Version of Record: https://www.sciencedirect.com/science/article/pii/S0022249622000487
Manuscript_8136a361812b358fe044ee0268411616

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, and a Wald diffusion process to account for reaction times.
- Application to two empirical studies. One shows the mechanisms underlying age differences in memory updating performance; the other reveals potential training and transfer effects from working memory training.

Abstract

We propose a hierarchical Bayesian model for working memory updating. This model accounts for both the accuracy of the responses and the 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 the model 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 Wald diffusion process to characterize RT, where the drift rate 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. The first data set, from Oberauer & Kliegl (2001), examined age differences in working memory. Results from our model reveal an increased level of mutual interference, less use of memory trace information, and potentially less pre-activation of memorized items in older adults compared to younger adults. The second data set, from De Simoni & von Bastian (2018), investigated transfer effects of working memory training. Results from our model reveal a potential transfer effect in the speed of information accumulation, where training in one working memory task may improve the information processing speed in another.

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 (Ecker et al., 2010, 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 verbal-numerical versions of the task, which test the ability to maintain and manipulate numbers and letters.

We demonstrate the memory updating paradigm with a numerical version. In this paradigm, the memory updating task 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

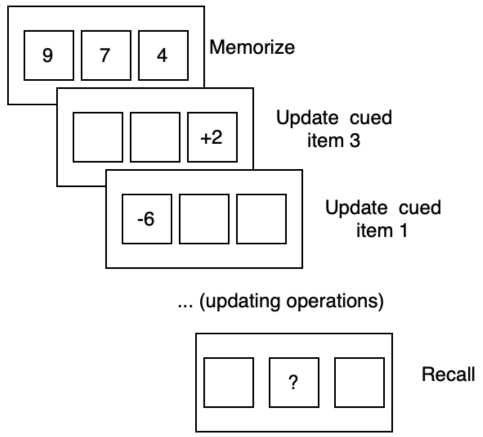

Figure 1: An example of a numerical memory updating task trial. The trial is composed of a memorizing period, an updating period containing multiple updating steps, and a final recall period where participants recall all the items in the order determined by the cue.

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. The performance of participants commonly decreases as the memory demand increases in this task. Other versions often have a similar task structure with a variety of stimuli and updating operations. Based on specific requirements, the memory updating paradigm can require intermediate responses after each updating step (e.g. De Simoni & von Bastian, 2018), or require no intermediate responses but impose time limits for each updating step (e.g. Oberauer & Kliegl, 2001).

In this paper, we use joint modeling to link responses and RTs from the memory updating tasks. We examined potential modeling approaches that allow such a link while providing theoretical explanations of the working memory process. A potential simple model that allows such a link is the speed-accuracy trade-off (SAT) model which can characterize the inverse relation between processing time and accuracy (Wickelgren, 1977; Heitz, 2014). This model was used in Oberauer & Kliegl (2006) for the memory updating task. However, due to its simplicity, the SAT function has only a limited ability to incorporate theory about cognitive mechanisms and corresponding working memory processes.

We base our model on established theories that explain how working memory performance declines with increasing memory demand. The most notable of these theories are resource theories hypothesizing limited working memory resources (e.g. Anderson et al., 1996; Cowan, 2010), time-based decay theories hypothesizing memory decay (e.g. Schweickert & Boruff, 1986; Barrouillet & Camos, 2001; Camos, 2017), and mutual interference theories hypothesizing interference between items in working memory (e.g. Nairne, 1990; Oberauer & Kliegl, 2006; Oberauer & Lin, 2017). Corresponding theory-based statistical models include the time-based resource-sharing model for decay theory (Barrouillet et al., 2004; Oberauer & Lewandowsky, 2011), and the activation-framed models for interference theory (Oberauer & Kliegl, 2001, 2006; Oberauer & Lin, 2017). In this paper, we base our model on the mutual interference theory and its related modeling, as there has been an increasing amount of evidence supporting the existence of interference from the literature (e.g. Oberauer et al., 2016; Farrell et al., 2016; Souza & Oberauer, 2015; Barrouillet et al., 2018).

In this paper, we propose a hierarchical Bayesian model for the memory updating task. Our model builds on a working memory interference framework from 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 the 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 jointly characterize the RTs at each memory updating step under the mutual interference framework. We use a Wald diffusion process to account for RTs, and associate the process that yields the RT 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 some interference mechanisms in the interference theory and the interference-based 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 model. 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 potential mechanisms underlying interference in working memory, with a focus on the mechanism of lack of distinctiveness in cue-based retrieval and the mechanism of feature overwriting. We then describe the mutual interference model proposed by Oberauer & Kliegl (2006), which is based on the theory of feature overwriting but not exclusive to this theory in its statistical form. We modify and extend the model to a hierarchical Bayesian framework which incorporates information from both responses and RTs. This model is able to quantify the level of interference and the speed of processing with model parameters, and the joint modeling framework of responses and RT allows interference parameters to be informed by RT information, and vice versa. The model's parameter recovery ability is evaluated with a simulation study in Appendix 1.

2.1 Potential mechanisms of interference

We focus on mechanisms explaining mutual interference caused by similarities between target items and competitors in working memory. We mainly describe the interference mechanisms from lack of distinctiveness in cue-based retrieval and feature overwriting, along with some evidence for each in the literature. However, these different mechanisms are not mutually exclusive and may jointly cause mutual interference in working memory.

The lack of distinctiveness in cue-based retrieval is a potential mechanism of interference most active during the retrieval period (Brown et al., 2007; Oberauer et al., 2012; Surprenant & Neath, 2013; Ecker et al., 2015). This theory assumes that memory traces are laid down for each item stored in working memory. During retrieval, the retrieval cues activate memory traces of the target items, thus retrieving the target item. However, when there is a lack of distinctiveness between different items stored in working memory, traces from different items may be associated with the same cues. Therefore, these cues may activate traces from both the target item and competitor items that lack distinctiveness with the target, resulting in interference between items and potential erroneous retrieval. This mechanism is consistent with various findings showing a link between item similarity and lowered retrieval accuracy (e.g. Oberauer et al., 2012; Ecker et al., 2015; Villata et al., 2018; Park et al., 2006).

The mechanism of feature overwriting assumes that mutual interference is caused by the shared features of items stored in working memory (Nairne, 1990; Oberauer & Kliegl, 2001, 2006; Oberauer, 2009; Cowan, 1988; Nairne, 2006). According to this theory, the representation of each item in working memory is composed of a number of features. When different items share the same features, each item would lose some of these shared features to the other items during encoding, resulting in feature overwriting and mutual interference. Thereby the mechanism of feature overwriting is potentially most active during the encoding period. This mechanism is consistent with some findings showing similarity-based interference effects in the encoding period (e.g. Oberauer, 2009; Hofmeister & Vasishth, 2014; Guitard et al., 2021).

Besides these mechanisms, there are a number of other mechanisms meant to explain interference in memory, including activation leveling (Villata et al., 2018; Smith et al., 2021) and superposition (Rumelhart et al., 1988; Oberauer et al., 2012, 2016). Because the nature of interference in working memory is still largely unclear (Li & Cowan, 2021), we do not intend to build a model to evaluate the plausibility of each potential mechanism, but instead quantify the level of interference affecting a person's final recall performance regardless of the mechanism.

2.2 Original model framework

Oberauer & Kliegl's 2006 model is based on the mechanism of feature overwriting. It 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 mean proportion C of features ($0 < C < 1$) and items in the pair compete for shared features. As a result of this competition, half of the features shared between two items are assumed to be allocated to each item[1]. Thus, if there is one interfering item present, the target is left with a mean 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 mean 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, this framework models the upper limit of the target's activation level with the formula[2]

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

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

[1] Although the actual proportions of shared features may differ for different pairs of items, the mean proportion C was used in this model. This was because the memory updating task featured a homogeneous set of stimuli (without grouping within stimuli), thus the mean proportion C was considered a reasonable approximation to the actual shared proportions (Oberauer & Kliegl, 2001).

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

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 θ and time limit T, the maximum activation level for the target is

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

and for a competitor is

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

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[3]. It is assumed that participants tend to 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 choosing the target 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

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

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. Because the exact mechanisms of interference are unclear, we do not assume interference to result solely from feature overwriting. We use the interference parameter to quantify the level of interference affecting the final recall regardless of interference mechanisms.

2.3 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 framework so the interference mechanism also explains RTs. We give the model a hierarchical Bayesian structure, so that it can fit data sets composed of groups of individuals representing the experimental groups to be compared. We first describe our model in the case when the paradigm requires intermediate responses during the updating period, then describe the alterations needed in the cases when intermediate responses are not required, or when the task includes pre/post-test conditions.

2.3.1 Response

To construct the hierarchical Bayesian model, we denote the group identifier as c and the participant identifier as i. We use the parameter C_{ic} and σ_{ic} to quantify mutual interference and noise, respectively. For Trial j with a memory demand of n_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 step are

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

When trials in the updating task have n plausible responses, for example, $n = 9$ for the numerical version shown in Figure 1, the corresponding probabilities of choosing the target, competitors and extralist items are

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

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

and

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

respectively. To obtain the probabilities of choosing each type of item at each step, we use a Markov chain structure with transition matrix

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

Assume that an item is encountered at Step x followed by the next encounter at Step $x + z$, this matrix shows the probability of choosing items of types 1, 2, and 3 at Step $x + z$ given the choice 1, 2, or 3 at Step x. 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 $x + z$, 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, as 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 $x+z$ is based on the wrong item from Step x, and we assume that this participant can still obtain the correct result by chance, as 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 probabilities of responding with each type of item are

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

where m_j is the number of encounters of the target item up to the current trial. Therefore, denoting the response to be $R_{ic,j}$ ($R_{ic,j} \in \{1, 2, 3\}$), the probability of making each response, denoted as $\mathrm{P}_{\mathrm{resp}}$, is

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

The probability $p^*_{k,ic,j}$ reflects the proportions of activated traces linked to items of type k, and correspondingly, the probabilities of responding with each type of item.

In this structure, we do not restrict the mutual interference to be solely the result of feature overwriting, but assume it to be the overall interference affecting the final performance that could be the result of different mechanisms. It could be understood as follows: during each encoding period, a number of memory traces are laid down for each item, and because of interference mechanisms such as feature overwriting in this period, each item loses some of its traces to the competitor items. During the retrieval period, each memory cue activates traces from both the target and competitor items. It is partly because some overlapping memory traces are attributed to competitors during encoding, but is also partly because each cue can activate both a number of traces from the target and a small number of traces from competitors due to mechanisms such as the lack of distinctiveness in cue-based retrieval. Based on the activation levels of all items, the participant determines one item as the target, then proceeds to retrieve, (possibly) update, and respond with that item. Overall, we quantify the final amount of interference with C_{ic} without distinguishing its source.

2.3.2 Response times

We model the RTs using a diffusion model based on the Wald distribution (Burbeck & Luce, 1982). We selected the Wald diffusion model because it is theoretically motivated, can fit RT data well, and can be easily applied to the framework inspired by mutual interference. The Wald diffusion model proposes that, in each trial, a participant samples information from the display and memory, then stores this information in a neural accumulator. To determine the end of each accumulation process, this participant sets a decision boundary determined by a certain amount of information: when the accumulated information reaches the decision boundary, the process is terminated and a response is made. The Wald diffusion model characterizes information accumulation as a Wiener diffusion process with drift and a single absorbing boundary. Therefore, the time of each process follows a Wald distribution.

To integrate the Wald diffusion process into the mutual interference framework, we consider the process as such: first, based on the cues and memory traces, a participant determines the item to be retrieved from working memory using the mechanism from Section 2.3.1. Then, the participant accumulates information about the final response from the retrieved item and (possibly) updating of that item. The speed of accumulation depends on the participant's speed of processing, the number of traces linked to the chosen item, and the difficulty of the (potential) updating process. When the participant accumulates enough information to reach the decision boundary a response is made.

To formulate the RT model, we denote the group identifier as c, the participant identifier as i, and the trial identifier as j. We characterize each individual's decision boundary with the parameter b_{ic} ($b_{ic} > 0$). Denote the information accumulation rate as $V_{k,ic,j}$ ($V_{k,ic,j} > 0$) for Trial j when the response is k, where $k = 1, 2, 3$ corresponds to targets, competitors, and extralist items respectively, we model $V_{ic,j}$ in the updating process as

$$V_{k,ic,j} = \exp(v_{u,ic} + \kappa_{ic} p^*_{k,ic,j}),$$

and we model $V_{k,ic,j}$ in the recall process as

$$V_{k,ic,j} = \exp(v_{r,ic} + \kappa_{ic} p^*_{k,ic,j}).$$

The parameter $v_{u,ic}$ is the speed of accumulation in the updating period with the subscript "u" standing for "updating", while the parameter $v_{r,ic}$ is the speed of accumulation in the recall period with the subscript "r" standing for "recall". As is shown in Equation (3) from Section 2.3.1, $p^*_{k,ic,j}$ is the probability of responding with items of Type k, and reflects the number of traces laid down for this type of item. Correspondingly, the type-to-RT parameter κ_{ic} characterizes the accumulation rate differences between different response item types. In the task, a participant is likely to lay down more memory traces for the target than non-targets, leading to a larger $p^*_{1,ic,j}$. As such, the accumulation rate shall be higher for targets than non-targets when $\kappa_{ic} > 0$. The rationale is that when more memory traces are laid for an item, the participant can collect information from that item at a faster rate due to the larger amount of information from traces. We justify the inclusion of this type-to-RT parameter κ_{ic} using a model comparison, discussed in Appendix 2.

With a drift rate v and a decision boundary b, the Wald distribution density is given by

$$f_w(t|v,b) = \frac{b}{\sqrt{2\pi t^3}} \exp\Big(-\frac{(vt-b)^2}{2t}\Big), \quad t > 0.$$

We include a non-decision time τ_{ic} for each participant that accounts for processes outside of the information accumulation process, including times needed for perception and motor execution. Denote the RT as $t_{ic,j}$, when the response $R_{ic,j} = k$, the RT has a distribution f_{rt} as

$$f_{\mathrm{rt}}(t_{ic,j}|\tau_{ic}, V_{k,ic,j}, b_{ic}, R_{ic,j} = k) = f_w(t_{ic,j} - \tau_{ic}|V_{k,ic,j}, b_{ic}).$$

2.3.3 Ancillary processes and pre-activation processes

In addition to the Wald diffusion processes, we consider the impacts from two different types of ancillary processes and a potential pre-activation process that may also affect the observed responses and RTs. Figure 2 shows the pooled RT histograms from one of De Simoni & von Bastian's 2018 data set. It shows three processes in addition to the Wald diffusion process that we identified through pre-analysis in empirical data sets.

The two ancillary processes are sub-cognitive processes and supra-cognitive processes characterized by very short and long RTs (Kim et al., 2017). The sub-cognitive processes correspond to very fast responses that may result from guessing. In Figure 2, sub-cognitive processes may lead to fast RTs from updating steps and some very fast RTs from recall steps.[4] The supra-cognitive processes correspond to very slow responses, which may result from distraction or mind-wandering. The RTs in the tails shown by Figure 2 may be partly due to supra-cognitive processes. Although it is common to discard responses and RTs associated with these processes, we keep these observations in our analyses and model them with mixture distributions, because the relative proportions of these processes may be meaningful (Province & Rouder, 2012).

Another process, shown in Figure 2 in the RTs of the recall period, is characterized by fast RTs peaked around 400 milliseconds but clearly distinct from the RTs of the main Wald diffusion processes. This process is also observed at the individual level, as is shown in Figure 3. We hypothesize that these fast RTs at the smaller mode may be a result of pre-activation. 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 sufficient working memory capacity might pre-activate more than one of 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

[4]We demonstrate with empirical data sets that these fast RTs are likely to be a result of sub-cognitive guessing in Sections 3.1.1 and 3.2.1.

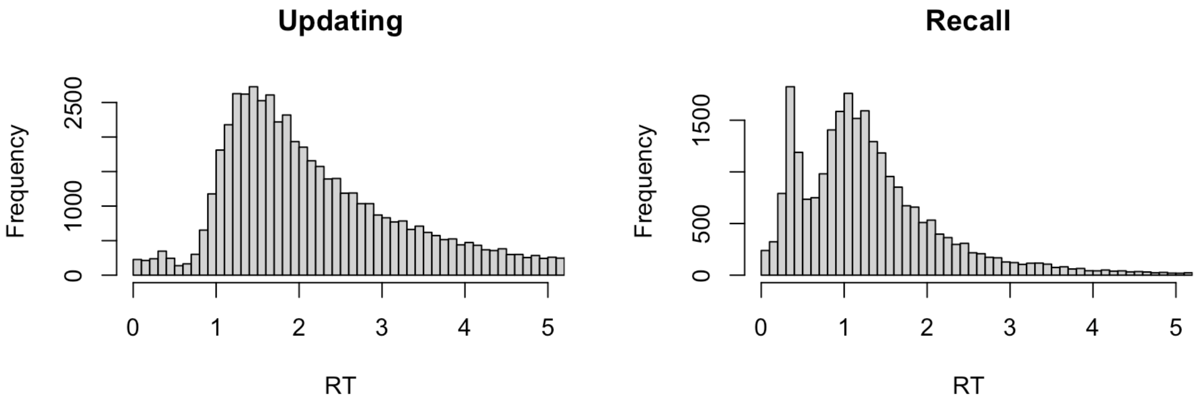

Figure 2: The histogram of the RTs of all 197 participants from De Simoni & von Bastian (2018), from the updating process (left) and the recall process (right). The bin width is taken as 0.1 seconds. The recall process clearly consists of both short sub-cognitive process RTs (close to 0) and supra-cognitive RTs in the tail. Responses from algorithmic Wald diffusion 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 fewer fast RTs around the smaller mode and slower RTs in the main peaks.

the algorithmic cognitive process operating on the stored items (see Figure 2). We consider pre-activation to be a more plausible mechanism for these fast RTs, because patterns from empirical data are consistent with pre-activation, and are inconsistent with alternative mechanisms such as guessing and the recency effect. These patterns are described in Sections 3.1.1 and 3.2.1. We also use a mixture component to model these fast pre-activation RTs. Because RTs from these pre-activation processes are difficult to distinguish from sub-cognitive processes, we integrate them into the same mixture component in the recall period.

