HHS Public Access

Author manuscript

Psychol Aging. Author manuscript; available in PMC 2023 March 31.

Published in final edited form as:

Psychol Aging. 2022 February ; 37(1): 10–29. doi:10.1037/pag0000665.

Transparency, replicability, and discovery in cognitive aging research: A computational modeling approach

Kevin P. Darby,

Per B. Sederberg

Department of Psychology, University of Virginia

Abstract

Healthy aging is associated with deficits in performance on episodic memory tasks. Popular verbal theories of the mechanisms underlying this decrement have primarily focused on inferred changes in associative memory. However, performance on any task is the result of interactions between different neurocognitive mechanisms, such as perceptuomotor, memory, and decision-making processes. As a result, age-related differences in performance could arise from multiple processes, which could lead to incomplete or incorrect conclusions about the sources of aging effects. In addition, standard statistical comparisons of group-level summary statistics, such as mean accuracy, may not provide sufficient information to allow detailed mechanistic explanations of age-related change. We argue that these and other drawbacks of relying exclusively on verbal theories can hamper replicability, transparency, and scientific progress in aging research and psychological science more generally, and that computational modeling is a tool that can address many of these limitations. Computational models make mathematically transparent claims about how latent processes give rise to observed behavior, and decompose an individual's performance into model parameters governing hypothesized mechanisms. In this work, we present a short memory task designed for and analyzed with mechanistic model-based approaches. We provide an example of a computational model, and fit the model to data from young and older adults with hierarchical Bayesian techniques in order to (1) detect differences in latent cognitive processes between young and older adults (as well as individual participants), (2) quantitatively compare models to assess different processes that could underlie performance, and (3) simulate data to make predictions for future experiments based on model mechanisms. We argue that computational modeling is a powerful tool to examine age differences in latent processes, make theories more transparent, and facilitate discovery in cognitive aging research.

Keywords

Computational modeling; Replicability; Cognitive aging; Episodic memory; Temporal context model

Correspondence: Per B. Sederberg pbs5u@virginia.edu.

The young adult data and an earlier version of the computational model included in this work were reported in a previously published article (Weichart et al., 2021). The experimental stimuli, data, and model code are freely available at https://osf.io/g5tbh/. This manuscript was prepared with Pandoc and Markdown (https://pandoc.org; see Tenen & Wythoff, 2014 for a tutorial).

The field of psychological science has received criticism for difficulty replicating many findings in different participant samples, and in some cases even difficulty reproducing the results of statistical analyses of the same sample (Anvari & Lakens, 2018; Maassen, Assen, Nuijten, Olsson-Collentine, & Wicherts, 2020; Open Science Collaboration, 2015). A popular recommendation for addressing these concerns is to increase transparency by pre-registering study predictions, publishing null findings, and by making experimental paradigms, data, and analyses publicly available (Asendorpf et al., 2013; Nosek et al., 2015). Despite these efforts, however, challenges to replicability remain. Some researchers have suggested that an additional way to improve replicability and make psychological science more rigorous may be to increase the transparency of psychological *theories* by mathematically instantiating them in generative computational models of latent neurocognitive mechanisms (Farrell & Lewandowsky, 2010, 2018; Guest & Martin, 2021; Haines et al., 2020; Jolly & Chang, 2019; Oberauer & Lewandowsky, 2019a; Smaldino, 2020).

Generative computational models, often called explanatory models or cognitive process models, are systems of equations that mathematically instantiate an explicit theory of latent mechanisms. Such models are called generative because they provide an explanation for how the brain could generate observed behavior, albeit at varying levels of abstraction. A more common approach is to design experiments and base hypotheses on a purely verbal theory of a proposed process or how a process might differ within or between individuals, e.g., due to aging or the onset of a disease. As we argue below, verbal theories are often lacking in mechanistic specificity and can be interpreted differently by researchers. Generative computational models, by contrast, make theories transparent in that the mathematical equations that instantiate a model specify the operations of latent processes and remove the need for subjective interpretations of a theory that could vary between researchers (Guest & Martin, 2021).

In addition to making theories more transparent, more widespread application of computational modeling may increase replicability by creating stronger links between theory and experimental hypotheses (Muthukrishna & Henrich, 2019; Oberauer & Lewandowsky, 2019a; P. L. Smith & Little, 2018). In some cases, a theory may be too vague to provide sufficient guidance on what experiments would provide strong support either for or against it (see Oberauer & Lewandowsky, 2019a for discussion). Relatedly, research is sometimes conducted in an exploratory way, without strong ties to a specific theory. Some exploratory hypotheses are faulty but may still result in statistically significant results by chance that are unlikely to replicate. Findings that were hypothesized as a way to test the predictions of an explicit mechanistic theory, however, may be more likely to replicate because the hypotheses are born out of a line of reasoning and previous research in a principled way. By implementing a theory as a computational model, the researcher is (a) obliged to specify mechanistic details of the theory that would likely remain vague in a verbal theory, and (b) able to simulate data for different experiments that provide explicit hypotheses based directly on the model that can be compared to observed data. Cases when model predictions are not compatible with observed data, or when a predicted finding does not replicate, provide opportunities to reassess the model and iterate towards better theories (Guest &

Martin, 2021). For these reasons, computational modeling may help make psychological theories more robust to replication failure.

Models are most often employed for the purpose of developing and testing mechanistic theories of cognitive processes in healthy, and typically young, adults (Weichart et al., 2021). Increasingly, however, modeling frameworks have been applied to jointly test theories and gain insight into mechanistic differences between individuals from distinct groups. For example, a central goal of the nascent field of computational psychiatry (Huys, Maia, & Paulus, 2016; Maia, Huys, & Frank, 2017; Wiecki, Poland, & Frank, 2015) is to apply computational models to gain insights into how neurocognitive processes are affected by various neurological and developmental disorders (Braver, Barch, & Cohen, 1999; Cockburn & Holroyd, 2010; Cohen et al., 1996; Frank, 2008; Frank, Santamaria, O'Reilly, & Willcutt, 2007).

One domain in which computational modeling is still relatively rare is research on cognitive aging, despite long-standing arguments from influential researchers in the field on the potential benefits of this approach (Salthouse, 1988). Thankfully, however, a number of exceptions have been presented (Benjamin, 2010; Healey & Kahana, 2016; Howard, Kahana, & Wingfield, 2006; Kliegl & Lindenberger, 1993; Li, Naveh-Benjamin, & Lindenberger, 2005; Myerson, Hale, Wagstaff, Poon, & Smith, 1990; Ratcliff & McKoon, 2015; Ratcliff, Thapar, Gomez, & McKoon, 2004; Starns & Ratcliff, 2010; Stephens & Overman, 2018; Surprenant, Neath, & Brown, 2006), in which researchers have applied models to make their theories transparent and detect mechanistic changes between young and older adults, and in some cases have quantitatively compared mechanistic theories of aging.

In the current work, we discuss advantages of a computational modeling framework and provide examples of specific modeling techniques that we believe would particularly benefit research on aging. Our aim in this work is to encourage further adoption of computational modeling in the study of cognitive aging. We present a computational model of episodic memory, and apply hierarchical Bayesian techniques to fit the model to data from young and aging adults. We apply this approach to (1) examine mechanistic differences between age groups and individual participants in latent processes proposed to underlie task performance, (2) compare different models to test specific mechanistic ideas, and (3) generate data to make model-based quantitative predictions for unobserved experimental outcomes. We begin by discussing benefits of computational modeling in the context of aging effects on episodic memory.

Verbal versus computational models of episodic memory and aging

A great deal of experimental work has suggested that episodic memory declines with age (Cansino, 2009; Mitchell, Brown, & Murphy, 1990; Nilsson, 2003; Tromp, Dufour, Lithfous, Pebayle, & Després, 2015). Performance on episodic memory tasks such as free recall (Craik, 1968), serial recall (Golomb, Peelle, Addis, Kahana, & Wingfield, 2008), cued recall (A. D. Smith, 1977; Taconnat, Clarys, Vanneste, Bouazzaoui, & Isingrini, 2007), source memory (Dodson, Bawa, & Slotnick, 2007; Schacter, Kaszniak, Kihlstrom, & Valdiserri,

1991), and associative recognition (Castel & Craik, 2003; Naveh-Benjamin, Guez, Kilb, & Reedy, 2004; Old & Naveh-Benjamin, 2008) is typically less accurate in older adults than young adults. However, performance tends to be less affected by age on other memory tasks, such as item recognition (Naveh-Benjamin, 2000; Naveh-Benjamin, Guez, Kilb, & Reedy, 2004) and vocabulary tests (Verhaeghen, 2003).

A number of verbal theories have been proposed to explain this pattern of deficits, including developmental changes in attention (Craik, Luo, & Sakuta, 2010), inhibition (Hasher & Zacks, 1988; Healey, Hasher, & Campbell, 2013), and processing speed (Salthouse, 1996). One of the most popular theories of episodic memory change is the associative deficit account, which suggests that the age-related decline in episodic memory is primarily due to deficits in forming associations between items, whereas item familiarity is relatively intact (Naveh-Benjamin, 2000). Verbal theories such as these help guide the field and our conceptualization of how cognition is affected by aging. However, this approach has a number of important limitations that we argue can be addressed with a computational modeling approach (for similar arguments, see Farrell & Lewandowsky, 2018; Guest & Martin, 2021; Haines et al., 2020; Oberauer & Lewandowsky, 2019a; Salthouse, 1988; Smaldino, 2020).

One limitation of verbal theories is that they can be understood in different ways, as communicating a verbal theory necessarily involves some amount of interpretation and inference, which may differ between researchers (Farrell & Lewandowsky, 2010, 2018). As pointed out by Castel and Craik (2003), for example, the word “association” could mean different things in relation to memory, including integrating features together to form a conjunctive object or scene, integrating an item with its context, or forming a link between separate items. It is not difficult to imagine that even these more specific descriptions could be understood differently between researchers. Relatedly, verbal theories often do not provide mechanistically detailed explanations (Haines et al., 2020). For example, one can theorize that older adults have difficulty with associative memory, but it is unclear how associations are learned or retrieved, or even what aspects of an experience are associated together. Computational modeling, however, forces the researcher to make their theory mathematically transparent, without relying on language that could be interpreted in different ways, and to consider the specific implementation of proposed mechanisms.

For example, a number of computational models have instantiated specific mechanisms that could underlie episodic memory (for a review, see Sederberg & Darby, under review). One model that has proven especially successful at explaining episodic memory phenomena is the temporal context model (TCM; Howard & Kahana, 2002; Sederberg, Gershman, Polyn, & Norman, 2011; Sederberg, Howard, & Kahana, 2008). In TCM, temporal context is defined as a recency-weighted representation of experience, and is instantiated as a vector of feature activations that change across time. This contextual change is driven primarily by items that are presented to participants in the memory task, such that when an item is presented it becomes strongly activated in context, whereas the activations of less recently presented items decay. Critically, associative encoding in TCM occurs through binding each item to the state of temporal context at the time when the item is presented, and retrieval is driven by reinstating past states of context that have been associated with tested items.

These processes are mathematically defined according to the equations that make up the model. Computational models such as TCM, then, can provide a rich and transparent theory of memory processes such as recall and recognition. In the current work, we apply TCM to better understand age-related changes in episodic memory processes.

In addition to transparency and level of mechanistic detail, another difference between verbal and computational models is that the former typically address a proposed process, such as associative memory, in isolation, without considering other processes that could affect performance. However, no task is process-pure, and performance will necessarily be the result of interactions between multiple mechanisms, such as memory, perceptuomotor, and decision-making processes, all of which can vary between individuals and even within individuals across time or development. These additional sources of variability decrease our ability to make mechanistic inferences: if both memory and decision-making processes contribute to accuracy on a given task, for example, it may be difficult to assess the extent to which an age difference is due to changes in memory, decision-making, or both. Computational models, by contrast, decompose task performance into proposed latent mechanisms that are typically governed by model parameters that may be allowed to vary between individuals and/or groups. This is a particularly useful feature for research on aging: by comparing parameter values, the researcher can gain insight into specific processes that may be expected to differ (or not) between age groups, while accounting for other mechanisms with separate parameters. As an example, Howard, Kahana, & Wingfield (2006) applied TCM to examine age differences in free recall and found evidence that (1) older adults were less able to form new associations between items and temporal context, and that (2) older adults' associations were noisier, with a greater tendency to include irrelevant information in the associations that they did form. By contrast, they did not find evidence that older adults differed in item-context associations that had been formed prior to the experiment, analogous to pre-existing semantic knowledge.

Another benefit of generative computational models is that they can be employed to simulate data, such as trial-level choices and RTs, which can be compared to observed data. While model-based simulations that closely match the pattern of observed data can provide evidence in favor of a theory's legitimacy, a poor fit provides evidence that a theory, as implemented in the model, should be revised or even abandoned (Guest & Martin, 2021; Oberauer & Lewandowsky, 2019a). Crucially, model-generated data can be examined to quantitatively adjudicate between theories through formal model comparison. This process consists of comparing how well two or more models quantitatively fit the same observed set of data, while accounting for potential differences in model complexity (Myung, 2000; Pitt, Myung, & Zhang, 2002). An example of this process in the cognitive aging literature was provided by Healey & Kahana (2016), who applied a variant of TCM (Lohnas, Polyn, & Kahana, 2015; Polyn, Norman, & Kahana, 2009) to develop a larger theory of aging deficits in episodic memory. These authors systematically compared models that instantiated differences between young and older adults in inhibition, attention, associative binding, and processing speed, and found that age differences in a subset of model parameters related to all four of these processes were necessary and jointly sufficient to account for aging effects on free recall performance. Formal comparison of computational models is relatively

straightforward, but it is more difficult to compare different verbal theories, which lack quantitative ways of assessing model fit, prediction, and complexity.

