Probing the origins of subjective confidence in source memory decisions in young and older adults: A sequential sampling account

Kevin P. Darby

Department of Psychology, Florida Atlantic University

Jessica N. Gettleman, Chad S. Dodson, Per B. Sederberg

Department of Psychology, University of Virginia

Last updated: September 6, 2024

Author Note

Correspondence: Kevin P. Darby Email: darbyk@fau.edu The data included in this work were reported in a previously published article (Dodson, Bawa, & Slotnick, 2007). The results of this project were partially reported in oral presentations at the 2021 meeting of the Psychonomics Society and the 2021 Context and Episodic Memory Symposium (CEMS). 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. The data and analyses presented in this work are publicly available at https://osf.io/7yxpb/. The study materials are available upon reasonable request.

Abstract

Subjective confidence is an important factor in our decision-making, but how confidence arises is a matter of debate. A number of computational models have been proposed that integrate confidence into sequential sampling models of decision-making, in which evidence accumulates across time to a threshold. An influential example of this approach is the relative balance of evidence (RBOE) hypothesis, in which confidence is determined by the amount of evidence for the choice that was made compared to the evidence for all possible choices. Here, we modify this approach by mapping distance from a decision threshold to confidence via a sigmoid function. This allows for individual differences in bias toward lower or higher levels of confidence, as well as sensitivity to differences in evidence between choices. We apply several variants of the model to assess potential age differences between young and older adults in source memory decision-making in an existing dataset (Dodson, Bawa, & Slotnick, 2007). We compare our model to the RBOE approach, and the results indicate that the sigmoidal method substantially improves model fit. We also consider models in which memory errors can arise from a misrecollection process that involves associating items with the incorrect source, a process that has been proposed to account for age differences in source memory confidence and accuracy, but find no evidence that misrecollection is necessary to account for the results. This work provides a viable model of subjective confidence that is integrated with well-established models of decision-making, and provides insights into effects of aging on source memory decisions.

Public significance statement: This study presents a theory of how subjective levels of confidence in memory decisions are determined in young and older adults by comparing evidence for different choice options. The findings suggest that the way confidence is calculated differs between people and may change across the lifespan.

Keywords: Confidence, Source Memory, Aging, Sequential Sampling Models, Leaky Competing Accumulator Model

Probing the origins of subjective confidence in source memory decisions in young and older adults: A sequential sampling account

Introduction

Subjective confidence is an aspect of metacognitive monitoring that plays a key role in many areas of life. For example, the confidence of eyewitnesses can be predictive of lineup identification accuracy and can influence juror perceptions of guilt (Gettleman et al., 2021; Slane & Dodson, 2022; Wixted & Wells, 2017). In addition, the confidence of a student in their knowledge of class material is likely to influence how much they study the material (Robey et al., 2017), which may in turn impact their performance on examinations. More generally, research has demonstrated that our subjective feelings of confidence can shape how we learn and explore the world around us (Carlebach & Yeung, 2020; Dautriche et al., 2021; Destan & Roebers, 2015), and even how we make financial decisions (Samanez-Larkin et al., 2023). It is important to understand, then, how we determine the confidence we have in our decisions.

In addition to understanding how subjective confidence is generated, an important consideration is whether and how these mechanisms vary across the lifespan. As an example, prior work has suggested that older adults are often less able than young adults to accurately *calibrate* their confidence so that it closely matches their objective performance, and are more prone to making errors with high confidence, on some episodic memory tasks (see Dodson, 2017, for a review), suggesting the possibility of age differences in the mechanisms underlying confidence. In the current work, we present a computational model of subjective confidence in the domain of source memory and apply this framework to better understand developmental differences between young and older adults. We begin by discussing aging effects on source memory and confidence. We then examine existing models of subjective confidence before introducing our computational framework, which we apply to a dataset probing source memory and confidence in young and older adults that was presented previously in published research (Dodson, Bawa, & Slotnick, 2007).

Aging effects on source memory and confidence

A great deal of work has found that older adults tend to perform less accurately than young adults on tasks that probe episodic memory, or memory for specific, personally experienced events and their spatiotemporal contexts (Cansino et al., 2020; Castel & Craik, 2003; Craik, 1968; Darby & Sederberg, 2022; Golomb et al., 2008; Greene et al., 2022; Korkki et al., 2020; Naveh-Benjamin, 2000; A. D. Smith, 1977). For example, older adults are typically less able to remember the *source* of information they have learned, such as who told them a particular fact (Dodson, Bawa, & Slotnick, 2007; Dodson, Bawa, & Krueger, 2007; Schacter et al., 1991). Older adults often show less of a deficit, if any, on other kinds of memory tasks, such as tests of item recognition (Fraundorf et al., 2019; Rhodes et al., 2019), vocabulary (Verhaeghen, 2003), or general knowledge (Baltes et al., 1999; Dodson, Bawa, & Krueger, 2007). Although a number of theories have been proposed to explain episodic-specific memory deficits in older adults (Amer et al., 2022; Benjamin, 2010; Craik et al., 2010; Greene & Naveh-Benjamin, 2023; Salthouse, 1996; Stephens & Overman, 2018), a popular idea is that older adults have a deficit in *associative* memory, in which multiple elements of an experience are bound, or linked, whereas their memory for individual elements or items is largely intact (Chalfonte & Johnson, 1996; Naveh-Benjamin, 2000; Oberauer & Lewandowsky, 2019; Ratcliff & McKoon, 2015). For example, Naveh-Benjamin and colleagues (2004) presented young and older adults with face-name pairs before testing recognition memory for individual faces, individual names, and the associative pairings between them. The authors found a greater difference in performance between the age groups for the associative test compared to the face or name recognition tests, consistent with the hypothesis of a deficit specific to associative memory in older adults.

Other evidence suggests that older adults may differ from young adults not only in their episodic memory performance, but also in their ability to calibrate their confidence to their performance in episodic memory tasks. Indeed, even when performance is equated between age groups through various manipulations (e.g., number of repetitions during encoding, length of delay, etc.), older adults may exhibit inferior calibration between their subjective level of confidence and objective level of episodic memory performance compared to younger adults (Dodson, Bawa, & Krueger, 2007). In other words, older adults are often less able to match their level of confidence to the accuracy of their episodic memory. Often, older adults are more overconfident in their responses, and are more likely to express high confidence for memory errors (Dodson, Bawa, & Krueger, 2007; Fandakova et al., 2013; Greene et al., 2022; Shing et al., 2009). This monitoring deficit suggests that differences in confidence may not be due only to differences in associative binding ability between young and older adults, since they persist even when memory performance is equated between these two groups (Dodson, Bawa, & Krueger, 2007).

One theory that has been proposed to account for age differences in episodic memory performance and confidence calibration is the *misrecollection* account (Delhaye et al., 2018; Dodson, Bawa, & Slotnick, 2007; Dodson, Bawa, & Krueger, 2007; Dodson & Krueger, 2006; Shing et al., 2009), which suggests that older adults may not have a deficit in the ability to form associations per se, but may instead have a stronger tendency than younger adults to bind together the *wrong* information. In a source memory task, for example, in which participants are presented with items that should each be associated with one of two sources, older adults may be more likely than young adults to incorrectly form an association between an item and a source that was presented with different items. We will test this hypothesis in the current work while presenting a novel model that accounts for memory decisions, response times (RTs), and confidence judgments, with or without a misrecollection mechanism, as we will explain in the Results section. Before presenting this model, we discuss existing models of decision-making and confidence.

Mechanistic models of decisions and confidence

Signal detection models. One popular framework for modeling confidence is signal detection theory (SDT) (Green & Swets, 1966). SDT models assume that both memory performance and confidence judgments arise from distributions corresponding to levels of signal and noise, which are situated along an axis representing a decision space. SDT models have often been applied to recognition and source memory decisions (Dodson, Bawa, & Slotnick, 2007; Slotnick et al., 2000; Wixted, 2007). For example, for “New” versus “Old” decisions in a recognition memory task, the signal distribution is typically assumed to correspond to memory strengths for studied items, whereas the noise distribution would correspond to baseline familiarity for unstudied items due to prior experience. Each tested item is assigned to the “New” or “Old” response depending on a criterion estimated from the data. SDT models may be extended to account for confidence by assuming that the signal and noise distributions are further divided into segments corresponding to different levels of confidence. For example, a criterion may be placed that differentiates between medium and high confidence “old” decisions.

An SDT model of source memory and confidence was applied by Dodson, Bawa, and Slotnick (2007), who presented young and older adults with a task in which they studied a list of statements, each of which was made by either a male or female speaker. On each trial of a subsequent memory test, participants were asked to indicate in a three-alternative forced choice task whether a given statement was novel (“New”), or whether the statement had been presented by the male source or the female source (Source 1 or Source 2). Following each memory decision, participants provided a confidence rating.

The authors analyzed source memory performance and confidence judgments with an SDT model. The model was designed to test the aforementioned misrecollection account of age differences in source memory performance and confidence. In a standard SDT model, the two sources were associated with distributions of memory strength, such that strengths for the male source would be higher for statements that had been associated with that source, and likewise for the female source. In a model of misrecollection, an additional distribution was included for each source that corresponded to the *incorrect* response. For example, a subset of statements from the male source were presumed to have higher memory strength for the *female* source. Whereas the standard SDT model was able to account for the young adults' performance, the misrecollection model provided a superior fit to the data of older adults, suggesting that older adults were more likely to form associations between items and the incorrect source.

In the current work, we revisit the misrecollection account of the findings of Dodson, Bawa, and Slotnick (2007) with a different modeling framework that provides an account of the latent processes underlying RTs in addition to choices (i.e., "New," Source 1, or Source 2) and confidence values. In what follows, we discuss existing approaches for modeling subjective confidence before introducing our own framework.

Sequential sampling models. Although SDT models can account for memory decisions and confidence ratings, they are not typically designed to account for RTs (although see DeCarlo, 2021 for work incorporating RTs into SDT models), which, along with confidence judgments, often provide important information to adjudicate between theories of memory and cognition more generally (see Ratcliff & Starns, 2013, for discussion). A modeling framework that is often applied to account for both choices and RTs, and sometimes confidence, as we discuss below, is the sequential sampling model (SSM) framework. SSMs instantiate decision-making as a process of noisy evidence accumulation, whereby information about a stimulus is sampled across time. This information provides evidence for different response options, which is integrated across time and stored in so-called evidence accumulators. Once evidence in support of a choice crosses an estimated threshold, that decision is made; the amount of time it took to cross the threshold determines the decision time, which is added to an estimate of time required for perceptual and motor processing (called nondecision time, or t_0) to determine the simulated RT. Evidence accumulates across time steps according to a constant "drift rate" (estimating the quality of evidence provided by a stimulus), which provides a mean level of evidence accumulation across time, as well as random noise. Each possible answer choice may be assigned its own accumulator, which stores the state of evidence across time, or in cases in which there are two response options, there may be a single accumulator that could cross one of two thresholds (Ratcliff, 1978; Stone, 1960). SSMs have often been applied to two alternative forced choice tasks (Falbén et al., 2020; Kirkpatrick et al., 2021; Ratcliff & McKoon, 2008; Weichart et al., 2020), including “old” versus “new” recognition judgments (Darby & Sederberg, 2022; Ratcliff & McKoon, 2015), as well as more complex decision spaces with many possible answers (Lohnas et al., 2015; Sederberg et al., 2008), or even continuous decision spaces (P. L. Smith et al., 2020; Zhou et al., 2021). An illustration of evidence accumulating separately for two response options up to a single threshold is provided in Figure 1, along with distributions of simulated RTs for each option.

One of the most influential SSMs is the diffusion decision model or DDM (Ratcliff, 1978; Ratcliff & McKoon, 2008), in which a single accumulator may cross one of two thresholds, corresponding to two different choices such as “new” or “old.” Because there is a single accumulator, positioned between two thresholds, gaining evidence for one choice is equivalent to losing evidence for the second choice. One study (Ratcliff & McKoon, 2015) applied this type of model to item and associative memory in young and older adults. To do so, the researchers defined drift rates that estimated memory strength for items, as well as the strength of associations between items. The study found larger differences between age groups for the drift rates corresponding to associative memory compared to item memory, supporting the hypothesis discussed above that older adults have a memory deficit specific to associative binding (Naveh-Benjamin, 2000). Older adults also showed evidence of slower non-decision times, and higher decision thresholds; both of these findings are consistent with a great deal of other work applying diffusion models to young and older adults in a variety of task domains (see Theisen et al., 2021 for a meta-analysis), suggesting domain-general age differences in these decision-making processes.

Although the DDM has been highly influential and remains a popular model in a variety of cognitive domains, many other models within the general SSM framework have been proposed. One particularly powerful SSM is the leaky, competing accumulator (LCA) model (Usher & McClelland, 2001). Instead of a single accumulator that can cross one of two

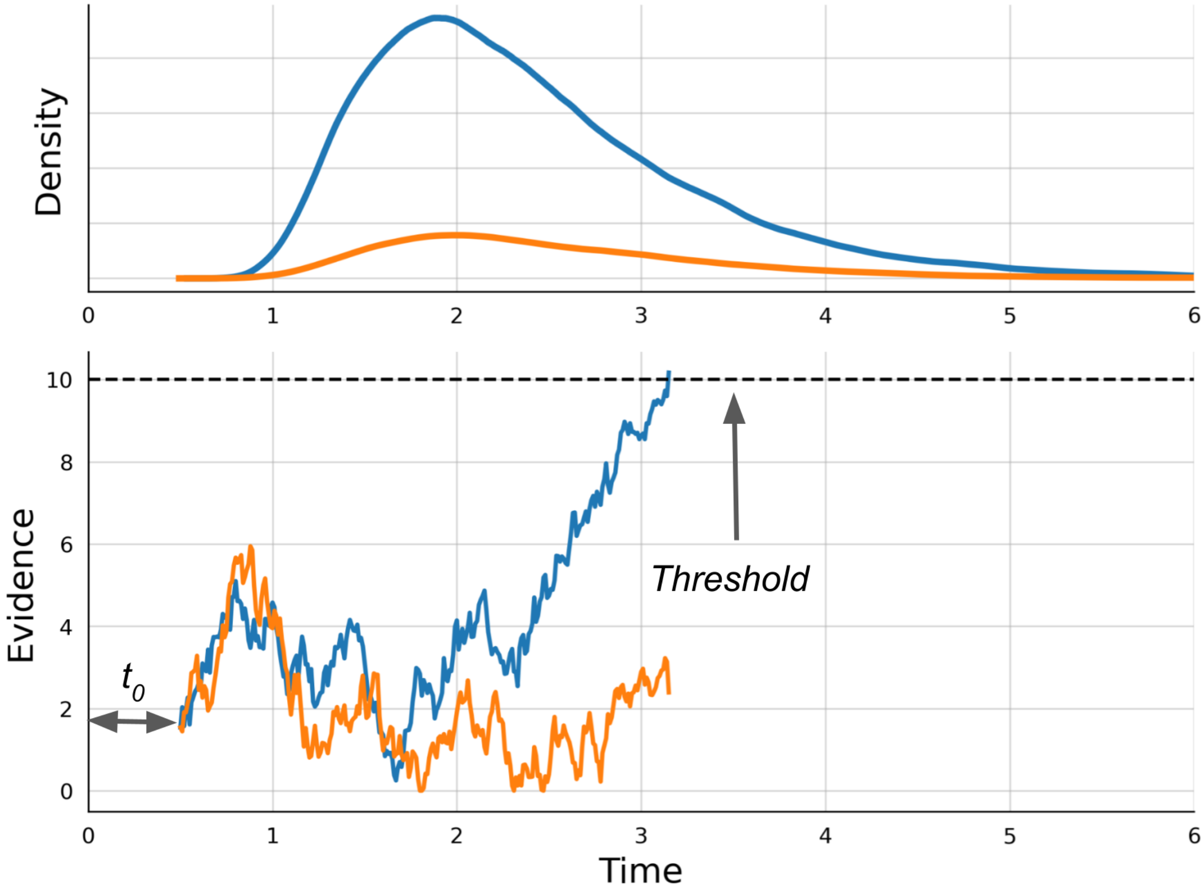

Figure 1. *Example of evidence accumulation in a sequential sampling model.* In this model, evidence stochastically accumulates separately for each possible response option, illustrated as blue and orange lines in the lower plot. Each response time (RT) is defined as a summation of the decision time, corresponding to the time taken for an accumulator to reach an estimated decision threshold, and nondecision time (t_0), which estimates the perceptuomotor processes required to perceive the stimulus and execute the response. The distributions in the upper plot show the distributions of RTs across many thousands of simulated choices, and are scaled by the proportion of each response.

thresholds, in LCA, each possible choice is associated with its own accumulator; whichever accumulator reaches the decision threshold first becomes the choice. In Figure 1, for example, the decision corresponding to the blue line reached the threshold, ending the decision-making process. Perhaps more importantly, LCA includes two sources of inhibitory dynamics that are not present in most SSMs: leak and lateral inhibition. Leak is a self-inhibitory mechanism, in which each accumulator encounters more resistance as it gains more evidence, such that eventually evidence will plateau given a sufficiently high threshold. In addition, due to lateral inhibition, the accumulators partially inhibit each other based on the amount of evidence that has been accumulated for each choice, such that as one accumulator gains more evidence, it has a greater inhibitory impact on the other accumulators. In LCA, therefore, accumulators exhibit partial dependence on each other, but are not perfectly anticorrelated as they are in the DDM (see Kirkpatrick et al., 2021 for discussion). In the current study, we apply LCA as the basis of our confidence framework, although our approach to determining confidence could be applied to many other SSMs as well.