To model each of these processes, we denote the group identifier as c, the participant identifier as i, and the trial identifier as j. In the updating period, the fast responses may result mostly from sub-cognitive processes. We denote the RT distribution as g_1 and account for it using the Log-normal distribution,

$$g_1(t_{ic,j}|\mu_{u,ic}) = f_{ln}\left(t\middle|\mu_{u,ic}, 1\right), \tag{4}$$

when $t_{ic,j}$ is the RT from fast ancillary processes, f_{ln} is the Log-normal density, and $\mu_{u,ic}$ is the mean parameter of the Log-normal distribution with the subscript "u" standing for "updating".

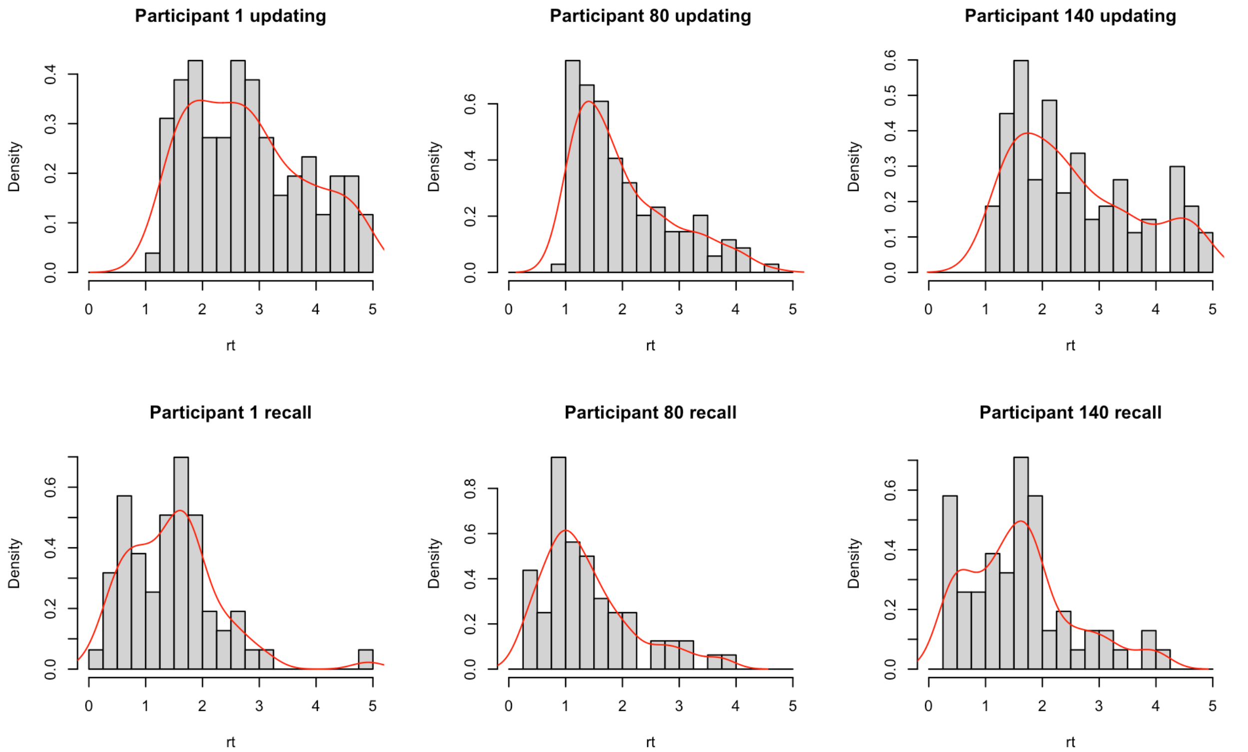

Figure 3: The histograms and densities of the RTs of Participants 1, 80, and 140 from De Simoni & von Bastian (2018), from the updating process (upper) and the recall process (lower).

We fix the standard deviation to 1 for these processes.

In the recall period, the fast responses may result from either sub-cognitive and pre-activation processes. We denote the RT distribution as g_2 and account for it using the Log-normal distribution,

$$g_2(t_{ic,j}|\mu_{r,ic}, \sigma_{\mu,r,ic}) = f_{ln}\left(t \middle| \mu_{r,ic}, \sigma^2_{\mu,r,ic}\right), \tag{5}$$

where the parameters $\mu_{r,ic}$ and $\sigma_{\mu,r,ic}$ are the mean and standard deviation respectively. The subscript "r" stands for "recall".

We model supra-cognitive RTs also with the Log-normal distribution, denoting the RT distribution as g_3,

$$g_3(t_{ic,j}|\mu_{s,ic}, \tau_{ic}) = f_{ln}\left(t - \tau_{ic} \middle| \mu_{s,ic}, 1\right),$$

when $t_{ic,j}$ is the RT from slow ancillary processes, $\mu_{s,ic}$ is the mean, and the standard deviation is fixed at 1. The subscript "s" stands for "supra-cognitive". To avoid identifiability problems in mixture estimation, we adopt informative priors for $\mu_{u,ic}$, $\mu_{r,ic}$, $\sigma_{\mu,r,ic}$, and $\mu_{s,ic}$ so that these distributions are appropriate for each process. Selections of prior distributions are explained in Section 2.3.4.

2.3.4 Priors and likelihood

In this section, we formulate the hierarchical Bayesian structure of the model. Because the participants' response accuracies can be higher than chance in the processes mentioned in Section 2.3.3, we use the parameter $q_{k,ic}$ ($k = 1, 2, 3$) to indicate the probabilities of responding with targets, competitors, and extralist items in these ancillary and pre-activation processes, so that the probabilities of response $R_{ic,j}$ are

$$\mathrm{P}_{\text{non-diff}}(R_{ic,j} = k) = q_{k,ic}, \quad k = 1, 2, 3. \tag{6}$$

Denote the mixture proportions for sub-cognitive/pre-activation, Wald diffusion, and supra-cognitive processes as ($\phi_{u,1,ic}$, $\phi_{u,2,ic}$, $\phi_{u,3,ic}$) respectively for updating, and ($\phi_{r,1,ic}$, $\phi_{r,2,ic}$, $\phi_{r,3,ic}$) respectively for recall. Then the RT $t_{ic,j}$ and response $R_{ic,j}$ have a joint distribution with density

$$\begin{aligned} f(t_{ic,j}, R_{ic,j} | \phi_{u,\cdot,ic}, \mu_{u,ic}, \tau_{ic}, V_{\cdot,ic,j}, b_{ic}, \mu_{s,ic}) =& \phi_{u,1,ic} g_1(t_{ic,j}|\mu_{u,ic}) \mathrm{P}_{\text{non-diff}}(R_{ic,j}) \\ &+ \phi_{u,2,ic} f_{\text{rt}}(t_{ic,j}|\tau_{ic}, V_{\cdot,ic,j}, b_{ic}, R_{ic,j}) \mathrm{P}_{\text{resp}}(R_{ic,j}) \\ &+ \phi_{u,3,ic} g_3(t_{ic,j}|\mu_{s,ic}, \tau_{ic}) \mathrm{P}_{\text{non-diff}}(R_{ic,j}) \end{aligned}$$

in the updating period, and

$$\begin{aligned} f(t_{ic,j}, R_{ic,j} | \phi_{r,\cdot,ic}, \mu_{r,ic}, \sigma_{\mu,r,ic}, \tau_{ic}, V_{\cdot,ic,j}, b_{ic}, \mu_{s,ic}) =& \phi_{r,1,ic} g_2(t_{ic,j}|\mu_{r,ic}, \sigma_{\mu,r,ic}) \mathrm{P}_{\text{non-diff}}(R_{ic,j}) \\ &+ \phi_{r,2,ic} f_{\text{rt}}(t_{ic,j}|\tau_{ic}, V_{\cdot,ic,j}, b_{ic}, R_{ic,j}) \mathrm{P}_{\text{resp}}(R_{ic,j}) \\ &+ \phi_{r,3,ic} g_3(t_{ic,j}|\mu_{s,ic}, \tau_{ic}) \mathrm{P}_{\text{non-diff}}(R_{ic,j}) \end{aligned}$$

in the recall period, where $\mathrm{P}_{\text{resp}}(R_{ic,j})$ and $\mathrm{P}_{\text{non-diff}}(R_{ic,j})$ are informed by Equations (3) and (6).

In the hierarchical Bayesian framework, we select the priors and hyper-priors as shown in

Priors and hyper-priors		
Interference	$\text{logit}(C_{ic}) \sim N(C_c, \delta_{C,c})$	$C_c \sim N(C_0, 0.2)$
	$C_0 \sim N(0, 0.2)$	$\log(\delta_{C,c}) \sim N(-1, 0.2)$
Noise	$\log(\sigma_{ic}) \sim N(\log(\sigma_c), \delta_{\sigma,c})$	$\log(\sigma_c) \sim N(\log(\sigma_0), 0.2)$
	$\log(\sigma_0) \sim N(0, 0.2)$	$\log(\delta_{\sigma,c}) \sim N(-1, 0.2)$
Accumulation speed	$v_{\cdot,ic} \sim N(v_{\cdot,c}, \delta_{v,\cdot,c})$	$v_{\cdot,c} \sim N(v_{\cdot,0}, 0.2)$
	$v_{\cdot,0} \sim N(0, 0.2)$	$\log(\delta_{v,\cdot,c}) \sim N(-1, 0.2)$
Decision boundary	$\log(b_{ic}) \sim N(b_c, \delta_{b,c})$	$b_c \sim N(b_0, 0.2)$
	$b_0 \sim N(0, 0.2)$	$\log(\delta_{b,c}) \sim N(-1, 0.2)$
Type-to-RT	$\kappa_{ic} \sim N(\kappa_c, \delta_{\kappa,c})$	$\kappa_c \sim N(\kappa_0, 0.2)$
	$\kappa_0 \sim N(0, 0.2)$	$\log(\delta_{\kappa,c}) \sim N(-1, 0.2)$
Mixture proportion	Non-informative	
Non-decision time	$\tau_{ic} \sim N(\tau_{ic}^{(0)}, 0.001)$	
Ancillary processes	$\mu_{u,ic} \sim N(-2, 0.05)$	$\mu_{r,ic} \sim N(-1, 0.05)$
	$\mu_{s,ic} \sim N(3, 0.05)$	$\text{logit}(\sigma_{\mu,r,ic}) \sim N(0, 1)$
Activation rate	$\log(\theta_{ic}) \sim N(\theta_c, \delta_{\theta,c})$	$\theta_c \sim N(\theta_0, 0.2)$
	$\theta_0 \sim N(0, 0.2)$	$\log(\delta_{\theta,c}) \sim N(-1, 0.2)$

Table 1: The model's priors and hyper-priors. If a parameter η has a dot in its subscript, that parameter has both updating and recall conditions, η_u and η_r.

Table 1. The mean of non-decision times $\tau_{ic}^{(0)}$ is fixed at an arbitrary value of 0.15 seconds. Denoting the entirety of the model parameters by η, the likelihood of the model is given by

$$\mathcal{L}(\eta|R, t) = \prod_{i,c,j} f(t_{ic,j}, R_{ic,j}|\phi_{\cdot,\cdot,ic}, \mu_{\cdot,ic}, \tau_{ic}, V_{\cdot,ic,j}, b_{ic}, \mu_{s,ic}, \mu_{r,ic}, \sigma_{\mu,r,ic}).$$

Figure 4 shows the model in graphical form.

2.3.5 Model fitting

In this section, we summarize some alterations needed for the model when the task does not require intermediate responses, or when the data include results from multiple conditions such as a pre-test and post-test. We performed simulation studies to test the model's parameter recovery ability, which is presented in Appendix 1.

We first consider the case where the task only requires recall responses and limits the times spent on updating steps, for example, as in the aforementioned paradigm from Oberauer & Kliegl

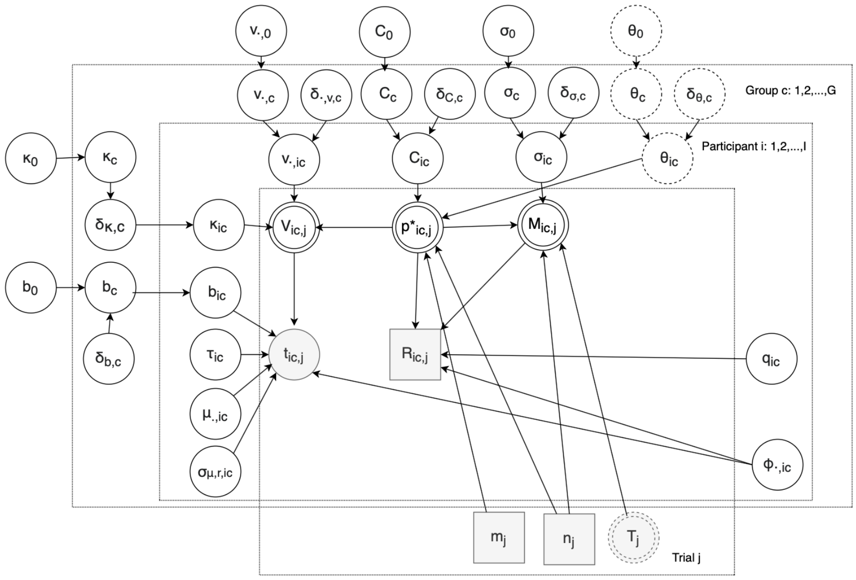

Figure 4: The diagram of the hierarchical Bayesian model. The rectangular boxes contain integers, the round boxes contain real values, and the double-edged boxes contain computed values. The observed variables have gray backgrounds. Parameters to be estimated are unshaded. The arrows indicate dependence. 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.

(2001). Denoting the updating time limit for Trial j as T_j, we use the formulas

$$\begin{aligned} a_{1,ic,j} &= (1 - C_{ic}/2)^{n_j-1}(1 - \exp(-\theta_{ic}T_j)), \quad \text{and} \\ a_{2,ic,j} &= (C_{ic}/2)(1 - C_{ic}/2)^{n_j-2}(1 - \exp(-\theta_{ic}T_j)) \end{aligned} \tag{7}$$

instead of those from Equation (1). Parameter θ_{ic} is the activation rate parameter with priors and hyper-priors shown in Table 1. We compute the probabilities of responding with each item as

$$(p^*_{1,ic,j}, p^*_{2,ic,j}, p^*_{3,ic,j}) = (p_{1,ic,j}, p_{2,ic,j}, p_{3,ic,j})M^{m_j}_{ic,j}M^*_{ic,j}, \tag{8}$$

where m_j is the number of updates performed on each item on Trial j. Because the updating time limit for Trial j, denoted as T_j, is variable in these designs, 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 into 4 segments according to the 3 quartiles (25%, 50%, and 75%), and used the mean in each of the bins as the T_j values in Equation (7).

In the case when the data include pre-test and post-test conditions, we permitted interference parameters C, noise parameters σ, RT parameters v, κ, b, and the mixture proportions ϕ to differ for pre-test and post-test conditions as in De Simoni & von Bastian (2018).

3 Empirical data analysis and results

In this section, we report the results of fitting the hierarchical Bayesian model to data from two studies, 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 fit our model to these data and show that the model accounts for these data by examining the posterior predictive distributions and using out-of-sample validation. We present the modeling results and discuss the theoretical implications of these results, with a focus on the potential cognitive mechanisms underlying group and individual differences.

3.1 Application 1: Age differences in working memory

The influence of age on 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.

Oberauer & Kliegl (2001) studied the effect of aging on working memory using the memory updating task. They mainly focused on the modeling of response accuracy. Their results revealed that older adults have a higher level of mutual interference compared with younger adults, but there were no significant differences between younger and older adults in noise and activation rates in the updating processes.