Model-generated data can also be employed to make formal, quantitative hypotheses for new experiments. Simulating data with a model could serve as a powerful basis for pre-registering a future experiment, as this would not only allow for qualitative predictions (e.g., older adults are expected to perform worse than young adults in Condition A, but not Condition B), but also quantitative predictions (e.g., the expected magnitude of performance measures in each condition and age group) before any data are collected. This process could also assist the researcher in creating an experimental design with the goal of adjudicating between theories that predict maximally different patterns of performance. The ability to quantitatively simulate data based on a theory is a tool that is unavailable to researchers relying purely on verbal theories.

Current work

Despite notable exceptions, the great majority of work on cognitive aging has relied on verbal theories and standard inferential statistical models (Benjamin, 2010). In the current work, we provide an example of a generative model to investigate episodic memory in young and older adults in an effort to demonstrate various ways a model-based approach may be beneficial to research on cognitive aging.

The model we apply is a variant of TCM. We chose TCM because it is a well-established model of episodic memory that is able to provide a principled and mechanistic explanation of many behavioral findings (Lohnas, Polyn, & Kahana, 2015; Sederberg, Gershman, Polyn, & Norman, 2011; Sederberg, Howard, & Kahana, 2008). In addition, two studies have applied TCM and related models to provide compelling explanations for deficits in episodic memory in older adults, as described above (Healey & Kahana, 2016; Howard, Kahana, & Wingfield, 2006). We expand on prior work with this model in important ways. First, TCM has been primarily applied as a model of free recall, including in the work on aging. We extend the model to account for a different episodic memory task – associative recognition – which has been one of the most widely used paradigms in research on aging deficits in episodic memory (Castel & Craik, 2003; Chen & Naveh-Benjamin, 2012; Greene & Naveh-Benjamin, 2020; Light, Patterson, Chung, & Healy, 2004; e.g., Naveh-Benjamin, 2000; Naveh-Benjamin, Guez, Kilb, & Reedy, 2004; Overman & Becker, 2009; Ratcliff & McKoon, 2015; Stephens & Overman, 2018).

We also extend TCM to implement a new learning mechanism. Most work with TCM has instantiated what is known as Hebbian learning (Hebb, 1949), by which a simple association is formed between the presented item and current state of context to the extent that item and context features are co-active. However, Hebbian associations can grow without bound when items are presented multiple times, as items become associated with more and more contextual features (see Gershman, Moore, Todd, Norman, & Sederberg, 2012, for discussion). Because the associative recognition task we created contains multiple repetitions of items, as we describe below, we employ a different learning mechanism in which associations are formed based on prediction errors. Specifically, we examine the

possibility that the memory system makes predictions about what stimuli will be presented next based on learned associations with temporal context, and that learning occurs to the extent of a mismatch between observed and predicted stimuli. Recent work has found evidence of prediction error signals during human retrieval of episodic memories (Haque, Inati, Levey, & Zaghloul, 2020), and it has been suggested that changes in prediction-driven learning may play an important role in age-related memory decline (Ofen & Shing, 2013). In addition, work on prediction error-based reinforcement learning has suggested neural asymmetries between positive prediction errors, which occur when an unexpected event takes place, and negative prediction errors, which occur when expected events do *not* take place (Cavanagh, Frank, Klein, & Allen, 2010). In the current work, we explore whether asymmetries may occur between the magnitude of positive and negative prediction error learning in young and older adults.

More broadly, the model makes our theoretical account of the mechanisms underlying associative recognition mathematically transparent. We fit the model with hierarchical Bayesian methods, allowing estimation of group-level age differences in latent cognitive processes, as well as differences in these processes between individual participants. We also provide an example of employing model comparison techniques to test different mechanistic explanations, and an example of generating data to make predictions for future experimental manipulations. We emphasize that these examples are not meant to provide an exhaustive search for the best model of associative recognition or aging differences therein. Our goal is to provide an example of what a generative model looks like and a roadmap of some ways that such models may be applied to benefit research on cognitive aging. After presenting these examples, we close by discussing implications of this work for episodic memory and aging research in relation to transparency, replicability, and discovery.

Method

Participants

Eighty-two young adults (M_{age} = 19.8 years, SD_{age} = 1.7, $range_{age}$ = 18 – 29) and 52 healthy older adults (M_{age} = 71.5 years, SD_{age} = 5.1, $range_{age}$ = 63 – 83) participated in the experiment. Four additional young adults participated but were excluded from the analysis because of failure to record their age. The young adults were recruited via flyers at The University of Virginia, and the older adults were recruited via phone and email from the existing participant pool of the Virginia Cognitive Aging Project (VCAP). Young adults received $10/hr for participating, and older adults received $50 per session, which included an MRI scan and other neuropsychological tests unrelated to the associative recognition task we focus on below. This project was approved by the Institutional Review Board of the University of Virginia.

Stimuli

Participants were presented with images of real-world objects. These images were part of a “massive memory” database of 2400 images (Brady, Konkle, Alvarez, & Oliva, 2008), which is publicly available (http://olivalab.mit.edu/MM/uniqueObjects.html). The collection of images was pruned to exclude human faces, text, and images deemed inappropriate,

including weapons and religious symbols; 1996 images were included in the stimulus set. A total of 24 different objects were presented to the participant in each block of the task; the images presented in a block were selected randomly, with the constraint that images could not be seen in more than one block for any participant, and could not be presented in multiple sessions. Objects were presented to participants side-by-side on white squares located at the center of the screen with a grey background.

Design and Procedure

To probe episodic memory, we developed a variant of an associative recognition task. Typically, associative recognition tasks include a study phase, in which pairs of items are encoded, and a separate test phase, in which some pairs are presented intact, and others are recombined into new pairs (Castel & Craik, 2003; Cox & Criss, 2020; Craik, Luo, & Sakuta, 2010; Gallo, Sullivan, Daffner, Schacter, & Budson, 2004; Greene & Naveh-Benjamin, 2020; Hockley & Consoli, 1999; Naveh-Benjamin, 2000; Ratcliff & McKoon, 2015). In our variant, participants were asked to respond “old” or “new” to every presentation of a pair in a continuous stream of trials, without separate study and test phases (see Chen & Naveh-Benjamin, 2012; Hockley, 1992 for other continuous variants of associative recognition). All items were presented in both intact and recombined pairs. In addition, intact pairs were presented multiple times, before and after the items were recombined, allowing for analysis of repetition and interference effects between different pairings, and for assessment of how well a prediction error-based learning mechanism could account for these effects. We refer to this paradigm as the continuous associative recognition (CAR) task.

On every trial of the CAR task, participants were shown a pair of object images, and were asked to respond whether the pair was “new” or “old.” A pair was considered new either if the objects were novel, or if they were repeated, but in a novel pairing; a pair was old only if the objects had been presented previously in the same pairing. There were four types of pairs: *New* (the first presentation of two new objects), *Intact 1* (the first repetition of the same pair), *Intact 2* (the second repetition of the pair), and *Recombined* (a new pair made up of objects from two different previously presented pairs). Each object was presented four times, once for each pair type, such that each object was repeated in both intact and recombined pairs. See Figure 1 for a visualization of each pair type.

As in other associative recognition tasks, the most critical manipulation in this task is the distinction between Recombined and repeated Intact pairs. In both cases, the presented items have been seen previously, and so are familiar to the participant. Despite this, the participant’s task is to determine whether the two familiar items were previously presented in the same pair, as in Intact 1 and Intact 2 pairs, or in different pairs, as in Recombined pairs. This distinction requires associative memory for the specific pairings of objects.

The point at which a Recombined pair was presented varied across three within-subject conditions (Table 1). In the *Weak* condition, recombined objects had each been seen once (i.e., in New pairs), whereas in the *Medium* and *Strong* conditions, recombined objects had been seen two or three times, respectively, in the initial pairings. The strength manipulation was inspired by prior work suggesting that young adults may make fewer false alarms to Recombined pairs following multiple repetitions of the original pairings, whereas older

adults may make comparable or even greater numbers of false alarms under those conditions (Gallo, Sullivan, Daffner, Schacter, & Budson, 2004; Light, Patterson, Chung, & Healy, 2004; Overman & Becker, 2009). Thus, we expected that false alarms would decrease from the Weak to the Strong condition in young adults, but not older adults. We were also interested in how the recombined pairs would affect memory for repeated pairs. Given that items that were recombined are later seen again in their initial intact pairings in the Weak and Medium conditions, we expected that memory for Intact 1 pairs in the Weak condition and Intact 2 pairs in the Medium condition would be reduced due to interference from the Recombined pairs (Darby & Sloutsky, 2015; Postman & Underwood, 1973).

On each trial, a pair of objects was presented for 2.5 seconds. If the participant responded before this amount of time, a black rectangle appeared behind both objects to indicate that a response had been made. Whether or not a participant made a response in 2.5 seconds, the objects disappeared after that time before the start of the next trial. The screen was empty during the interstimulus interval, which was jittered between 0.5 and 1.0 s.

Young adults completed between two and four blocks of the CAR task in a single session (most participants completed four, with a few exceptions due to computer or experimenter error). Older adults completed between two and five blocks across either one or two sessions (most completed two blocks per session). Each block of the task contained 48 trials (four trials for each of the 12 conditions listed in Table 1), and lasted approximately 2.5 minutes. In addition to the CAR task, all participants completed other tasks from a larger cognitive battery (Weichart et al., 2021), which we do not address in the current manuscript. At least one intervening task was presented between each block of the CAR task. In addition, the objects presented in the CAR task did not repeat across sessions or blocks, and in all analyses below we make the simplifying assumption that performance in each block was independent.

Results

To avoid including data from participants who did not understand or were unable to perform the CAR task, or were not paying attention, we assessed participants' performance in each block separately. We excluded blocks in which overall accuracy was not above chance (50%). We also excluded blocks in which responses were severely biased toward either "new" or "old" responses; to do so, we required that the percentage of "old" responses be significantly below 90% and significantly above 10%, as determined by one-tailed binomial tests. These criteria resulted in exclusion of 0.30% of data from young adults, and 4.25% of data from older adults. All participants performed at least one block that qualified for analysis (M_{young} = 3.9 blocks; M_{older} = 3.5 blocks).

In addition, we excluded individual trials from the analysis that (1) we deemed too fast to be deliberate (i.e., faster than 0.35 s), which resulted in exclusion of 1.0% of trials for young adults, and 0.3% of trials for older adults, or (2) were outliers based on RTs. Outliers were detected for each participant separately by first applying a Box-Cox transformation to all RTs for that participant, in order to approximate a normal distribution. Then, the mean and standard deviation of the transformed distribution were calculated, and any trials with

a response time (RT) greater than three standard deviations away from the mean in either direction were discarded. This process was repeated until no outliers were detected for that participant. This resulted in exclusion of 0.1% of the remaining trials for young adults, and 0.1% of trials for older adults.

We took a two-pronged approach to analyze the data. We performed conventional regression models on the responses and RTs, as well as a generative computational model. For both approaches, we fit the models to the data with hierarchical Bayesian techniques, which simultaneously estimated group-level hyper-parameters and individual participant-level parameters. To examine potential differences between age groups, we focused our analyses on the group-level hyper-parameters. For each hyper-parameter, we calculated the 95% highest posterior density (HPD) of the posterior distribution, reflecting the most probable parameter values. To assess age differences in parameters, we applied an index of distributional similarity based on probability density function estimates of the hyper-parameter posteriors (Pastore & Calcagnì, 2019). This index, $\hat{\eta}$, may be thought of as the proportion of overlapping regions compared to the total area of the two distributions, such that $\hat{\eta} = 1$ would indicate identical distributions, and $\hat{\eta} = 0$ would indicate completely non-overlapping distributions. Although this is a continuous measure, for ease of exposition we consider $\hat{\eta} < .05$ to constitute strong evidence of a difference in parameters between age groups.

Measures of choices (hits and false alarms) and RTs for each task condition in young and older adults are presented in Figure 2. To assess performance with standard statistics, we performed a series of hierarchical Bayesian regression models. The methods and results of these analyses are presented in the Supplementary Material. Overall, the results suggest that older adults' responses in the CAR task were generally slower and less accurate, while replicating previous work demonstrating older adults' greater difficulty discriminating between Intact and Recombined pairs (Castel & Craik, 2003; Naveh-Benjamin, 2000). We also found evidence of a repetition effect resulting in higher hit rates to Intact 2 compared to Intact 1 pairs, in both young and older adults, as well as interference effects in both age groups resulting in lower hit rates for Intact pairs after they were recombined (specifically, Weak Intact 1 and Medium Intact 2 pairs). However, standard statistical estimation of these effects provide limited insights into the cognitive processes underlying task performance and how these processes may differ due to age. Many mechanistic questions remain unanswered, such as, what information is associated together, how does this learning transpire, how do item repetitions affect encoding processes, and how does age affect these mechanisms? To provide more transparent and mathematically explicit hypotheses about these latent processes, we developed a computational model.

Computational model

We now provide a conceptual overview of the mechanisms of the model, the mathematical details of which are presented in the Supplemental Material. Our model is a variant of TCM (Howard & Kahana, 2002; Sederberg, Gershman, Polyn, & Norman, 2011; Sederberg, Howard, & Kahana, 2008), in which memory retrieval is tied to states of temporal context. Temporal context is primarily made up of features corresponding to the items presented

during the CAR task, which become activated when an item is initially presented in the task. This activation then decays as other items are presented, such that context changes across the course of the task. The extent to which the current context decays when a pair of items is presented is modulated by a change rate parameter ρ, as well as the extent to which the items are already activated in context (see the Supplemental Material for details).