Confidence in sequential sampling models. Although LCA and other SSMs provide a well-supported computational framework for the dynamics of decision-making that produce choices and RTs, a critical component of decision-making is metacognitive monitoring. One aspect of metacognitive monitoring that is often studied in the literature is subjective levels of confidence in one's decisions. A number of models of confidence within the SSM framework have been proposed (Desender et al., 2021; Hellmann et al., 2023; Kiani et al., 2014; Moreno-Bote, 2010; Pereira et al., 2020; Pleskac & Busemeyer, 2010; Ratcliff & Starns, 2009, 2013; Yu et al., 2015).

An idea that has been especially influential in SSM-based models of confidence has been that confidence depends on the state of accumulated evidence for different possible choices. If there is a great deal of evidence in support of both the "winning" response option (i.e., the accumulator that reached the threshold) as well as the "losing" option(s), confidence might be expected to be low. If there is strong evidence for the winning option and little evidence for other options, in contrast, confidence in the decision would be expected to be high. This idea is called the balance of evidence hypothesis (Vickers, 1978; Vickers, 1979). Note that, in practice, confidence is a function of the evidence for the *losing* response option(s). To see why, recall that a decision is made when an accumulator reaches a threshold. As a result, the evidence for the winning response option is always equal to the threshold at the time of the decision, which is typically constant across trials. Unless the threshold is allowed to vary, or evidence is allowed to continue to accumulate after it reaches the threshold (Pleskac & Busemeyer, 2010), the evidence for the winning response option will always be equal to the threshold for every decision. As a result, evidence for the winning option has no variability across decisions and is uninformative with regard to confidence. However, there is variability in the evidence accumulated for the other response option(s), which allows for variability in confidence judgments across decisions.

The balance of evidence hypothesis was originally implemented for a two-choice task by simply taking the difference between evidence for Choice A compared to Choice B: $C_A = x_A - x_B$, where C_A is confidence for A, and x_A and x_B are the final states of evidence for choices A and B, respectively. For example, if the Choice A accumulator has reached the threshold at $x_A = \alpha = 5$ (where α is the threshold), and at that moment the evidence for choice B is $x_B = 1.5$, confidence would be calculated as $C_A = 5 - 1.5 = 3.5$.

One limitation of this approach is that the threshold value will determine the confidence values that are possible, making it difficult to make comparisons in confidence between individuals or groups that may be estimated to have different thresholds. If one individual were estimated to have a threshold value of 5, for example, confidence values could range from 0 (if the evidence for both responses were equal) to 5 (if there was no evidence for the alternative choice). By contrast, an individual with a threshold value of 15 could have confidence values anywhere between 0 and 15. It was therefore unclear how to compare confidence values across individuals with different thresholds. Merkle and Van Zandt (2006) proposed a solution to this problem, called the *relative* balance of evidence (RBOE) approach, in which confidence values are standardized by the total amount of evidence across all accumulators: $C_i = \frac{x_i}{\sum_i^j x_i}$. In the example above, $C_A = \frac{5}{5+1.5} \approx 0.77$. This standardization ensures that confidence values lie between 0 and 1 regardless of the threshold. Note that the RBOE approach is not limited to 2-choice tasks, and could be applied to tasks with any number of possible responses.

A potential drawback of the RBOE approach, however, is that it does not allow for individual or group differences in the *function* mapping evidence onto confidence. Although RBOE allows for comparing confidence values despite individual differences in the decision threshold, this framework assumes that the same relative values of evidence will always result in the same values of confidence. It seems reasonable to suppose, however, that individuals could differ in how evidence is mapped onto confidence. Consider the case of psychometric functions in the literature on psychophysics. Often, psychometric functions contain parameters estimating an individual's or group's *sensitivity* to some psychophysical stimulus attribute, such as brightness, as well as the possibility of *bias* toward making one response over another (Gold & Ding, 2013; McCarley & Yamani, 2021; Morgan et al., 2012). Similarly, in the current work, we assume that individuals may vary in the mapping between evidence and confidence. We do so by applying a sigmoid function that includes parameters to control the sensitivity of the function to different levels of evidence, as well as bias toward higher or lower levels of confidence overall. Similar to RBOE, the function compares evidence for the winning choice to all other choice options, but it allows for potential individual differences in the function itself. We describe this model in detail in the Results section.

Current work

In this work, we propose a computational model of how subjective confidence could arise from the decision-making process used to identify the source of an item or determine its novelty. The theory is built upon LCA and extends the RBOE framework by assuming that a sigmoid function is used to map differences in evidence between accumulators onto confidence values. We compare the sigmoid-based method to an RBOE-based model that does not allow for individual differences in sensitivity or bias. Because misrecollection of incorrect associations has been proposed as a mechanism for older adults' deficits in source memory in addition to miscalibration between source memory accuracy and subjective confidence (Dodson, Bawa, & Slotnick, 2007; Dodson, Bawa, & Krueger, 2007; Shing et al., 2009), we include model variants in which items may be bound to the incorrect source on a proportion of trials. We test how well these models are able to account for young and older adults' source memory performance and confidence responses in an existing source memory dataset (Dodson, Bawa, & Slotnick, 2007). We emphasize, however, that the models we instantiate are not intended to test specific mechanistic hypotheses about the memory process, and merely specify that memory decisions in the source memory task are based on a combination of memory for items and their sources, the strengths of which we estimate with free parameters. We test the misrecollection account by allowing source memory strength to support the incorrect source, as we describe in detail below.

Method

The data we used to assess our models of confidence were presented by Dodson, Bawa, and Slotnick (2007). This paper presented two experiments, each of which was designed to equate source memory performance between young and older adults, as described below. In what follows, we describe the methods and results of the two experiments concurrently. Following this, we present a series of models designed to account for participants' choices, RTs, and confidence judgments in both experiments.

Participants

Eighteen participants participated in each of five datasets across two experiments conducted by Dodson, Bawa, and Slotnick (2007), for a total of 90 participants overall. The young adult participants were between 18 and 23 years of age, with mean ages between 20 and 21 across the experiments, whereas the older adults were between 60 and 80 years of age, with mean ages between 66 and 70 (see Dodson, Bawa, & Slotnick, 2007 for details).

Stimuli & Procedure

In what follows, we present an overview of the source memory paradigm before describing the specific manipulations between datasets. All participants were presented with statements (e.g., "Al Capone's business card said he was a used furniture dealer") in an encoding phase. Half of the statements were associated with the same male source, and half were associated with the same female source, randomly intermixed with the constraint that no more than two consecutive statements could be from the same source. Each statement was presented as visual text for 7 s, and was simultaneously spoken aloud by either the male or female source. Each statement was accompanied by a picture of the speaker, along with his/her name. Participants were told to pay attention to the speaker of each statement as they would later be tested on their memory. Each participant was presented with a total of 88 statements, half of which were associated with the male source and half with the female source. The first four and the last four statements were discarded to attenuate potential primacy and recency effects.

Following a delay (the length of which differed between experimental conditions, as described below), the studied statements were presented to participants again in the test phase. The memory test included an additional 40 statements that were not presented in the study phase. On each test trial, participants were given a 3 alternative forced-choice task, in which they were asked to respond whether the statement was “New” (i.e., not presented during study), or had been spoken by the male source, or by the female source. No time limit was imposed for this decision, and participants were instructed to take as much time as needed to answer each question. Following each memory judgment, participants were asked to rate their confidence in their response on a 6-point scale between 50 and 100 (i.e., 50, 60, 70, 80, 90, or 100), with 50 indicating a guess and 100 indicating complete confidence.

Although the general procedure was the same for all participants, there were variations designed to equate source memory accuracy between young and older adults. In Experiment 1, both young and older adults were presented with each statement once during the study phase, but young adults were tested after a 24 h delay, whereas older adults were tested after just 5 minutes. We refer to these as the Y_d and O_1 groups, respectively. In Experiment 2, all participants were tested after a 5 m delay, but young adults only studied each statement once, whereas one group of older adults studied each statement twice, and another group studied each statement three times. We refer to these Experiment 2 groups as Y_1, O_2, and

O_3, respectively. Note that the O_1 and Y_1 groups performed the same experimental procedure across experiments. See Table 1 for a summary of each experiment and dataset. See Dodson et al. (2007) for full details of the experimental procedure.

Table 1

Summary of experiments and datasets. h = hours; m = minutes

Experiment	Age	Label	Delay	Stimulus presentations
1	Young	Y_d	24 h	1
	Older	O_1	5 m	1
2	Young	Y_1	5 m	1
	Older	O_2	5 m	2
	Older	O_3	5 m	3

Transparency and openness

The data analyzed in this research, along with model and analysis code, are available via the Open Science Framework at https://osf.io/7yxpb/. The data were analyzed using Python 3.917. The primary Python libraries used to analyze the data were Pingouin 0.5.3 (for conventional analyses) and RunDEMC 01.0 (for model-fitting; https://github.com/compmem/RunDEMC). This study was not preregistered.

Results

Because we were interested in modeling RTs in addition to choices and confidence judgments, we removed data from trials with very long or outlier RTs before analyzing the results. As stated above, there was no time limit for participants to respond to stimuli in the test phase, such that some very long RTs ($>$ 20 s) were observed, which we reasoned could bias outlier detection. We assumed that such long RTs were due to noise, and hence discarded trials with RTs above 20 s. This resulted in the removal of 1.4% of trials in the Y_d group, 3.0% of trials in the O_1 group, 0.7% of trials in the Y_1 group, 2.5% of trials in the O_2

group, and 1.7% of trials in the O_3 group. Following this, we removed trials with RTs that could be considered outliers. Specifically, for each participant we performed a Box Cox transformation of that participant's RTs, and then removed trials that were more than 2.5 standard deviations away from the mean of the transformed data. We then repeated this procedure until no outliers were detected for that participant. This procedure resulted in the removal of 1.1% of Y_d data, 0.1% of O_1 data, 0.7% of Y_1 data, 0.6% of O_2 data, and 0.3% of O_3 data. Following these pruning procedures, between 96% and 99% of all trials for each group were included in the analysis.

Conventional statistical comparisons

In their original paper, Dodson, Bawa, and Slotnick (2007) analyzed differences between young and older adults within each experiment by assessing hit rates, false alarm rates, and the SDT recognition metric *d*‘, along with SDT models to test the misrecollection account. Here, to assess the data with conventional statistical approaches, we conducted a series of analyses of variance (ANOVAs) and post-hoc comparisons of choices, RTs, and confidence judgments. We did so with the Python library Pingouin (Vallat, 2018). Our general approach was to compare performance between groups within the two experiments separately. For each analysis we additionally compared the Y_1 and O_1 groups across the experiments, as the procedure for young and older adults was identical in these datasets. Because the primary goal of the current manuscript was to develop and test a computational model of confidence in memory-guided decisions, we do not focus on these more conventional analyses here, but we include the details of the analyses and results in the Supplemental Material. However, the general trends in the data are important to establish, so here we provide a verbal summary of these statistical findings.

We first assessed correct rejection rates (i.e., accurate "new" responses to novel items), and hit rates (i.e., correctly identifying an item as studied by responding with one of the sources, regardless of whether the source was accurately identified). These metrics provide measures of how well participants could identify item-based novelty and identify repetition of items. Each group's performance on these metrics is shown in Figure 2. The statistical comparisons between groups, presented in the Supplemental Material, suggested that older adults were not impaired in detecting novel items or recognizing studied ones. Indeed, as was reported by Dodson, Bawa, and Slotnick (2007), older adults outperformed young adults in terms of correct rejection rates in Experiment 1, and in terms of hit rates in Experiment 2. However, these differences were not significant when comparing the Y_1 and O_1 groups that completed the same procedure, suggesting that the differences were likely due to the methodological advantages provided to older adults that were designed to equate source memory between the groups.