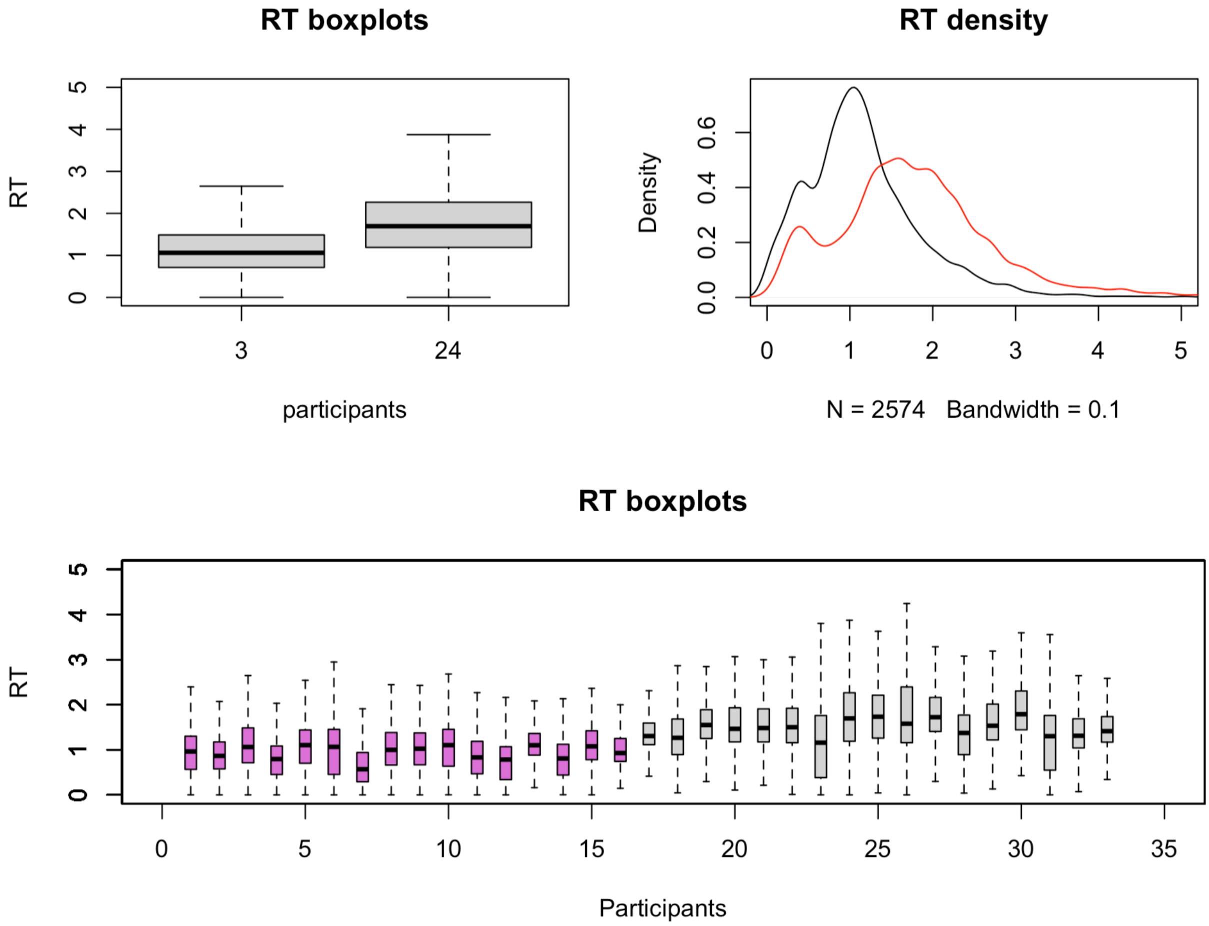

Figure 5: Contrast between RTs from the younger and older groups. The upper figures display the distributions of Participant 3 (Younger) and Participant 24 (Older). The black and red density lines correspond to Participants 3 and 24 respectively. Despite having similar response accuracy (0.656 for Participant 3 and 0.657 for Participant 24), the RTs are overall much slower for Participant 24. The lower figure shows the distributions for all participants (black 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).

However, in addition to response accuracy, RTs may also serve as an important, meaningful source of information reflecting the capabilities of working memory. Figure 5 shows that the older group displays overall longer RTs compared to those of the younger group. Even when an older adult (like Participant 24) has a similar response accuracy as a younger adult (like Participant 3, shown in Figure 5), 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.

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. We first describe the data set and test some assumptions made about the RT distributions and ancillary processes in this data set. We apply the model to this data set and evaluate its goodness of fit, and present modeling results and discuss their implications.

3.1.1 Data

Oberauer & Kliegl (2001) tested participants on the numerical version of the memory updating task, where participants memorize digits from "1" to "9" and update them by addition and subtraction as shown in Figure 1. This data set 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 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 from Oberauer & Kliegl (2001). Table 2 shows the summary statistics of the response accuracies and RTs. The older adults had overall lower response accuracies, longer RTs, and larger RT variances compared with the younger adults, indicating that the older adults recall the items with longer times and lower accuracies. Table 2 also shows the mean RTs from responses to targets, competitors, and non-competitors from both groups. Overall, responses to targets take a shorter time than responses to non-target items. This result is consistent with a mechanism in which more memory traces (linked to targets) can improve the speed of information accumulation in the Wald diffusion process.

We then evaluate whether the fast RTs around the smaller mode may be a result of chance performance, the recency effect, or pre-activation in this data set. Table 2 shows the response accuracies in the RT ranges of 0-0.2 seconds, 0.2-0.6 seconds, and over 0.6 seconds. These statistics show that the response accuracies in the 0.2-0.6 seconds are well above chance, and are also higher than the accuracies in the 0-0.2 seconds range and are closer to the accuracies in the over 0.6

Table A		Younger	Older		Younger	Older
	Accuracy	0.68 (0.04)	0.61 (0.05)	RT	1.01 (0.66)	1.62 (1.41)
Table B	Type	Target	Competitor	Non-competitor		
	RT (Younger)	0.96	1.06	1.12		
	RT (Older)	1.54	1.69	1.75		
	RT (Younger, >0.6s)	1.18	1.45	1.43		
	RT (Older, >0.6s)	1.64	1.90	1.95		
Table C	RT	0-0.2s	0.2-0.6s	>0.6s		
	Younger	0.58	0.68	0.69		
	Older	0.38	0.53	0.62		
Table D			Sequential place in recall			
	Memory demand	1	2	3	4	
	1 (Younger)	19.4%				
	1 (Older)	9.4%				
	2 (Younger)	8.0%	42.4%			
	2 (Older)	2.6%	31.3%			
	3 (Younger)	4.2%	18.3%	39.0%		
	3 (Older)	1.7%	11.2%	26.2%		
	4 (Younger)	3.6%	11.0%	16.3%	33.9%	
	4 (Older)	1.6%	8.3%	11.5%	23.3%	

Table 2: Statistics of the data set from Oberauer & Kliegl (2001). "Table A" displays the group means and standard deviations (in brackets) of response accuracies and RTs from the groups of younger and older adults. "Table B" displays the mean RTs for the target, competitor, and non-competitor responses. "Table C" displays the response accuracies in the RT ranges of 0-0.2 seconds, 0.2-0.6 seconds, and over 0.6 seconds for each group respectively. "Table D" displays the proportions of fast RTs around the smaller mode in the 0.2-0.6 seconds range from each item to be recalled. The columns "Sequential place in recall" shows the sequential order of the item to be recalled.

seconds range. We therefore conclude that although there is fast guessing at chance performance, it is not likely the main reason for the fast RTs around the smaller mode. Table 2 also shows the proportions of fast RTs around the smaller mode (0.2-0.6 seconds) from the first to last recall items in each trial. For both the younger and older adults, the proportions of fast RTs around the smaller mode are the smallest for the first item to be recalled then gradually increase to the largest for the last item to be recalled. This pattern does not show evidence for or against a recency effect, but it is consistent with the assumption that fast RTs around the smaller mode may be a result of pre-activation: the later items to be recalled, which have a larger chance of being pre-activated, correspond to a higher proportion of fast RTs compared with earlier items to be recalled.

3.1.2 Model fit

We fit the hierarchical Bayesian model to the data set using Stan (Stan Development Team, 2018) with the adjustments for no-intermediate response paradigm described in Section 2.3.5. We fixed the mean of non-decision time priors $\tau_{ic}^{(0)}$ to 0.15 for all participants. We obtained 3 chains, each containing 500 warm-up samples and 2000 iterations. 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 parameters) suggested reliable posterior estimates and satisfactory convergence.

To determine goodness of fit, we examined the posterior predictive distributions, and performed an out-of-sample validation analysis to evaluate how well the model can generalize to new data. Because we binned the updating limiting times during model fitting (see Section 2.3.5), for a more generalized evaluation we generated posterior predictive distributions with unbinned limiting times. Figure 6 shows the posterior predictive summaries contrasted against the observed accuracies and RTs. Apart from small divergences for a few participants, the observed accuracies are overall consistent with the estimated posterior predictive accuracies despite the binning in modeling. The posterior predictive RTs are quite consistent with the empirical RTs.

To evaluate whether the model can correctly recover the association between the types of responses and RT, for each type of response we show the posterior predictive mean RTs in contrast to the empirical mean RTs in Figure 7 from selected participants. We computed means from RTs larger than 0.6 seconds to minimize the confounding from fast ancillary processes. Shown in Figure 7, most participants have higher mean RTs when responding with non-targets than targets, which is consistent with the mechanism that more memory traces (in targets) lead to an increase in the speed of evidence accumulation. The posterior predictive RTs are able to recover these RT differences, and are overall consistent with empirical data, indicating a good fit of the model.

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 and RTs from the posterior predictive distributions for other data. Figure 8 shows the simulated accuracies and RTs compared to the true accuracies and RTs of the other 50% of data. Despite a larger divergence than those shown in Figure 6, most of the simulated accuracies are close to the true accuracy values, and the simulated RTs generally show patterns consistent with the true RTs. These results indicate that the estimation results from a subset can be generalized to the remainder of the data set. Thus, we consider the model to have a satisfactory fit to the data and a reasonable ability to generalize.

3.1.3 Results and implications

To analyze the mechanism and implications of the age data, we present the group and individual estimated posterior results. Figure 9 displays the estimated posterior distributions of the group-level parameters, including the interference parameter C_c, the noise parameters σ_c, the RT accumulation rate parameter v_c, the type-on-RT parameter κ_c, the decision boundary parameter b_c, and the activation rate parameter θ_c. From posterior samples, the estimated posterior probability that the younger group has a lower interference C_c is 0.96; the estimated posterior probability that the younger group has a lower noise σ_c is 0.98; the estimated posterior probability that the younger group has a higher RT accumulation rate v_c is 0.63; the estimated posterior probability that the younger group has a higher type-on-RT parameter κ_c is 0.81; the estimated posterior probability that the younger group has a lower decision boundary parameter b_c is 1.00; and the estimated posterior probability that the younger group has a lower activation level θ_c is 0.91. Based on these results, all parameters except the accumulation rate v_{ic} have clear group differences.

Figure 10 shows the estimated posteriors for the individual-level parameters. The interference parameters C_{ic} and boundary parameters b_{ic} display a clear difference in the majority of younger and older participants, further supporting the existence of their group-level differences. The noise parameters σ_{ic}, the type-to-RT parameters κ_{ic},, and the activation rate parameters θ_{ic}, however, are similar for most participants. The accumulation rate parameters v_{ic} appear to be more divergent in the younger group than in the older group, but like the group-level results, these parameters do not display consistent group differences. The younger participants have overall higher proportions of sub-cognitive and pre-activation processes, shown by higher $\phi_{1,ic}$s, and have overall lower proportions of algorithmic Wald diffusion processes and supra-cognitive processes, shown by lower

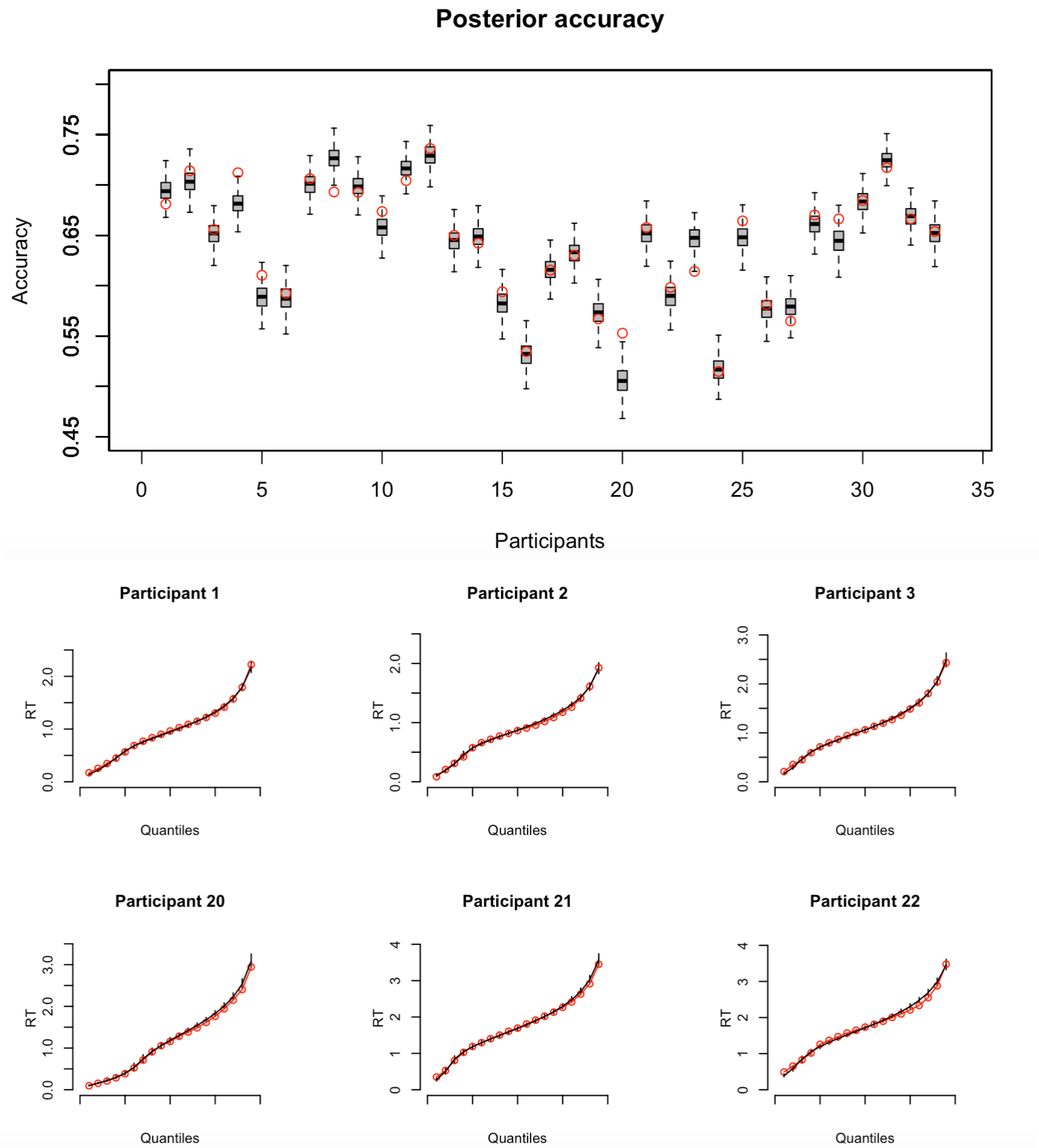

Figure 6: Posterior predictive results of response accuracy and RT. The upper figure shows the posterior accuracy for each participant in the data set from Oberauer & Kliegl (2001). The red points are the observed accuracies from the participants, and the corresponding box-plots are the posterior predictive accuracies. The lower figure shows the posterior predictive RT distributions contrasted with the empirical RT distributions for Participants 1-3 (younger) and 20-22 (older). Results of the other participants can be found in the supplemental materials. The red points are the 5% to 95% RT quantiles incremented by 5% from empirical data, connected by red lines; and the black bars show the 5% to 95% RT quantiles incremented by 5% from posterior predictive samples, connected by black lines. The x-axis corresponds to the quantiles, and the y-axis corresponds to RT values at each quantile.

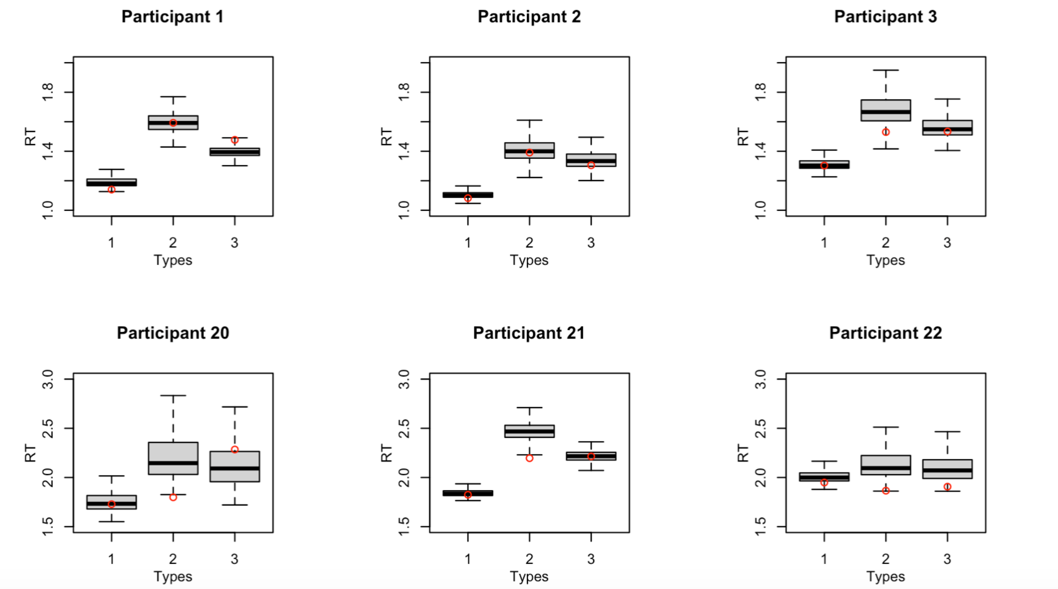

Figure 7: Posterior predictive mean RTs (box-plots) and empirical mean RTs (red points) for Participants 1-3 (younger) and 20-22 (older). The x-axes show the types of response, where 1, 2, and 3 correspond to targets, competitors, and non-competitors, respectively.

$\phi_{2,ic}$s and $\phi_{3,ic}$s.