Associative learning takes place by binding the pair of items to the state of context when the pair was presented; these associations are stored in a matrix $\boldsymbol{M}$. Item-context binding occurs via prediction error learning, a process thought to play a key role in episodic memory (Bar, 2009; Mizumori, 2013), and aging-related changes in memory and other cognitive processes (Federmeier, Kutas, & Schul, 2010; Ofen & Shing, 2013; Samanez-Larkin, Worthy, Mata, McClure, & Knutson, 2014). The idea is that participants predict what items are likely to be presented on each trial based on past learning, and associations are modified to the extent that these predictions are incorrect. Specifically, the model uses $\boldsymbol{M}$ to make a prediction of what items it expects based on the current state of context. These predictions are then compared to the items that are actually presented on that trial. Positive prediction error learning increases the association between current context and the presented items to the extent that these items could not be predicted by the model. By contrast, negative prediction error learning *decreases* the association between current context and items that were predicted but not actually presented. A free parameter α controls the extent of prediction error learning overall. A second parameter κ controls the extent of negative prediction error learning as a proportion of α, such that if $\kappa = 1$ positive and negative prediction error learning are symmetric, whereas $\kappa < 1$ would indicate greater positive prediction error learning, and $\kappa > 1$ would indicate greater negative prediction error learning.

The model assesses the activation of items in context and the item-context associations stored in $\boldsymbol{M}$ to estimate the "strengths" of memory supporting the new versus old choice for each presented pair of items. Some strength supporting an "old" response is provided by familiarity with each item, estimated as the extent of the item's activation in the current state of temporal context. The familiarity signal is constrained by a parameter λ controlling the maximal or asymptotic level of familiarity, as well as a parameter τ controlling the sensitivity of familiarity strength to differences in item activations in context due to factors like recency. The model also probes associations learned in $\boldsymbol{M}$ by retrieving the states of context previously bound to each of the two objects. The extent of *match* (or overlap) of the two retrieved contexts provides evidence that the pair is old. In addition, the *mismatch* (or difference) between the contexts provides evidence that the pair is new. These match and mismatch mechanisms are analogous to the idea of recall-to-accept and recall-to-reject processes in associative recognition (Rotello & Heit, 2000). Recall-to-accept has been proposed as a process that contributes evidence that a pair is old when retrieved associations match the presented pairing, whereas a recall-to-reject process has been proposed to contribute evidence that a pair is "new" when the presented pairing is different from the associations stored in memory. In the current model, the extent of overlap between the *contexts* retrieved from the presented items provides evidence for "old," whereas the extent to which they are different provides evidence for "new." Given suggestions that older adults, and especially those with Alzheimer's disease, may be less

proficient at recall-to-reject processes (Cohn, Emrich, & Moscovitch, 2008; Gallo, Sullivan, Daffner, Schacter, & Budson, 2004), we allowed the context mismatch strengths to be scaled by a free parameter γ. Finally, the model estimates a baseline level of novelty-based strength supporting a "new" response with a free parameter ν.

For every trial of the experiment, these sources of strength (familiarity, match, mismatch, and baseline novelty strength) are combined into an overall memory strength, which is then passed to a decision-making component that simulates a response ("new" or "old") and RT for that trial. This decision-making component is instantiated as a sequential sampling model (Navarro & Fuss, 2009; Ratcliff & McKoon, 2008; Stone, 1960), in which evidence accumulates across time toward a threshold for a "new" or "old" response, determined by a parameter a, whereas a potential bias toward "new" or "old" responses is estimated by a parameter w. Whether the accumulated evidence reaches the threshold for "new" or "old" first determines which response is made by the model, and the time required to reach one of these thresholds determines the decision time. This decision time is added to an estimate of time required for perceptuomotor processes unrelated to the decision itself, called non-decision time, determined by parameter t_0. Sequential sampling models have been applied in a number of studies of cognitive aging, and typically provide evidence of a higher threshold for evidence accumulation in older adults (i.e., a in the current model), as well as longer non-decision times (t_0; see Theisen, Lerche, von Krause, & Voss, 2020 for a meta-analysis). A schematic of the processes in our model is presented in Figure 3; we provide the mathematical implementation of these processes in the Supplemental Material.

Bayesian model-fitting.—Most computational models, including ours, involve free parameters that must be "fit" to observed data, which is akin to turning knobs to reduce or heighten effects of different cognitive mechanisms. One popular method for fitting the parameters of a model is maximum likelihood estimation (Myung, 2003), which attempts to find the single set of parameter values that best fit a set of observed data. However, this approach does not allow for assessment of uncertainty, as each parameter is estimated as a single value. Bayesian methodologies, by contrast, enable estimation of entire posterior distributions of parameter values that could have generated the observed data. This allows for proper assessment of uncertainty, which is an invaluable feature because it provides information about how confident researchers should be in the parameter estimates. We therefore chose to fit our model to the observed data with Bayesian techniques implemented within the RunDEMC Python library (https://github.com/compmem/RunDEMC). By fitting the model hierarchically we were able to estimate age group-level parameter distributions while accounting for participant-level variability within each group.

We fit the model to all trial-level choices and RTs simultaneously. The model contained a total of ten free parameters, which are summarized in Table 2. We fit all of the model parameters to each participant's data hierarchically, with the exception of t_0, which we fit independently for each participant, as this parameter was constrained by each participant's minimum RT. The other subject-level parameters were constrained by hyper-parameters governing the mean and standard deviation across participants within each age group. To assess age differences in latent mechanisms we compared the mean hyper-parameters

between age groups. See the Supplemental Material for details on model priors and hyper-priors for each parameter.

Computational model fit.—To assess model fit, we first computed the maximum a posteriori, or MAP, estimates of each parameter, for each participant. We employed these MAP estimates to generate choices and RTs for every trial for every participant. If the model is able to capture the latent cognitive processes underlying participants' performance on the CAR task, there should be a close correspondence between model-predicted and observed performance.

Figure 2 shows a close qualitative fit between the choices and RTs in each condition of the CAR task. We also quantitatively assessed model fit in terms of its ability to capture individuals' overall performance, taking into account both accuracy and RTs. To assess both aspects of performance simultaneously, we calculated each participant's rate correct score, or RCS (Woltz & Was, 2006), which is calculated as:

$$RCS = \frac{c}{\sum RT},$$

where c is the total number of correct responses in the task, which is divided by the sum of all RTs (in seconds). This metric may be interpreted as the number of correct responses per second of effort, and has performed favorably in studies comparing different speed-accuracy measures (Vandierendonck, 2017). To assess how well the model could account for participants' overall performance, we calculated each participant's overall RCS for both observed and model-generated performance. Figure 4 shows the correlations between observed and model-predicted RCS values for young and older adults ($rs >= .93$). We also calculated d' for the observed as well as model-generated performance: $d' = z(H) - z(F)$, where H is the hit rate averaged across all Intact 1 and Intact 2 pairs, F is the false alarm rate across all Recombined pairs, and $z(.)$ designates the inverse of the Gaussian cumulative distribution function. This metric estimates the ability of the participant to discriminate "old" responses between repeated Intact and Recombined pairs. Similar to RCS, there was high correspondence between the observed and model-predicted d' scores ($rs >= .94$). We conclude that the model was able to capture overall performance on the CAR task very well for both age groups.

Given that the model provides an adequate fit to the data, how can a generative computational model provide greater insights into effects of aging, while supporting transparency, replicability, and scientific discovery? We argue that aging research could particularly benefit from three aspects of modeling: (1) examining hierarchical hyper-parameter distributions to gain insight into cognitive mechanisms affected by aging, (2) formally comparing different hypothesized mechanisms via model comparison techniques, and (3) generating data independent from the observed data used to fit the model in order to either validate the model or make formal hypotheses for future experiments. We provide examples of these approaches below.

Age group comparisons.—A central goal of this work was to decompose observed performance into latent cognitive mechanisms via model parameters. With a hierarchical model-fitting approach, we could compare the posterior hyper-parameter distributions to examine differences in hypothesized latent processes between age groups. We expected to find the strongest evidence of age differences in parameters governing learning of associations between items and context. Figure 5 presents the posterior distributions of the hyper-parameters governing the mean parameter values for each age group, along with $\hat{\eta}$, a measure of the overlap between distributions. There was strong evidence of age differences for a number of parameters. First, older adults showed reduced item–context associative learning, as estimated by the a parameter, $\hat{\eta} < .001$, which is consistent with prior studies applying TCM to examine age differences in free recall (Healey & Kahana, 2016; Howard, Kahana, & Wingfield, 2006). Interestingly, although young adults' item-context binding was stronger overall, the parameter controlling negative prediction error learning, κ, was higher in older adults than young adults ($\hat{\eta} = .042$). This suggests that older adults' learning was characterized more by forgetting associations between context and items that were incorrectly predicted based on prior learning compared to young adults' learning. This age difference was unexpected, and should be interpreted with caution, although it does suggest an interesting potential mechanistic source of age differences between young and older adults. We also found evidence that young adults were more sensitive to the mismatch between retrieved contexts, as estimated with the parameter γ ($\hat{\eta} = .005$), which may be thought of as a recall-to-reject mechanism (Rotello & Heit, 2000), as discussed above. Therefore, these results support prior work hypothesizing that older adults may have a deficit in recall-to-reject processes (Cohn, Emrich, & Moscovitch, 2008).

Additionally, we found evidence that memory strengths in general were higher in young adults, including the maximum level of familiarity, λ ($\hat{\eta} < .001$), and the baseline novelty strength supporting a "new" response, ν ($\hat{\eta} < .001$). Consistent with prior work (Theisen, Lerche, von Krause, & Voss, 2020), we also found evidence of a higher decision threshold in older adults, a, $\hat{\eta} = .028$, suggesting that older adults required more memory-driven evidence to make a decision than did young adults. Despite these differences between age groups, we found little evidence of differences in rate of contextual change, estimated by parameter ρ ($\hat{\eta} = .372$), differences in sensitivity to variations in familiarity (e.g., due to recency), estimated by τ ($\hat{\eta} = .189$), or differences in decision bias, estimated by w($\hat{\eta} = .589$). Because we did not fit the t_0 parameter hierarchically, instead allowing the parameter to vary independently between all participants, we do not make inferences about age differences in non-decision time. Overall, these results point to a number of age-related differences in cognitive processes, including differences in associative learning, with weaker overall item-context binding, but proportionally stronger negative prediction error learning, in older adults. At the same time, the model results point to processes that were comparable between age groups, including context change rate, changes in familiarity due to recency, and old-new decision bias.

The memory and decision-making processes these parameters represent provide evidence of mechanistically interpretable differences between young and older adults. By contrast, the regression models presented in the Supplemental Material provide evidence of differences

in performance between the age groups, but require additional inferences on the part of the researcher to provide mechanistic insights. The insights that we might gain from those statistics, such that older adults are generally slower and less accurate, and even that older adults have greater difficulty with associative memory (as could be inferred from differences in the d' statistic between Intact and Overlapping pairs), provide very little insight into *why* these effects were found.

Individual participant comparisons.—It is generally very difficult to make inferences about differences between individuals' task performance because conventional metrics such as task accuracy are typically summarized by a single number, with no ability to assess uncertainty. However, Bayesian approaches allow us to gain insight into the uncertainty of model estimates through the posterior distributions of parameters. As a result, we can make comparisons between parameter estimates for individuals and make inferences about differences between participants in the mechanisms generating the observed behaviors.

Recall that we can generate data with our model, given a set of parameter values. Because we have entire posterior distributions for the parameters, we can generate data using many parameter estimates drawn from the posterior distributions for each participant, creating a distribution of predicted overall performance scores, called a posterior predictive distribution (PPD). The PPD estimates how the participant would be expected to perform given the model and uncertainty in the parameter estimates, were the participant to perform the identical task many times. In the upper-left (boxed) plot of Figure 6, we present the observed d' as dots, along with the corresponding PPDs as split-violin plots, for four sample participants: an older adult and a young adult with low overall performance, and a different older adult and young adult with high performance. Within each of these dyads, the two participants performed similarly, with PPDs almost entirely overlapping, whereas across the two dyads, performance was quite different, with little overlap between PPDs. Critically, we can examine the posterior distributions of each model parameter to estimate the cognitive processes that may have differed (or not differed) between individuals.

Despite the very similar performance within each low- and high-performing dyad, clear differences emerge in model parameters. Values of the item-context association (α) parameter were consistently higher in the high-performing dyad than the low-performing dyad. In these particular participants, values of the maximum item familiarity (λ) parameters tended to be higher in the two young participants than the two older participants, whereas the the decision threshold (a) and non-decision time (t_0) tended to be higher in the older participants, despite very similar overall performance within the two participant dyads. This example illustrates the power of computational models coupled with Bayesian approaches to make inferences about latent processes, even in individual participants. This approach could be particularly useful in case studies or clinical settings in which cognitive mechanisms within the individual are of interest, or could be applied to track longitudinal changes within individuals as they age, identifying new cognitive aging phenotypes that summary statistics are not sensitive enough to detect.

Model comparisons.—We have presented a model that is able to account for age differences in performance on the CAR task, but there could be other models that account

for these differences as well or better by applying different mechanisms than the model just described. Theoretical progress through computational modeling depends to a large degree on a process of model comparison. In what follows, we describe two variants of the prediction error learning mechanisms that could be reasonably supposed to underlie performance on this task.

Associative learning in our variant of TCM involves prediction error learning, as described above. This process includes positive prediction error learning, in which the presented items are bound to the current context to the extent that those items could not be predicted from associations already formed with features in context. Perhaps a more controversial aspect of this learning is negative prediction error learning, in which features that were predicted, but are not presented in the given pair of items, are *unbound* from features to the extent they are active in context. This may be considered a mechanism for unlearning, a concept that is often maligned in theories of memory, which often prefer to ascribe forgetting of learned information to interference from other learned information, rather than unlearning of the information itself (Slamecka, 1966).

One possibility is that negative prediction error learning is not necessary: perhaps participants learn based only on positive prediction errors, without the unlearning process based on negative prediction errors. To assess this possibility we simply fixed the negative prediction error learning rate parameter κ to zero and otherwise left the model intact. A second possibility is that negative prediction error learning does occur, but that positive and negative prediction error learning are symmetric, such that it is unnecessary to allow for a second learning rate (κ) in our model. We assessed this possibility by fixing κ to 1.