Although older adults performed well in these experiments in terms of item-based performance, older adults are known to struggle with source memory (Dodson, 2017; Old & Naveh-Benjamin, 2008; Schacter et al., 1991). Dodson, Bawa, and Slotnick (2007) specifically designed Experiments 1 and 2 to equate source memory performance in young and older adults by disadvantaging young adults with a 24-hour delay compared with a 5-minute delay (Experiment 1) or by providing them with only a single stimulus presentation compared to 2 or 3 presentations (Experiment 2). To quantify source memory performance, the researchers calculated the paired-source conditional source-identification measure or PSCSIM (Murnane & Bayen, 1996). This metric is calculated as the proportion of studied-item trials in which the correct source was identified, compared to the overall hit rate (i.e., the proportion of studied statements for which either the correct or incorrect source was identified). In other words, given that the participant correctly identified the statement as studied, the PSCSIM represents the probability of correctly identifying the statement's source. The PSCSIMs for each group are presented in Figure 2. As was reported by Dodson et al. (2007), the PSCSIMs did not differ between groups within each experiment, suggesting that source memory was successfully equated between young and older adults. However, older adults' PSCSIMs in the O_1 group were significantly lower than those of young adults in the Y_1 group, replicating past demonstrations of reduced source memory accuracy due to aging when young and older

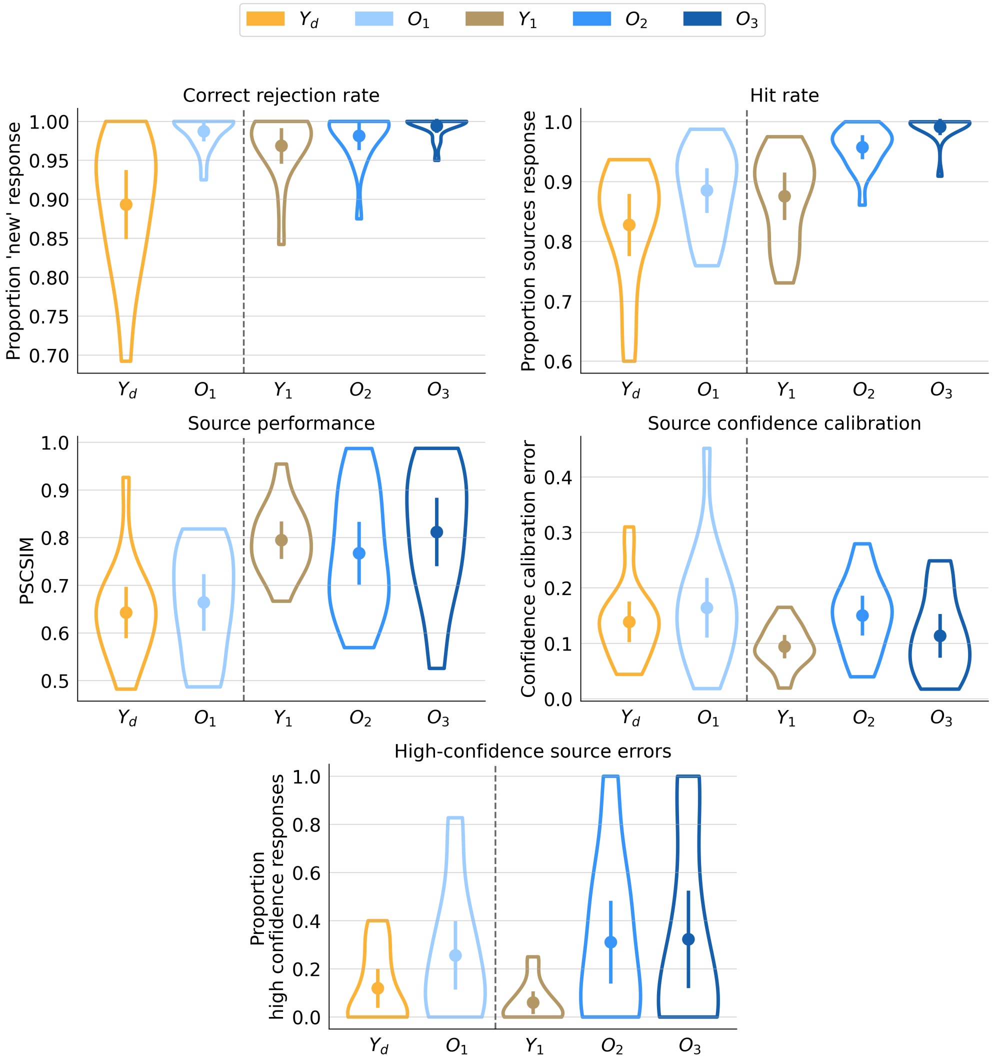

Figure 2. *Response accuracy and calibration with confidence.* Each violin shows the distribution of participants' observed performance, along with the mean level of performance for each group. The vertical dashed line on each plot separates the datasets of Experiments 1 and 2. The error bars indicate the 95% confidence interval for each distribution. PSCSIM = paired-source conditional source-identification measure.

adults completed the same procedure (Old & Naveh-Benjamin, 2008; Schacter et al., 1991).

We next examined RTs, which were not analyzed by Dodson and colleagues (2007). We anticipated that there could be differences in RTs between statements that were novel versus studied, and between trials that were correct versus incorrect. We therefore analyzed mean

RTs for correct rejections (i.e., correct “new” responses to new items), correct source responses (i.e., identifications of the correct source of studied items), and incorrect source responses (i.e., identification of the incorrect source for studied items). We did not assess RTs for false alarms to novel items or “misses” of studied items because these responses were relatively rare. The distributions of participants’ mean log-transformed RTs for each type of response are shown in Figure 3. The results of the ANOVAs and post-hoc comparisons presented in the Supplemental Material indicated that RTs tended to be slower in older adults compared to young adults. In addition, responses were generally faster for correct compared to incorrect responses. Older adults were faster to make correct rejection responses than they were to make correct source responses, suggesting that older adults may have had a bias to respond that an item was “new” more easily than to identify a studied item’s source. Interestingly, however, young adults did not show this pattern, and responded equally as quickly on correct rejection and correct source trials.

To assess subjective confidence, for each participant, we calculated the mean level of confidence for different types of responses in a series of analyses not reported by Dodson and colleagues (2007). Similar to the analysis of RTs, we assessed participants’ mean confidence for correct rejections, correct source responses, and incorrect source responses, as shown in Figure 3. If participants are able to effectively monitor their performance, they should be more confident for correct responses than for errors, and we found that this was the case for both young and older adults, in both experiments. Similar to differences in RTs, however, we found that each group of older adults was more confident in their correct rejection responses than they were in their correct source responses, whereas young adults did not show differences in confidence between correct rejections and correct source responses.

We also examined how well young and older adults were able to monitor their source judgments by calculating calibration error scores (Oskamp, 1962), which were not analyzed by Dodson and colleagues (2007):

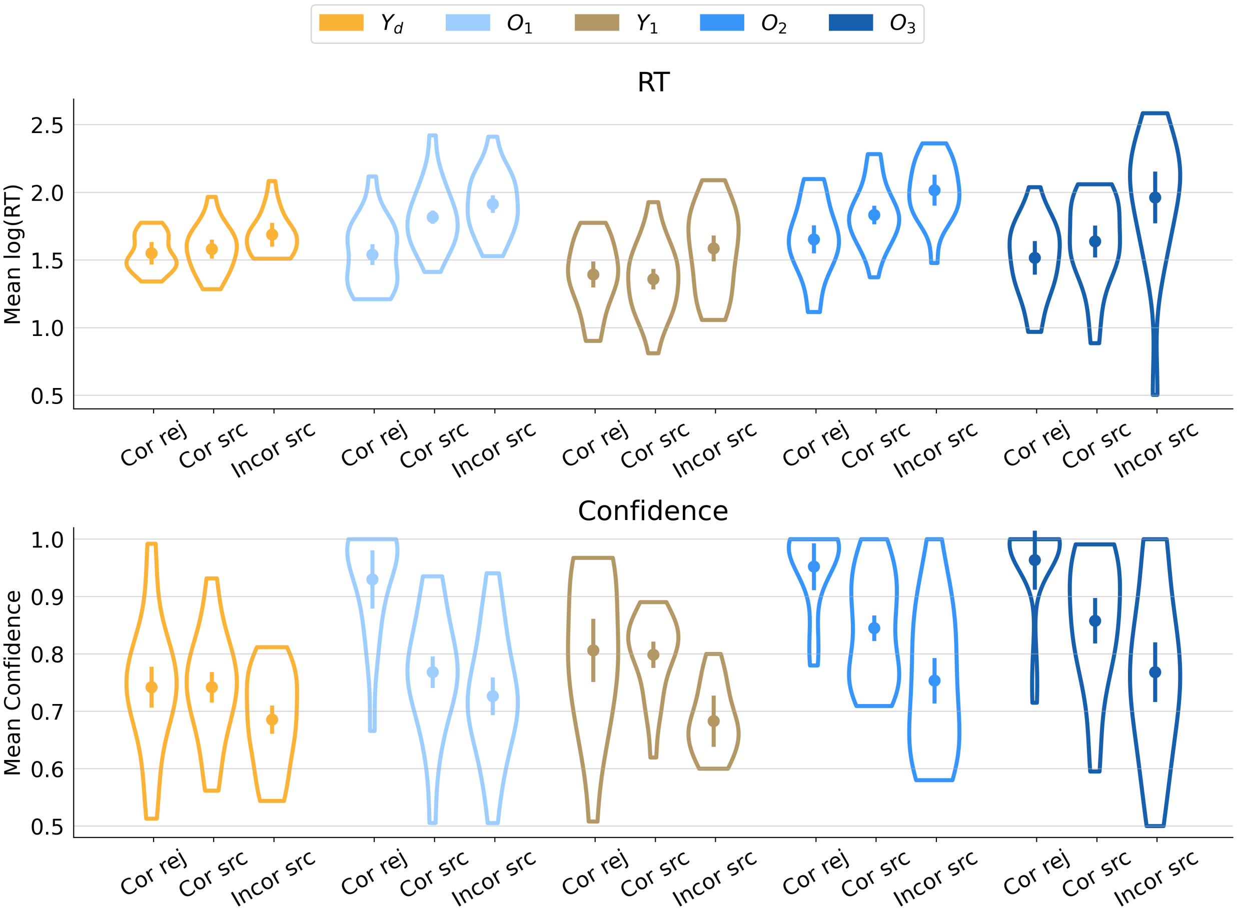

Figure 3. *Mean response times (top) and confidence judgments (bottom) for correct rejections, correct source responses, and incorrect source responses.* Each violin shows the distribution of individuals' observed performance, along with the mean level of performance for each group. The error bars indicate the 95% confidence interval for each distribution, corrected for within-subject comparisons within each group. Cor rej. = correct rejections; Cor src. = correct source; Incor src. = incorrect source.

$$E = \frac{\sum n_{H_i}|c_i - a_i|}{N_H}. \tag{1}$$

The goal of this metric is to assess to what extent participants' subjective levels of confidence differ from their objective task performance. According to Equation 1, calibration error (E) is calculated by summing the absolute value of the difference between confidence

(c) and source accuracy (a) for each confidence level i, scaled by the number of hits at that level of confidence (n_{H_i}). This scaled sum is then divided by the total number of hits (N_H). If participants are perfectly calibrated, each level of subjective confidence that they report would be matched by their objective accuracy across all trials with the same reported confidence. To measure calibration for source memory specifically, we applied this metric to trials in which participants chose either the correct or incorrect source in response to studied statements. Because participants chose one of the two sources in these trials, chance accuracy would correspond to a proportion of 0.5. Recall that participants were told that the lowest level of confidence should correspond to guessing. We therefore assumed that perfectly calibrated responses would have a 0.5 proportion of accuracy when responding with the 0.5 level of confidence; a 0.6 proportion of accuracy when responding with the 0.6 level, et cetera, resulting in $E = 0$. By contrast, E values above zero indicate calibration error. Note that this metric does *not* take into account the direction of calibration error, such that accuracy levels of 0.2 or 0.8 would result in the same level of error for a confidence value of 0.5. The calibration scores for each group are shown in Figure 2. As we present in the Supplemental Material, we compared this calibration score between groups, and found that calibration error was lower in the Y_1 group compared to both the O_1 and O_2 groups, providing support for the idea that older adults have a deficit in calibrating subjective confidence to objective source memory accuracy. However, there was no difference between the Y_d and O_1 groups in Experiment 1, or between Y_1 and O_3 in Experiment 2, suggesting that procedural differences likely impacted calibration differences. Together, these results suggest that age and experimental manipulations were likely both important modulators of confidence-accuracy calibration in this study.

Finally, we also examined the incidence of source errors made with high confidence (i.e., confidence levels 9 or 10), which was not reported by Dodson and colleagues (Dodson, Bawa, & Slotnick, 2007). The theory of misrecollection predicts that older adults will be more likely to respond with high confidence when making incorrect source responses compared to younger adults. This difference was generally found, particularly when comparing age groups in Experiment 2 and when comparing age groups who performed the same procedure (i.e., O_1 vs. Y_1). Together, the findings on confidence-accuracy calibration and high-confidence errors demonstrate that confidence for source correct and source incorrect responses is less discriminating in older adults than in younger adults, particularly those who experience the same procedure (i.e., O_1 vs. Y_1).

To summarize, older adults did not exhibit a deficit in detecting novel items or recognizing studied items. In addition, the experimental manipulations successfully equated source memory between young and older adults by providing advantages to the latter group, although young adults' source memory was superior when the groups completed the same procedure. Older adults exhibited faster RTs and greater confidence for correct rejections compared to correct source responses, suggesting a potential bias in how older adults identify novel items compared to the source of studied items. Finally, older adults were more likely to have high confidence in incorrect source decisions than were young adults, and there was some evidence that older adults were less able to calibrate their confidence in their source memory decisions.

These differences in performance suggest that there are likely differences between young and older adults in processes related to memory, decision-making, and subjective confidence. However, the patterns in performance alone do not provide a mechanistic understanding of confidence judgments or of developmental change. In order to better understand how subjective confidence is determined in memory-based decisions, and how these processes could change due to aging, we sought to account for young and older adults' memory choices, RTs, and confidence judgments within a single computational framework.

Modeling framework

We implemented a series of computational models to better understand the mechanisms underlying subjective confidence in young and older adults' memory-guided decisions. Each of these models utilized the LCA framework of decision-making (Usher &

McClelland, 2001). As noted in the Introduction, LCA is a type of SSM, in which decision-making is based on the gradual and noisy accumulation of evidence, which is integrated across time up to a threshold. The rate of evidence accumulation for each choice i across time is driven by a “drift rate” ρ_i, which reflects the quality of evidence in favor of that choice. Importantly, LCA also includes passive “leak” of evidence and lateral inhibition between accumulators. The following differential equation is applied to integrate evidence for the accumulators on each time step:

$$dx_i = (\rho_i - \kappa x_i - \beta \sum_{j \neq i} x_j)\frac{dt}{\Delta t} + \eta\sqrt{\frac{dt}{\Delta t}}. \tag{2}$$

According to Equation 2, evidence for a given choice (x_i) changes at every time step based on the corresponding drift rate (ρ_i), which is modulated by leak and lateral inhibition, as well as noise. The leak of evidence, modulated by a free parameter κ, increases as the evidence x_i grows (i.e., as the evidence approaches the threshold). As was discussed in the Introduction, leak may be thought of as self-inhibition or passive decay of each accumulator. LCA also allows accumulators to impact each other through lateral inhibition. The inhibition of accumulator i as a function of the evidence supporting the other choices is modulated by parameter β. This mechanism creates a negative correlation between accumulators, such that as evidence for one accumulator increases, evidence for the other accumulators is more likely to decrease due to inhibition. The interested reader is referred to Usher & McClelland (2001) for detailed explanations of leak and lateral inhibition.

The combination of the drift rate, leak, and lateral inhibition drives the evidence accumulation process for each choice across time. In keeping with prior work (Brown et al., 2006; Kirkpatrick et al., 2021; Weichart et al., 2020), we applied the Euler method to discretize the function into time steps with step size dt, which was fixed to .01 seconds, and a time constant Δt, fixed to 0.1. At each time step, noise is added to the accumulation process via a draw from a normal distribution with a mean of 0 and standard deviation of 1, symbolized by η in Equation 2.

The evidence accumulation process stops when any accumulator reaches the decision threshold, at which point the choice option corresponding to that accumulator becomes the response. The number of time steps it took to reach the threshold determines the decision time, which is added to a free parameter estimating perceptual and motor processing time, t_0, called nondecision time, in order to simulate an RT. An illustration of this model is presented in Figure 4.

Bias in Decision Thresholds. Recall that the evidence accumulation process is completed whenever an accumulator reaches the threshold. A bias may be implemented in SSMs by including a different threshold for each response, such that more or less evidence is needed to make a given choice compared to other choices. In the current work, we estimated two thresholds: one corresponding to the "New" response and the other corresponding to both source responses. We allowed the thresholds for these responses to be different because, as mentioned above, the RT and confidence results analyzed in this study (Dodson, Bawa, & Slotnick, 2007) suggest that older adults may have adjusted their decision-making process between "new" and source responses.

We allowed for a potential bias by estimating one free parameter for the threshold for "New" responses, α_{new}, while a second free parameter, α_{prop}, determined the threshold for the two source accumulators as a proportion of α_{new}. Therefore, the threshold for source responses was calculated as $\alpha_{source} = \alpha_{new} \times \alpha_{prop}$. When $\alpha_{prop} = 1$, the two thresholds were the same, whereas when $\alpha_{prop} < 1$ the threshold for the sources was lower than for the "New" response, and when $\alpha_{prop} > 1$, the threshold for the source responses was higher. A higher threshold for the two sources would result in slower RTs overall when responding with the two sources, as well as a reduced probability of making source responses overall, as more evidence would be needed to make these responses.

