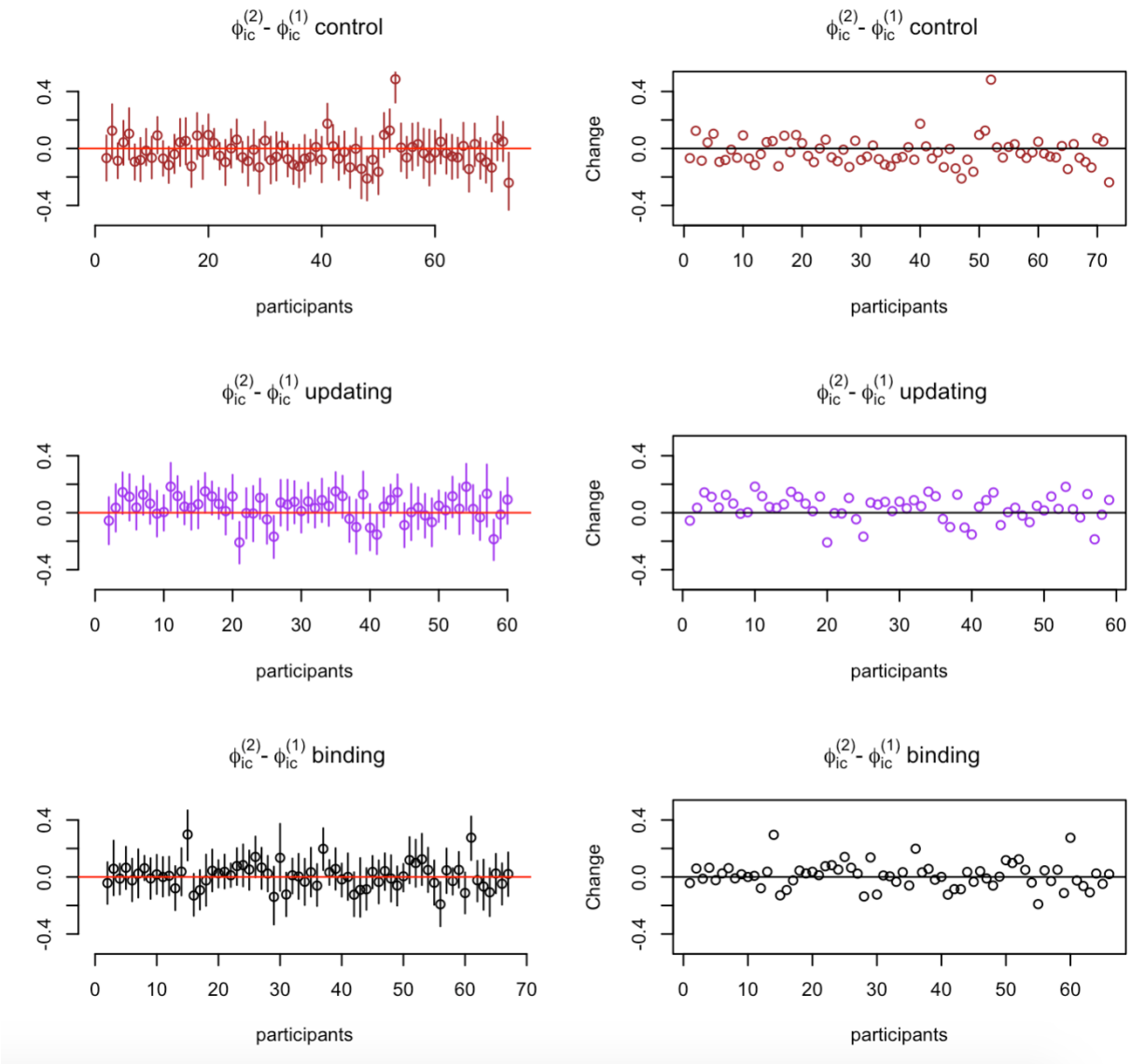


Figure 18: The whiskerplots of pre-post test differences (left) and the change in means (right) of individual posteriors for the proportion of algorithmic processes ϕ_1. The brown plots in the top panel are for the control group, the purple plots in the center are for the memory-updating group, and the gray plots on the bottom are for the binding group. The left column shows whiskerplots, where the points are placed at the posterior medians, and the whiskers are the 95% equal-tailed credible intervals. The right column shows the means of the estimated posterior distribution of $\phi_{ic}^{(2)} - \phi_{ic}^{(1)}$.

features are processed. It may be the result of familiarizing and speeding up passive working memory processes, where the participants learned to use the cognitive capacity more efficiently. The change in the binding group may be due to the shared passive components in binding and memory updating, so the benefit from binding training can also benefit related passive components in the memory updating task.