The results of the interference parameter 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 lowered ability to resist mutual interference between items. Oberauer & Kliegl found no evidence of differences in noise σ or activation rate θ. Results from this model indicate that the older group might have an overall higher level of noise than the younger group because of some individuals: some younger participants may have less cognitive noise than most of the others, and some older participants may have higher noise than most of the others. Similarly, these results suggest that some older participants may have a higher activation rate than most of other participants, which may have resulted in an overall higher group-level activation rate in the older group. It is also noteworthy that the posterior means of parameters σ and θ are positively correlated in both groups, with $r(14) = 0.82, p < 0.001$ in the younger group of 16 participants, and with $r(15) = 0.87, p < 0.001$ in the older group of 17 participants. This may indicate that a higher activation level is related to a higher level of noise in mental representation, so that fast activation may not always indicate better performance.

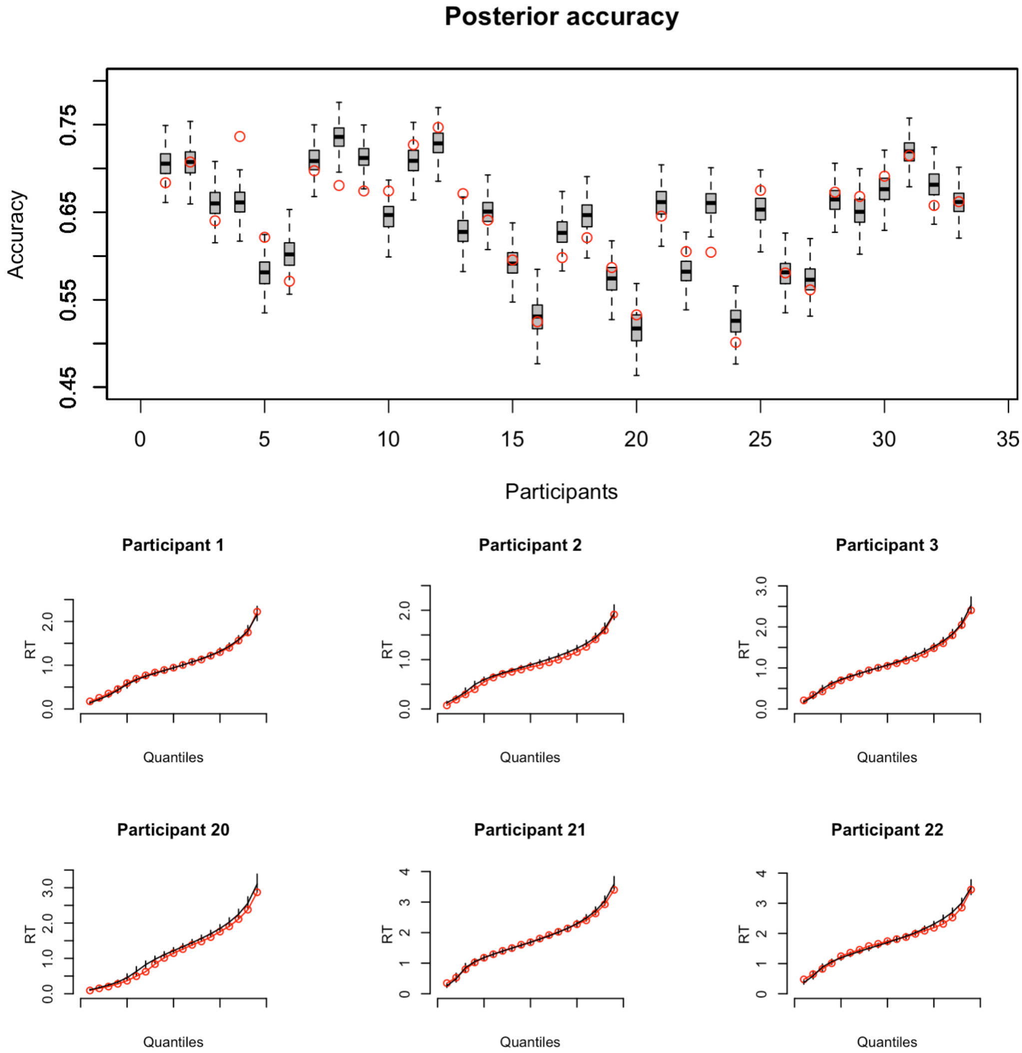

Figure 8: Posterior predictive results of response accuracy and RT based on the generated out-of-sample validation parameters (black) from test data, compared to the observed accuracies and RTs from the other data (red). These figures are plotted in the same way as those shown in Figure 6.

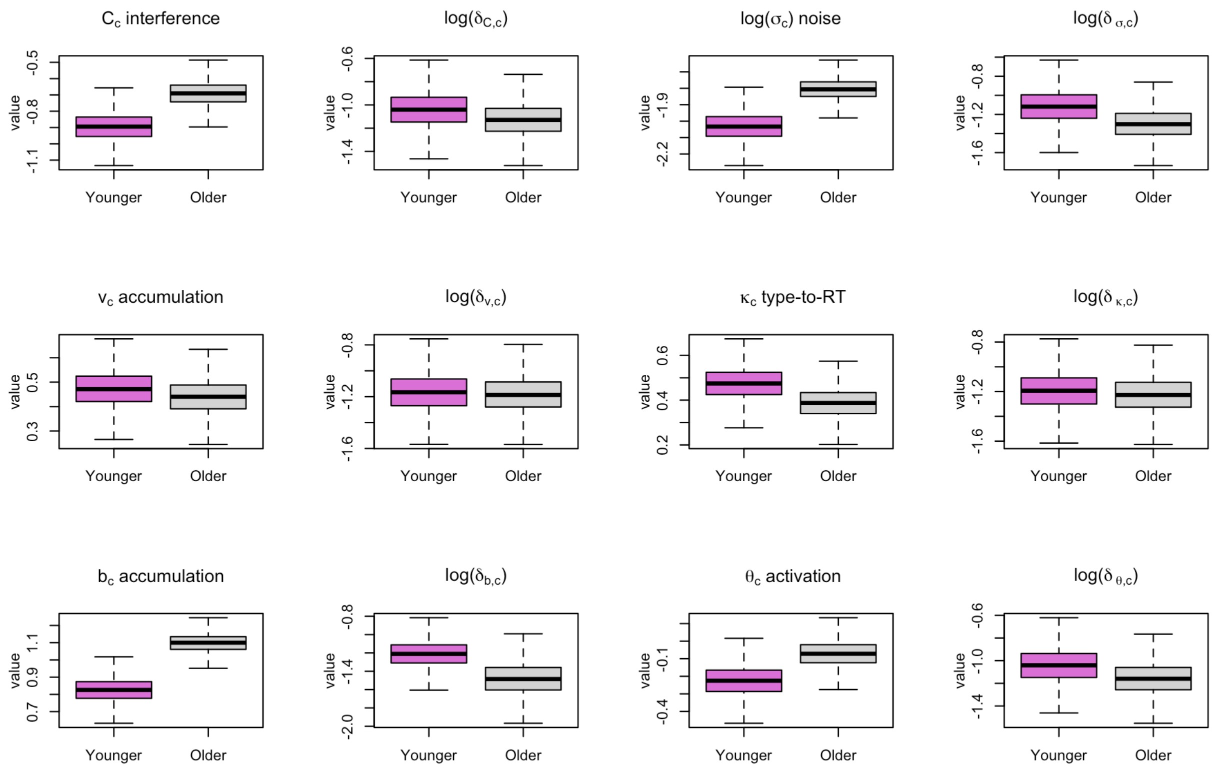

Figure 9: The box-plots for the estimated posteriors of the group-level parameters. The black and gray box-plots correspond to the younger and older groups, respectively. These parameters can have negative values because of the transformations shown in Table 1. The priors for each parameter are shown in Table 1.

In the RT results, the younger participants have relatively lower decision boundary parameters b_{ic} than the older participants. This may indicate that younger participants need less information to make a response. Some younger participants also have higher type-to-RT parameters κ_{ic} than other participants, which may indicate that they have a better ability to use memory trace information to guide their accumulation processes

The mixture proportions of the ancillary and pre-activation processes show some results worthy of notice. The mixture components for sub-cognitive/pre-activation processes are higher for younger than older participants, even though the younger participants have generally higher response accuracies. Despite being indistinguishable from sub-cognitive processes, the pre-activation processes may have contributed the most to these proportions judging from response accuracies well above chance (see Table 2). This may imply that the younger participants are more capable

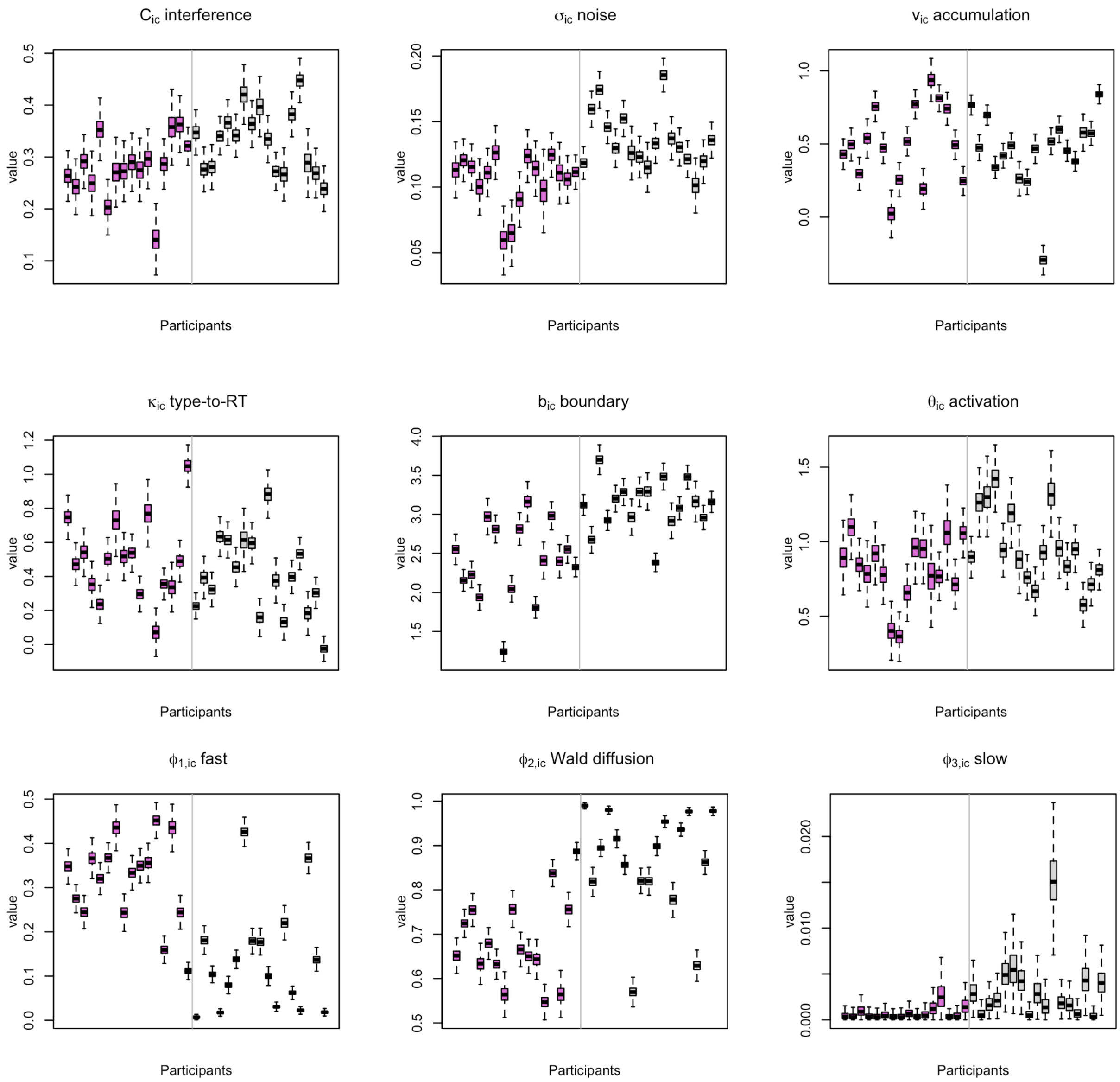

Figure 10: Box-plots for estimated posteriors of the individual-level parameters. Younger participants are shown in black (left of the vertical gray line) and older participants are shown in gray (right of the vertical line).

of using cognitive resources for pre-activation processing than the older participants, where they pre-activate items, store them in a "batch", and read items out upon seeing cues. In contrast, the older participants may rely more on algorithmic processing, where they retrieve only one item after seeing a cue. This difference may be a result of decline in working memory abilities in some older adults.

3.2 Application 2: Transfer effects of working memory training

De Simoni & von Bastian (2018) performed 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). It is assumed 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). 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. In this application, we focus on the near transfer effects (Shipstead et al., 2010), where 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 De Simoni & von Bastian (2018)'s design to search for near transfer effects, 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 were pre-tested in all three types of tasks before training, then received training of their specific allocated task across five weeks, followed by post-tests in all three types of tasks. De Simoni & von Bastian (2018) used measurement statistics and a latent-variable confirmatory factor analysis to investigate whether these data embed transfer effects. In their analysis, despite improved performances in the trained tasks, little evidence was found to support the existence of near transfer effects. 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. We first describe the data sets from De Simoni & von Bastian (2018) and test some assumptions made to the RT distributions and ancillary processes in these data sets. We then apply the model to these data, and evaluate its goodness of fit to these data sets. We then present the model fitting results and summarize their implications.

3.2.1 Data

The De Simoni & von Bastian (2018) study investigated multiple versions of the memory updating tasks using different stimuli. We focus our analyses on the numerical and verbal versions, as they are both linked to the verbal-numerical aspect of working memory, and may correspond well to the assumptions of our model. In this section, we describe the numerical and verbal updating tasks used in the De Simoni & von Bastian (2018) study, summarize some characteristics of the data sets from each task, and show that the characteristics of these data sets are consistent with RT assumptions made in Sections 2.3.3 and 2.3.4.

The data sets 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 excluding one with missing information in the verbal version. The memory updating training group had 59 participants, the binding training group had 66 participants (65 in the verbal version), 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 in the numerical version, or 2 or 4 recall steps in the verbal version. The different numbers of recall steps correspond to different memory demands in each trial. As such, each participant has 416 and 394 observations overall in the numerical and verbal versions, respectively.

In the numerical version of the task, the participants were required to memorize digits from "1" to "9", and update them by addition and subtraction as is shown by Figure 1. In the verbal version, the participants were required to memorize alphabetic letters “A” to “H”, and update them by shifting the letters forward and backward the alphabet according to the cues provided. The updating cues in the verbal version are similar to those in Figure 1, and indicate the direction and amounts to move in the alphabet, for example, if the memorized item is “A” and the cue is “+3”, the correct response would be the letter that is 3 places after “A”, which is “D”.

We show some summary statistics of the response accuracies and RTs in Table 3. All groups show an increase in the response accuracy and a decrease in RT after training, where the updating-trained group shows the largest improvement in response accuracy and the largest decrease in RT in both versions. The binding-trained group overall shows a slightly larger increase in response accuracy and a slightly larger decrease in RT compared with the control group. However, because these differences are relatively small and could be due to participant variability, it is impossible to determine whether they are the result of training by simply inspecting the summary statistics.

To evaluate whether more memory traces may lead to a faster speed of responding, we show the mean RTs from responses of each type (targets, competitors and non-competitors) in Table 3. We computed mean RTs from RTs larger than 0.6 seconds to minimize the confounds of fast ancillary processes. It is shown that non-targets require overall longer mean RTs than targets, and this trend is persistent in both tasks, all groups, and both the updating and recall periods of each task. This is consistent with the mechanism that more memory traces, linked to targets, can improve the speed of information accumulation in the Wald diffusion process.

We then evaluate whether the characteristics of these data sets are consistent with some RT assumptions made in Sections 2.3.3 and 2.3.4, especially whether the RT bi-modality is likely a result of pre-activation. To evaluate whether the fast RTs at the smaller mode are a result of fast guessing, we examine the response accuracies of trials with RTs in the range of 0-0.2 seconds, corresponding to potential sub-cognitive processes; and in the range of 0.2-0.6 seconds, corresponding to the first peak (Figure 2), and larger that 0.6 seconds, corresponding to the main process. Table 3 shows the response accuracies in each RT range. In the updating processes, the response accuracies in trials with RTs under 0.6 seconds are around chance performance. In the recall processes, while response accuracies with RTs under 0.2 seconds are lower than the overall accuracies, the response accuracies in the 0.2-0.6 seconds RT range are close to the overall accuracy. Therefore, the fast RTs at the smaller mode in the recall process are unlikely a result of fast guessing that should result in chance performance.

To evaluate whether the fast RTs around the smaller mode are a result of recency effect, we evaluate the proportions of RTs in the range of 0.2-0.6 seconds in the first recall response in each trial. When the first recall item is in the same location as the last updated item, there are 6.8% and 17.1% of RTs in the range of 0.2-0.6 seconds from the numerical and verbal tasks, respectively. When the first recall item is in a different location from the last updated item, there are 1.9% and 3.9% of RTs in the range of 0.2-0.6 seconds from the numerical and verbal tasks, respectively. This indicates that a recency effect may be present. However, in all recall responses, there are 14.9% and 20.1% of RTs in the 0.2-0.6 seconds range from the numerical and verbal tasks respectively, which are larger than the 6.8% and 17.1% in the first recall response when the first recall item is in the same location as the last updated item. These statistics indicate that the recency effect is unlikely to be the main cause of fast RTs around the smaller mode, as the proportions of overall fast RTs are larger than what a recency effect may induce.