Figure S4 presents the fit of these alternative models to the observed data. The model fixing κ to 0 over-predicted false alarms to Recombined pairs, and predicted higher false alarm rates for higher strength conditions. Inspection of the model-generated sources of memory strength indicated that without negative prediction error learning, there was a strong retrieved context match signal for recombined pairs, which increased with more repetitions of the intact pairings. In addition, the modified model was unable to sufficiently capture interference effects (i.e., the drop in performance for Weak Intact 1 compared to the other Intact 1 conditions and the drop for Medium compared to Weak and Strong Intact 2 pairs), suggesting that negative prediction error learning was necessary in order to account for these effects. This was somewhat surprising, as we expected interference between non-matching contexts due to recombining the pairs would likely be sufficient to explain the drop in performance. Although fixing κ to zero resulted in a qualitatively poor fit, the model fixing κ to 1 qualitatively fit the data relatively well, as shown in Figure S4.

To assess the fit of these models more quantitatively, we calculated the Bayesian predictive information criterion (BPIC), a metric of model fit that accounts for differences in the number of parameters, for each model and each participant. Models with lower BPIC values are preferred. For each participant, we mean-centered the BPIC values for the three models, including the full model, and found that the centered BPIC values were on average lowest for the full model in young adults ($M_{BPIC} = -12.7$), with the full model preferred for 69.5% of participants, followed by the model with κ set to 1 ($M_{BPIC} = -9.1$), which best

accounted for the performance of 26.8% of participants. Interestingly, however, the same was not true of older adults, 92.3% of whom were best fit by the model with κ set to 1 (M_{BPIC} = −8.8), with only 1.9% of participants best fit by the full model (M_{BPIC} = −4.6). The model without negative prediction error learning, such that κ was set to zero, performed poorly for both young (M_{BPIC} = 21.8), and older adults (M_{BPIC} = 13.4). Overall, these results suggest that negative prediction error learning was necessary to fit the data, given the other mechanisms of the model. For young adults, prediction error learning was asymmetric, such that positive prediction error learning was stronger than negative prediction error learning, whereas in older adults the two learning processes were symmetric. This pattern of differential development of positive and negative prediction error learning would be very difficult to conclude without a generative mechanistic model, although we emphasize the possibility that other model frameworks or other variants of TCM could fit the data as well or better than the models we have presented.

Novel model-generated hypotheses.—An important advantage of a generative model is that it allows one to make precise quantitative predictions about independent data. This can be an important way to validate the model, if a model is fit to one set of data and then is able to accurately account for performance on a different set of data, particularly if the independent data is from a different task (for an example, see Healey and Kahana's, 2016, demonstration of fitting a model to free recall data in young and older adults and accurately predicting the same individuals' performance on a recognition task). This process can also be applied to generate formal hypotheses for future experiments. In other words, by generating data for a new experiment, one is able to qualitatively and quantitatively hypothesize what the results will be, according to the theory implemented in the computational model, which one could use to design a task that is better able to distinguish between alternative models.

We illustrate this process by examining model predictions for a hypothetical experiment, in which Recombined pairs are repeated a second time. In this case, the first time a Recombined pair is seen, the correct response is "new," whereas upon the second presentation the correct response is "old." To simplify the design, we removed the second repetition of intact pairs (i.e., Intact 2 pairs), and all trials in the Strong condition (see Table 3). To examine predictions of the model for this design, we created 10 lists of item pairs according to the design, and simulated data for each of these lists according to the best-fitting model parameters for every young and older participant in our sample.

The model-generated predictions for young and older adults for this hypothetical experiment are presented in Figure 7. Several general observations may be made. First, the model was more likely to identify the Recombined pairs as "old" the second time they were observed, as we expected, since these pairs were indeed repeated. However, the model was less likely to correctly identify a repeated Recombined pair as "old" compared to a repeated Intact pair, especially in the Medium condition. This is because the Intact pairs were always observed at least once (as New pairs) before the Recombined pairs, such that contexts reinstated from Recombined pairs, repeated or not, would be partially mismatching, and therefore more likely to be judged as "new." In addition, the model predicts an effect of the strength conditions, such that hit rates are lowest for repeated Recombined 2 pairs in the Medium

condition, at least in young adults, because the Intact pairs were associated with more states of temporal context prior to being recombined, which generates a strong mismatch signal. Finally, the model predicts a very strong interference effect for Weak Intact pairs, compared to Medium Intact pairs, especially for young adults. This is because at the time these pairs are presented, the Intact pair would have only been seen once, whereas the Recombined pairs would have been seen twice, generating strong interference. The model predicts greater interference for these pairs in young adults, due to the higher values of context mismatch sensitivity parameter γ. These model-generated data are, of course, only predictions, and it remains to be seen whether performing the experiment would result in the predicted patterns.

Discussion

In this work, we have provided examples of generative computational modeling approaches that we believe could facilitate greater transparency, replicability, and discovery in cognitive aging research. To do so, we applied a model to decompose associative recognition performance of young and older adults into latent mechanisms of episodic memory and decision-making. This approach made our hypotheses and assumptions mathematically transparent, with no need for other researchers to make subjective inferences about what is meant by a verbal theory's explanation of the mechanisms proposed to underlie performance. We fit the model with hierarchical Bayesian techniques, allowing examination of differences in parameters between age groups and individual participants, such that we could make inferences about latent cognitive processes underlying performance and how they may be affected by aging. We also compared the model with variants implementing alternative assumptions about associative learning processes. Finally, we demonstrated the application of a generative model to simulate data for a hypothetical new experiment, making quantitative predictions of future data we would expect to observe from young and older adults. In what follows, we discuss these approaches and how a generative computational modeling framework could increase transparency, replicability, and discovery in cognitive aging research.

We have argued that a computational model provides greater transparency and mechanistic specificity to theory than is typically provided by purely verbal accounts. The model we applied as an example is a variant of TCM, a well-established model of episodic memory that instantiates encoding and retrieval mechanisms that are centered around a representation of context that varies across time as items are presented in a memory task. These mechanisms are defined mathematically and controlled by parameters, allowing for estimation of how latent processes may differ between individuals and age groups. Therefore, the model goes beyond demonstrating that young and older adults differ in episodic memory (as has been demonstrated many times, e.g. Castel & Craik, 2003; Chen & Naveh-Benjamin, 2012; Gallo, Sullivan, Daffner, Schacter, & Budson, 2004; Golomb, Peelle, Addis, Kahana, & Wingfield, 2008; Healey & Kahana, 2016; Howard, Kahana, & Wingfield, 2006; Li, Naveh-Benjamin, & Lindenberger, 2005; Naveh-Benjamin, 2000; Naveh-Benjamin, Guez, Kilb, & Reedy, 2004; Oberauer & Lewandowsky, 2019b; Old & Naveh-Benjamin, 2008; Ratcliff & McKoon, 2015), and provides a mechanistic explanation of *why* they differ.

The model's explanation of why young and older adults' performance differs on episodic memory, as measured by the CAR task, hinges on contextually mediated associative learning and retrieval. In our TCM-based model, associations are not formed directly between items within a pair, but between individual items and the state of context when the items are presented. However, because each item was presented multiple times over the course of the CAR task, we did not instantiate Hebbian associations between items and context, as in a number of prior versions of TCM, as this could result in unbounded associative strengths as items are repeated (see Gershman, Moore, Todd, Norman, & Sederberg, 2012, for discussion). In the current work, we explored a different mechanism: prediction error learning, which only modifies associations to the extent of discrepancies between what items the model predicted based on the current context and the actually presented items. Others have suggested that predictions and prediction errors play important roles in learning and memory (Haque, Inati, Levey, & Zaghloul, 2020), and researchers have suggested that aging differences in prediction error learning could play a role in changes in memory (Ofen & Shing, 2013), as well as processes like reinforcement learning (Samanez-Larkin, Worthy, Mata, McClure, & Knutson, 2014) and language processing (Federmeier, Kutas, & Schul, 2010).

The current work explored effects of aging on both positive and negative prediction error learning, investigations that would not have been straightforward without a computational model. When we allowed parameters for both kinds of learning to vary, we found evidence of stronger positive prediction error learning in young adults than in aging adults. Interestingly, however, we also found that older adults exhibited evidence of proportionally greater negative prediction error learning. This suggests that while young adults were better able to learn new associations based on the presence of unexpected items, older adults were more likely to *unlearn* associations between contextual features and items that were predicted but not presented. In addition, through formal model comparison we found evidence that negative prediction error learning was a necessary mechanism to account for the data of both age groups, given the other processes of the model. These results are suggestive that prediction error learning may play an important role in associative recognition, at least in situations like the CAR task in which items are repeated during learning. The results also suggest the exciting possibility that positive and negative components of prediction error learning may differ across development, with negative prediction error having a greater impact on memory in aging adults than in young adults. Although others have suggested that prediction is an integral aspect of episodic memory across the lifespan (Ofen & Shing, 2013), we are unaware of other work suggesting potential developmental dissociations between positive and negative prediction error learning in episodic memory. The model-based pattern of aging effects on prediction error learning is intriguing and suggests avenues for future research. However, the estimation of proportionally stronger negative prediction error learning in older adults was not expected a priori, and we emphasize that caution is needed in interpreting these results.

In addition, we acknowledge the possibility that models employing other associative mechanisms could capture the pattern of observed results in our task as well or better than the model we have presented here. A clear advantage of computational models over verbal theories is that computational models can be quantitatively compared with formal

model comparison techniques. We demonstrated the application of such techniques in the current work to assess different ways that prediction error learning could be implemented. Beyond these, there may be other variants of TCM that could capture the results with alternative mechanisms of item-context binding. It is also possible that models of associative recognition that are not based on TCM could account well for the currently presented data. For example, most models of associative recognition instantiate direct associations or conjunctions formed between the items in each pair (Gillund & Shiffrin, 1984; Hintzman, 1984; Li, Naveh-Benjamin, & Lindenberger, 2005), unlike the indirect associations between items and context in TCM. For example, Cox and Criss (2020) recently presented an associative recognition model in which conjunctions of paired items are encoded slowly in working memory based on features of each separate item, which are encoded more rapidly. These working memory representations are stored in long-term memory, and retrieval is a process of comparing representations of tested pairs to all representations stored in long-term memory. Although neither TCM nor the model of Cox and Criss were specifically designed to investigate aging differences, comparing the ability of these models to account for CAR task performance and age differences therein would be an interesting avenue of future research.

Some computational models of associative recognition have been proposed specifically to account for effects of aging. A neural network model by Li, Naveh-Benjamin and Lindenberger (2005) was designed to provide a neuromodulatory instantiation of the associative deficit account, which proposes that aging results in greater impairments in associative compared to item-based memory. In this model, associations are formed between features within each item, and across each item in a pair. Aging is modeled by varying the distinctiveness of internal representations, such that representations are less discriminable in simulations of older adults, which affects associative binding between items to a greater degree than memory for individual items. An alternative model was proposed by Benjamin (2010), who suggested that older adults do not have a deficit that is selective to associative memory per se, but instead have a global deficit in memory fidelity, such that all memory representations are generally more sparse and less valid in older adults. Interestingly, this model (the density of representations yields age-related dissociations model, or DRYAD) makes no distinction between item and associative information at all. Smyth and Naveh-Benjamin debated with Benjamin in an exchange in *Psychology and Aging* on whether older adults' memory deficits should be considered global or specific to associative aspects of memory (Benjamin, 2016; Naveh-Benjamin & Smyth, 2016; Smyth & Naveh-Benjamin, 2016). This debate has not been settled in a satisfactory way, and we do not attempt to do so here. TCM instantiates explicit associations between items and context, which is perhaps more conceptually compatible with an associative deficit account. Our current model is not well-suited to examine the potential age differences in representations proposed by DRYAD, as the representations that we instantiated are quite simple. However, it would be possible to explore potential age effects on representations within the TCM framework in the future. As we have noted, an advantage of computational modeling is the ability to quantitatively compare different theories, and an exciting avenue of future work is to formally compare models implementing aging deficits in different ways to iterate toward a better understanding of the mechanistic sources of behavioral deficits.

In addition to conducting model comparison, an important way for the field to better understand latent cognitive processes and how they change due to aging is to assess how well models can provide an explanation of performance across multiple tasks. Although our model was able to account for performance on the CAR task, a stronger test of the model would be to assess how well it could also account for young and older adults' performance on a variety of other memory tasks like item recognition or free recall (see Healey & Kahana, 2016 for a demonstration of fitting item recognition performance with a different TCM-based model of free recall). This is especially important as effects of aging on performance tend to vary a great deal between paradigms (with greater deficits on associative compared to item-based tasks, for example), such that models of episodic memory and effects of aging therein should be able to provide an explanation of performance on a variety of paradigms.

In this work, we have focused on benefits of computational modeling to cognitive aging research in terms of theory development, such as how generative models provide theoretical transparency, mechanistic specificity, and the ability to formally compare competing models. An additional benefit of modeling may be to increase the replicability of psychological findings. Some researchers have argued that the field's difficulty with replication may be due to weak links between a psychological theory and the experimental hypotheses that are tested (Oberauer & Lewandowsky, 2019a; P. L. Smith & Little, 2018). Theories provide a systematic and experimentally supported framework to understand cognition, and, as a result, hypotheses that are derived from a theory already have support from our prior understanding of neurocognitive processes. Hypotheses that are more exploratory, without strong links to a theory, lack this support. As a result, statistically significant results of more exploratory studies may be more likely to be spurious, resulting in replication failures (see Oberauer & Lewandowsky, 2019a for a systematic and quantitative approach to this issue). Computational models make specific claims about cognitive processes, as well as quantitative hypotheses about expected results given a set of parameters and experimental design. In addition, established models like TCM have been strongly supported by past findings. As a result, designing experiments and making hypotheses based on computational models may be a way to improve replicability. Although we did not directly investigate replicability in the current work, we suggest that more widespread use of computational modeling may be a way to improve replicability in cognitive aging research, and encourage future research to investigate this possibility systematically.