Truncating Evidence, Baseline Drift Rate, and Starting Point of Evidence. Although evidence in this model is integrated across time up to a threshold, evidence values can potentially fall below zero due to noise and lateral inhibition. This is often considered

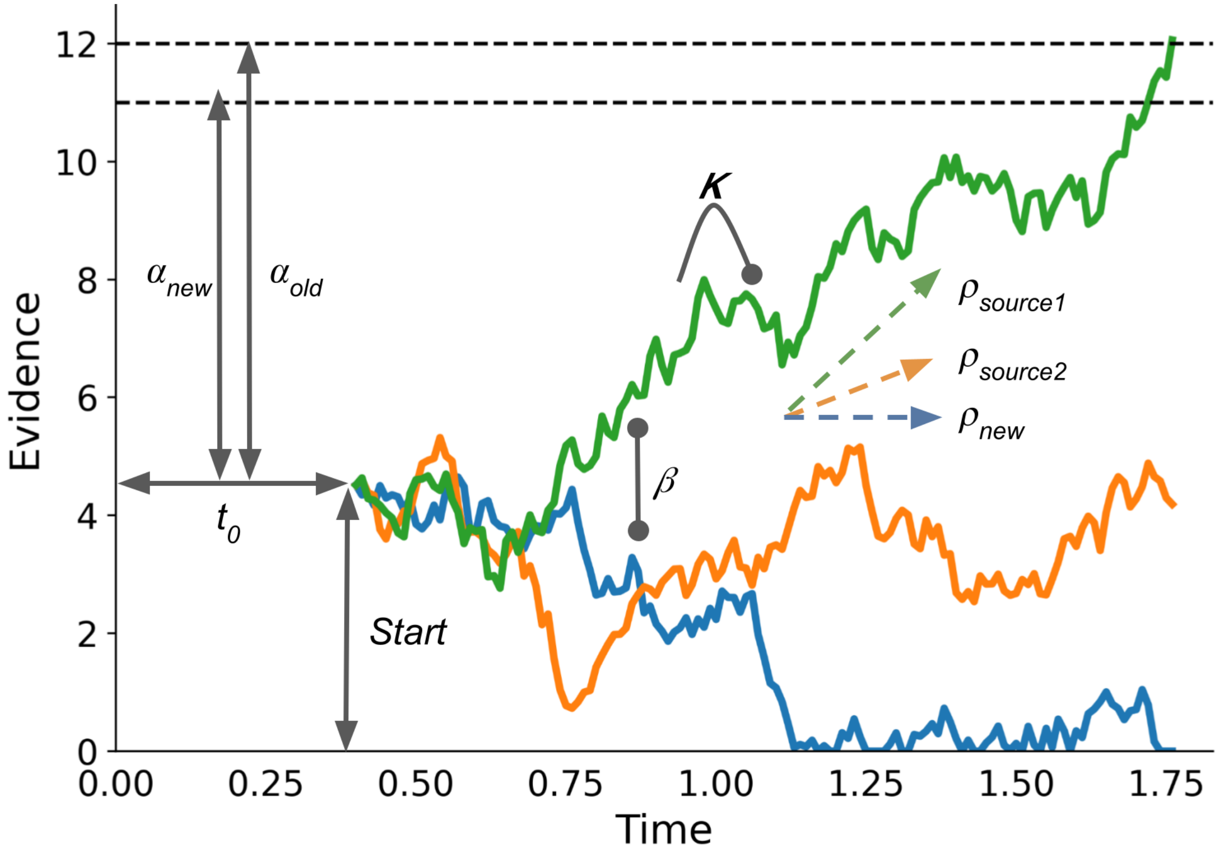

Figure 4. *Example of evidence accumulation in the LCA model framework as applied to the data of (2007).* The figure shows evidence stochastically accumulating separately for the "New" response (blue), Source 1 (green), and Source 2 (orange) across time (in seconds). Each response time (RT) is defined as a summation of the decision time, corresponding to the time taken for an accumulator to reach an estimated decision threshold, and nondecision time (t_0), which estimates the perceptuomotor processes required to perceive the stimulus and execute the response. Evidence accumulation for each response option is driven by a corresponding drift rate, along with noise, evidence leak (κ), and lateral inhibition between the accumulators (β).

suboptimal for multiple reasons. One reason is that negative evidence values introduce non-linearities due to lateral inhibition. Specifically, lateral inhibition could *facilitate* the activation of an accumulator if the sum of the evidence across the other accumulators were negative, driving the accumulator toward the threshold at a faster rate. In addition, the accumulation of evidence is often understood as analogous to neural firing rates, and because neurons cannot have negative firing rates, it may not be tenable to allow negative evidence values (Usher & McClelland, 2001).

We took two precautions against negative evidence values. First, we followed past work by simply truncating evidence such that it cannot be negative: $x_i \rightarrow max(x_i, 0)$ (Kirkpatrick et al., 2021; Usher & McClelland, 2001; Weichart et al., 2020). In addition, we followed past work by van Ravenzwaaij and colleagues (2012) to minimize negative evidence values in the first place by providing a baseline drift rate (ρ_b) for all accumulators, and by adjusting the starting point of evidence accumulation, $start_{x_i}$, according to the following equation:

$$start = \frac{\rho_b}{\kappa + (n-1)\beta},$$

where n is the number of accumulators (in this case, 3). This equation offsets effects of inhibition, such that the starting point of evidence accumulation for all choices is lowered as inhibition increases, which minimizes the probability of negative evidence values (see van Ravenzwaaij et al., 2012 for discussion). In order to reduce complexity of the model and facilitate model-fitting, we did not fit ρ_b as a free parameter, but fixed this value to 0.5 for all participants (fixing the value to 1.0 instead of 0.5 resulted in comparable findings).

To ensure straightforward interpretation of the bias in response threshold across choices, the starting point of evidence *start* was the same for all accumulators. The threshold parameters α_{new} and α_{source} were then added onto the starting point to define the thresholds $thresh_{new}$ and $thresh_{source}$: $thresh_{new} = start + \alpha_{new}$, and $thresh_{source} = start + \alpha_{source}$.

Memory strengths. Recall that in LCA, the drift rate of each accumulator supplies a mean drive toward a threshold, which reflects the quality of information available for that choice. In the present work, we assume that the drift rates are determined by the strength of memory supporting each response. We applied a simple measurement model to estimate memory for items as well as associations linking each item to its source, without attempting to specify trial-level encoding and retrieval processes (see Oberauer & Lewandowsky, 2019 for a similar approach to dissociating item and associative memory in young and older adults).

We first provide a verbal description of this approach before providing more formal equations. Recall that in the experiments analyzed here, there were three choice options on every trial: "New," Source 1, and Source 2. As noted above, we assumed that each of these response options would always have a baseline level of strength, parameterized as ρ_b, which we fixed at 0.5 for all participants, as mentioned above. This baseline strength drives evidence for all choices equally, so other sources of strength are needed in order for some response options to be more likely than others. We assumed that when presented with a studied statement in the test phase, the participant would respond based on a combination of memory for observing that item, as well as the strength of association between the statement and its source. Item-based strength, estimated with parameter ρ_f, is indicative of memory that the item was observed in the experiment. This knowledge would presumably allow the participant to be less likely to choose the "New" response option and be more likely to choose between the two source options. However, memory that one had seen the statement before would not help the participant determine the *correct* source. Additional strength for the correct source is needed, and this is supplied by the strength of association with the correct source, estimated with free parameter ρ_s. For the correct source option, the item and source-based strengths were summed together to determine the drift rate, whereas for the incorrect source, only the item-based strength contributed to the drift rate. Neither the item-based nor source-based memory strength was allowed to contribute to the drift rate for the "New" response to studied statements, such that this drift rate was determined solely by the baseline strength.

The drift rates for the three response options changed, however, when participants were presented with a novel statement that had not been presented during study. In this case, we assumed that evidence for the "New" response would be driven by the ability to detect novelty, estimated with another free parameter, ρ_n. However, for novel statements, participants should have no item-based or association-based memory strength for either of the two sources, since the item had not been presented during study. As a result, the drift rate for each of the two sources was limited to ρ_b.

As just described, the drift rate for each of the three choices (“New”, Source 1, Source 2) included a minimum value of ρ_b, which was fixed at 0.5, as well as sources of strength which depended on which choice was correct. The drift rates for each answer choice (i.e., “New”, Source 1, or Source2) were then modulated by free parameters estimating memory strength for items (ρ_f), associative memory strength for sources (ρ_s), and novelty-based strength (ρ_n), as follows:

$$\rho_{new} = \begin{cases} \rho_b + \rho_n & \text{if "New" correct} \\ \rho_b & \text{if either source correct} \end{cases}$$

$$\rho_{source1} = \begin{cases} \rho_b & \text{if "New" correct} \\ \rho_b + \rho_i + \rho_s & \text{if Source 1 correct} \\ \rho_b + \rho_i & \text{if Source 2 correct} \end{cases}$$

$$\rho_{source2} = \begin{cases} \rho_b & \text{if "New" correct} \\ \rho_b + \rho_i & \text{if Source 1 correct} \cdot \\ \rho_b + \rho_i + \rho_s & \text{if Source 2 correct} \end{cases}$$

Modeling confidence. We have thus far discussed how we applied LCA, in conjunction with a simple measurement model of item- and source-based memory strengths, in order to simulate choices and RTs. In what follows, we describe how we integrated confidence judgments into this model.

Recall that previous work has modeled confidence within the SSM framework by applying the RBOE hypothesis (Merkle & Van Zandt, 2006; Vickers, 1978), according to the following equation:

$$C_i = \frac{x_i}{\sum_i^j x_i},$$

where x_i is the amount of evidence for the winning accumulator, which is divided by the sum of evidence for all accumulators.

This approach has been influential for models of confidence within the SSM framework (Pleskac & Busemeyer, 2010). Importantly, RBOE assumes that the mapping between evidence and confidence is constant, such that the same proportions of evidence values for different choices will always lead to the same level of confidence. However, an interesting hypothesis is that this mapping could differ between individuals, and could perhaps change across the lifespan. This hypothesis was inspired by work in psychophysics mapping differences in physical stimulus properties to individuals' ability to perceive these differences. Often, psychometric functions are applied to estimate sensitivity to stimulus changes, as well as perceptual biases (Gold & Ding, 2013; McCarley & Yamani, 2021; Morgan et al., 2012). In the current work, we apply a similar approach to allow for individual differences in how participants map evidence for different possible decisions onto confidence for those decisions. Specifically, we propose that individual differences may be found in the sensitivity of confidence values to different levels of evidence, and the level of bias toward higher or lower confidence judgments overall.

To instantiate this idea, we used sigmoid functions, which are controlled by two free parameters. The first parameter, τ, controls the steepness of the function, such that higher values of τ result in a more step-like function between low and high confidence responses, whereas smaller values result in a more gradual transition from low to high confidence (see Figure 5). This parameter, then, allows for individual differences in the extent to which one is sensitive to differences in evidence when making confidence judgments. The other parameter, δ, does not change the shape of the function but allows it to shift, such that the same evidence values result in higher or lower confidence values with different values of δ (Figure 5). This parameter, then, may be thought of as instantiating a bias toward more or less confident responding overall. In the current work, we will compare this sigmoid-based approach to confidence derived from the standard RBOE approach. Formal model comparison allows quantitative assessment of whether the sigmoid approach is better able to fit the observed data while taking into account the increase in complexity due to the addition of two free parameters, τ and δ.

In addition to a different functional form, our approach differs from RBOE in that confidence is not based on values of evidence directly, but on the *distance* d_j of each "losing" accumulator from the corresponding threshold, where $d_j = thresh_j - x_j$. This change is useful due to the possibility of different thresholds for the "New" and source responses. Although one could construct a similar equation based on the magnitude of evidence for choice j (x_j), this would ignore the potential difference in thresholds between accumulators. The distance-based model assumes that it is not the value of evidence per se that drives confidence, but how close or far each losing accumulator is from the corresponding threshold at the time of the decision.

Because there were three response options in this task (New, Source 1, or Source 2), there are always two accumulators corresponding to "losing" choices. The central hypothesis of both the RBOE- and sigmoid-based approaches is that evidence for these choices will determine subjective judgments of confidence. Because there is more than one losing choice, however, a key question is how the evidence for the choices will impact the mapping between the decision space and confidence space. We instantiated different possibilities in a series of model variants.

The first model variant assumes that the distance of each accumulator contributes to confidence judgments separately via the following sigmoid function, which we refer to as the *separate sigmoid model*:

$$C_i = \frac{1}{1 + \sum_j^k e^{-(d_j+\delta)\tau}}. \tag{3}$$

Equation 3 derives the confidence supporting choice i, C_i, based on the distances of

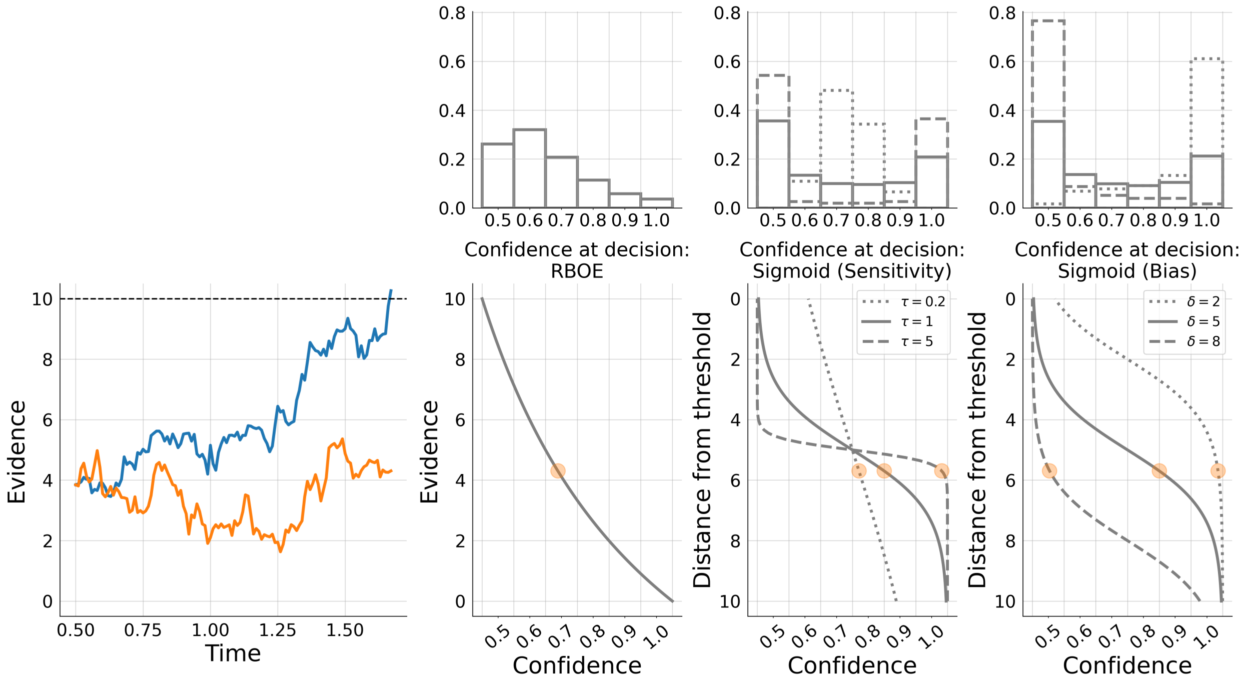

Figure 5. *RBOE and sigmoid functions mapping evidence accumulation onto confidence values for a decision with two choices.* The bottom row illustrates how, at the time when an accumulator crosses a decision threshold (bottom left), confidence may be calculated with RBOE based on the magnitude of evidence for the winning choice compared to the total amount of evidence in the system, including evidence for the losing choice. Alternatively, confidence may be calculated with the sigmoid approach, which is based on the distance of the losing choice from the threshold. Individual differences in evidence-confidence mapping can be accounted for with parameters determining the sigmoid function's shape (sensitivity) of the function, as well as its bias for higher or lower values of confidence. The top row shows the distributions of confidence values using the RBOE and sigmoid-based approaches. To create this figure, drift rates for the winning and losing accumulators were set to 0.4 and 0.2 respectively, plus a baseline drift rate of 0.5; the threshold was set to 10; κ was set to .05; β was set to .08; t_0 was set to 0.5.

each of the losing accumulators, d_j, from the corresponding threshold. Each losing accumulator j contributes separately to confidence through its own exponential function

$e^{-(d_j+\delta)\tau}$. A consequence of this is that if any distance d_j is small (i.e., if any losing accumulator is approaching the threshold), confidence will generally be low, even if the distance of the other accumulator is high. See Figure S1 of the Supplemental Material for an illustration of this model and the three alternative models described below.