The mixture proportion of the algorithmic cognitive process ϕ_1 (Figure 18) is slightly larger for the memory-updating group but not for other groups. In the memory-updating group, 25 out of 59 participants have an estimated posterior probability larger than 0.8 that $\phi_{ic}^{(2)} > \phi_{ic}^{(1)}$; the control group has 11 such participants out of 72, and the binding group has 12 such participants out of 66. This increase for some participants in the memory-updating group may be related to the reduction of extralist errors. It is likely that training reduces extralist errors by reducing the proportion of sub/supra cognitive processes, as extralist errors may be more likely to occur in these processes.

The posterior parameter estimates related to interference show no near transfer effects from binding training consistent with the findings reported by De Simoni & von Bastian (2018), as the interference parameter for the binding-trained group does not have a decrease. From the perspective of active manipulation abilities, the effect of training on interference appears to be limited to the specific task that was trained, so that memory updating training reduces mutual interference but not binding training. Thus the transfer may not occur between the active components of working memory, as they are specific to the tasks that use them and do not appear to affect the tasks that do not. However, some degree of near transfer effect may be present in the passive components shared by all working memory components, shown by a larger increase in the capacity parameter in the binding group than the control group, and the sharing of these components mean that training of one may indirectly improve another.

4 Discussion

In this paper, we developed a hierarchical Bayesian model for working memory updating based on mutual interference. The model adapted the activation framework of interference theory (Oberauer

& Kliegl, 2006) and the Weibull race framework (Logan, 1992), allowing it to incorporate the accuracies of responses and RTs into the mutual interference framework. The hierarchical Bayesian model yielded reasonable fits to memory updating data, thus we conclude that it is a feasible model for the memory updating task. Compared with previous models, the hierarchical Bayesian model can be used to investigate the probabilities of choosing targets, competitors and extralist items, making it easier to study extralist errors and other potential error mechanisms. The estimated model parameters characterize the cognitive properties of each individual, and identify potential cognitive deficits and task-specific strategies.

For working memory, the mutual interference model appears to be plausible for the memory updating construct, where it fits data well and can explain many higher-level behavioral patterns and group differences, such as the decrease of working memory performance for older adults. We saw that patterns of the model parameters can reflect potential working memory deficits in the older population in the first application. The data also show changes of means and variances of the RT distributions that may corresponds to changes in the processing speed related to the passive maintenance and active manipulation components of working memory. In the second application, model parameters show that improved performance in the passive components in one working memory task may be transferred to other working memory tasks that share the same components.

The model can potentially be applied to other types of memory updating tasks in future studies. Because some paradigms use types of stimuli other than digits, such as shapes, letters and locations, the model's interference structure and the probability transition matrix can be altered to suit each condition, and the number of RT racers can be adjusted accordingly.

It is also possible to adapt the model so that it can explain more complicated effects from the memory updating task. For example, when an error is made in arithmetic operations, the digits closer to the correct answer may be more likely to be selected than those far away from the answer. The arrangement of the boxes and the approximation of locations may also affect mutual interference and the outcomes of the memory updating task. Therefore, further investigation and additional modeling may be helpful for understanding related phenomena. The role played by pre-activation processing of responses may also be worth investigation, where its cause and the

underlying mechanism may influence working memory processing and performance.

Acknowledgement

We thank Professors Klaus Oberauer and Claudia von Bastian for generously sharing their data sets and providing information about their studies. This material is based upon work supported by the National Science Foundation under Grants No. SES-1424481 and No. SES-1921523. This material is based upon work performed while Van Zandt was serving at the National Science Foundation. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Supplemental materials

The supplemental materials can be found at https://github.com/Van-Zandt-Lab-at-OSU.