Limitations and alternative approaches

While there are many strengths of computational modeling, one limitation of this endeavor is that every model is necessarily "wrong" to some extent (Box, 1976), just as a map could never be a perfect representation of a geographical area. A model attempts to provide a useful simplification of the brain's mechanisms underlying observable behavior, but it is important to remember that simplifications have necessarily been made (McClelland, 2009). In addition, since latent processes are not observable, it is impossible to directly verify that a modeled process exists in the brain, although the recent emergence of joint modeling approaches show promise for helping researchers constrain computational theories with both behavior and neural activity (Palestro et al., 2018; Turner, Rodriguez, Norcia, McClure, &

Steyvers, 2016; Turner, Sederberg, Brown, & Steyvers, 2013). Even in the absence of neural data, we argue that because generative models provide precise, quantitative predictions about behavior, which can be formally compared between different models, it is possible to iterate toward more and more useful approximations of the computations taking place in the brain and how they differ due to aging.

Relatedly, it is important to bear in mind that the interpretation of parameter differences between young and older adults depends in part on the specification of mechanisms in the model. Modifying the mechanisms of the model will likely lead to changes in how parameters act and interact, which can have an effect on how model results are interpreted. An important advantage of model comparison, however, is that it is possible to test different models to assess what model, and by extension what interpretation of model parameters, is best able to account for the data.

We also note that while we have focused on generative computational models in this paper, there are alternative ways to address mechanistic questions. One approach is to apply *measurement models*, which are informed by a theory but, in contrast to generative models, do not attempt to directly instantiate explanatory mechanisms. These can be very helpful tools, particularly in terms of estimating latent processes via parameters. A recent example of this was provided by Oberauer and Lewandowsky (2019b), who developed a series of simple measurement models to estimate contributions of different facets of working memory, such as memory for individual elements, associative memory between elements, and filtering of irrelevant information in young and older adults (note, however, that although the model estimated the contributions of these processes, the processes themselves were not mechanistically instantiated, as in a generative model). These authors found evidence that older adults' working memory was impaired for associations, but not for individual elements. In another recent study, Greene and Naveh-Benjamin (2020) applied a multinomial processing tree (MPT) model to estimate the contribution of specific and gist-like associative processes to memory in young and older adults, and found that older adults were only impaired in retrieving more specific associations. We applaud these efforts and believe that measurement models such as these can be quite useful. However, we argue that where it is possible to develop a more-complete theory, generative computational models are even more powerful, as they take the extra step of instantiating the mechanisms needed to explain behavior as a function of the representations, associations and dynamics of cognition.

We also acknowledge that it is possible to provide support for or against verbal theories with carefully constructed experiments and conventional inferential statistical models like regression. A verbal theory that older adults have a selective deficit in associative memory to a greater degree than item memory is supported by an inferential statistic demonstrating a greater aging deficit in performance on an associative compared to an item memory task. Thus, inferential statistics and careful experimental design are important and necessary tools to make theoretical progress. However, as we have argued in this paper, we believe that a generative computational model is a powerful tool because it actually instantiates how behavior could be generated from latent processes, making the theory mathematically transparent while allowing for direct model comparison to test different theories of

cognition and developmental change, as well as allowing for model simulations to construct quantitative hypotheses directly based on the model.

Finally, it is important to note that *how* one fits a model also affects the potential interpretation of the results. In this work, we fit all models with hierarchical Bayesian approaches in order to properly account for uncertainty in our parameter estimates as well as variability within age groups. Although Bayesian methods are increasing in popularity (Van De Schoot, Winter, Ryan, Zondervan-Zwijnenburg, & Depaoli, 2017), the vast majority of standard statistical work in psychology relies on the *p*-value, a single number used to determine the significance of an effect. As others have pointed out, the *p*-value can be easily abused (Wagenmakers, 2007), such as by "*p*-hacking," or continuing to collect data until $p<.05$ for the statistical test of interest. Part of the problem with the *p* value is that it gives no indication of uncertainty, with convention dictating that $p<.05$ indicates significance. Over-reliance on the all-or-nothing *p* value has been offered as a contributor to the replication crisis in psychology (Anderson, 2020). An alternative to this approach is Bayesian model-fitting, by which effects of interest may be tested through *distributions* of statistical parameter values that inherently indicate uncertainty, in that more dispersion of parameter values indicates more uncertainty. This approach provides much more information that can inform scientists' confidence in the robustness of a finding, which in turn can help scientists better assess the amount of evidence for a given psychological effect. This may help the field avoid spurious findings that are unlikely to replicate. An extreme example of this was provided by Wagenmakers and colleagues (2011), who showed that claims of extra-sensory perception, supported by significant *p*-values in many experiments (Bem, 2011), received very little support in a Bayesian re-analysis of the data.

Pathways forward

As noted in the Introduction, other researchers have argued that generative computational modeling is a powerful tool and have encouraged more widespread application of these methods in aging research (Benjamin, 2010; Salthouse, 1988). Despite this, the great majority of studies on cognitive aging, and in psychology in general, rely on verbal theories and standard statistical approaches. An important question is why this is the case. We suspect that a major barrier to computational modeling is that this approach requires more time and effort compared to the typical pipeline of making a hypothesis based on a verbal theory and analyzing experimental results with conventional models like regression. Although it is true that computational modeling typically requires additional time and effort on the part of the researcher, we argue that the benefits gained, including theoretical transparency, estimation of latent processes (along with comparison of such processes between groups or individuals), and quantitative model comparison, far outweigh the costs. In addition, we consider the additional investment of time and cognitive resources to be a benefit in many cases, as it forces the researcher to think deeply about the cognitive mechanisms underlying the behavior of interest, and to make mathematically explicit the theory that is guiding the research endeavor.

We also suspect that many researchers do not feel they have the necessary skills to develop or apply potentially complex computational models, or to make use of techniques like

Bayesian model-fitting methods. Thankfully, many resources exist to guide the novice modeler (Epstein, 2008; Farrell & Lewandowsky, 2018; Forstmann & Wagenmakers, 2015; Guest & Martin, 2021; Heathcote, Brown, & Wagenmakers, 2015; McClelland, 2009; Smaldino, 2020; Wilson & Collins, 2019; see Sederberg & Darby, under review, for a review and example implementations of popular episodic memory models). Some easy-to-use programming packages exist that can help beginners to make use of models, such as the Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python project, or HDDM (Wiecki, Sofer, & Frank, 2013), which allows researchers to apply sequential sampling decision-making models similar to the one applied in the current work simply by providing their data in a tabular form. In addition, many introductory resources are available on Bayesian model-fitting procedures, in general (Gelman, 2006; Gelman et al., 2013; Kruschke, 2013; Lee, 2011; Shiffrin, Lee, Kim, & Wagenmakers, 2008).

Conclusions

In this work, we have argued that computational models make theories transparent by instantiating proposed neurocognitive mechanisms in explicit mathematical equations, and that research on the cognitive effects of aging would benefit from more widespread use of model-based approaches, including comparing model parameters to assess developmental changes in cognitive mechanisms, formally comparing models to adjudicate between theories, and generating data to make quantitative hypotheses. Although computational modeling is not an easy process, especially for beginners, we suggest that the field of cognitive aging would greatly benefit from more widespread adoption of these methods. These techniques may be applied to systematically and quantitatively iterate toward better theories of cognitive change due to aging. In addition, computational modeling allows the researcher to make quantitative hypotheses based directly on the theory as instantiated in the model. As other researchers have pointed out (Oberauer & Lewandowsky, 2019a; P. L. Smith & Little, 2018), strengthening the relation between theory and hypotheses, and reducing reliance on exploratory studies not driven by theory, may be a powerful way to increase replicability. Computational models, in sum, are a powerful tool that can help us to develop more robust and transparent theories to improve replicability and scientific discovery in psychological science and aging research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by National Institute on Aging Grant RO1AG024270 to Timothy A. Salthouse, and by Air Force Research Labs Grant FA8650–16-1–6770 to Per B. Sederberg.

References

Anderson SF (2020). Misinterpreting p: The Discrepancy Between p Values and the Probability the Null Hypothesis is True, the Influence of Multiple Testing, and Implications for the Replication Crisis. PSYCHOLOGICAL METHODS, 25(5), 596–609. 10.1037/met0000248 [PubMed: 31829657]

Anvari F, & Lakens D (2018). The replicability crisis and public trust in psychological science. Comprehensive Results in Social Psychology, 3(3), 266–286. 10.1080/23743603.2019.1684822

Asendorpf JB, Conner M, Fruyt FD, Houwer JD, Denissen JJA, Fiedler K, … Wicherts JM (2013). Recommendations for Increasing Replicability in Psychology. European Journal of Personality, 27(2), 108–119. 10.1002/per.1919

Bar M (2009). The proactive brain: Memory for predictions. Philosophical Transactions of the Royal Society B: Biological Sciences, 364(1521), 1235–1243. 10.1098/rstb.2008.0310

Bem DJ (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. Journal of Personality and Social Psychology, 100(3), 407–425. 10.1037/a0021524 [PubMed: 21280961]

Benjamin AS (2010). Representational explanations of "process" dissociations in recognition: The DRYAD theory of aging and memory judgments. Psychological Review, 117(4), 1055–1079. 10.1037/a0020810 [PubMed: 20822289]

Benjamin AS (2016). Aging and associative recognition: A view from the DRYAD model of age-related memory deficits. Psychology and Aging, 31(1), 14–20. 10.1037/pag0000065 [PubMed: 26866587]

Box GEP (1976). Science and Statistics. Journal of the American Statistical Association, 71(356), 791–799. 10.1080/01621459.1976.10480949

Brady TF, Konkle T, Alvarez GA, & Oliva A (2008). Visual long-term memory has a massive storage capacity for object details. Proceedings of the National Academy of Sciences, 105(38), 14325–14329. 10.1073/pnas.0803390105

Braver TS, Barch DM, & Cohen JD (1999). Cognition and control in schizophrenia: A computational model of dopamine and prefrontal function. Biological Psychiatry, 46(3), 312–328. 10.1016/S0006-3223(99)00116-X [PubMed: 10435197]

Cansino S (2009). Episodic memory decay along the adult lifespan: A review of behavioral and neurophysiological evidence. International Journal of Psychophysiology, 71(1), 64–69. 10.1016/j.ijpsycho.2008.07.005 [PubMed: 18725253]

Castel AD, & Craik FIM (2003). The Effects of Aging and Divided Attention on Memory for Item and Associative Information. Psychology and Aging, 18(4), 873–885. [PubMed: 14692872]

Cavanagh JF, Frank MJ, Klein TJ, & Allen JJB (2010). Frontal theta links prediction errors to behavioral adaptation in reinforcement learning. NeuroImage, 49(4), 3198–3209. 10.1016/j.neuroimage.2009.11.080 [PubMed: 19969093]

Chen T, & Naveh-Benjamin M (2012). Assessing the Associative Deficit of Older Adults in Long-Term and Short-Term/Working Memory. PSYCHOLOGY AND AGING, 27(3), 666–682. 10.1037/a0026943 [PubMed: 22308997]

Cockburn J, & Holroyd CB (2010). Focus on the positive: Computational simulations implicate asymmetrical reward prediction error signals in childhood attention-deficit/hyperactivity disorder. Brain Research, 1365, 18–34. 10.1016/j.brainres.2010.09.065 [PubMed: 20875804]

Cohen JD, Braver TS, OReilly R, Roberts AC, Robbins TW, & Weiskrantz L (1996). A computational approach to prefrontal cortex, cognitive control and schizophrenia: Recent developments and current challenges. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 351(1346), 1515–1527. 10.1098/rstb.1996.0138 [PubMed: 8941963]

Cohn M, Emrich SM, & Moscovitch M (2008). Age-related deficits in associative memory: The influence of impaired strategic retrieval. Psychology and Aging, 23(1), 93–103. 10.1037/0882-7974.23.1.93 [PubMed: 18361659]

Cox GE, & Criss AH (2020). Similarity leads to correlated processing: A dynamic model of encoding and recognition of episodic associations. Psychological Review, No Pagination Specified–No Pagination Specified. 10.1037/rev0000195

Craik FIM (1968). Two components in free recall. Journal of Verbal Learning and Verbal Behavior, 7(6), 996–1004. 10.1016/S0022-5371(68)80058-1

Craik FIM, Luo L, & Sakuta Y (2010). Effects of aging and divided attention on memory for items and their contexts. Psychology and Aging, 25(4), 968–979. 10.1037/a0020276 [PubMed: 20973605]

Darby KP, & Sloutsky VM (2015). The cost of learning: Interference effects in memory development. Journal of Experimental Psychology: General, 144(2), 410–431. 10.1037/xge0000051 [PubMed: 25688907]

DeCarlo LT (1998). Signal detection theory and generalized linear models. Psychological Methods, 3(2), 186–205. 10.1037/1082-989X.3.2.186

Dodson CS, Bawa S, & Slotnick SD (2007). Aging, source memory, and misrecollections. Journal of Experimental Psychology: Learning, Memory, and Cognition, 33(1), 169–181. 10.1037/0278-7393.33.1.169 [PubMed: 17201560]

Epstein JM (2008). Why Model? Journal of Artificial Societies and Social Simulation, 11(4), 12.