An alternative hypothesis is that the most important factor is the total distance of all accumulators from their corresponding thresholds. We instantiated this idea with the following sigmoid, which we refer to as the *summed sigmoid model*:

$$C_i = \frac{1}{1 + e^{-(\sum_j^k d_j+\delta)\tau}}. \tag{4}$$

An important difference between this model and the separate sigmoid model is that if one accumulator is close to the threshold (i.e., has a small distance value d_j), it is still possible to have high confidence if the distance of the other accumulator is high (see Figure S1).

Another possibility is that only one distance is considered when determining confidence. For example, it may be the case that only the smallest distance is taken into account. If the distance of the accumulator nearest to the corresponding threshold is small, confidence will be low, irrespective of evidence for the other losing accumulator. In this model, then, both distances must be high in order to attain high confidence. This idea was implemented in what we refer to as the *minimum sigmoid model*:

$$C_i = \frac{1}{1 + e^{-(d_{min}+\delta)\tau}}, \tag{5}$$

where d_{min} is the smallest of the two distances. In this case, the accumulator with the greatest distance is ignored when it comes to calculating confidence (see Figure S1).

Finally, it is possible that only the maximum distance is taken into account, such that confidence is only low if both accumulators are close to the threshold. We implemented this idea in what we call the *maximum sigmoid mode*l:

$$C_i = \frac{1}{1 + e^{-(d_{max} + \delta)\tau}}. \tag{6}$$

Here, the accumulator with the smallest distance is ignored when calculator confidence (see Figure S1).

These four models are all similar in that they use a sigmoid as the mapping between decision distance and confidence, but they vary in what and how distances are taken into account, as just described. In addition to these models, we also included a standard RBOE model. Importantly, the RBOE and all of the sigmoid confidence functions result in continuous but bounded values of confidence. Confidence values stemming from any of the sigmoid approaches could be any value between 0 and 1. For the RBOE approach, the lowest values of confidence would result from evidence that is near the threshold for all three accumulators, such that the lowest value of confidence would approach 1/3[1], whereas confidence could be as high as 1 if the losing accumulators both have zero evidence. Although simulated confidence values are continuous between bounds, in the experiments considered here (Dodson, Bawa, & Slotnick, 2007), participants had to choose between six confidence values: 50, 60, 70, 80, 90, or 100 (in our figures and analyses we divided these numbers by 100 so that confidence would range from 0.5 to 1.0). To discretize confidence values simulated with the sigmoid or RBOE approaches, we then constructed six bins of confidence values that were spaced equally between the minimum and maximum values for each model, such that any simulated confidence value within a particular bin would result in the corresponding value of a six-point confidence scale, as implemented in the experiment. These bins are visualized in Figure 5.

[1] Because we allow for different thresholds for "new" and source responses, confidence values below 1/3 are possible in the RBOE framework. This can be the case if the "losing" accumulators have higher evidence values than the winning accumulator due to a higher threshold. In these cases, we assign the lowest confidence value.

Testing the Misrecollection Account. By fitting the models described above, we can estimate via free parameters whether novelty-based, item-based, and source-based memory strengths may differ between young and older adults, depending on the experimental manipulations implemented by Dodson, Bawa, and Slotnick (2007). However, this simple measurement model of memory does not test the *misrecollection* account of memory aging that was discussed in the Introduction. As a brief reminder, this account stipulates that older adults may not have a deficit in the ability to form associations per se, but may be more likely than young adults to form associations that are incorrect. In the source memory task by Dodson et al. (2007), for example, older adults could form an association between an item and an incorrect source that was also presented in the experiment. In the current modeling framework, we reasoned that this idea could be instantiated by switching which source's drift rate would be boosted by ρ_s for some proportion of repeated item trials. Because the drift rates for the two sources are always identical except for ρ_s, we simply took the simulated trials in which one of the two sources was selected (i.e., trials in which one of the source accumulators first crossed the threshold) and switched the response to the other source on some proportion of trials, controlled by a free parameter ϕ that could vary between 0 and 1. Importantly, while this changes the source judgment, it does not affect the RT or confidence in the response. When ϕ is above zero, some source responses will be switched. This allows for less accurate responding while retaining high levels of confidence (and fast RTs), as the memory strength for source information, ρ_s, supports the incorrect source accumulator on these trials. We added this additional mechanism to the RBOE model, as well as the best-fitting sigmoid model, which was the summed sigmoid model, as we report below. We included these models to test whether this additional mechanism improves the fit of the model compared to a standard model with no misrecollection.

Summarizing models and parameters. To examine effects of misrecollection and different confidence functions, we implemented seven different models. We included models to compare our sigmoid approach with RBOE for generating confidence, which allowed us to investigate the possibility of individual and group-level differences in the function mapping decision evidence onto confidence judgments. In addition, we fit models with and without the misrecollection mechanism described above, allowing us to investigate whether older adults' confidence judgments may be influenced by incorrectly binding items to the wrong source. All of the parameters included in this modeling framework are described in Table 2. In addition, a summary of the free parameters fit for each model is provided in Table 3.

Table 2

Summary of model parameters.

Process	Parameter	Description	Range of values
Memory	ρ_b	Baseline drift rate	0.5
	ρ_n	Novelty strength for correct "new" responses	(0,∞)
	ρ_f	Item-based memory strength for studied items	(0,∞)
	ρ_s	Source-based associative memory strength for studied items	(0,∞)
	ϕ	Probability of misrecollection	(0,1)
Decision	κ	Passive leak of evidence	(0,∞)
	β	Lateral inhibition or competition between accumulators	(0,∞)
	α_{new}	Decision threshold for "new" responses	(0,∞)
	α_{prop}	Bias (threshold for source responses as proportion of α_{new})	(0,∞)
	t_0	Non-decision time	(0, min RT)
Confidence	τ	Sensitivity to distance for confidence	(0,∞)
	δ	Bias for low vs. high confidence	(0,∞)

Table 3

Summary of the free parameters for each model variant. RBOE = relative balance of evidence; Sig. = sigmoid; misrec. = misrecollection; sep. = separate; max. = maximum; min = minimum; sum. = summed.

	RBOE	RBOE misrec.	Min. sig.	Max. sig.	Sep. sig.	Sum. sig.	Sum. sig. misrec.
ρ_n	X	X	X	X	X	X	X
ρ_f	X	X	X	X	X	X	X
ρ_s	X	X	X	X	X	X	X
ϕ	-	X	-	-	-	-	X
κ	X	X	X	X	X	X	X
β	X	X	X	X	X	X	X
α_{new}	X	X	X	X	X	X	X
α_{prop}	X	X	X	X	X	X	X
t_0	X	X	X	X	X	X	X
τ	-	-	X	X	X	X	X
δ	-	-	X	X	X	X	X
N. params	8	9	10	10	10	10	11

Model-fitting. We fit all models with hierarchical Bayesian techniques. To do so, we wrote custom scripts for the Python library RunDEMC (https://github.com/compmem/RunDEMC) (Turner et al., 2013), which implements a differential evolution Markov chain Monte Carlo (DE-MCMC) genetic algorithm. Gibbs sampling was applied to update hyper-prior distributions, and DE-MCMC was applied to update individual participant distributions (Turner & Van Zandt, 2014). We fit the model across 100 Markov chains, each of which included 5,000 iterations, the first 500 of which were fit with burn-in procedures. We implemented a “thinning” procedure by including every 7th sample of the final 2,000 iterations of each chain in the final posterior distributions for analysis. The purpose of this thinning was to reduce effects of autocorrelation within chains.

In order to implement our Bayesian model-fitting procedures it was necessary to be able to calculate the likelihood of the data given each set of parameter proposals. Unfortunately, a tractable likelihood was not available for integrating confidence judgments with LCA, so we approximated the likelihood of each proposal with probability density approximation (PDA) techniques (Turner & Sederberg, 2014). The PDA method involves simulating data many thousands of times for each parameter proposal to generate an approximated joint density for the choices, RTs, and confidence values, which we could then evaluate for each observed data point. Specifically, for every set of parameter proposals (i.e., a proposed value for each free parameter), we simulated 25,000 choices, RTs, and confidence values for each condition (i.e., the novel and studied conditions of the test phase), resulting in distributions of values for each of these dependent variables. By normalizing these responses (continuous for RT, and discrete for choice and confidence) within each condition, we could then jointly estimate a likelihood for each memory choice, RT, and confidence level by evaluating the observed data points in their matching estimated joint probability density functions.

The choice, RT, and confidence likelihoods were log-transformed and summed across conditions and choices to calculate the overall estimated log-likelihood for each parameter proposal, separately for every participant in a hierarchical model. Note that similar PDA-based likelihood estimation for choices and RTs have been conducted for the LCA model in prior work (Kirkpatrick et al., 2021; Weichart et al., 2020).

Model comparison

In order to quantitatively compare the different models, we calculated the Bayesian predictive information criterion (BPIC) (Ando, 2007) for each model and each participant [2]. The BPIC is a measure of model fit that takes into account the complexity of the model, similar to the Bayesian information criterion (BIC) or the Akaike information criterion

[2] The same overall pattern of results is obtained by employing the Bayesian information criterion (BIC), an alternative model-comparison technique.

(AIC). In order to compare different models, we mean-centered the BPIC values for each model within each participant, such that more negative values indicate a better model fit.

The results are presented in Table 4. Averaging the mean-centered BPIC values across participants within each dataset, it is clear that the sigmoid-based models were preferred over the RBOE models for every dataset. This suggests that the greater flexibility afforded by the parameters of the sigmoid provided a much better fit to the data, even when accounting for the greater complexity of the model due to the addition of two parameters. In addition, the summed sigmoid model fit slightly better in the majority of participants overall. We therefore conclude that this model was the preferred model overall for these datasets, but caution that the differences in fit between the sigmoidal model variants were relatively small, and that the maximum sigmoid model fit slightly better in older adults. Importantly, the misrecollection mechanism did *not* provide a substantially better fit to the data, such that both the RBOE and summed sigmoid models without this mechanism were preferred because the misrecollection models were penalized for the additional parameter. Across datasets, only one participant was best fit by either misrecollection model, and only four participants were best fit by either RBOE model across the five datasets We conclude that the sigmoid approaches were strongly supported over the standard RBOE approach, and that the misrecollection component was not supported by formal model comparison.

Table 4

Model comparison for each dataset. The values indicate the average mean-centered BPIC value (and number of best-fit participants) for each model. Lower BPIC values are preferred.

Model	Y_d	O_1	Y_1	O_2	O_3
RBOE	51.84 (0)	118.64 (0)	58.76 (1)	74.59 (1)	51.82 (1)
RBOE misrec.	56.24 (1)	128.01 (0)	60.79 (0)	82.95 (0)	57.73 (0)
Sep. sig.	-24.21 (5)	-47.14 (3)	-26.30 (3)	-31.11 (6)	-22.67 (5)
Min. sig.	-23.86 (5)	-48.83 (4)	-23.80 (3)	-33.00 (3)	-22.40 (4)

Model	Y_d	O_1	Y_1	O_2	O_3
Max. sig.	-13.23 (1)	**-53.71** (7)	-14.16 (3)	**-33.89** (5)	-22.40 (3)
Sum. sig.	**-25.07** (6)	-51.32 (4)	**-30.44** (8)	-31.18 (3)	**-22.97** (5)
Sum. sig. misrec.	-21.71 (0)	-45.66 (0)	-24.85 (0)	-28.35 (0)	-19.10 (0)

Table 5

Model comparison for all young adults, all older adults, and all participants overall. The values indicate the average mean-centered BPIC value (and number of best-fit participants) for each model. Lower BPIC values are preferred.

Model	Y_{all}	O_{all}	*All*
RBOE	55.30 (1)	81.68 (2)	71.13 (3)
RBOE misrec.	58.52 (1)	89.56 (0)	77.14 (1)
Sep. sig.	-25.25 (8)	-33.64 (14)	-30.28 (22)
Min. sig.	-23.83 (8)	-34.74 (11)	-30.38 (19)
Max. sig.	-13.69 (4)	**-36.67** (15)	-27.48 (19)
Sum. sig.	**-27.76** (14)	-35.16 (12)	**-32.20** (26)
Sum. sig. misrec.	-23.28 (0)	-31.04 (0)	-27.93 (0)

Model fits

We focus our discussion of model fit and parameters on the summed sigmoid model, as it was preferred overall by model comparison. As may be seen in Figures 6 and 7, the summed sigmoid model was able to qualitatively capture the overall trends in the data very well. There were a few places, however, where the model somewhat missed the mark, particularly for the young adult groups. The model underestimated correct rejection rates in young adults (Figure 6) , and incorrectly assumed higher confidence ratings, and faster RTs, for correct rejection responses compared to correct source responses in these groups (Figure

7). Overall, however, the model closely fit the distributions of data across participants. It successfully accounted for general differences between datasets, such as stronger source memory in Experiment 2 compared to Experiment 1 and overall slower RTs in older adults. This model also accounted for lower confidence for incorrect compared to correct source responses in all groups, and, in older adults, differences in confidence and RTs between correct source and correct rejection responses.

Although this sigmoid-based model provided close fits to the data, the RBOE approach was less successful, as shown in Figure 8, as well as Figures S1 and S2 of the Supplemental Material. The most notable difference was that the RBOE model was less able to fit the observed distributions of confidence responses between conditions (Figure 8). Specifically, RBOE generally overestimated confidence responses in the middle of the confidence scale and underestimated more extreme confidence responses, especially low confidence responses. The additional flexibility of the sigmoid approach was needed to account for the full range of confidence judgments and differences in confidence across conditions.

Although there were substantial differences in model predictions between the RBOE- and sigmoid-based models, there were only minor differences in model predictions between the standard and misrecollection models (not shown). This suggests that the addition of the misrecollection mechanism was not useful to the model. Overall, we conclude that the mechanisms of the summed sigmoid model could provide a reasonable approximation of the latent cognitive processes that are responsible for memory-guided decision-making and metacognitive monitoring in young and older adults.