We then examine the proportions of fast RTs around the smaller mode from responses to the first to last recall items in each trial. Table 3 displays the proportions in the range of 0.2-0.6 seconds from both versions. These statistics show that the proportions of fast RTs around the smaller mode are smallest from the first item to be recalled, then gradually increase to the largest from the last item to be recalled. It is in support of the assumption that fast RTs around the smaller mode may be a result of pre-activation, where the later items to be recalled are more likely to be pre-activated when recalling the previous items, and it may result in the larger proportions of faster RTs from the later items to be recalled.

To summarize, the data sets show clear effects of updating training on the performance in memory updating tasks, but it is unclear whether binding training can have transfer effects. The RTs from these data sets show patterns consistent with the assumption of pre-activation made in the model. We then fit the hierarchical Bayesian model to the data set to investigate whether binding training induces transfer effects that are reflected in model parameters, or if the differences between the control group and the binding-trained group is purely a result of participant variability.

3.2.2 Model fit

We fit the hierarchical Bayesian model using Stan to numerical and verbal data using methods described in Section 2.3. To accommodate the small sample size of the verbal data set, especially in the recall period, we adjusted the standard deviation of the hyper-priors of $v_{r,c}$ and $v_{r,0}$ from 0.2 to 0.1. We fixed the mean of non-decision time priors $\tau_{ic}^{(0)}$ to 0.1 for the updating-trained group and to 0.15 for other groups. For each data set, we generate 3 chains, each consisting of 500 warm-up samples and 2000 iterations. The Gelman-Rubin statistic $\hat{R}$ ($\hat{R} < 1.01$ for all parameters) and the effective sample size were both reasonable, indicating satisfactory convergence and reliable posteriors.

To evaluate the model's ability to fit the data, we examined the estimated posterior predictive distributions and performed out-of-sample validations for numerical and verbal data. We present the posterior predictive accuracies and RTs in contrast with empirical accuracies and RTs in Figure 11 and Figure 12, respectively. Figure 11 shows that the posterior predictive accuracies are overall consistent with empirical accuracy regardless of the value of accuracy. Figure 12 shows that the posterior predictive RTs have consistent fits with empirical RTs. For some participants (such as Participant 140), the posterior predictive RT distributions have small divergences from the empirical distributions, but this is reasonable considering the small sample sizes in this data set.

To evaluate whether the model can correctly recover the association between the types of responses and RT, for each type of response, we show the posterior predictive RT distributions in contrast with the empirical RT distributions in Figure 13. We computed quantiles from RTs larger than 0.6 seconds and less than 8 seconds to minimize the confounding from ancillary processes. Shown in Figure 13, the model is able to reflect the differences in RTs from different types of response. For competitors and non-competitors, the RTs are sometimes overestimated, but the differences are relatively small.

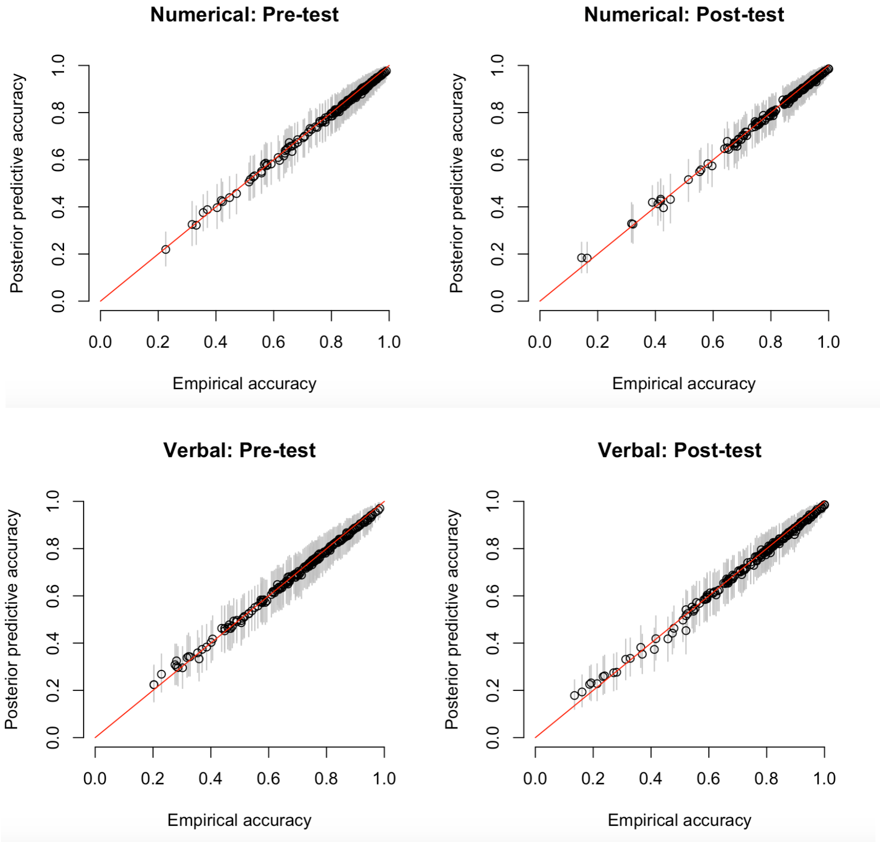

Figure 11: Posterior predictive accuracies for each participant from De Simoni & von Bastian (2018). The upper plots show results from the pre-test and post-test of the numerical version, and the lower plots show results from the verbal version. The red lines are identity lines. The x-axis reflects each participant's empirical response accuracy. The y-axis reflects the posterior predictive accuracy of each participant. The black points are the mean of posterior predictive accuracies from all samples, and the gray whisker plots are the 95% equal-tailed credible intervals from all samples.

To perform out-of-sample validation, we drew 50% of the data set at random. We applied the model to the subset, simulated accuracies and RTs from the posterior predictive distributions and plotted the mean predictive accuracies against the observed accuracies for the rest of the data, shown in Figure 14 and Figure 15, respectively. Figure 14 shows some divergences but overall linear patterns following the identity line. Figure 15 shows that there are larger divergences between the simulated RT distributions and the empirical ones, but the patterns are overall consistent considering the small sample size. This indicates that our model has a reasonable ability to generalize with these data sets.

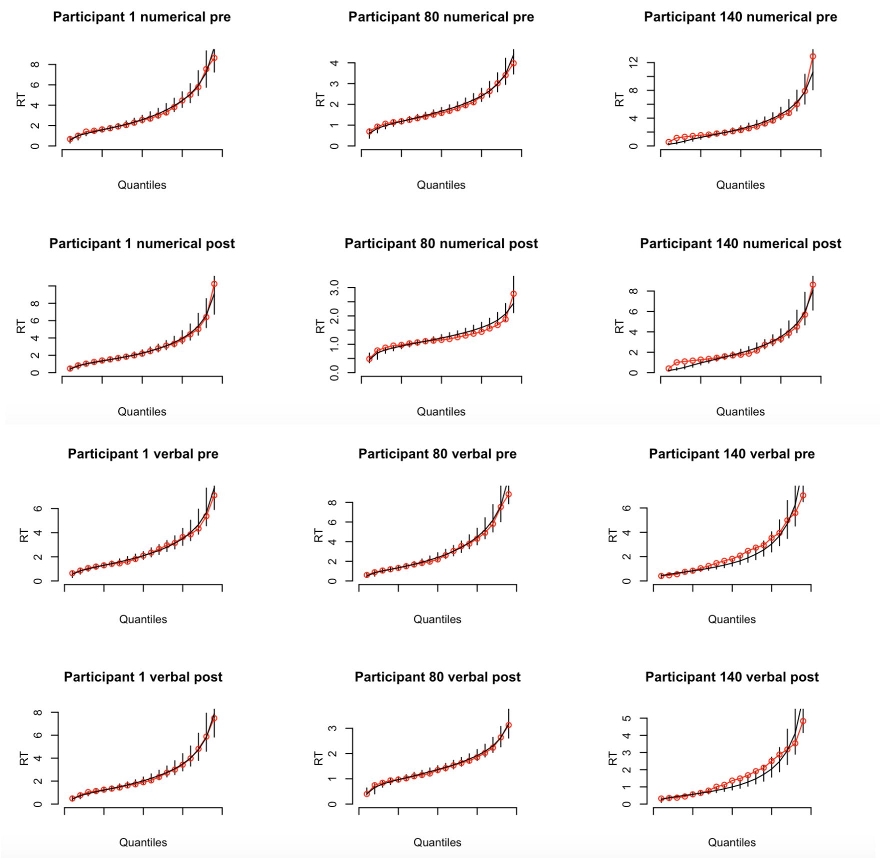

Figure 12: The posterior predictive RT distributions contrasted with the empirical RT distributions for Participants 1, 80, and 140 from the control, updating, and binding groups, respectively. Results of the other participants can be found in the supplemental materials. The red points are the 5% to 95% RT quantiles incremented by 5% from empirical data, connected by red lines. The vertical black bars show the 5% to 95% RT quantiles incremented by 5% from posterior predictive samples, connected by black lines. The length of black bars reflect the size of 95% equal-tailed credible intervals from all samples. The x-axis corresponds to the quantiles, and the y-axis corresponds to RT values at each quantile.

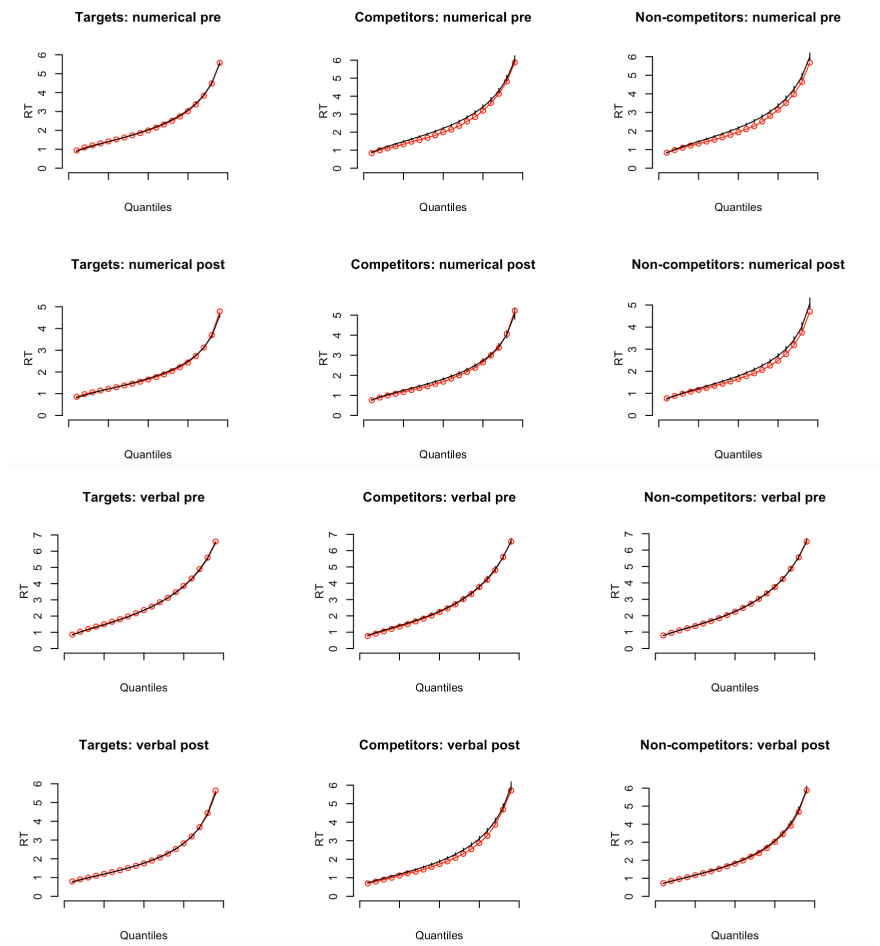

Figure 13: The posterior predictive RT distributions contrasted with the empirical RT distributions for responses from targets, competitors and non-competitors. The red points are the 5% to 95% RT quantiles incremented by 5% from empirical data, connected by red lines. The vertical black bars show the 5% to 95% RT quantiles incremented by 5% from posterior predictive samples, connected by black lines. The length of black bars reflect the size of 95% equal-tailed credible intervals from all samples. The x-axis corresponds to the quantiles, and the y-axis corresponds to RT values at each quantile.

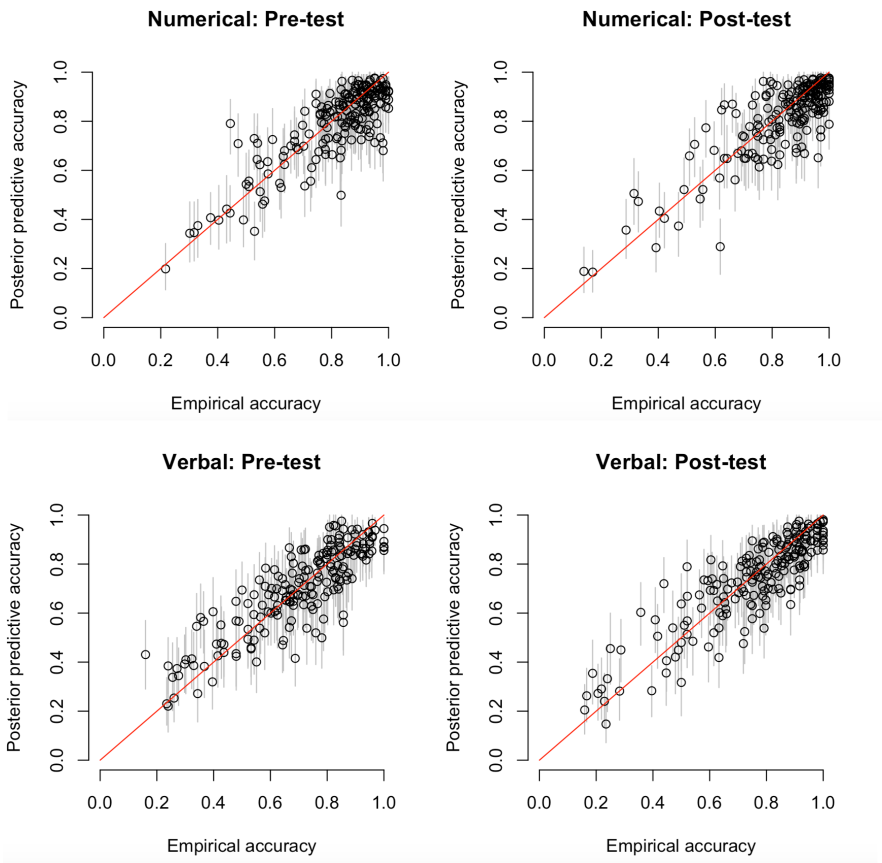

Figure 14: Posterior predictive results of response accuracy based on the generated out-of-sample validation parameters from test data (y-axis), compared to the observed accuracies and RTs from the other data (x-axis). These figures are plotted in the same way as those shown in Figure 11.

3.2.3 Results and implications

In this section, we display a subset of estimated group and individual posterior parameters, and discuss their patterns and implications.

Figure 16 shows the estimated posterior distributions of the pre/post-test differences of group-level parameters from the numerical version (upper) and the verbal version (lower). From inspection of the plot, the updating-trained participants have a larger decrease in the level of interference than the other groups, reflected by C_c. However, the binding-trained participants do not show a similar decrease in the level of interference: in contrast, they have less decrease than the control group in both tasks. Among other parameters, the noise parameters, σ_c, and the accumulation rate parameters in the recall period, $v_{r,c}$, appear to show potential transfer effects, as both the updating

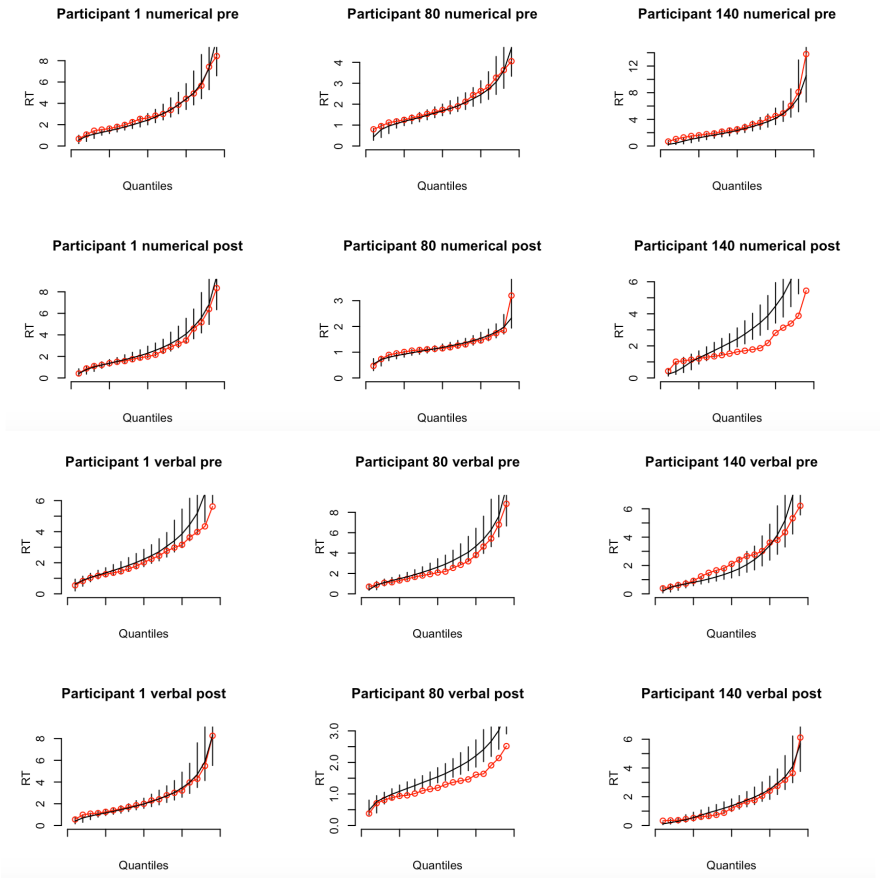

Figure 15: Posterior predictive results of RT based on the generated out-of-sample validation parameters from test data (black), compared to the observed accuracies and RTs from the other data (red). These figures are plotted in the same way as those shown in Figure 12.

and binding groups have a similar trend of changes compared with the control group. Based on these results, we discuss the results of each of the parameters C_c, σ_c and $v_{r,c}$ by investigating their posterior distributions. For each parameter, we compute the estimated posterior probability that the pre/post-test differences are numerically smaller in the control group than in the other groups. When the estimated posterior probability is less than 0.3 or more than 0.7, we consider that there may be a meaningful divergence between the control group and another group.