Farrell S, & Lewandowsky S (2010). Computational Models as Aids to Better Reasoning in Psychology. Current Directions in Psychological Science, 19(5), 329–335. 10.1177/0963721410386677

Farrell S, & Lewandowsky S (2018). Computational Modeling of Cognition and Behavior: (First). Cambridge University Press. 10.1017/CBO9781316272503

Federmeier KD, Kutas M, & Schul R (2010). Age-related and individual differences in the use of prediction during language comprehension. Brain and Language, 115(3), 149–161. 10.1016/j.bandl.2010.07.006 [PubMed: 20728207]

Forstmann BU, & Wagenmakers E-J. (Eds.). (2015). An introduction to model-based cognitive neuroscience (pp. xi, 354). New York, NY, US: Springer Science + Business Media. 10.1007/978-1-4939-2236-9

Frank MJ (2008). Schizophrenia: A Computational Reinforcement Learning Perspective. Schizophrenia Bulletin, 34(6), 1008–1011. 10.1093/schbul/sbn123 [PubMed: 18791075]

Frank MJ, Santamaria A, O'Reilly RC, & Willcutt E (2007). Testing Computational Models of Dopamine and Noradrenaline Dysfunction in Attention Deficit/Hyperactivity Disorder. Neuropsychopharmacology, 32(7), 1583–1599. 10.1038/sj.npp.1301278 [PubMed: 17164816]

Gallo DA, Sullivan AL, Daffner KR, Schacter DL, & Budson AE (2004). Associative Recognition in Alzheimer's Disease: Evidence for Impaired Recall-to-Reject. Neuropsychology, 18(3), 556–563. [PubMed: 15291733]

Gelman A (2006). Multilevel (Hierarchical) Modeling: What It Can and Cannot Do. Technometrics, 48(3), 432–435. 10.1198/004017005000000661

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, & Rubin DB (2013). Bayesian Data Analysis. Chapman and Hall/CRC. 10.1201/b16018

Gershman SJ, Moore CD, Todd MT, Norman KA, & Sederberg PB (2012). The Successor Representation and Temporal Context. Neural Computation, 24(6), 1553–1568. 10.1162/NECO_a_00282 [PubMed: 22364500]

Gillund G, & Shiffrin RM (1984). A retrieval model for both recognition and recall. Psychological Review, 91(1), 1–67. 10.1037/0033-295X.91.1.1 [PubMed: 6571421]

Golomb JD, Peelle JE, Addis KM, Kahana MJ, & Wingfield A (2008). Effects of adult aging on utilization of temporal and semantic associations during free and serial recall. Memory & Cognition, 36(5), 947–956. 10.3758/MC.36.5.947 [PubMed: 18630201]

Greene NR, & Naveh-Benjamin M (2020). A Specificity Principle of Memory: Evidence From Aging and Associative Memory. Psychological Science, 31(3), 316–331. 10.1177/0956797620901760 [PubMed: 32074021]

Guest O, & Martin AE (2021). How Computational Modeling Can Force Theory Building in Psychological Science. Perspectives on Psychological Science, 1745691620970585. 10.1177/1745691620970585

Haines N, Kvam PD, Irving LH, Smith C, Beauchaine TP, Pitt MA, … Turner B (2020). Theoretically Informed Generative Models Can Advance the Psychological and Brain Sciences: Lessons from the Reliability Paradox. PsyArXiv. 10.31234/osf.io/xr7y3

Haque RU, Inati SK, Levey AI, & Zaghloul KA (2020). Feedforward prediction error signals during episodic memory retrieval. Nature Communications, 11(1), 6075. 10.1038/s41467-020-19828-0

Hasher L, & Zacks RT (1988). Working Memory, Comprehension, and Aging: A Review and a New View. In Bower GH (Ed.), Psychology of Learning and Motivation (Vol. 22, pp. 193–225). Academic Press. 10.1016/S0079-7421(08)60041-9

Healey MK, Hasher L, & Campbell KL (2013). The role of suppression in resolving interference: Evidence for an age-related deficit. Psychology and Aging, 28(3), 721–728. 10.1037/a0033003 [PubMed: 23957222]

Healey MK, & Kahana MJ (2016). A Four-Component Model of Age-Related Memory Change. PSYCHOLOGICAL REVIEW, 123(1), 23–69. 10.1037/rev0000015 [PubMed: 26501233]

Heathcote A, Brown SD, & Wagenmakers E-J (2015). An Introduction to Good Practices in Cognitive Modeling. In Forstmann BU& Wagenmakers E-J (Eds.), An Introduction to Model-Based Cognitive Neuroscience (pp. 25–48). New York, NY: Springer. 10.1007/978-1-4939-2236-9_2

Hebb DO (1949). The organization of behavior: A neuropsychological theory. Wiley.

Hintzman DL (1984). MINERVA 2: A simulation model of human memory. Behavior Research Methods, Instruments, & Computers, 16(2), 96–101. 10.3758/BF03202365

Hockley WE (1992). Item Versus Associative Information - Further Comparisons of Forgetting Rates. JOURNAL OF EXPERIMENTAL PSYCHOLOGY-LEARNING MEMORY AND COGNITION, 18(6), 1321–1330.

Hockley WE, & Consoli A (1999). Familiarity and recollection in item and associative recognition. Memory & Cognition, 27(4), 657–664. 10.3758/BF03211559 [PubMed: 10479824]

Howard MW, & Kahana MJ (2002). A distributed representation of temporal context. Journal of Mathematical Psychology, 46(3), 269–299.

Howard MW, Kahana MJ, & Wingfield A (2006). Aging and contextual binding: Modeling recency and lag recency effects with the temporal context model. Psychonomic Bulletin & Review, 13(3), 439–445. 10.3758/BF03193867 [PubMed: 17048728]

Huys QJM, Maia TV, & Paulus MP (2016). Computational Psychiatry: From Mechanistic Insights to the Development of New Treatments. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 1(5), 382–385. 10.1016/j.bpsc.2016.08.001 [PubMed: 29560868]

Jolly E, & Chang LJ (2019). The Flatland Fallacy: Moving Beyond Low–Dimensional Thinking. Topics in Cognitive Science, 11(2), 433–454. 10.1111/tops.12404 [PubMed: 30576066]

Kliegl R, & Lindenberger U (1993). Modeling intrusions and correct recall in episodic memory: Adult age differences in encoding of list context. Journal of Experimental Psychology: Learning, Memory, and Cognition, 19(3), 617–637. [PubMed: 8501432]

Kruschke JK (2013). Bayesian estimation supersedes the t test. Journal of Experimental Psychology: General, 142(2), 573–603. 10.1037/a0029146 [PubMed: 22774788]

Lee MD (2011). How cognitive modeling can benefit from hierarchical Bayesian models. Journal of Mathematical Psychology, 55(1), 1–7. 10.1016/j.jmp.2010.08.013

Li S-C, Naveh-Benjamin M, & Lindenberger U (2005). Aging Neuromodulation Impairs Associative Binding: Neurocomputational Account. Psychological Science, 16(6), 445–450. 10.1111/j.0956-7976.2005.01555.x [PubMed: 15943670]

Light LL, Patterson MM, Chung C, & Healy MR (2004). Effects of repetition and response deadline on associative recognition in young and older adults. Memory & Cognition, 32(7), 1182–1193. 10.3758/BF03196891 [PubMed: 15813499]

Lohnas LJ, Polyn SM, & Kahana MJ (2015). Expanding the scope of memory search: Modeling intralist and interlist effects in free recall. Psychological Review, 122(2), 337–363. 10.1037/a0039036 [PubMed: 25844876]

Maassen E, Assen MALM van, Nuijten MB, Olsson-Collentine A, & Wicherts JM (2020). Reproducibility of individual effect sizes in meta-analyses in psychology. PLOS ONE, 15(5), e0233107. 10.1371/journal.pone.0233107 [PubMed: 32459806]

Maia TV, Huys QJM, & Frank MJ (2017). Theory-based computational psychiatry. Biological Psychiatry, 82(6), 382–384. 10.1016/j.biopsych.2017.07.016 [PubMed: 28838466]

McClelland JL (2009). The Place of Modeling in Cognitive Science. Topics in Cognitive Science, 1(1), 11–38. 10.1111/j.1756-8765.2008.01003.x [PubMed: 25164798]

Mitchell DB, Brown AS, & Murphy DR (1990). Dissociations between procedural and episodic memory: Effects of time and aging. Psychology and Aging, 5(2), 264–276. 10.1037/0882-7974.5.2.264 [PubMed: 2378692]

Mizumori SJY (2013). Context Prediction Analysis and Episodic Memory. Frontiers in Behavioral Neuroscience, 7. 10.3389/fnbeh.2013.00132

Muthukrishna M, & Henrich J (2019). A problem in theory. Nature Human Behaviour, 3(3), 221–229. 10.1038/s41562-018-0522-1

Myerson J, Hale S, Wagstaff D, Poon LW, & Smith GA (1990). The information-loss model: A mathematical theory of age-related cognitive slowing. Psychological Review, 97(4), 475–487. [PubMed: 2247538]

Myung IJ (2000). The Importance of Complexity in Model Selection. Journal of Mathematical Psychology, 44(1), 190–204. 10.1006/jmps.1999.1283 [PubMed: 10733864]

Myung IJ (2003). Tutorial on maximum likelihood estimation. Journal of Mathematical Psychology, 47(1), 90–100. 10.1016/S0022-2496(02)00028-7

Navarro DJ, & Fuss IG (2009). Fast and accurate calculations for first-passage times in Wiener diffusion models. Journal of Mathematical Psychology, 53(4),222–230. 10.1016/j.jmp.2009.02.003

Naveh-Benjamin M (2000). Adult age differences in memory performance: Tests of an associative deficit hypothesis. Journal of Experimental Psychology: Learning, Memory, and Cognition, 26(5), 1170–1187. 10.1037/0278-7393.26.5.1170 [PubMed: 11009251]

Naveh-Benjamin M, Guez J, Kilb A, & Reedy S (2004). The Associative Memory Deficit of Older Adults: Further Support Using Face-Name Associations. Psychology and Aging, 19(3), 541–546. 10.1037/0882-7974.19.3.541 [PubMed: 15383004]

Naveh-Benjamin M, & Smyth AC (2016). DRYAD and ADH: Further Comments on Explaining Age-Related Differences in Memory. PSYCHOLOGY AND AGING, 31(1), 21–24. 10.1037/pag0000066 [PubMed: 26866588]

Nilsson L-G (2003). Memory function in normal aging. Acta Neurologica Scandinavica, 107(s179), 7–13. 10.1034/j.1600-0404.107.s179.5.x [PubMed: 12542507]

Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, … Yarkoni T (2015). Promoting an open research culture. Science, 348(6242), 1422–1425. 10.1126/science.aab2374 [PubMed: 26113702]

Oberauer K, & Lewandowsky S (2019a). Addressing the theory crisis in psychology. Psychonomic Bulletin & Review, 26(5), 1596–1618. 10.3758/s13423-019-01645-2 [PubMed: 31515732]

Oberauer K, & Lewandowsky S (2019b). Simple Measurement Models for Complex Working-Memory Tasks. PSYCHOLOGICAL REVIEW, 126(6), 880–932.10.1037/rev0000159 [PubMed: 31524425]

Ofen N, & Shing YL (2013). From perception to memory: Changes in memory systems across the lifespan. Neuroscience & Biobehavioral Reviews, 37(9, Part B), 2258–2267. 10.1016/j.neubiorev.2013.04.006 [PubMed: 23623983]

Old SR, & Naveh-Benjamin M (2008). Differential effects of age on item and associative measures of memory: A meta-analysis. Psychology and Aging, 23(1), 104–118. 10.1037/0882-7974.23.1.104 [PubMed: 18361660]

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251), aac4716–aac4716. 10.1126/science.aac4716 [PubMed: 26315443]

Overman AA, & Becker JT (2009). The associative deficit in older adult memory: Recognition of pairs is not improved by repetition. Psychology and Aging, 24(2), 501–506. [PubMed: 19485666]

Palestro JJ, Bahg G, Sederberg PB, Lu Z-L, Steyvers M, & Turner BM (2018). A tutorial on joint models of neural and behavioral measures of cognition. Journal of Mathematical Psychology, 84, 20–48. 10.1016/j.jmp.2018.03.003

Pastore M, & Calcagnì A (2019). Measuring Distribution Similarities Between Samples: A Distribution-Free Overlapping Index. Frontiers in Psychology, 10. 10.3389/fpsyg.2019.01089

Pitt MA, Myung IJ, & Zhang S (2002). Toward a method of selecting among computational models of cognition. Psychological Review, 109(3), 472–491. 10.1037/0033-295X.109.3.472 [PubMed: 12088241]

Polyn SM, Norman KA, & Kahana MJ (2009). A context maintenance and retrieval model of organizational processes in free recall. Psychological Review, 116(1), 129–156. [PubMed: 19159151]

Postman L, & Underwood BJ (1973). Critical issues in interference theory. Memory & Cognition, 1(1), 19–40. 10.3758/BF03198064 [PubMed: 24214472]

Ratcliff R, & McKoon G (2008). The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. Neural Computation, 20(4), 873–922. 10.1162/neco.2008.12-06-420 [PubMed: 18085991]

Ratcliff R, & McKoon G (2015). Aging effects in item and associative recognition memory for pictures and words. Psychology and Aging, 30(3), 669–674. 10.1037/pag0000030 [PubMed: 25985326]

Ratcliff R, Thapar A, Gomez P, & McKoon G (2004). A Diffusion Model Analysis of the Effects of Aging in the Lexical-Decision Task. Psychology and Aging, 19(2), 278–289. 10.1037/0882-7974.19.2.278 [PubMed: 15222821]

Rotello CM, & Heit E (2000). Associative recognition: A case of recall-to-reject processing. Memory & Cognition, 28(6), 907–922. 10.3758/BF03209339 [PubMed: 11105517]

Salthouse TA (1988). Initiating the formalization of theories of cognitive aging. Psychology and Aging, 3(1), 3–16. 10.1037/0882-7974.3.1.3 [PubMed: 3077318]

Salthouse TA (1996). The processing-speed theory of adult age differences in cognition. Psychological Review, 103(3), 403–428. 10.1037/0033-295X.103.3.403 [PubMed: 8759042]

Samanez-Larkin GR, Worthy DA, Mata R, McClure SM, & Knutson B (2014). Adult age differences in frontostriatal representation of prediction error but not reward outcome. Cognitive, Affective, & Behavioral Neuroscience, 14(2), 672–682. 10.3758/s13415-014-0297-4

Schacter DL, Kaszniak AW, Kihlstrom JF, & Valdiserri M (1991). The relation between source memory and aging. Psychology and Aging, 6(4), 559–568. 10.1037/0882-7974.6.4.559 [PubMed: 1777144]

Sederberg PB, & Darby KP (under review). Computational models of episodic memory.