References

Anderson, J. R., Reder, L. M., & Lebiere, C. (1996). Working memory: Activation limitations on retrieval. *Cognitive Psychology*, *30*(3), 221–256.

Barrouillet, P., Bernardin, S., & Camos, V. (2004). Time constraints and resource sharing in adults' working memory spans. *Journal of Experimental Psychology: General*, *133*(1), 83.

Barrouillet, P., & Camos, V. (2001). Developmental increase in working memory span: Resource sharing or temporal decay? *Journal of Memory and Language*, *45*(1), 1–20.

Berger, J., Bayarri, M., & Pericchi, L. (2014). The effective sample size. *Econometric Reviews*, *33*(1-4), 197–217.

Bigorra, A., Garolera, M., Guijarro, S., & Hervás, A. (2016). Long-term far-transfer effects of working memory training in children with adhd: a randomized controlled trial. *European Child & Adolescent Psychiatry*, *25*(8), 853–867.

Borella, E., Carretti, B., Riboldi, F., & De Beni, R. (2010). Working memory training in older adults: evidence of transfer and maintenance effects. *Psychology and Aging*, *25*(4), 767.

Busemeyer, J. R., & Diederich, A. (2010). *Cognitive Modeling*. Sage.

Camos, V. (2017). Domain-specific versus domain-general maintenance in working memory: Reconciliation within the time-based resource sharing model. *Psychology of Learning and Motivation*, *67*, 135–171.

Camos, V., & Barrouillet, P. (2011). Developmental change in working memory strategies: From passive maintenance to active refreshing. *Developmental Psychology*, *47*(3), 898.

Caplan, D., DeDe, G., Waters, G., Michaud, J., & Tripodis, Y. (2011). Effects of age, speed of processing, and working memory on comprehension of sentences with relative clauses. *Psychology and Aging*, *26*(2), 439.

Colonius, H. (1995). The instance theory of automaticity: Why the weibull? *Psychological Review*, *102*(4), 744.

Covey, T. J., Shucard, J. L., & Shucard, D. W. (2019). Working memory training and perceptual discrimination training impact overlapping and distinct neurocognitive processes: Evidence from event-related potentials and transfer of training gains. *Cognition*, *182*, 50–72.

Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why? *Current Directions in Psychological Science*, *19*(1), 51–57.

Cragg, L., Keeble, S., Richardson, S., Roome, H. E., & Gilmore, C. (2017). Direct and indirect influences of executive functions on mathematics achievement. *Cognition*, *162*, 12–26.

Craigmile, P. F., Peruggia, M., & Van Zandt, T. (2012). A bayesian hierarchical model for response time data providing evidence for criteria changes over time. In *Current issues in the theory and application of latent variable models* (pp. 42–61). Taylor and Francis.

De Simoni, C., & von Bastian, C. C. (2018). Working memory updating and binding training: Bayesian evidence supporting the absence of transfer. *Journal of Experimental Psychology: General*, *147*(6), 829.

Dobbs, A. R., & Rule, B. G. (1989). Adult age differences in working memory. *Psychology and Aging*, *4*(4), 500.

Ecker, U. K., Oberauer, K., & Lewandowsky, S. (2014). Working memory updating involves item-specific removal. *Journal of Memory and Language*, *74*, 1–15.

Gelman, A., Rubin, D. B., et al. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

Hasher, L., & Zacks, R. T. (1988). Working memory, comprehension, and aging: A review and a new view. *The Psychology of Learning and Motivation*, *22*, 193–225.

Heitz, R. P. (2014). The speed-accuracy tradeoff: history, physiology, methodology, and behavior. *Frontiers in Neuroscience*, *8*, 150.

Hovik, K. T., Saunes, B.-K., Aarlien, A. K., & Egeland, J. (2013). Rct of working memory training in adhd: long-term near-transfer effects. *PLoS One*, *8*(12), e80561.

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: a latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General*, *133*(2), 189.

Kim, S., Potter, K., Craigmile, P. F., Peruggia, M., & Van Zandt, T. (2017). A bayesian race model for recognition memory. *Journal of the American Statistical Association*, *112*(517), 77–91.