In the difference of interference parameters, $C_c^{(2)} - C_c^{(1)}$, the posterior probabilities that the control group has a higher $C_c^{(2)} - C_c^{(1)}$ than the updating and binding groups are 0.18 and 0.68 respectively in the numerical version, and are 0.14 and 0.78 respectively in the verbal version. Therefore, results from the interference parameters provide evidence supporting a general training effect, because the updating group consistently show a larger decrease in the level of interference than the control group. However, there is no evidence supporting a transfer effect in mutual interference, as the binding-trained group shows either no contrast or less decrease in the level of interference than the control group.

In the differences of noise parameters, $\log(\sigma_c^{(2)}) - \log(\sigma_c^{(1)})$, the posterior probabilities that the control group has a lower $\log(\sigma_c^{(2)}) - \log(\sigma_c^{(1)})$ than the updating and binding groups are 0.34 and 0.32 respectively in the numerical version, and are 0.01 and 0.13 respectively in the verbal version. This may indicate that, regardless of whether the training is in updating or binding tasks, training in verbal tasks may reduce noise in mental representation in the verbal version of the memory updating task, because both the updating and the binding group have a decrease in noise compared with the control group. However, training in numerical tasks may not reduce the noise in the numerical version of the memory updating task as effectively.

In the differences of RT accumulation rate parameters from the recall period, $v_{r,c}^{(2)} - v_{r,c}^{(1)}$, the posterior probabilities that the control group has a lower $v_{r,c}^{(2)} - v_{r,c}^{(1)}$ than the updating and binding groups are 0.98 and 0.85 respectively in the numerical version, and are 0.98 and 0.88 respectively in the verbal version. This parameter shows a clear effect of training, where the updating group has an increased accumulation speed after training in both numerical and verbal versions, and this group improves more than the other groups. There may also be a transfer effect associated with this parameter, as the binding-trained group also have consistent increases in the accumulation rate in both tasks, and the increases are larger than those from the control group.

Based on group-level results, for individual-level parameters, we examine the interference parameters C_{ic}, as they are tightly linked with the interference mechanism of interest, and the accumulation rate parameters $v_{r,ic}$ in the recall period, as they may embed possible transfer effects. Figures 17 and 18 display individual pre/post-test differences of these parameters.

As displayed by the numerical version of Figure 17, in the updating group, 36 out of 59 participants have an estimated posterior probability larger than 0.7 that $C_{ic}^{(2)} < C_{ic}^{(1)}$; the control group has 32 such participants out of 72, and the binding group has 26 such participants out of 66. In the verbal version, the updating group has 26 such participants out of 59; the control group has 18 out of 72; and the binding group has 15 out of 65. From the individual level, a larger proportion

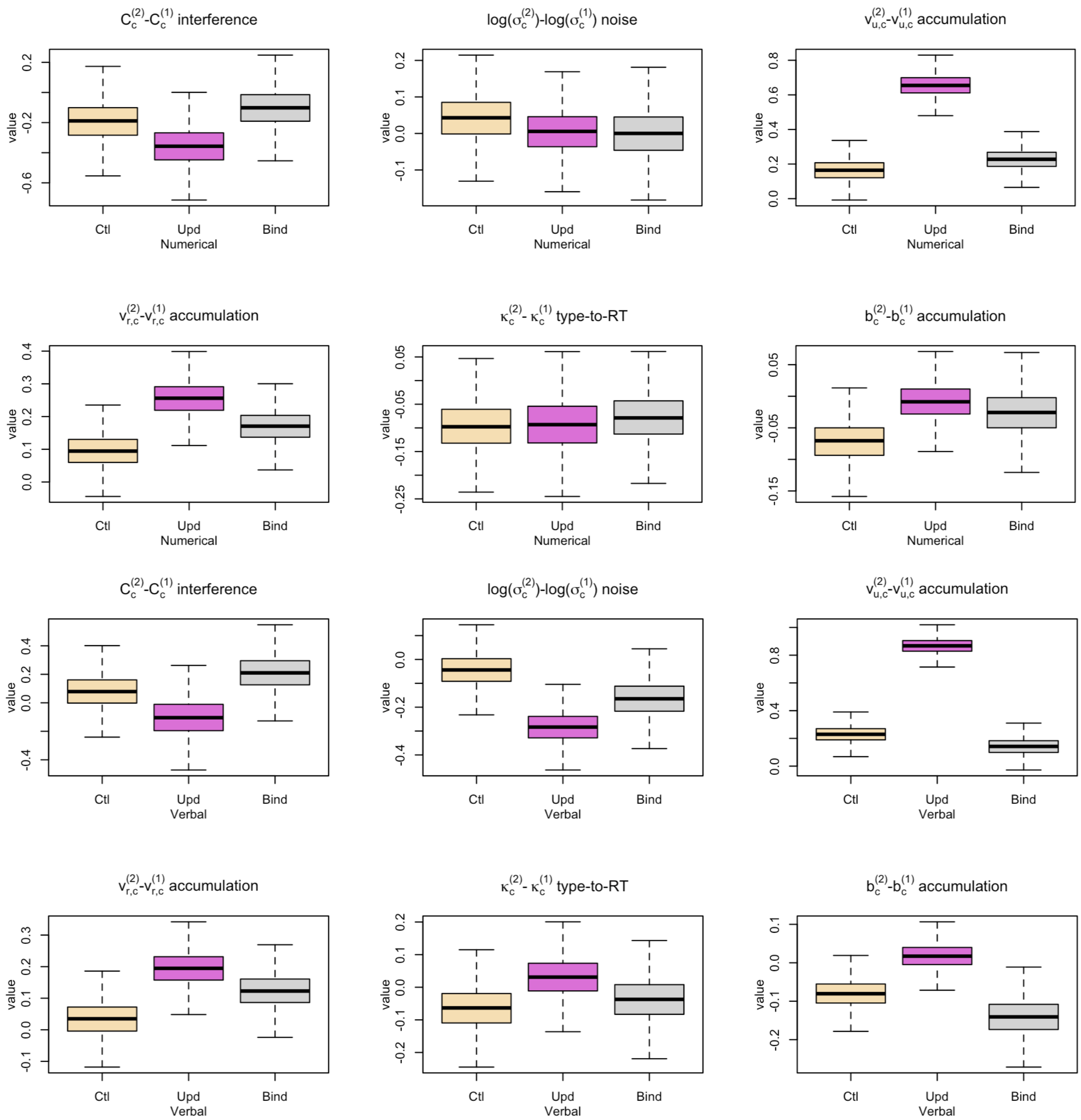

Figure 16: Posterior box-plots for the pre-post test differences of group-level parameters from the numerical version (upper) and the verbal version (lower). Parameters $\theta^{(1)}_{\cdot,c}$ ($\theta \in (C, \sigma, \kappa, \alpha)$) corresponds to the pre-test condition, and $\theta^{(2)}_{\cdot,c}$ corresponds to the post-test condition. The subscripts u and r indicate that the parameters are from the updating period and the recall period, respectively. The control group ("Ctl"), memory-updating group ("Upd"), and binding group ("bind") are each colored in brown, black and gray, respectively.

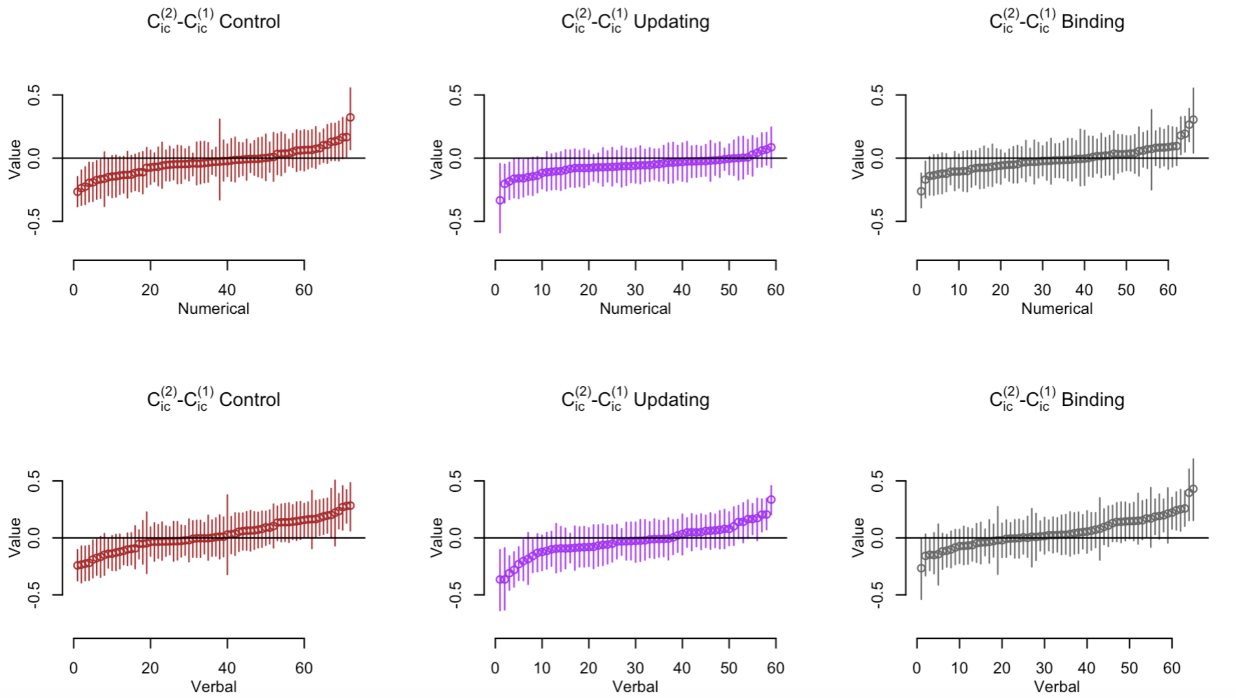

Figure 17: The whisker plots of pre-post test differences of individual posteriors for the interference parameter C in the recall period. The upper plots are from the numerical task and the lower plots are from the verbal task. The control, updating, and binding groups are colored in brown, black, and gray, respectively. In the whisker plots, the points are placed at the posterior medians, and the whiskers are the 95% equal-tailed credible intervals. We re-ordered participants in each group by estimated parameter values so that their results are shown in an increasing order.

of participants from the updating group have a reduction in the level of interference, which may indicate that memory-updating training could potentially reduce the degree of mutual interference in working memory processing for some individuals in the same task. However, similar to results from the group level, there is no evidence that binding training can help to reduce the level of interference in memory updating tasks.

Figure 18 shows the individual-level accumulation rate parameters in the recall period. Most participants in the updating and binding groups have an increase in the RT speed after training. In the updating group, 49 out of 59 participants has an estimated posterior probability larger than 0.7 that $v_{r,ic}^{(2)} > v_{r,ic}^{(1)}$ in the numerical version, and 46 out of 59 in the verbal version. The binding group has 45 out of 66 such participants in the numerical version and 33 out of 65 participants in the verbal version. The control group has 37 out of 72 such participants in the numerical version

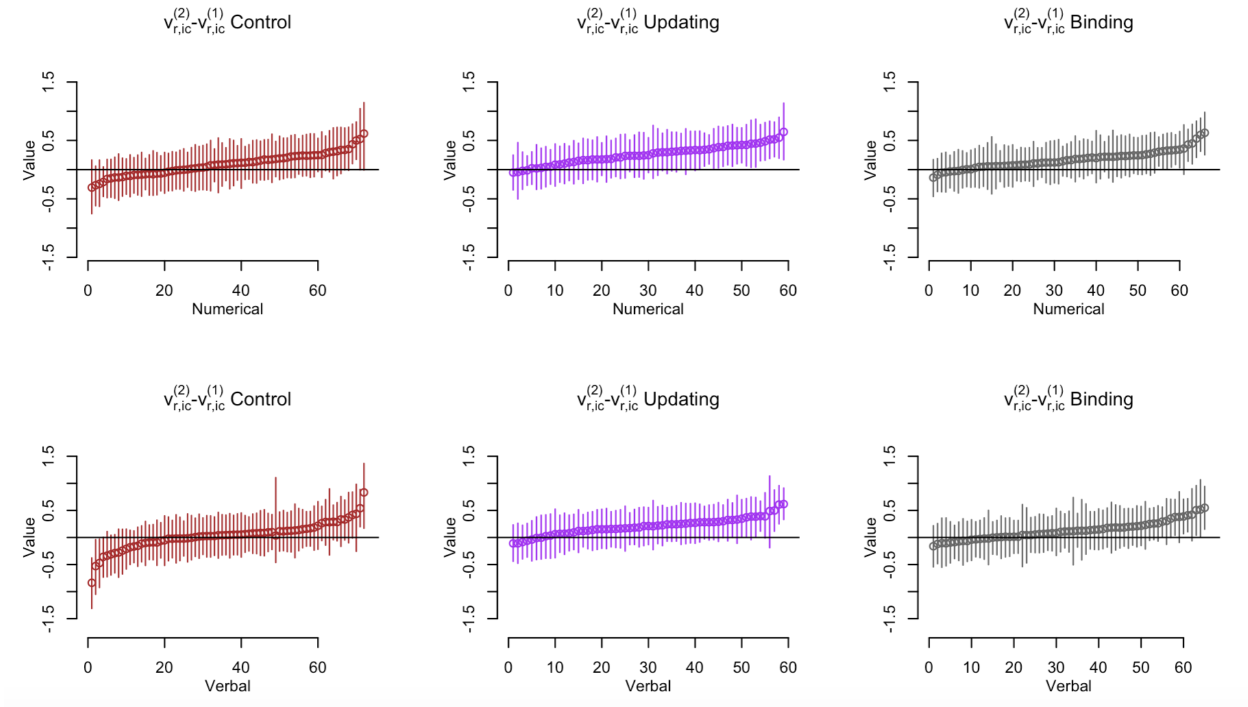

Figure 18: The whisker plots of pre-post test differences of individual posteriors for the RT accumulation rate parameter v in the recall period. The upper plots are from the numerical task and the lower plots are from the verbal task. The control, updating, and binding groups are colored in brown, black, and gray, respectively. In the whisker plots, the points are placed at the posterior medians, and the whiskers are the 95% equal-tailed credible intervals. We re-ordered participants in each group by estimated parameter values so that their results are shown in an increasing order.

and 25 out of 72 participants in the verbal version. These results may indicate that both updating and binding training are more helpful in improving the speed of information accumulation in the recall period than the visual search tasks. It may be the result of familiarizing and speeding up passive working memory processes, where the participants learned to use the cognitive resources 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.

Consistent with the findings reported by De Simoni & von Bastian (2018), we found no evidence of transfer effects in the level of interference. However, our results indicate that there may be training and transfer effects in the speed of information accumulation in these memory tasks. As is shown by the RT accumulation parameter in the recall period, the updating group has a large increase in the accumulation speed after training in both numerical and verbal versions of the task. The binding group does not have as large an increase as that of the updating group, but still have a larger increase in the speed of information accumulation than the control group in both versions of the task. At the individual level, larger proportions of participants have increases in the speed of information accumulation in updating and binding groups than in the control group. This may indicate that some degree of near transfer effects may be present in the passive components shared by all working memory components. Because these components are shared, training of them in one task may indirectly improve them in another task.

4 Conclusion and future directions

In this paper, we developed a hierarchical Bayesian model for working memory updating based on mutual interference. The model adapted the activation framework based on mutual interference (Oberauer & Kliegl, 2006) and the Wald diffusion model, allowing it to jointly model the response accuracies and RTs. This hierarchical Bayesian model yielded reasonable fits to several memory updating data sets, thus we conclude that it is a feasible model for the memory updating task. Compared with previous models, the joint modeling framework in this model allows each of response and RT information to inform and potentially improve parameter estimation of the other. Because of the inclusion of RT parameters, results of this model may reflect possible RT-related effects that are otherwise not shown by parameters related to response accuracy. This model also used mixtures to account for ancillary processes such as pre-activation, which could otherwise confound the modeling results.

In this section, we discuss potential future directions in working memory studies. Some are related to possible improvements to our current modeling methods, and the others are related to extensions of working memory studies based on findings in Section 3.

4.1 Discussion: Modeling

We discuss potential work related to the modification and improvement of our modeling approach.

Firstly, we constructed the hierarchical Bayesian model with the purpose of quantifying the level of interference and RT differences in groups and individuals. Our focus is not on the evaluation and comparison of different interference mechanisms, or on the mechanisms limiting working memory capacity in general. In future studies, it may be plausible to build joint models or further improve existing models based on different specific mechanisms, and evaluate the plausibility of each mechanism by model comparison (e.g. Oberauer & Kliegl, 2001; Ecker et al., 2015; Tan et al., 2017). Because RT distribution information is shown to be informative in this study, it may also be beneficial to incorporate RT distributions in potential models and perform estimations jointly. With several models, model comparisons may also be used to evaluate the flexibility of different models, which regards their abilities to fit the data with their numbers of effective parameters evaluated by the deviance information criterion (Spiegelhalter et al., 2002) or the Watanabe-Akaike information criterion (Gelman et al., 2013).