Model parameter comparisons

To assess potential differences in model-defined latent processes between groups, we compare the posterior distributions of the group-level hyperpriors for those parameters that were fit hierarchically (i.e., all parameters except for t_0). We focus our analysis on the group-level distributions governing the *mean* of parameter values across participants. Although we fit these group-level distributions with normal distributions, all of the

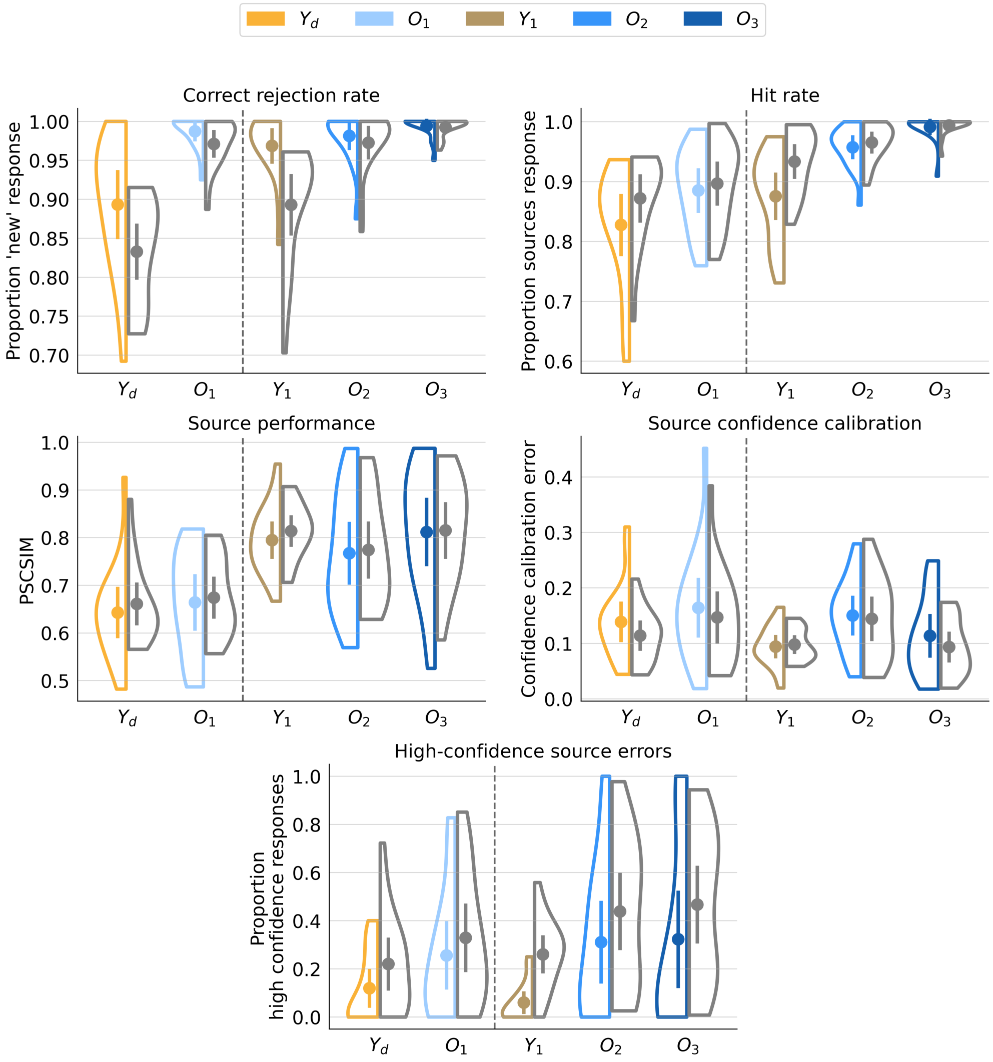

Figure 6. *Comparison of observed and standard sigmoid model-predicted response accuracy and accuracy-confidence calibration results.* Each split violin shows the distribution of observed performance on the left, along with mean level of performance, for each group. The gray distribution on the right side of each split violin shows the performance simulated by the winning computational model for each participant, along with the mean across participants. The vertical dashed line on each plot separates the datasets of Experiments 1 and 2. The error bars indicate the 95% confidence interval for each distribution. PSCSIM = paired-source conditional source-identification measure.

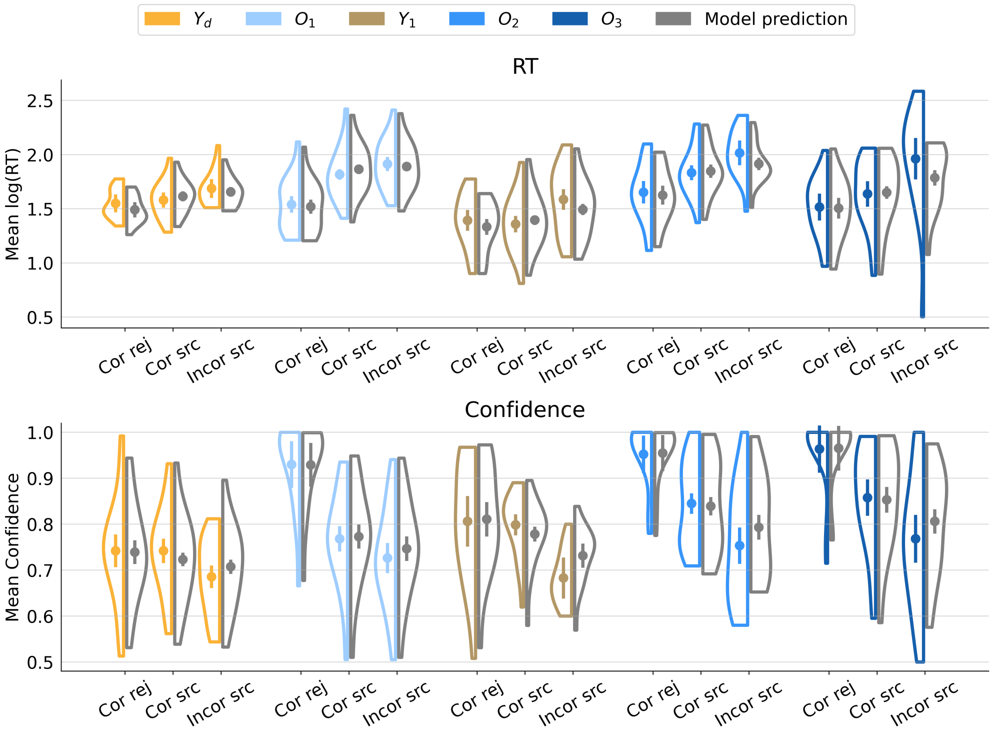

Figure 7. *Comparison of observed and summed sigmoid model-predicted mean response times (top) and confidence judgments (bottom) for correct rejections, correct source responses, and incorrect source responses.* Each split violin shows the distribution of observed performance on the left, along with mean level of performance, for each group. The gray distribution on the right side of each split violin shows the performance simulated by the winning computational model for each participant, along with the mean across participants. The error bars indicate the 95% confidence interval for each distribution, corrected for within-subject comparisons within each group. Cor rej = correct rejections; Cor src = correct source; Incor src = incorrect source.

participant-level parameters were limited to positive values by applying an exponential transform. To assess group differences, we took the exponential transform of each value in the hyper-mean posterior distributions. This provides the median values of the log-normal distributions acting as the priors for each group. These transformed posterior distributions are presented in Figure 9, and form the basis of the following parameter comparisons.

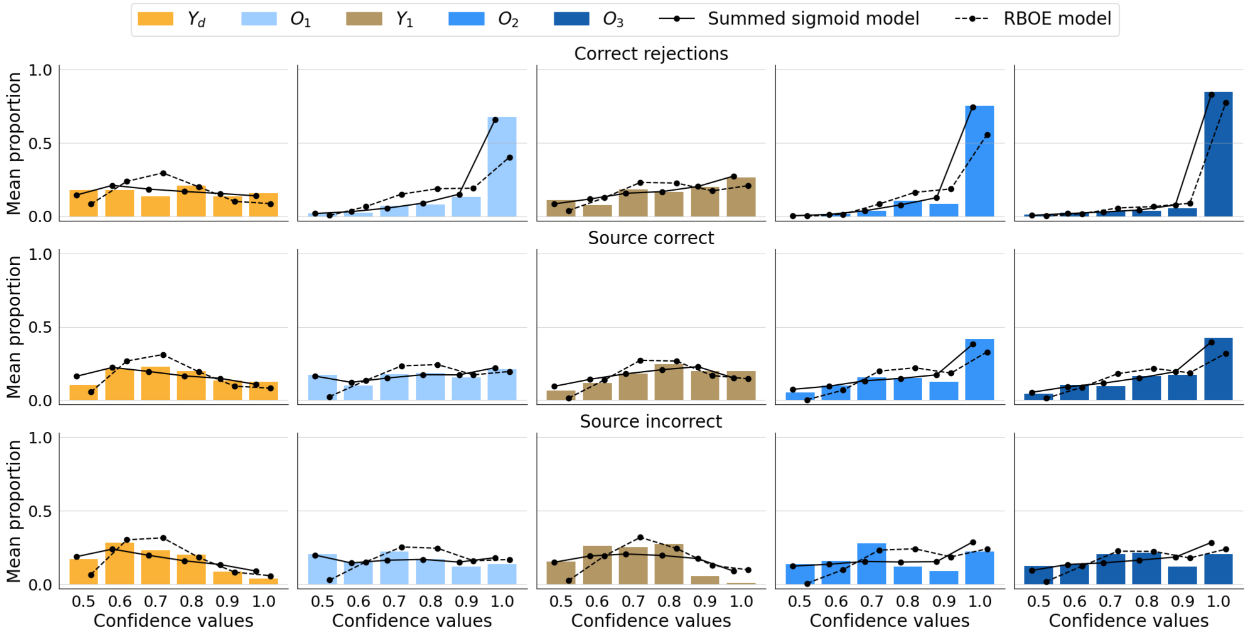

Figure 8. *Comparison of the distributions of observed and model-predicted mean confidence judgments for correct rejections (top), correct source responses (middle), and incorrect source responses (bottom).* The observed mean proportion of each confidence value for the different response types are shown as bars, and the corresponding data predicted by the summed sigmoid and RBOE models are shown as solid and dotted lines, respectively.

To compare the posterior distributions between groups, we adopted an approach from previous work (Darby et al., 2022; Weigard et al., 2020; Winkel et al., 2016), in which odds ratios are calculated for differences between distributions. Specifically, for each parameter, we took 1 million random samples from each posterior, with replacement, and then took the pairwise differences between the samples from two posteriors at a time. If the median of these differences was above zero, we calculated the odds ratio as the proportion of sample differences above zero divided by the proportion of sample differences below zero. Conversely, if the median was below zero, the odds ratio was calculated as the proportion below zero divided by the proportion above zero. Each odds ratio can be interpreted similarly to a Bayes Factor, based on the guidelines set forth by Jeffreys (1961). Specifically, odds ratios $<$ 3 are interpreted as providing no evidence of a difference; odds ratios $>$ 3 and $<$ 10 are

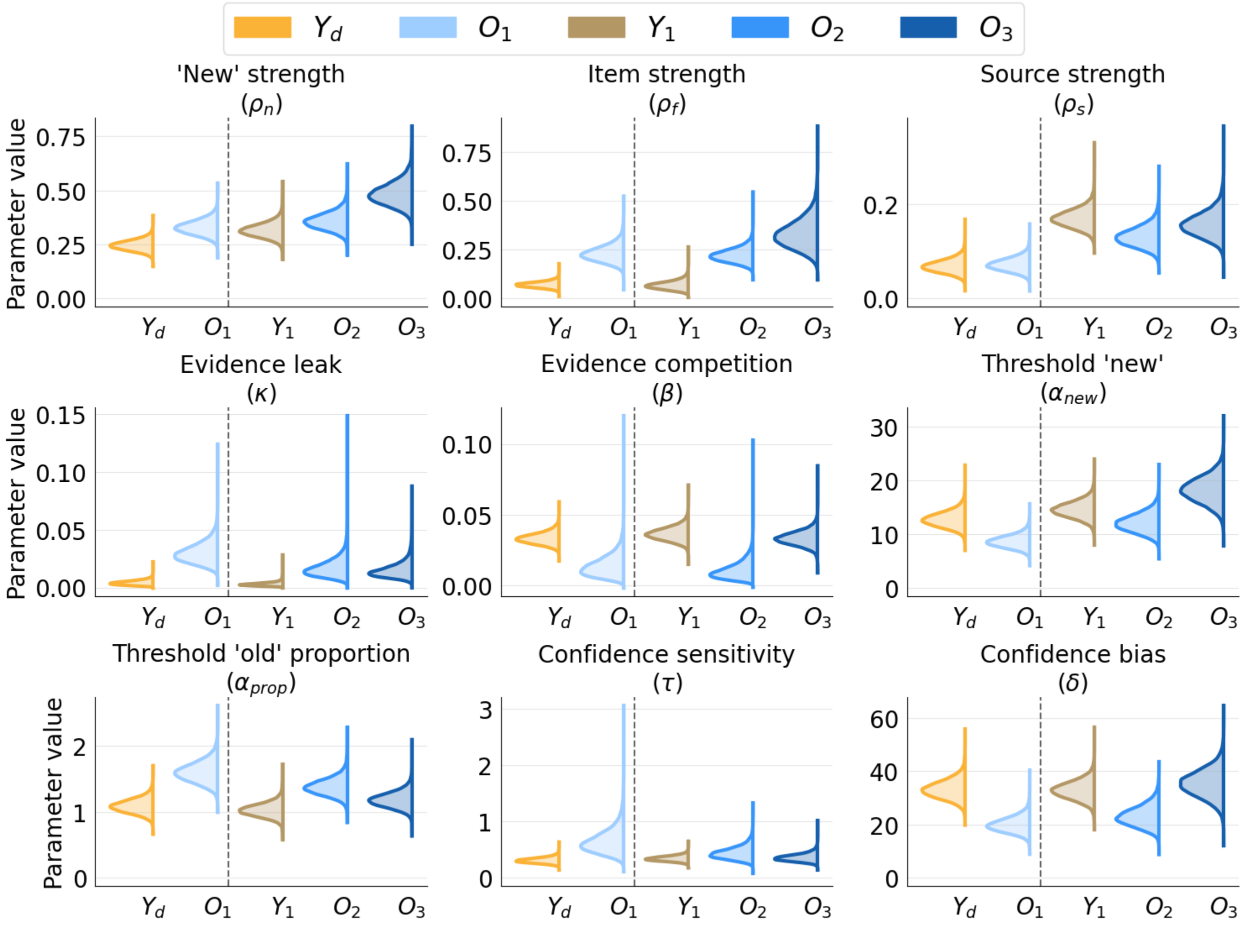

Figure 9. *Group-level posterior distributions of the hierarchical parameters for the standard sigmoid model.* Each half-violin shows the posterior of the hyper-parameter distribution for each hierarchical parameter. The values of each distribution have undergone an exponential transform to correspond to the participant-level parameters that underwent the same transformation. See the text for details.

interpreted as providing “positive” evidence; odds ratios > 10 and < 30 are interpreted as “substantial” evidence; odds ratios > 30 and < 100 are interpreted as “strong” evidence; and odds ratios > 100 are interpreted as “decisive” evidence of a difference between groups.

Similar to the conventional statistical models reported above, we focus our comparisons on the groups within Experiment 1 (Y_d and O_1) and Experiment 2 (Y_1, O_2, and O_3), as well as the comparison between Y_1 and O_1, in which young and older adults completed the same procedure. In brief, we found consistent evidence of age differences in item and source memory strengths, and some, albeit less consistent, evidence of other differences, including in confidence sensitivity and bias. Below, we describe the key differences between groups (see Table 6 for a full list of difference odds ratios between groups).