Logan, G. D. (1992). Shapes of reaction-time distributions and shapes of learning curves: a test of the instance theory of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(5), 883.

Masse, N. Y., Yang, G. R., Song, H. F., Wang, X.-J., & Freedman, D. J. (2019). Circuit mechanisms for the maintenance and manipulation of information in working memory. *Nature Neuroscience*, *22*(7), 1159–1167.

McClelland, J. L. (1979). On the time relations of mental processes: an examination of systems of processes in cascade. *Psychological Review*, *86*(4), 287.

McGill, W. J., & Gibbon, J. (1965). The general-gamma distribution and reaction times. *Journal of Mathematical Psychology*, *2*(1), 1–18.

Minear, M., Brasher, F., Guerrero, C. B., Brasher, M., Moore, A., & Sukeena, J. (2016). A simultaneous examination of two forms of working memory training: Evidence for near transfer only. *Memory & Cognition*, *44*(7), 1014–1037.

Nairne, J. S. (1990). A feature model of immediate memory. *Memory & Cognition*, *18*(3), 251–269.

Oberauer, K. (2005). Control of the contents of working memory–a comparison of two paradigms and two age groups. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(4), 714.

Oberauer, K. (2006). Is the focus of attention in working memory expanded through practice? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(2), 197.

Oberauer, K., & Kliegl, R. (2001). Beyond resources: Formal models of complexity effects and age differences in working memory. *European Journal of Cognitive Psychology*, *13*(1-2), 187–215.

Oberauer, K., & Kliegl, R. (2006). A formal model of capacity limits in working memory. *Journal of Memory and Language*, *55*(4), 601–626.

Oberauer, K., & Lewandowsky, S. (2011). Modeling working memory: A computational implementation of the time-based resource-sharing theory. *Psychonomic Bulletin & Review*, *18*(1), 10–45.

Oberauer, K., & Lin, H.-Y. (2017). An interference model of visual working memory. *Psychological Review*, *124*(1), 21.

Oberauer, K., Süβ, H.-M., Wilhelm, O., & Wittmann, W. W. (2008). Which working memory functions predict intelligence? *Intelligence*, *36*(6), 641–652.

Oberauer, K., Süß, H.-M., Schulze, R., Wilhelm, O., & Wittmann, W. W. (2000). Working memory capacity—facets of a cognitive ability construct. *Personality and Individual Differences*, *29*(6), 1017–1045.

Reuter-Lorenz, P. A., & Sylvester, C.-Y. C. (2005). The cognitive neuroscience of working memory and aging. *Cognitive Neuroscience of Aging: Linking Cognitive and Cerebral Aging*, 186–217.

Rouder, J. N., Sun, D., Speckman, P. L., Lu, J., & Zhou, D. (2003). A hierarchical bayesian statistical framework for response time distributions. *Psychometrika*, *68*(4), 589–606.

Sala, G., & Gobet, F. (2017). Does far transfer exist? negative evidence from chess, music, and working memory training. *Current Directions in Psychological Science*, *26*(6), 515–520.

Salthouse, T. A., & Babcock, R. L. (1991). Decomposing adult age differences in working memory. *Developmental Psychology*, *27*(5), 763.

Salthouse, T. A., Babcock, R. L., & Shaw, R. J. (1991). Effects of adult age on structural and operational capacities in working memory. *Psychology and Aging*, *6*(1), 118.

Schmiedek, F., Oberauer, K., Wilhelm, O., Süß, H.-M., & Wittmann, W. W. (2007). Individual differences in components of reaction time distributions and their relations to working memory and intelligence. *Journal of Experimental Psychology: General*, *136*(3), 414.

Schwaighofer, M., Fischer, F., & Bühner, M. (2015). Does working memory training transfer? a meta-analysis including training conditions as moderators. *Educational Psychologist*, *50*(2), 138–166.

Schweickert, R., & Boruff, B. (1986). Short-term memory capacity: Magic number or magic spell? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*(3), 419.

Shipstead, Z., Redick, T. S., & Engle, R. W. (2010). Does working memory training generalize? *Psychologica Belgica*, *50*(3), 245–276.

Shipstead, Z., Redick, T. S., & Engle, R. W. (2012). Is working memory training effective? *Psychological Bulletin*, *138*(4), 628.