Sederberg PB, Gershman SJ, Polyn SM, & Norman KA (2011). Human memory reconsolidation can be explained using the temporal context model. Psychonomic Bulletin & Review, 18(3), 455–468. 10.3758/s13423-011-0086-9 [PubMed: 21512839]

Sederberg PB, Howard MW, & Kahana MJ (2008). A context-based theory of recency and contiguity in free recall. Psychological Review, 115(4), 893–912. 10.1037/a0013396 [PubMed: 18954208]

Shiffrin RM, Lee MD, Kim W, & Wagenmakers E-J (2008). A Survey of Model Evaluation Approaches With a Tutorial on Hierarchical Bayesian Methods. Cognitive Science, 32(8), 1248–1284. 10.1080/03640210802414826 [PubMed: 21585453]

Siefke BM, Smith TA, & Sederberg PB (2019). A context-change account of temporal distinctiveness. Memory & Cognition, 47(6), 1158–1172. 10.3758/s13421-019-00925-5 [PubMed: 30912034]

Slamecka NJ (1966). Differentiation versus unlearning of verbal associations. Journal of Experimental Psychology, 71(6), 822–828. 10.1037/h0023223 [PubMed: 5939360]

Smaldino PE (2020). How to Translate a Verbal Theory Into a Formal Model. Social Psychology, 51(4), 207–218. 10.1027/1864-9335/a000425

Smith AD (1977). Adult age differences in cued recall. Developmental Psychology, 13(4), 326–331. 10.1037/0012-1649.13.4.326

Smith PL, & Little DR (2018). Small is beautiful: In defense of the small-N design. Psychonomic Bulletin & Review, 25(6), 2083–2101. 10.3758/s13423-018-1451-8 [PubMed: 29557067]

Smyth AC, & Naveh-Benjamin M (2016). Can DRYAD Explain Age-Related Associative Memory Deficits? PSYCHOLOGY AND AGING, 31(1), 1–13. 10.1037/a0039071 [PubMed: 25961878]

Starns JJ, & Ratcliff R (2010). The effects of aging on the speed–accuracy compromise: Boundary optimality in the diffusion model. Psychology and Aging, 25(2), 377–390. 10.1037/a0018022 [PubMed: 20545422]

Stephens JDW, & Overman AA (2018). Modeling age differences in effects of pair repetition and proactive interference using a single parameter. Psychology and Aging, 33(1), 182. 10.1037/pag0000195 [PubMed: 29494189]

Stone M (1960). Models for choice-reaction time. Psychometrika, 25(3), 251–260. 10.1007/BF02289729

Surprenant AM, Neath I, & Brown GDA (2006). Modeling age-related differences in immediate memory using SIMPLE. Journal of Memory and Language, 55(4), 572–586. 10.1016/j.jml.2006.08.001 [PubMed: 18172514]

Taconnat L, Clarys D, Vanneste S, Bouazzaoui B, & Isingrini M (2007). Aging and strategic retrieval in a cued-recall test: The role of executive functions and fluid intelligence. Brain and Cognition, 64(1), 1–6. 10.1016/j.bandc.2006.09.011 [PubMed: 17182162]

Tenen D, & Wythoff G (2014). Sustainable Authorship in Plain Text using Pandoc and Markdown. Programming Historian.

Theisen M, Lerche V, von Krause M, & Voss A (2020). Age differences in diffusion model parameters: A meta-analysis. Psychological Research. 10.1007/s00426-020-01371-8

Tromp D, Dufour A, Lithfous S, Pebayle T, & Després O (2015). Episodic memory in normal aging and Alzheimer disease: Insights from imaging and behavioral studies. Ageing Research Reviews, 24, 232–262. 10.1016/j.arr.2015.08.006 [PubMed: 26318058]

Turner BM, Rodriguez CA, Norcia TM, McClure SM, & Steyvers M (2016). Why more is better: Simultaneous modeling of EEG, fMRI, and behavioral data. NeuroImage, 128, 96–115. 10.1016/j.neuroimage.2015.12.030 [PubMed: 26723544]

Turner BM, Sederberg PB, Brown SD, & Steyvers M (2013). A method for efficiently sampling from distributions with correlated dimensions. Psychological Methods, 18(3), 368–384. 10.1037/a0032222 [PubMed: 23646991]

Turner BM, & Van Zandt T (2014). Hierarchical approximate Bayesian computation. Psychometrika, 79(2), 185–209. 10.1007/s11336-013-9381-x [PubMed: 24297436]

Van De Schoot R, Winter SD, Ryan O, Zondervan-Zwijnenburg M, & Depaoli S (2017). A systematic review of Bayesian articles in psychology: The last 25 years. Psychological Methods, 22(2), 217. [PubMed: 28594224]

Vandierendonck A (2017). A comparison of methods to combine speed and accuracy measures of performance: A rejoinder on the binning procedure. Behavior Research Methods, 49(2), 653–673. 10.3758/s13428-016-0721-5 [PubMed: 26944576]

Verhaeghen P (2003). Aging and vocabulary scores: A meta-analysis. PSYCHOLOGY AND AGING, 18(2), 332–339. 10.1037/0882-7974.18.2.332 [PubMed: 12825780]

Wagenmakers E-J (2007). A practical solution to the pervasive problems ofp values. Psychonomic Bulletin & Review, 14(5), 779–804. 10.3758/BF03194105 [PubMed: 18087943]

Wagenmakers E-J, Wetzels R, Borsboom D, & van der Maas HLJ (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). Journal of Personality and Social Psychology, 100(3), 426–432. 10.1037/a0022790 [PubMed: 21280965]

Weichart ER, Darby KP, Fenton AW, Jacques BG, Kirkpatrick RP, Turner BM, & Sederberg PB (2021). Quantifying mechanisms of cognition with an experiment and modeling ecosystem. Behavior Research Methods. 10.3758/s13428-020-01534-w

Wiecki TV, Poland J, & Frank MJ (2015). Model-Based Cognitive Neuroscience Approaches to Computational Psychiatry: Clustering and Classification. Clinical Psychological Science, 3(3), 378–399. 10.1177/2167702614565359

Wiecki TV, Sofer I, & Frank MJ (2013). HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. Frontiers in Neuroinformatics, 7. 10.3389/fninf.2013.00014

Wilson RC, & Collins AG (2019). Ten simple rules for the computational modeling of behavioral data. eLife, 8, e49547. 10.7554/eLife.49547 [PubMed: 31769410]

Woltz DJ, & Was CA (2006). Availability of related long-term memory during and after attention focus in working memory. Memory & Cognition, 34(3), 668–684. 10.3758/BF03193587 [PubMed: 16933773]

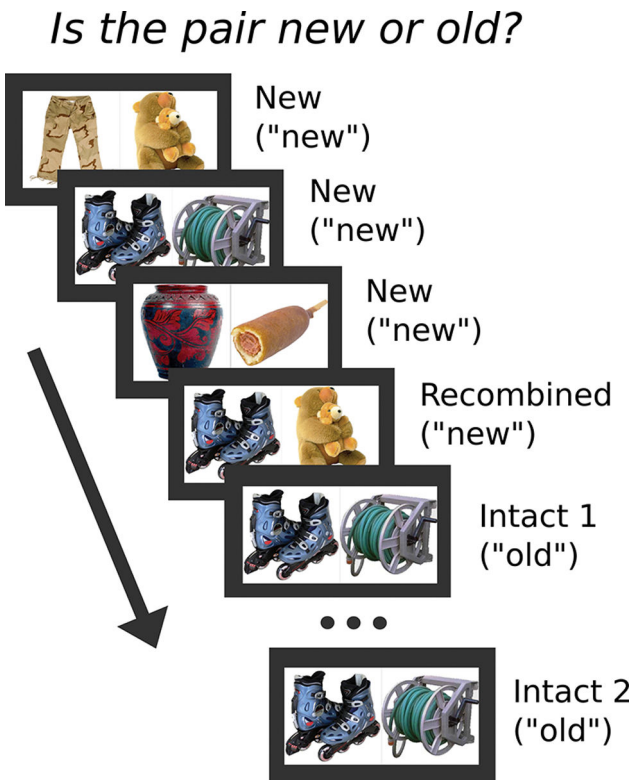

Figure 1. Types of object pairings presented in the continuous associative recognition task.
On every trial of the task, participants were presented with a pair of objects and asked to determine whether the pair was “new” or “old.” Each object was presented in four pairs: New, Intact 1, Intact 2, and Recombined pairs.

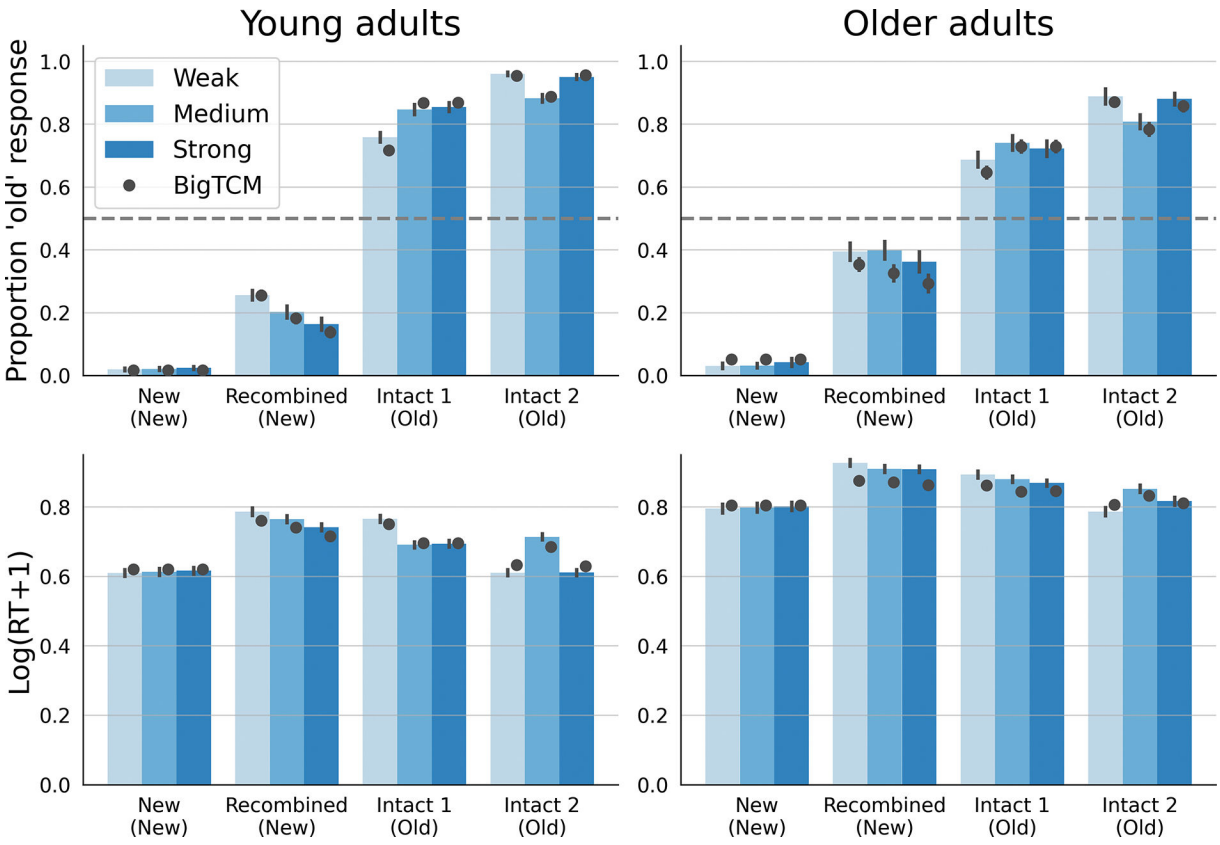

Figure 2. Observed and model-simulated CAR task performance.

False alarms (for New and Recombined pairs), and hits (for Intact 1 and Intact 2 pairs) for each strength condition are presented in the upper panels, and log-transformed RTs (in seconds) are presented in the lower panels. A value of 1 was added to each RT prior to the log-transform so that the lower-bound of transformed values would be 0. Performance measures are presented for young adults on the left panels, and for older adults on the right panels. Mean observed performance values are plotted by bars, and mean model-simulated values are presented as dots. Error bars represent standard errors.

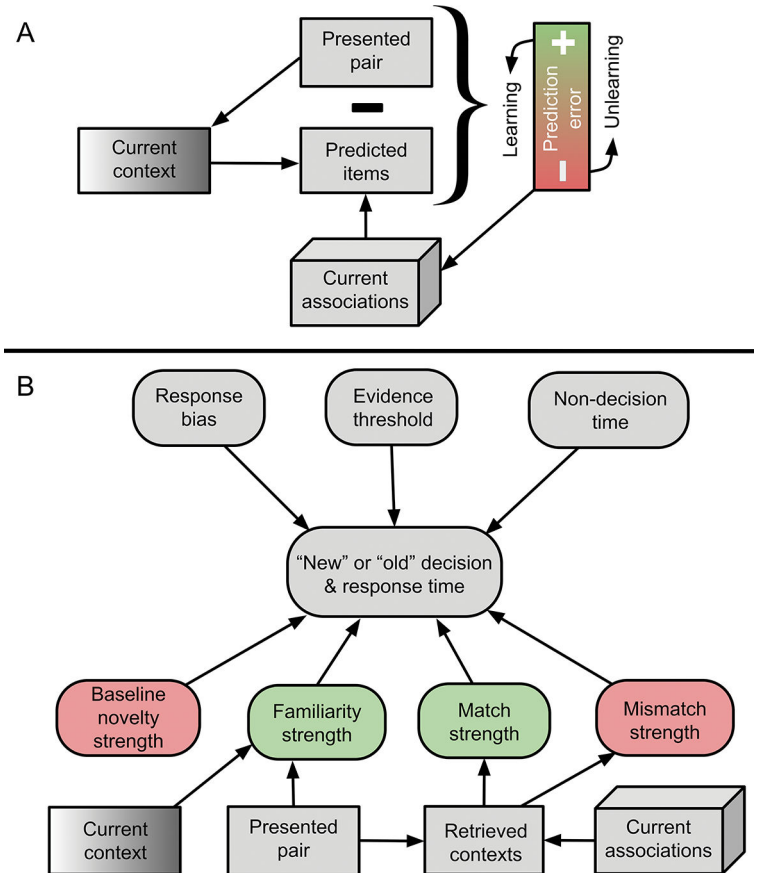

Figure 3. Schematic summaries of encoding and retrieval processes in BigTCM for associative recognition

Scalar values are indicated by rounded corners, whereas arrays have square corners; matrices are depicted as boxes. (A) Encoding processes. Items predicted from associations with context are compared to the presented pair to calculate prediction error. Positive prediction errors are the basis of learning, and negative prediction errors are the basis of unlearning. Both learning and unlearning are used to update the association matrix for the next trial, while temporal context is updated with the presented items. (B) Retrieval processes. A decision and response time are simulated for each trial based on memory strengths supporting "old" responses (green) and those supporting "new" responses (red), a response bias toward "new" or "old" responses, a threshold of evidence needed to make a decision, and an estimate of perceptuomotor (non-decision) processing time. See the text for additional details.