Table 6

*Differences in hyperparameter medians between groups (calculated as odds ratios). * > 10; ** > 30; *** > 100*

	Exp. 1	Exp. 2	Exp. 2	Same procedure
Parameter	Y_d v O_1	Y_1 v O_2	Y_1 v O_3	Y_1 v O_1
ρ_n	46.1$**$	4.0	>100$***$	1.6
ρ_f	>100$***$	>100$***$	>100$***$	>100$***$
ρ_s	1.3	10.0$*$	1.9	>100$***$
κ	>100$***$	>100$***$	>100$***$	>100$***$
β	36.4$**$	>100$***$	2.11	56.0$**$
α_{new}	77.8$**$	6.0	8.9	>100$***$
α_{prop}	>100$***$	72.4$**$	5.7	>100$***$
τ	50.6$**$	4.1	1.4	28.0$*$
δ	>100$***$	30.4$**$	1.9	>100$***$

Memory strength parameters. We found a number of differences between groups in the memory strength parameters.

A highly consistent age difference was that item memory strength (ρ_f) was estimated to be substantially higher in older adults compared to young adults for all comparisons (all $ORs > 100$). There was also decisive evidence that source memory strength (ρ_s) was lower in older adults compared to young adults who performed the same procedure (Y_1 v O_1; $OR > 100$), and substantial evidence that even when older adults had an advantage of being presented with the stimuli twice their source memory strength was reduced compared to young adults who saw the stimuli once (Y_1 v O_2; $OR = 10.0$). Together, these results suggest that older adults relied on item-based memory to a greater extent compared to young adults in these experiments. To quantify this relationship more directly, we calculated item reliance scores as $\frac{\rho_s}{\rho_s+\rho_f}$, which would equal 1 in the case of total reliance on item memory, such that item memory was present but no source memory, or would equal 0 in the case of total reliance on source memory. The posterior distributions of this metric are presented in Figure 10. The odds ratios indicated decisive evidence of greater reliance on item memory in older adults in every comparison within Experiment 1 and Experiment 2, as well as between experiments in participants who performed the same procedure ($ORs > 100$). We discuss how greater reliance on item memory could impact accuracy and confidence behaviors in the Discussion section.

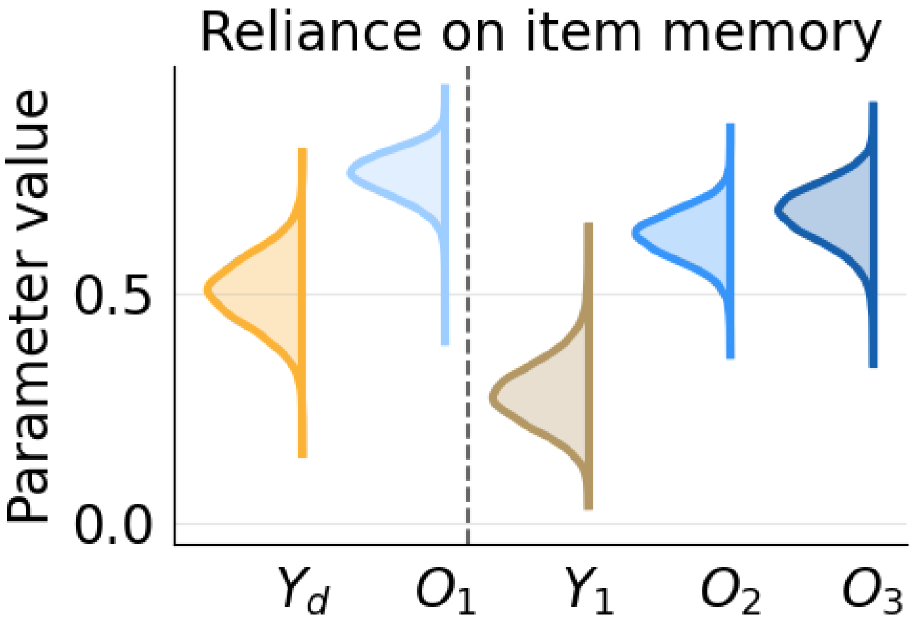

Figure 10. *Group-level posterior distributions of the reliance on item memory.* The values of each distribution were calculated as the proportion of item memory strength (i.e., ρ_f) compared to the sum of both item and source memory strength (i.e., ρ_s). See the text for details.

Leak and lateral inhibition. Next, we examined differences in the LCA parameters controlling leak (κ) and lateral inhibition (β). There were consistent age differences in leak, such that κ was higher in older adults for all comparisons within Experiment 1, Experiment 2, and across experiments in participants who performed the same procedure ($ORs > 100$). We also found largely consistent evidence of age differences in lateral inhibition, such that β was *lower* in older adults in Experiment 1 (Y_d v O_1 $OR = 36.4$), as well as in Experiment 2 in older adults who saw stimuli twice compared to young adults who only saw them once (Y_1 v O_2 $OR > 100$), although there was no difference in lateral inhibition between young adults and older adults who studied each stimulus three times (Y_1 v O_3 $OR = 2.1$). There was a difference between young and older adults who performed the same procedure (Y_1 v O_1 $OR = 56.0$). Overall, then, these results suggest that older adults experienced greater passive leak of evidence, which may be thought of as self-inhibition of the decision process, as well as reduced lateral inhibition between competing response options, compared to young adults.

Decision thresholds. We also examined group differences in the decision thresholds. Recall that we allowed separate thresholds for "new" and source responses. We found evidence of lower thresholds for "new" responses (α_{new}) in older adults in Experiment 1 (Y_d v O_1 $OR = 77.8$), but not in Experiment 2 (Y_1 v O_2 and Y_1 v O_3 $ORs < 10$). The threshold for "new" response was lower in older adults who performed the same procedure as young adults (Y_1 v O_1 $OR > 100$). The threshold for source responses was fit as a proportion of the threshold for "new" responses with a separate parameter (α_{prop}) in order to estimate a response bias. This bias parameter was higher in older adults in Experiment 1 (Y_d v O_1 $OR > 100$), and in older adults who were presented with the stimuli twice in Experiment 2 (Y_1 v O_2 $OR = 72.4$), and in older adults who only saw the stimuli once compared to young adults with the same procedure (Y_1 v O_1 $OR > 100$).

Confidence sigmoid sensitivity and bias. Finally, we examined potential differences in the parameters of the sigmoid function, applied to determine confidence judgments. The τ parameter, which determined the shape of the sigmoid, took on higher values for older adults in Experiment 1 (Y_d v O_1 $OR = 50.6$), indicating greater sensitivity to the distance between evidence for losing accumulators and the threshold when making confidence judgments. This pattern was also true of older adults who performed the same procedure as young adults (Y_1 v O_1 $OR = 28.0$), although no evidence of a difference between age groups in Experiment 2 were observed ($ORs < 10$). This suggests that there may be a difference in young and older adults' sensitivity to differences in evidence, although this appears to be impacted by procedural differences.

We also found evidence of age differences in bias toward lower or higher confidence judgments, as estimated with the parameter δ. There were lower levels of δ in older adults in Experiment 1 (Y_d v O_1 $OR > 100$), indicating that older adults were more biased toward high-confidence responses than young adults. This pattern was also found when comparing older adults who were presented with the stimuli twice compared to young adults who only saw the stimuli once in Experiment 2 (Y_1 v O_2 $OR = 30.4$), and in young and older adults who performed the same procedure across experiments (Y_1 v O_1 $OR > 100$), although there was no difference between young adults who saw the stimuli once and older adults who saw them three times in Experiment 2 (Y_1 v O_3 $OR = 1.9$).

Discussion

In this work, we have examined subjective confidence in memory-guided decisions in young and older adults with a computational modeling framework. We paired the LCA model of decision making (Usher & McClelland, 2001) with a sigmoid-based approach to confidence, in which confidence is determined from the distance between accumulated evidence for "losing" choices compared to a decision threshold. This confidence model is conceptually similar to the RBOE account, in which confidence is determined by comparing the evidence accumulated for a winning choice to the evidence accumulated for all choices (Merkle & Van Zandt, 2006; Vickers, 1979). The sigmoid function, however, allows for individual differences in the mapping between evidence and confidence by including parameters that control the sensitivity and bias of the mapping. We compared four sigmoid-based models to investigate whether confidence was based on all accumulators or either the accumulator with the maximum or minimum distance from the threshold. The results indicated that the sigmoid approaches were all strongly preferred to the RBOE framework in both young and older adults. This finding was consistent across manipulations designed to equate source memory performance between young and older adult groups in two experiments by Dodson, Bawa, and Slotnick (2007). In addition, we found that a model in which confidence was determined by the sum of the distances across accumulators was best able to account overall, although the differences in model performance between the sigmoid-based models were modest. Comparison of model parameters across age groups and experimental manipulations suggested consistent age differences in reliance on memory for individual items (versus association-based memory for sources). In addition, there were age differences in the sensitivity to the distance of evidence from the threshold when determining confidence, as well as bias in overall levels of confidence. We also tested a misrecollection account of aging effects on memory and confidence, which assumed that an association was formed between an item and the *incorrect* source on some proportion of trials. Contrary to this idea, we did not find that misrecollection was needed to account for the data in either young or older adults. In what follows, we discuss implications of these findings for our understanding of subjective confidence, as well as age differences in confidence judgments and source memory.

Implications for subjective confidence

In our model, one's subjective level of confidence in a decision is determined by the amount of evidence in favor of competing choice options, processed with a sigmoid function to map differences in evidence onto confidence values. The sigmoid allowed for individual differences in the metacognitive process of mapping decision evidence onto subjective confidence values, and allowed us to assess potential age differences in this process. As we discuss below, we tested four variants of the sigmoid model, and formal model comparison suggested that summing the distances between each losing accumulator and the corresponding threshold provided the best fit of the data overall.

The results of quantitative model comparison indicate that all of the sigmoid-based models were heavily preferred over the RBOE approach for a large majority of participants in all age and experimental groups. Examination of Figures 6 and 7 suggests a close qualitative fit between the winning model and the observed data overall. These results suggest that the sigmoid-based mapping between accumulated evidence and confidence is a viable model of confidence, one with significant advantages over the simpler (but less flexible) RBOE calculation.

The parameters of the sigmoid function allow for greater flexibility in the mapping between evidence and confidence. The substantial improvement in model fit for the sigmoid-based models compared to RBOE for most participants strongly suggests the presence of individual differences in this mapping that the latter approach has difficulty capturing. Importantly, however, the sigmoid-based framework is based on the same idea as RBOE: that confidence can be captured by comparing accumulated evidence for the different choices that are available to the decision maker. The sigmoid function and associated parameters simply allow for changes in this function to reflect individuals' level of bias toward low or high confidence values, and sensitivity to different distances between the threshold and accumulated evidence, similar to how psychophysical functions are often applied to account for individual differences in response bias and sensitivity to perceptual stimulus properties.

How do the sigmoid-based models improve our understanding of confidence? One way is that the sensitivity and bias parameters provide interpretable metrics of latent metacognitive monitoring processes. The sensitivity parameter, for example, quantifies the steepness of the mapping between the distance of losing accumulators from the threshold, or to what extent confidence is modulated by changes in evidence. Some individuals may require very large changes in this distance between accumulated evidence and the decision threshold in order to modulate confidence, whereas others may be much more sensitive to differences and oscillate drastically between low and high confidence depending on this distance. At the same time, individuals can vary depending on their bias toward high or low levels of confidence regardless of distance from the threshold.

One domain in which quantifying these aspects of confidence could be useful is in characterizing individual or group differences in confidence behaviors. One example of this was provided in the current work by characterizing differences in parameter values between young and healthy older adults, as we discuss in more detail below, but this approach could be helpful in other domains as well. For example, a number of psychiatric disorders are associated with abnormalities in subjective confidence judgments that could contribute to pathological decision-making (Hoven et al., 2019). Schizophrenia, for instance, has been associated with a weaker relationship between confidence and accuracy, along with greater overconfidence and more high-confidence errors (Eifler et al., 2015; Hoven et al., 2019; Moritz & Woodward, 2006). In addition, individuals with clinical depression or obsessive-compulsive disorder tend to be underconfident in their responses relative to controls (Dar et al., 2022; Hoven et al., 2019; Szu-Ting Fu et al., 2012). Characterizing how these populations differ in the sensitivity and bias of the mapping between decision and confidence, along with other aspects of the decision process such as drift rates and thresholds, could be a useful tool for better understanding pathological metacognition and decision-making.

Comparing sigmoid-based models of confidence. In this work, we have proposed a sigmoid-based mapping between decision evidence – specifically the distance between accumulated evidence and the decision threshold – and metacognitive confidence. To better understand this mapping, we implemented a series of four models that took into account the distance of losing accumulators in different ways. Specifically, we asked whether the evidence accumulated for both losing choices are taken into account, or if confidence is based only on either the nearest or farthest accumulator from the threshold.

Two models allowed both losing accumulators to impact confidence. In the separate sigmoid model, both losing accumulators contribute separately to confidence. A consequence of this model is that if one of the accumulators is near the threshold at the time of the decision, confidence values will be constrained towards lower values of confidence. By contrast, the summed sigmoid model also takes the evidence for each choice into account, but simply sums the distances together. As a result, only the total amount of distance matters, not whether either individual accumulator is close to the threshold.

Alternatively, we reasoned that confidence could be based on only a single accumulator. In the minimum sigmoid model, confidence is only based on whichever losing accumulator is closest to the threshold at the time of the decision, such that confidence can only be high if both losing accumulators are far from the threshold. Alternatively, in the maximum sigmoid model, confidence depends solely on whichever accumulator is farthest from the corresponding threshold, such that low levels of confidence are only possible when both accumulators are near the threshold.

The summed sigmoid model generally produced the best overall fit to the data across participants. However, we emphasize that this result should be interpreted with caution, as the differences in fit were not substantial and there was a great deal of variability in which model best fit individuals' data (see Table 4). In fact, there were at least 11 older adult participants best fit by each of the four sigmoid-based models (excluding the summed sigmoid model with the misrecollection component), and overall there was little difference in the fit of these models. We conclude, then, that the possibility for individual differences and the greater flexibility afforded by the sigmoid approach was needed to fit the data in older adults, although there was not enough information in their observed data to adjudicate between more specific implementations of the sigmoid and which accumulators are considered when mapping decision evidence onto confidence. There was somewhat more consensus for young adults, 14 of whom were best-fit by the summed sigmoid model, while only 4 were best-fit by the maximum sigmoid model. Although this is suggestive that young adults considered the evidence in both losing accumulators when determining confidence (which in the case of the summed sigmoid model were summed together), there was variability in the best-fitting model for young adults as well (with 8 participants being best-fit by both the minimum and separate sigmoid models).

The variability in model fit suggests that although these sigmoidal model variants provide intriguing hints about the mapping between evidence and confidence, the dataset we modeled here (Dodson, Bawa, & Slotnick, 2007) did not include some elements that would likely better constrain these models. For example, manipulating the number and characteristics of different response options would likely help adjudicate between these alternative mappings between evidence accumulation and subjective confidence. Specifically, the summed and maximum sigmoid models make the prediction that, keeping other parameters constant, adding additional choices should increase confidence. This is because adding accumulators will increase the likelihood of at least one accumulator having a large distance from the threshold (in the case of the maximum sigmoid model), or will increase the total amount of distance across all accumulators (in the case of the summed sigmoid model). Conversely, the minimum and separate sigmoid models make the opposite prediction, in that confidence would overall be expected to decrease with the addition of more response options, as this increases the likelihood of at least one losing accumulator being near its respective threshold. Manipulating the number of choice options, while keeping other aspects of an experiment constant as much as possible, would be expected to help disambiguate these models. Interestingly, there is a large literature on the impacts of choice overload, but the findings have been inconsistent, with some work suggesting that increasing the number of available choice options reduces choice satisfaction and confidence (Chernev et al., 2015), while other work has found no evidence of this (Scheibehenne et al., 2010). Unfortunately, however, we are unaware of any work investigating how the number of choices could impact confidence in memory-guided decisions specifically.