To accommodate the mixture structures in this model, we used strongly informative priors for some parameters to avoid potential identifiability problems and reduce computational time. However, if the ancillary processes are not the interest of a study, it may be feasible to discard fast and slow responses, model the process with only the algorithmic mutual interference and Wald diffusion processes, and relax the priors of these parameters. It may also be feasible to fit this model with another sampler other than the Hamiltonian Monte Carlo sampler from Stan, such as the Differential Evolution approach (Turner et al., 2013), so that the process has less computational cost.

4.2 Discussion: Working memory applications

We discuss potential future work based on our modeling assumptions and findings in empirical applications.

When constructing the model, we made several assumptions that can be tested or extended in future development. The first is the assumption that targets and competitor items share the same level of interference in each task. This assumption can be violated if a design involves stimuli of different levels of distinctiveness (e.g. Oberauer et al., 2012) or incentives (e.g. Strand et al., 2012), thus the modeling framework can be updated to incorporate these conditions in future development. The second is the assumption that items are perfectly encoded in the memorization period, which is a simplification and may be modified to include potential encoding effects such as elaborative encoding (Bradshaw & Anderson, 1982) in future works. The third is the assumption of pre-activation for the fast RTs around the smaller mode. This assumption is examined in this study but may need further evidence from future studies to validate.

In the application investigating aging effects in working memory based on data from Oberauer & Kliegl (2001), our results indicate that older adults may have higher levels of interference between items, less use of memory trace information to guide their information accumulation processes, and less use of potential pre-activation processes compared with younger adults. One possible extension to this work is related to the pre-activation processes. It may be helpful to further test the pre-activation assumption, and investigate the mechanisms underlying the processing differences between younger and older adults in relation to pre-activation in future research.

In the application investigating transfer effects in working memory training based on data from De Simoni & von Bastian (2018), we found strong evidence of training and transfer effects in the speed of processing. Because our analysis focused on the verbal-numerical aspect of working memory, effects in the visuo-spatial aspect may be investigated in future studies. For a better evaluation of the benefits of working memory training, it may also be helpful to study the reason for this transfer effect in future research. One possibility is to identify the specific working memory processes that benefit from the transfer effects, so that the scope and generalizability of the transfer effect can be understood.

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.

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.

Barrouillet, P., Uittenhove, K., Lucidi, A., & Langerock, N. (2018). On the sources of forgetting in working memory: The test of competing hypotheses. *Quarterly Journal of Experimental Psychology*, *71*(8), 1714–1733.

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

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.

Bradshaw, G. L., & Anderson, J. R. (1982). Elaborative encoding as an explanation of levels of processing. *Journal of Verbal Learning and Verbal Behavior*, *21*(2), 165–174.

Brown, G. D., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, *114*(3), 539.

Burbeck, S. L., & Luce, R. D. (1982). Evidence from auditory simple reaction times for both change and level detectors. *Perception & Psychophysics*, *32*(2), 117–133.

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.

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. *Psychological Bulletin*, *104*(2), 163.

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.

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., Brown, G. D., & Lewandowsky, S. (2015). Memory without consolidation: Temporal distinctiveness explains retroactive interference. *Cognitive Science*, *39*(7), 1570–1593.

Ecker, U. K., Lewandowsky, S., Oberauer, K., & Chee, A. E. (2010). The components of working memory updating: an experimental decomposition and individual differences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(1), 170.

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

Farrell, S., Oberauer, K., Greaves, M., Pasiecznik, K., Lewandowsky, S., & Jarrold, C. (2016). A test of interference versus decay in working memory: Varying distraction within lists in a complex span task. *Journal of Memory and Language*, *90*, 66–87.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis.* CRC press.

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

Guitard, D., Saint-Aubin, J., & Cowan, N. (2021). Asymmetrical interference between item and order information in short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(2), 243.

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

Hofmeister, P., & Vasishth, S. (2014). Distinctiveness and encoding effects in online sentence comprehension. *Frontiers in Psychology*, *5*, 1237.

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.

Li, Y., & Cowan, N. (2021). Attention effects in working memory that are asymmetric across sensory modalities. *Memory & Cognition*, 1–16.

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.

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.

Nairne, J. S. (2006). Modeling distinctiveness: Implications for general memory theory. *Distinctiveness and Memory*, 27–46.

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. (2009). Interference between storage and processing in working memory: Feature overwriting, not similarity-based competition. *Memory & Cognition*, *37*(3), 346–357.

Oberauer, K., Farrell, S., Jarrold, C., & Lewandowsky, S. (2016). What limits working memory capacity? *Psychological Bulletin*, *142*(7), 758.

Oberauer, K., Farrell, S., Jarrold, C., Pasiecznik, K., & Greaves, M. (2012). Interference between maintenance and processing in working memory: The effect of item–distractor similarity in complex span. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(3), 665.

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.

Park, H., Arndt, J., & Reder, L. M. (2006). A contextual interference account of distinctiveness effects in recognition. *Memory & Cognition*, *34*(4), 743–751.

Province, J. M., & Rouder, J. N. (2012). Evidence for discrete-state processing in recognition memory. *Proceedings of the National Academy of Sciences*, *109*(36), 14357–14362.

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.

Rumelhart, D. E., McClelland, J. L., Group, P. R., et al. (1988). *Parallel distributed processing* (Vol. 1). IEEE Massachusetts.

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.

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.

Smith, G., Franck, J., & Tabor, W. (2021). Encoding interference effects support self-organized sentence processing. *Cognitive Psychology*, *124*, 101356.

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.

Souza, A. S., & Oberauer, K. (2015). Time-based forgetting in visual working memory reflects temporal distinctiveness, not decay. *Psychonomic Bulletin & Review*, *22*(1), 156–162.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series b (Statistical Methodology)*, *64*(4), 583–639.

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

Strand, M. T., Hawk, L. W., Bubnik, M., Shiels, K., Pelham, W. E., & Waxmonsky, J. G. (2012). Improving working memory in children with attention-deficit/hyperactivity disorder: the separate and combined effects of incentives and stimulant medication. *Journal of Abnormal Child Psychology*, *40*(7), 1193–1207.

Surprenant, A. M., & Neath, I. (2013). *Principles of memory.* Psychology Press.

Tan, Y., Martin, R. C., & Van Dyke, J. A. (2017). Semantic and syntactic interference in sentence comprehension: A comparison of working memory models. *Frontiers in Psychology*, *8*, 198.

Turner, B. M., Sederberg, P. B., Brown, S. D., & Steyvers, M. (2013). A method for efficiently sampling from distributions with correlated dimensions. *Psychological Methods*, *18*(3), 368.

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.

Villata, S., Tabor, W., & Franck, J. (2018). Encoding and retrieval interference in sentence comprehension: Evidence from agreement. *Frontiers in Psychology*, *9*, 2.

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 1: Simulation and parameter recovery

We performed 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 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). In this section, we describe our methods to generate the simulated data sets, and the results of these simulations. Based on simulation results, we argue that the hierarchical Bayesian model has a reasonable parameter recovery ability for most parameters.

4.2.1 Simulation without intermediate responses

In this section, we examine whether the model can successfully recover the true parameter values when responses from the recall steps are recorded, and when the updating time limit is binned into four segments (see Section 2.3.5). To obtain simulated data sets, we simulated each data set from two groups, each consisting of 10 artificial participants. We used the experimental schemes similar to that from Oberauer & Kliegl (2001) to generate simulated data. We simulated two cases, where each participant completes 80 trials (around 250 observations depending on the memory demand) or 240 trials (around 700 observations). We generated the individual parameters with group-level parameters obtained from modeling results in Section 3.1.3, using distributions shown in Table 1. We obtained 30 different sets of parameters and 30 corresponding simulated data sets in each case.

We fit the model using the procedure described in Section 2.3. In particular, we computed quartiles for the limiting times during the updating period, and fit the model to data with binned limiting times. We fit the hierarchical Bayesian model to these data using the software Stan. For each simulated data set, we obtained a chain containing 500 warm-up samples and 1000 iterations.

We present parameter recovery results in Figure 19 and Figure 20. Figure 19 shows that when the sample size is relatively small, the interference parameter C_{ic} is slightly overestimated at lower values. Estimation of the activation rate parameter θ_{ic} is less accurate at higher values, which is likely due to the binning of updating limit times. Besides the main parameters of interest, the response probabilities in the ancillary processes, $q_{k,ic}$, and the Log-normal standard deviation of fast ancillary processes, $\sigma_{\mu,r,ic}$, have a relatively larger variability in their estimation. Figure 20 shows that the quality of estimation improves when the sample size is relatively large. The majority of the parameters are reasonably recovered and there are no major patterns of divergence from the identity lines. The ancillary processes parameters, especially $q_{k,ic}$ and $\sigma_{\mu,r,ic}$, have less variable estimation compared with the small sample size condition.

In these simulations, the parameter recovery results are reasonably good for model parameters, and the precision of estimation improves as the sample size increases. With a larger sample size, the model is able to recover all parameters well, including the activation rate parameter after binning. Therefore, the model is feasible for Oberauer & Kliegl (2001)'s data set.

4.2.2 Simulation with intermediate responses

In this section, we examine whether the model can successfully recover true parameter values for both the updating and recall steps, and when the number of observations per participant is relatively small. We describe in order the methods to generate the simulated data sets, the model fitting procedures, and simulation results.

To evaluate whether the model has a good parameter recovery ability when the data has a small number of observations like that in De Simoni & von Bastian (2018), we simulated each data set from 3 groups, each consisting of 10 artificial participants. We used experimental schemes identical to those of the numerical version of memory updating task from De Simoni & von Bastian (2018), described in detail in Section 3.2.1. This scheme consists of 208 trials in total, including 144 trials from the updating period and 64 trials from the recall period. In this simulation, each participant completes the task once (208 observations in total). In each simulation, we selected the group-level parameters as those obtained from Bayesian fits of the model to empirical data from Section 3.2.3, which embedded reasonable group differences. We generated the individual-level parameters from the group-level parameters using the distributions shown in Table 1. These values have considerable variability but are also within a reasonable range for an empirical scenario. For parameters outside the hierarchy, we randomly selected values out of all posterior samples for them. We generated 50 different sets of parameters and 50 corresponding simulated data sets. Although the first simulation shows that the model has a good parameter recovery ability when the sample size is large, for a more thorough comparison, we also performed 5 simulations where each participant performed the task 10 times, resulting in 2080 observations each. We fit the hierarchical Bayesian model to these

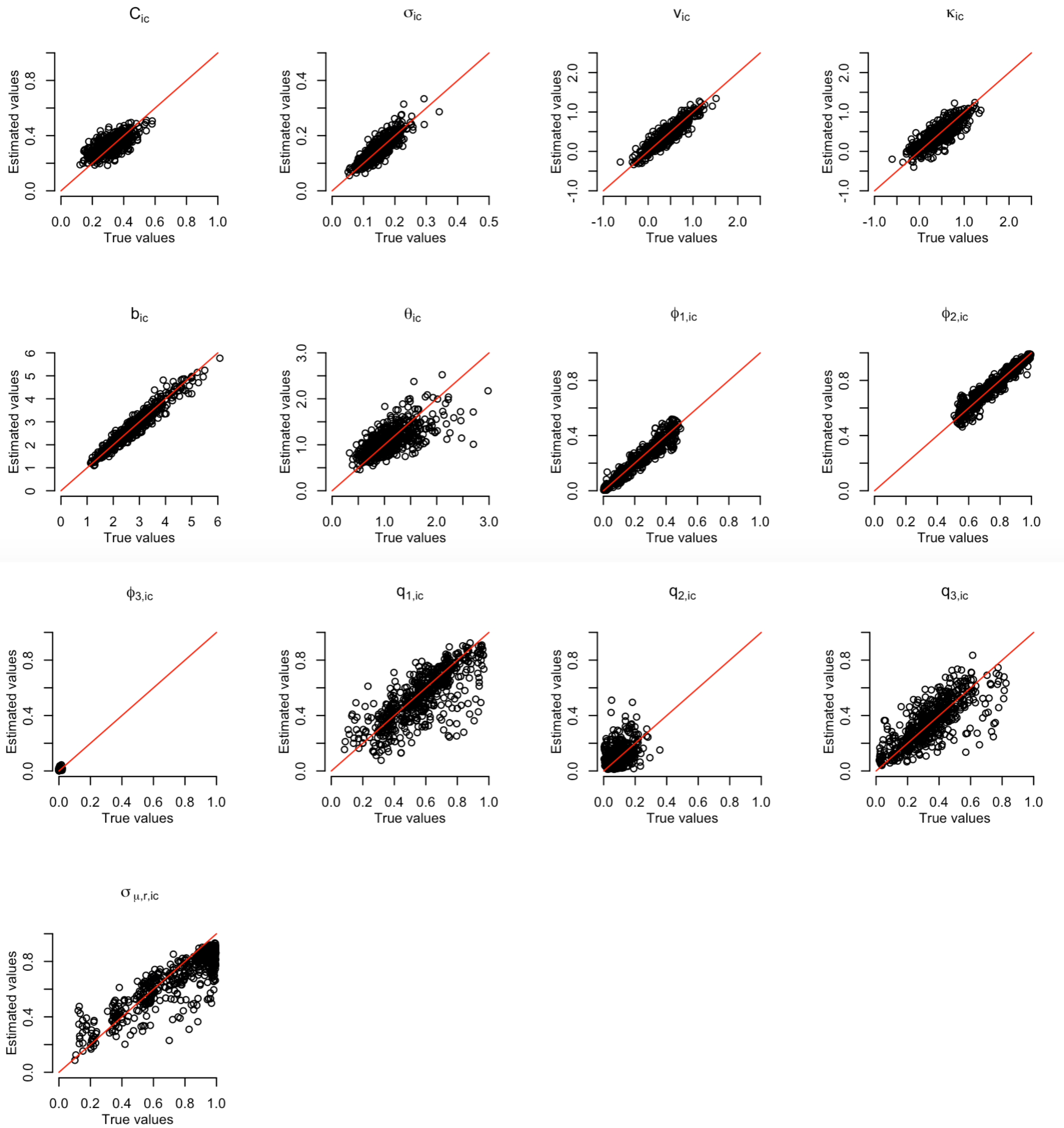

Figure 19: Results from simulations with 80 trials (about 250 observations per participant) where intermediate updating responses are not recorded. This figure shows the contrasts between true parameter values (x-axis) and parameter values recovered from simulations (y-axis) for individual-level parameters. Red lines are identity lines.

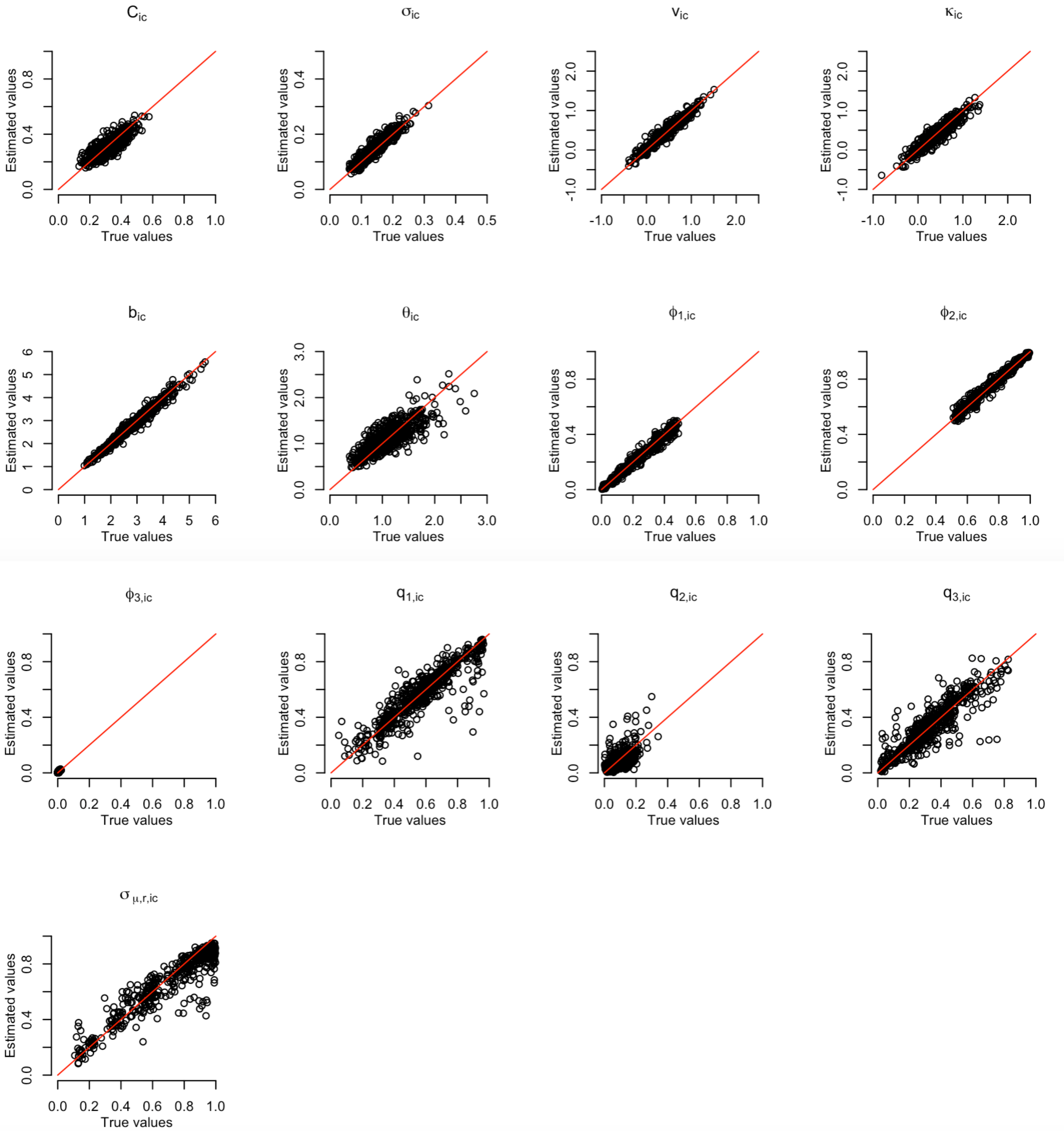

Figure 20: Results from simulations with 240 trials (about 700 observations per participant) where intermediate updating responses are not recorded. This figure shows the contrasts between true parameter values (x-axis) and parameter values recovered from simulations (y-axis) for individual-level parameters. Red lines are identity lines.

data using the software Stan (Stan Development Team, 2018). For each simulated data set, we obtained a chain containing 500 warm-up samples and 1000 iterations.