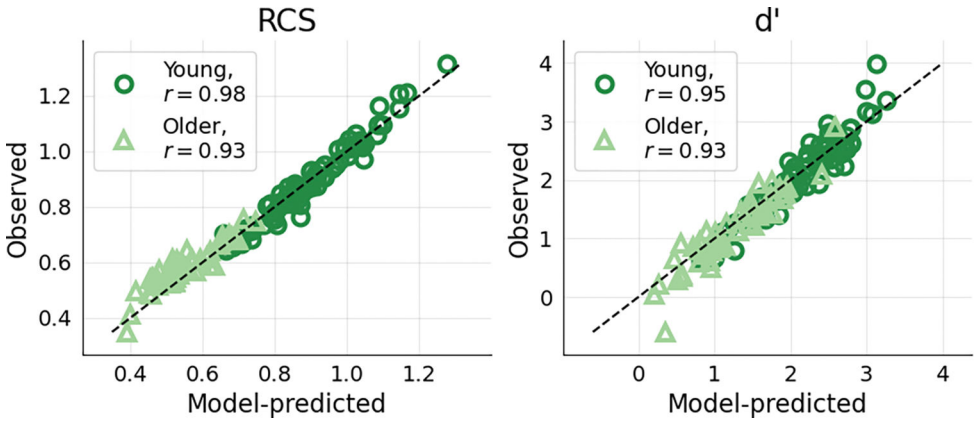

Figure 4. Correlations between observed and model-predicted performance.

The dark dashed lines indicate what would be a perfect correspondence between model-predicted and observed data. RCS = rate correct score.

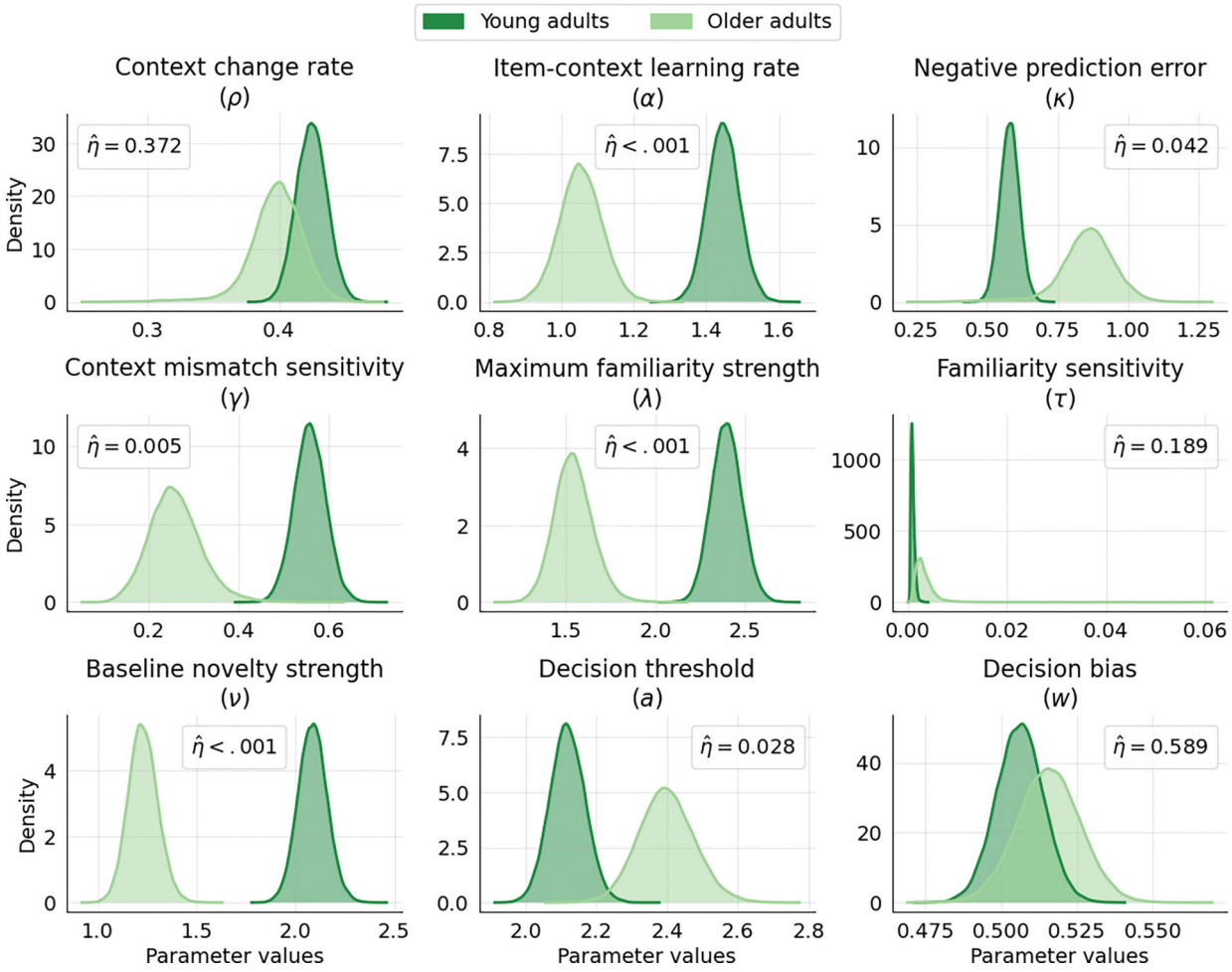

Figure 5. Posterior hyper-parameter distributions for young and older adults.

Each plot shows the posterior distributions for one parameter, along a metric of overlap between the two distributions.

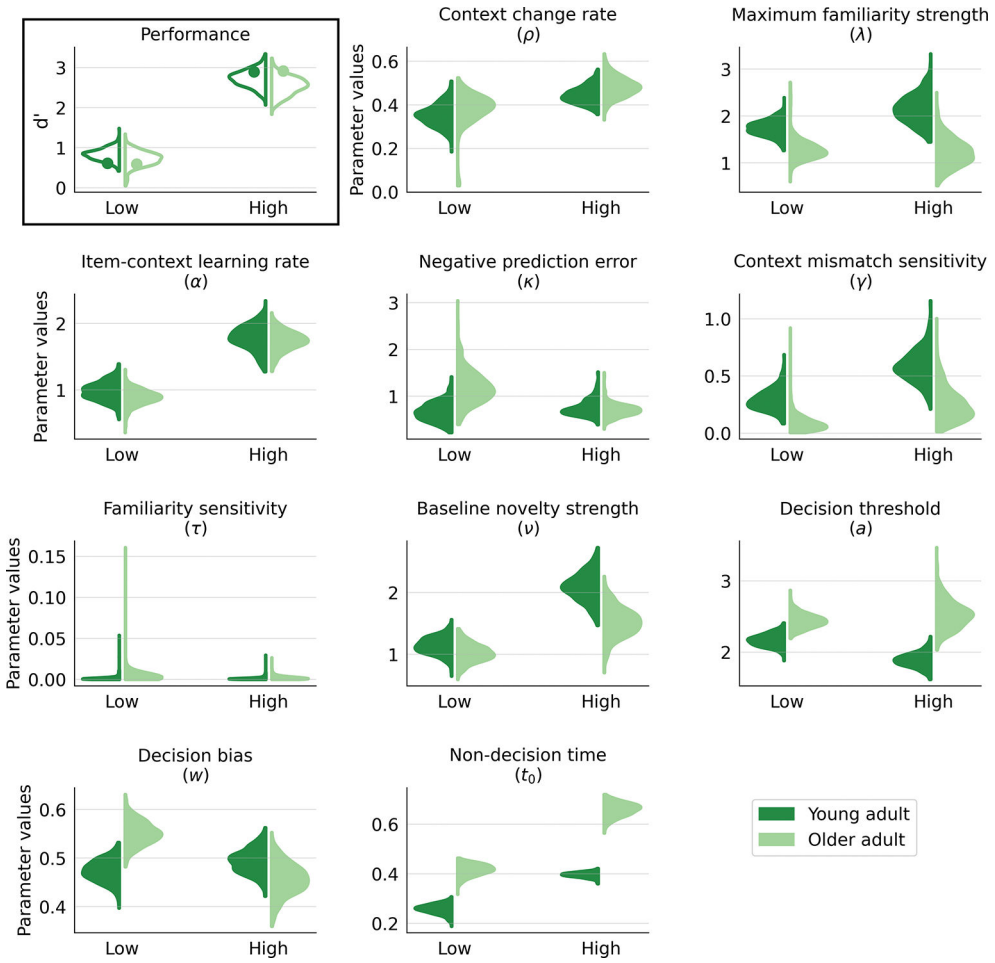

Figure 6. Posterior predictive distributions (PPDs) of performance scores and posterior parameter distributions for four example participants.

Dyads of low-performing and high-performing participants were selected for this example. The top-left (boxed) panel presents the PPDs for each participant, which were constructed by generating data with each sample participant's parameter posterior distributions (split violins plots), along with the observed d' score for each participant (dots). The other panels show the posterior parameter distributions for each participant.

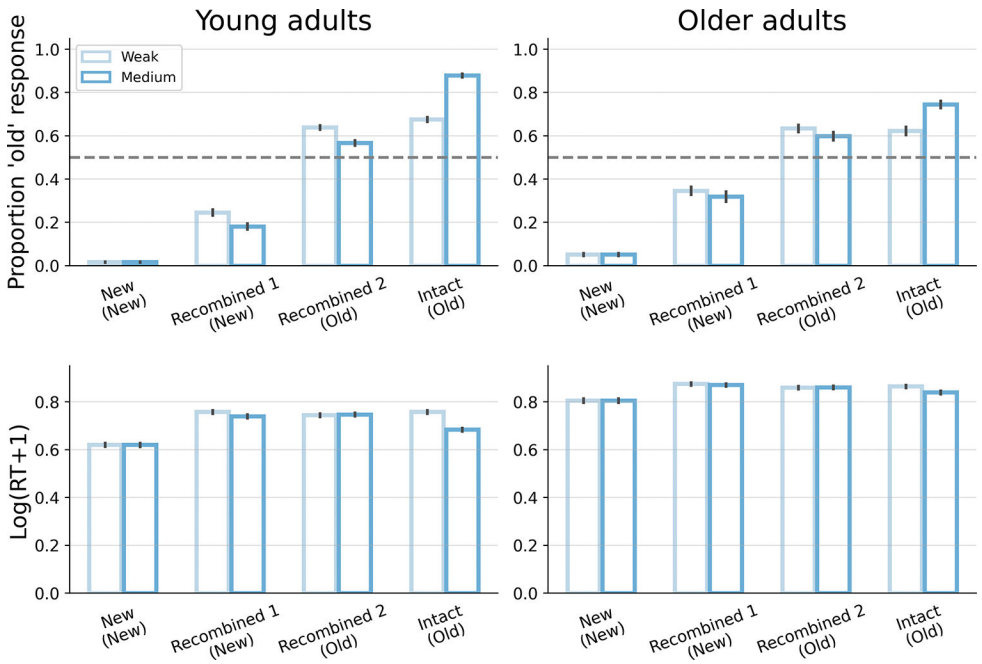

Figure 7. Model-generated performance for a hypothetical CAR task variant.

False alarms (for New and Recombined 1 pairs), and hits (for Recombined 2 and Intact pairs) for each strength condition are presented in the upper panels, and log-transformed RTs (in seconds) are presented in the lower panels. A value of 1 was added to each RT prior to the log-transform so that the lower-bound of transformed values would be 0. Performance measures are presented for young adults on the left panels, and for older adults on the right panels. Error bars represent standard errors.

Table 1

Strength conditions of the continuous associative recognition task. These within-subject conditions varied by the relative point at which intact pairs were recombined. Note that there could potentially be many trials between presentations of the same objects; this aspect of the task was not specifically constrained.

Weak	Medium	Strong
New	New	New
Recombined	Intact 1	Intact 1
Intact 1	Recombined	Intact 2
Intact 2	Intact 2	Recombined

Table 2

Summary of free model parameters.

Memory	ρ	Context change rate
	α	Item-context associative learning rate
	κ	Scales negative prediction error learning
	λ	Maximum familiarity strength of repeated items
	τ	Modulates familiarity sensitivity to recency
	γ	Scales retrieved context mismatch strength
	ν	Baseline novelty strength supporting a “new” response
Decision	a	Decision threshold
	w	Decision bias toward “new” v. “old” responses
	t_0	Non-decision time

Table 3

Strength conditions for a hypothetical variant of the CAR task. In this design, repetitions of the recombined pairs (i.e., Recombined 2 pairs) should receive an “old” response, similar to repeated Intact pairs.

Weak	Medium
New	New
Recombined 1	Intact
Recombined 2	Recombined 1
Intact	Recombined 2