Soto, D., Hodsoll, J., Rotshtein, P., & Humphreys, G. W. (2008). Automatic guidance of attention from working memory. *Trends in Cognitive Sciences*, *12*(9), 342–348.

Stan Development Team. (2018). *Rstan: the r interface to stan. r package version 2.17. 3.*

Vecchi, T., & Cornoldi, C. (1999). Passive storage and active manipulation in visuo-spatial working memory: Further evidence from the study of age differences. *European Journal of Cognitive Psychology*, *11*(3), 391–406.

Vecchi, T., Richardson, J., & Cavallini, E. (2005). Passive storage versus active processing in working memory: Evidence from age-related variations in performance. *European Journal of Cognitive Psychology*, *17*(4), 521–539.

Veltman, D. J., Rombouts, S. A., & Dolan, R. J. (2003). Maintenance versus manipulation in verbal working memory revisited: an fmri study. *Neuroimage*, *18*(2), 247–256.

von Bastian, C. C., & Oberauer, K. (2013). Distinct transfer effects of training different facets of working memory capacity. *Journal of Memory and Language*, *69*(1), 36–58.

Waris, O., Soveri, A., & Laine, M. (2015). Transfer after working memory updating training. *PloS One*, *10*(9), e0138734.

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica*, *41*(1), 67–85.

Wingfield, A., Stine, E. A., Lahar, C. J., & Aberdeen, J. S. (1988). Does the capacity of working memory change with age? *Experimental Aging Research*, *14*(2), 103–107.

Appendix

We present the models to obtain parameter estimates $\kappa_{ic}^{(s)}$ and $\lambda_{ic}^{(s)}$ of the sub-cognitive/pre-activation RTs in this Appendix. Oberauer & Kliegl's 2001 data set in Section 3.1 includes a relatively large number of observations per participant, thus we considered each individual to have

different parameters. We used a hierarchical Bayesian model to estimate $\kappa_{ic}^{(s)}$ and $\lambda_{ic}^{(s)}$, where short RTs $t_{ic,j}^{(s)}$ followed the Weibull distribution

$$t_{ic,j}^{(s)} \sim f_w(t|\kappa_{ic}^{(s)}, \lambda_{ic}^{(s)}).$$

The short RTs $t_{ic,j}^{(s)}$ were selected as RTs less than 0.6 seconds because 0.6 was approximately the location of the valley between the peak of sub-cognitive/pre-activation process RTs and the peak of algorithmic cognitive process RTs. We selected the following priors and hyper priors:

$$\log(\kappa_{ic}^{(s)}) \sim N(\kappa_c^{(s)}, 1), \quad \log(\lambda_{ic}^{(s)}) \sim N(\lambda_c^{(s)}, 1),$$

$$\log(\kappa_c^{(s)}) \sim N(\kappa_0^{(s)}, 1), \quad \log(\lambda_c^{(s)}) \sim N(\lambda_0^{(s)}, 1),$$

$$\log(\kappa_0^{(s)}) \sim N(0, 1), \quad \log(\lambda_0^{(s)}) \sim N(0, 1).$$

We used Stan to obtain a chain consisting of 1000 warm-up samples and 4000 iterations. The posterior means were plugged into the formal model.

De Simoni & von Bastian's 2018 data set in Section 3.2 includes a relatively small number of observations per participant, thus it is difficult to obtain estimates of $\kappa_{ic}^{(s)}$ and $\lambda_{ic}^{(s)}$ for each participant. We used common parameters $\kappa^{(s)}$ and $\lambda^{(s)}$ for all participants, where the short RTs $t_{ic,j}^{(s)}$ follow the Weibull distribution

$$t_{ic,j}^{(s)} \sim f_w(t|\kappa^{(s)}, \lambda^{(s)}).$$

The short RTs $t_{ic,j}^{(s)}$ were selected as RTs less than 0.65 seconds. The priors are

$$\log(\kappa^{(s)}) \sim N(0, 1), \quad \log(\lambda^{(s)}) \sim N(0, 1).$$

We used Stan to obtain a chain consisting of 1000 warm-up samples and 4000 iterations, and used posterior means as estimates for the formal model.