We present parameter recovery results in Figure 21 and Figure 22. Figure 21 shows that, even with a really small sample size, the model can recover parameters well. It is most noticeable that the interference parameter C_{ic} is slightly overestimated at low values under 0.1. In comparison, Figure 21 shows that when the sample size is large, C_{ic} is more precisely estimated. Based on these results, the overestimation of C_{ic} at small sample sizes is likely to occur because we used tighter priors (see Section 2.3.4) to accommodate the inclusion of mixtures and computational cost, and when the sample size is small, the priors are not properly shifted to the precise location. However, because the C_{ic} estimates positively relate to the true values, this overestimation is unlikely to lead to misinterpretation of modeling results. Therefore, the model is reasonable to use for parameter estimation and interpretation in data sets with small sample sizes like those from De Simoni & von Bastian (2018).

Appendix 2: Model comparison

We perform a model comparison to evaluate whether the participants' type of responses are linked to their information accumulation rate. From the mechanism in Section 2.3.1, the number of memory traces linked to each type of item (targets, competitors and non-competitors) are embedded in the probabilities $p^*_{k,ic,j}$ of responding by that type. More memory traces linked to an item (such as a target) can lead to a larger $p^*_{k,ic,j}$ for that item. In Section 2.3.2, we hypothesize that the number of memory traces can also affect the information accumulation rate in the Wald diffusion process while responding, reflected in the type-to-RT parameter κ_{ic}. When a participant performs the task well, we expect $\kappa_{ic} > 0$, so that as target items have more memory traces, this participant can collect information faster and respond earlier to targets than non-targets. In Sections 3.1.1 and 3.2.1, we showed that the relations of RTs and responses are consistent with this assumption in all data sets.

In this section, we use the Watanabe-Akaike information criterion (WAIC, Gelman et al., 2013)

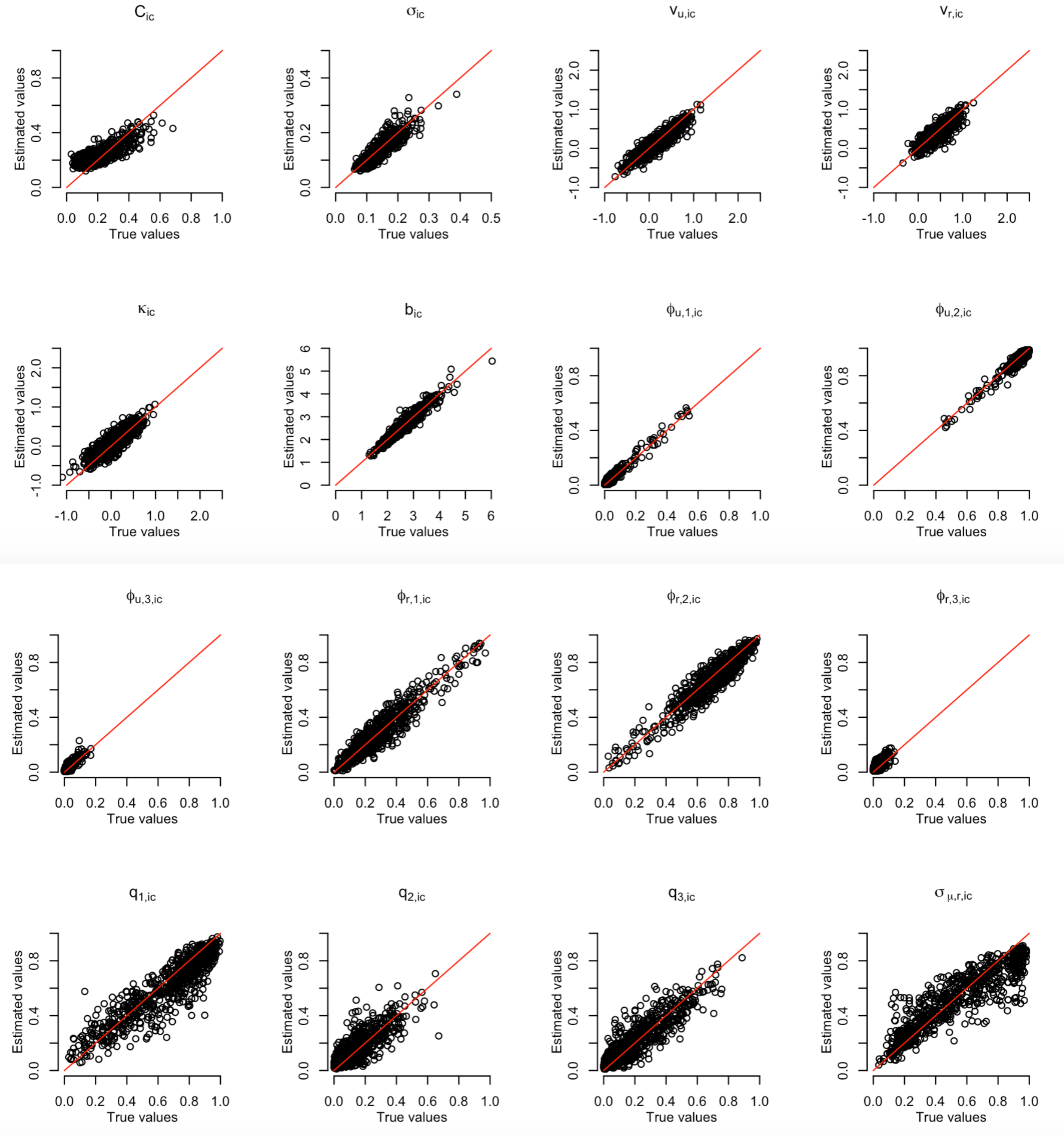

Figure 21: Results from simulations with 208 observations where intermediate updating responses are recorded. This figure shows the contrasts between true parameter values (x-axis) and parameter values recovered from simulations (y-axis) for individual-level parameters. Red lines are identity lines.

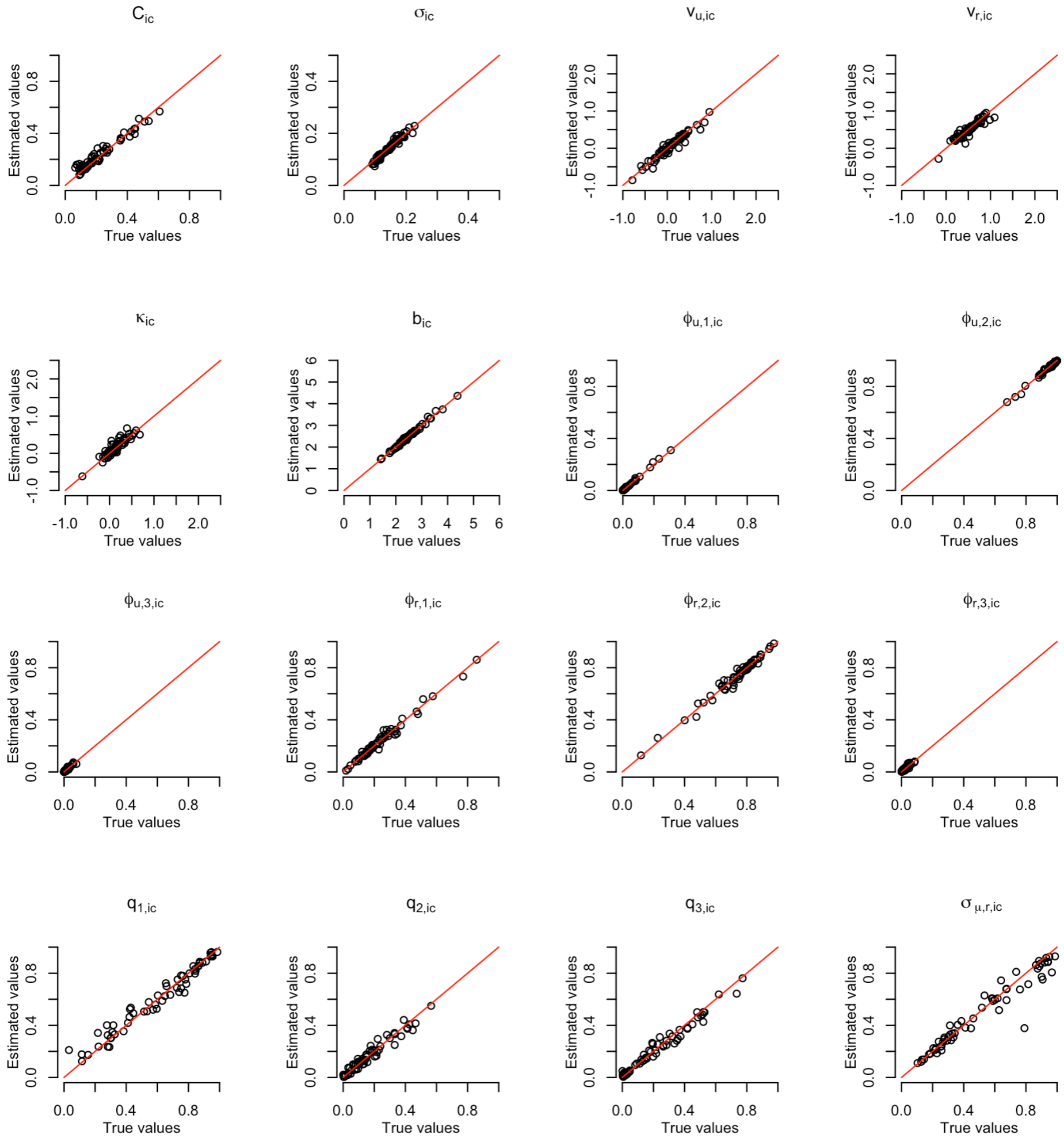

Figure 22: Results from simulations with 2080 observations where intermediate updating responses are recorded. This figure shows the contrasts between true parameter values (x-axis) and parameter values recovered from simulations (y-axis) for individual-level parameters. Red lines are identity lines.

to evaluate whether the inclusion of κ_{ic} improves model fit without over-fitting. For all data sets, we fit a model with κ_{ic} and a model without κ_{ic} to the data, generating a chain of 500 warm-up samples and 2000 iterations each. We then compute the WAICs for each model. Table 4 shows the WAICs and effective number of parameters from both models in each data set. Two versions of WAIC and effective numbers of parameters are used. In “WAIC1” and ”Eff1”, the effective number of parameters p_{WAIC1} is computed as

$$p_{\text{WAIC1}} = 2\sum_{j=1}^{n}\Big(\log\Big(\frac{1}{S}\sum_{s=1}^{S}p(y_j|\eta_s)\Big) - \Big(\frac{1}{S}\sum_{s=1}^{S}\log p(y_j|\eta_s)\Big)\Big),$$

where η denotes the posterior samples, $\hat{\eta}$ denotes the posterior sample means, and y denotes the data. With a total of n data points and S posterior samples, y_j denotes the jth data point and η_s denotes the sth posterior sample. In “WAIC2” and “Eff2”, the effective number of parameters p_{WAIC2} is computed as

$$p_{\text{WAIC2}} = \sum_{j=1}^{n}\text{var}_{s=1}^{S}(\log p(y_j|\eta_s)).$$

The log pointwise predictive density is computed as

$$\text{lppd} = \sum_{j=1}^{n}\log\Big(\frac{1}{S}\sum_{s=1}^{S}p(y_j|\eta_s)\Big),$$

and the WAICs are computed as

$$\text{WAIC} = -2\text{lppd} + 2p_{\text{WAIC}}.$$

From Table 4, for all data sets, both WAICs are lower for the model with κ_{ic} than the model without κ_{ic}. This indicates that the model with κ_{ic} is superior based on both measures for all data sets, and the type-to-RT parameter κ_{ic} should be retained in the model.

				Numerical version			
Table A1		Control	Updating	Binding	Control	Updating	Binding
			Accuracy			**Updating RT**	
	Pre-test	0.81	0.83	0.81	3.20 (3.27)	3.46 (3.43)	3.17 (2.77)
	Post-test	0.82	0.88	0.82	2.82 (2.91)	1.80 (1.26)	2.64 (2.29)
	Mean diff	0.01	0.05	0.01	-0.38	-1.66	-0.53
			Recall RT				
	Pre-test	1.51 (1.40)	1.57 (1.31)	1.49 (1.16)			
	Post-test	1.32 (1.12)	1.17 (1.20)	1.28 (0.90)			
	Mean diff	-0.19	-0.40	-0.21			
Table B1	RT	T1 (pre)	T1 (post)	T2 (pre)	T2 (post)	T3 (pre)	T3 (post)
	Control	3.25/1.65	2.89/1.52	3.67/2.20	3.04/1.96	3.40/2.05	2.91/1.76
	Updating	3.41/1.64	1.79/1.37	4.28/2.20	2.34/1.86	3.90/2.23	2.20/1.67
	Binding	3.17/1.67	2.72/1.46	3.61/2.06	3.20/1.88	3.64/1.88	2.60/1.72
Table C1			**Updating RT**			**Recall RT**	
	RT	0-0.2s	0.2-0.6s	>0.6s	0-0.2s	0.2-0.6s	>0.6s
	Accuracy	0.10	0.12	0.88	0.47	0.73	0.76
Table D1				**Sequential place in recall**			
	Memory demand	1	2	3	4	5	
	3	3.2%	18.6%	62.7%			
	5	2.4%	5.0%	11.7%	16.4%	24.0%	
				Verbal version			
Table A2		Control	Updating	Binding	Control	Updating	Binding
			Accuracy			**Updating RT**	
	Pre-test	0.72	0.73	0.66	4.52 (4.72)	4.55 (4.26)	4.33 (4.70)
	Post-test	0.72	0.86	0.69	3.48 (3.79)	1.82 (1.18)	3.31 (3.26)
	Mean diff	0.00	0.13	0.03	-1.04	-2.73	-1.02
			Recall RT				
	Pre-test	1.44 (1.72)	1.33 (1.18)	1.36 (1.31)			
	Post-test	1.24 (1.47)	0.94 (0.79)	1.07 (0.90)			
	Mean diff	-0.20	-0.39	-0.29			
Table B2	RT	T1 (pre)	T1 (post)	T2 (pre)	T2 (post)	T3 (pre)	T3 (post)
	Control	4.51/1.53	3.56/1.43	5.03/2.37	3.51/1.99	4.84/1.98	3.89/1.88
	Updating	4.51/1.43	1.80/1.16	5.01/1.97	2.19/1.74	4.98/1.89	2.21/1.51
	Binding	4.49/1.49	3.59/1.33	4.59/2.07	3.33/1.50	4.39/1.86	3.52/1.54
Table C2			**Updating RT**			**Recall RT**	
	RT	0-0.2s	0.2-0.6s	>0.6s	0-0.2s	0.2-0.6s	>0.6s
	Accuracy	0.14	0.16	0.79	0.39	0.68	0.62
Table D2				**Sequential place in recall**			
	Memory demand	1	2	3	4		
	2	15.7%	62.6%				
	4	3.6%	10.5%	20.3%	39.8%		

Table 3: Statistics from numerical and verbal versions of memory updating tasks from De Simoni & von Bastian (2018). "Table A"s display the summary statistics of response accuracies and RTs for the visual search (control) participants, updating-trained participants, and binding-trained participants. These tables show the mean response accuracies and the RT mean with standard deviations (in brackets) for each group. "Mean diff" rows show the mean differences of the post-test values from the pre-test values. "Table B"s display the mean RTs (calculated from RT larger than 0.6 seconds) of each type. T1, T2, and T3 correspond to targets, competitors, and non-competitors, respectively. The mean RTs before and after each "/" are from the updating and recall periods, respectively. "Table C"s display the response accuracies in the RT ranges of 0-0.2 seconds, 0.2-0.6 seconds, and more than 0.6 seconds. "Table D"s display the proportions of fast RTs around the smaller mode in the 0.2-0.6 seconds range in the recall period from each item to be recalled. The "Sequential place in recall" shows the sequential order of the item to be recalled.

	With κ_{ic}				No κ_{ic}			
	WAIC1	Eff1	WAIC2	Eff2	WAIC1	Eff1	WAIC2	Eff2
Age	**278072**	403	**278097**	415	282381	358	282404	370
Numerical	**348429**	5918	**350544**	6976	349406	5667	351458	6693
Verbal	**384085**	5452	**385574**	6197	385269	5233	386743	5970

Table 4: WAICs and effective number of parameters (Eff) of data sets from Oberauer & Kliegl (2001) ("Age") and De Simoni & von Bastian (2018) ("Numerical" and "Verbal").