Alternative sequential sampling-based models of confidence. In the current work, we have proposed an approach that, similar to the RBOE framework, bases confidence on the evidence for different choice alternatives at the time a decision is made. A number of other approaches to confidence within SSMs have been implemented in the literature. Below, we discuss several of these approaches and how they relate to the current framework.

One influential model, the two-stage dynamic signal detection (2DSD) model of Pleskac and Busemeyer (2010), can be applied to two-alternative-forced-choice tasks, and is similar to the DDM in that a single accumulator of evidence drifts between two thresholds corresponding to the two possible choices. The threshold that is crossed first determines the primary decision (e.g., “old” or “new” in a recognition memory paradigm). Critically, however, evidence continues to drift following the primary choice for a fixed amount of time after the threshold is crossed, at which point the decision maker maps the final state of evidence onto a confidence value. To do so, the model assumes different threshold values corresponding to different levels of confidence, the values of which are freely estimated as parameters of the model. In general, decisions in which evidence continues to drift more in favor of the chosen response result in higher levels of confidence, whereas confidence values will be lower if the evidence drifts more in favor of the other response.

The 2DSD model has been influential in the literature and has found success in accounting for a number of empirical phenomena (see Pleskac & Busemeyer, 2010 for discussion). An important difference between 2DSD and the current approach is that the latter can more easily scale to different numbers of primary response options as well as confidence judgment levels. In its standard form, the 2DSD model can only be applied to tasks with two response options for the primary decision, such that it could not be applied to the dataset we have analyzed here. In addition, because 2DSD typically treats the thresholds corresponding to different levels of confidence as free parameters, such that increasing the number of confidence levels (e.g. from 6, as in the currently analyzed data, to 10) would require additional parameters. The sigmoid approach can more easily scale to greater numbers of primary and secondary choices. In the sigmoid approach, we have mapped continuous confidence values between 0 and 1 onto confidence levels with evenly spaced bins, such that the number of values available to the participant has no impact on the number of parameters. In addition, because the LCA model allows evidence to accumulate separately for each primary choice, there is in principle no limit to the number of primary choice options available to the participant. We see the flexibility of the sigmoid approach in regards to the number of primary and secondary choice options as a significant strength.

Another important difference between the 2DSD and sigmoid-based approaches is that 2DSD bases confidence on additional evidence not available at the time of the primary decision, as evidence continues to accumulate after the decision has been made. A consequence of allowing additional accumulation of evidence in the 2DSD model is an opportunity to improve the calibration between confidence and accuracy. This is because, for correct responses, evidence for that response is more likely to increase with additional time, increasing confidence. At the same time, when an error is made, evidence is more likely to drift *away* from that response threshold, because the drift rate is pushing evidence toward the other threshold, lowering confidence. The model, then, predicts that as the time between the primary response and confidence response increases, confidence-accuracy calibration improves. Some work manipulating the time at which confidence is tested has confirmed this prediction for perceptual decisions (Desender et al., 2021). In addition, recent work has suggested that allowing for more time to evaluate one's confidence can strengthen the relationship between confidence and accuracy for memory-based decisions as well (Klopukh & Darby, 2024).

Although we did not allow for additional accumulation of evidence following the primary choice in the current work, it would be possible to implement this in the future by allowing evidence to continue to accumulate following the source memory decision for some amount of time before comparing the distances of each choice option to the corresponding thresholds. A possible aging-related hypothesis for future work would be that older adults may continue to accumulate evidence for less time following a memory decision compared to young adults, resulting in poorer confidence–accuracy calibration. The current results suggest, however, that calibration is partially dependent on experimental manipulations. For example, older adults' calibration improved following multiple presentations of each stimulus during the study phase. It is possible that such manipulations could impact the time taken for post-decision evidence accumulation, but it is not obvious why this would be the case.

A related model (Pereira et al., 2020) has also allowed evidence to accumulate for a postdecision interval, similar to 2DSD. In this model, evidence accumulated separately for a correct and incorrect perceptual response option. Following the postdecision interval, confidence was modeled with a sigmoid function with bias and sensitivity parameters, similar to the current approach. Importantly, however, the model of Pereira and colleagues (2020) only took into account evidence for the accumulator that had reached the primary decision threshold, ignoring evidence for the losing accumulator. This model, then, explicitly did *not* take into account a balance of evidence between alternatives, which is in contrast to the current approach whereby evidence is determined by the evidence for the losing accumulators at the time of the decision. Although this model allowed for individual differences in bias and sensitivity, this was not the primary focus of the paper by Pereira and colleagues (2020). The current work suggests that such individual differences play an important role in metacognitive monitoring, and that such processes may change with development.

As discussed above, the 2DSD model and the model of Pereira and colleagues (2020) assume that confidence is based on additional accumulation of evidence following the primary choice. Other models have instantiated other forms of evidence that could impact confidence as well. One influential SSM-based confidence framework assumes that participants base confidence not only on evidence corresponding to different choices, but also on the time it takes to make a decision (Hellmann et al., 2023; Kiani et al., 2014). In this model, shorter RTs are explicitly assumed to result in higher confidence values, whereas long RTs result in lower confidence values. Model comparisons have indicated that allowing RT to impact confidence improves model fit (Hellmann et al., 2023). Although we did not explicitly consider an impact of RTs on confidence in the current work, it would be possible to do so in future work. Another approach, which has been used to account for perceptual decisions, takes into account a separate evidence accumulation process, this one tracking the perceptual discriminability of the stimulus, in addition to postdecisional evidence accumulation (Hellmann et al., 2023). The idea is that easily visible or discriminable stimuli will induce higher states of confidence than stimuli that are more difficult to perceive. An interesting avenue of future work would be to investigate how stimulus perceivability could impact confidence in memory-based decisions, particularly in light of evidence that an important factor in older adults' inferior memory performance could be due in part to perceptual degradation (Davidson et al., 2019; Naveh-Benjamin & Kilb, 2014).

A final model we will discuss is RTCON2 (2013). This model was designed to account for simultaneous memory and confidence responses, such as making a "1" response for a high confidence "new" response, a "2" for a low confidence "new" response, a "3" for a low confidence "old" response, and a "4" for a high confidence "old" response. In this model, each joint memory-confidence response has a separate accumulator, with its own drift rate and decision threshold, and the simultaneous memory-confidence response is determined by which accumulator reaches its corresponding threshold first. In this model, when evidence accumulates in favor of one response option, evidence for the others recedes, such that the total amount of evidence across all choices remains constant. When this model was applied to account for age differences in item recognition memory performance between young and older adults (Voskuilen et al., 2018), the authors found evidence of age differences in nondecision time, as well as small changes in drift rates.

One drawback, in our view, of RTCON2 is that it has a large number of free parameters, owing to the inclusion of parameters determining drift rates and decision thresholds for each choice. Increasing the number of response options, such as by allowing for four levels of confidence within each level of memory response ("new" or "old") instead of two, could therefore increase the number of parameters substantially. As discussed above, in our sigmoid-based framework the number of confidence levels available to participants is arbitrary and would not change the number of parameters. In addition, RTCON2 was designed to account for simultaneous memory and confidence judgments, and it is unclear how the model could be extended to other task frameworks, such as a primary memory choice followed by a confidence judgment (as was the case for the data modeled in the current work). We suspect that the sigmoid-based model we have presented here could easily be extended to other task formats, such as simultaneous responding, but this is an avenue of future work.

Although our sigmoid-based approach excelled and was preferred over the less flexible RBOE approach, we did not compare our approach to the alternative modeling frameworks we have just discussed. Future work is needed to systematically compare different approaches to confidence (see Hellmann et al., 2023; Shekhar & Rahnev, 2024 for recent promising examples of quantitative comparison of confidence models).

Implications for age differences in subjective confidence. By applying the sigmoid-based approach in this work, we were able to assess whether the mapping between confidence and accumulated evidence may differ as a function of age. We found some evidence of age differences in the sensitivity of confidence values to the difference between the evidence for the winning and losing response options (τ), as well as evidence of differences in bias toward confidence values that are lower or higher overall (δ).

The τ parameter controlling the sensitivity or shape of the confidence sigmoid tended to be higher in older adults who were presented with each stimulus only once (i.e., those in the O_1 group) compared to both young adult groups. According to the model, this means that the magnitude of confidence judgments tended to change to a greater extent based on the magnitude of the difference in evidence between the winning and losing response options. A psychological interpretation of this result is that older adults in the O_1 group would be more likely to feel either very low or very high levels of confidence compared to young adults, who perhaps had a more graded sense of confidence. It must be noted, however, that there were no age differences between young adults and older adults who were presented with additional stimulus presentations (i.e., those in the O_2 and O_3 groups), suggesting that procedural differences play an important role, in addition to possible age differences. We also note that there were differences between groups in confidence bias (δ), such that older adults were more biased toward high-confidence responses overall. This is consistent with the observation that older adults typically reported higher confidence in their decisions, especially for correct rejections (see Figure 3). Overall, then, we note the possibility of age differences in the sensitivity and bias of the mapping between evidence and confidence, although more research is needed to assess and characterize these potential developmental changes.

Although we found evidence of age differences in the mapping between evidence and confidence, recent work has suggested a largely stable relationship between memory accuracy and confidence across the lifespan. Winsor and colleagues (2021) examined the relationship between confidence and memory accuracy across childhood in the context of eyewitness memory. These researchers found that children began showing a strong accuracy-confidence relationship by 8 years of age, and that even young children (aged 4-5 years) showed evidence of such a relationship with more implicit measures of confidence (i.e., looking behaviors). Other recent work (Greene et al., 2024) assessed the confidence-accuracy relationship in both working memory and long-term memory tests of object memory in children (aged 6-13 years), young adults (aged 18-27 years), and older adults (aged 65-77 years). The authors found that, in all groups, accurate responses tended to be made with higher levels of confidence in both working and long-term memory tests. Compared to young adults, however, children made more errors with high confidence in the working memory test, but not in the long-term memory test, whereas the opposite was true of older adults. Interestingly, the experiments analyzed by Greene and colleagues (2024) assessed object recognition, which typically are associated with less developmental differences in both accuracy and confidence, at least in older adults (see Dodson, 2017, for a review). Overall, then, these findings suggest that the mechanisms underlying confidence may develop relatively early and remain largely consistent across the lifespan, although contextual factors and memory abilities in specific domains almost certainly play a role. An important avenue of future work will be to investigate how the currently presented model accounts for memory and confidence behaviors across the lifespan in a range of tasks, particularly those that measure source memory and other forms of association-based episodic memory.

Implications for age differences in item and source memory. A great deal of past work suggests important age differences in memory. Deficits in episodic memory are often observed in older adults (Darby & Sederberg, 2022; Greene & Naveh-Benjamin, 2023; Naveh-Benjamin, 2000; Schacter et al., 1991; Tromp et al., 2015), whereas semantic memory and memory for individual items are less impacted by age (Dodson, Bawa, & Krueger, 2007; Fraundorf et al., 2019; Verhaeghen, 2003). This pattern is mirrored by findings that older adults show larger differences in calibration with confidence and high-confidence error rates in some episodic memory tasks than in other memory tasks (see Dodson, 2017, for a review)

In the episodic source memory task of Dodson, Bawa, and Slotnick (2007), which we have modeled in this paper, participants were asked to indicate whether statements were novel, or whether they had been associated before with Source 1 or Source 2. In our model, we assume that two types of memory strength come into play for studied statements. Item memory for the statement itself, estimated with parameter ρ_f, is assumed to drive evidence for *both* sources, whereas associative memory for which source was paired with each item, estimated with parameter ρ_s, is assumed to drive evidence for the correct source only. The current results align with the hypothesis of a deficit in associative memory, as source-based memory strength was estimated to be higher in young adults in the Y_1 group compared to older adults in both the O_1 and O_2 groups (i.e., in young and older adults who were presented with each stimulus once compared to older adults presented with each stimulus either once or twice). Interestingly, item-based memory strength was estimated to be substantially higher in *older* adults, and this difference was consistent across all comparisons.

Why would this be? One possibility is that older adults had better memory for the statements. Indeed, within each experiment, older adults showed some evidence of better statement recognition, as older adults were better able to (1) reject novel statements (in

Experiment 1), and to (2) recognize studied statements (in Experiment 2). Recall, however, that within each experiment older adults were provided with an advantage over young adults (a shorter delay period in Experiment 1 and more stimulus presentations in Experiment 2). When comparing young and older adults who performed the same procedure across the two experiments, there were no differences in either correct rejections or hit rates, suggesting that differences in recognition performance were likely due to procedural changes and not stronger item recognition abilities in older adults. Therefore, we do not interpret the higher item memory strength parameter values in older adults to reflect stronger memory for the statements per se. Rather, we suggest that older adults simply *relied on* item-based strength in making their decisions and confidence judgments than did young adults. In other words, we suggest that, in general, memory strength was proportionally more item-based in older adults. We confirmed this by calculating the proportion of item strength to the total amount of memory strength (item and source), and found that older adults' reliance on item strength was substantially higher than that of young adults (see Figure 10), who utilized associative source memory to a greater degree.

We propose that there are multiple consequences of older adults' greater reliance on item information when it comes to confidence judgments. According to the current computational model, one consequence of greater reliance on item-based memory in older adults would be relatively low confidence in source memory decisions, regardless of whether those decisions are correct or incorrect. The model predicts this finding because, for studied statements, item-based memory increases the drift rates for both sources, both correct and incorrect. This higher evidence for the incorrect source would decrease confidence, even when the memory response itself is correct. By contrast, for novel items, neither source would gain greater memory strength due to item-based or source-based memory, such that the drift rate would simply be equal to the baseline drift rate (i.e., 0.5). This asymmetry between conditions results in lower confidence overall for studied items compared to novel items, as long as the item-based strength for studied items is greater than zero. This pattern was observed in the data for older adults, who were substantially more confident in correct rejections of new statements compared to either correct or incorrect source judgments for repeated statements. Although this pattern was striking and consistent across experimental procedures in older adults, the same was not true of young adults, who reported comparable levels of confidence for correct rejections and correct source identifications. The model was partially able to account for this, in that it predicted a much smaller difference between confidence for these types of responses in young adults. The model did so by reducing item strength (i.e, ρ_f), such that the model relied more on source strength (ie., ρ_s), relative to older adults. Indeed, in the case when $\rho_f = 0$, there would be no difference between strength for the incorrect source and the "new" responses, such that the model would predict more similar confidence values between correct rejections and correct source responses. Although the model did predict a much smaller difference in confidence values between these response types in young adults, even in this age group the model predicted somewhat higher confidence for correct rejections, even though this pattern of data was not found for this age group. This suggests the intriguing possibility of a mechanistic difference between groups that the current model framework is unable to capture, which could be explored in future work.

Older adults' reliance on item strength may have also impacted their susceptibility to high-confidence source errors. To test this, we simulated data with different item and source memory strengths, making sure that the sum of drift rates for correct and incorrect sources would be kept constant. We tested how changes in reliance on item memory would modulate (1) response accuracy, (2) mean levels of confidence for correct and incorrect source responses, and (3) the proportion of source errors made with high confidence. The results are presented in Figure 11. As reliance on item memory increases, accuracy decreases, as well as confidence for correct source responses. Interestingly, the opposite is true of confidence for incorrect responses, which *increases* with greater reliance on item memory. The reason for this is that when reliance on item memory is low, the correct source response has a much higher drift rate than the incorrect source, such that when by chance the incorrect source crosses the threshold first (i.e., when an error is made), the evidence accumulated for the correct response is likely close to the threshold, leading to low confidence. When reliance on item memory is higher, however, the accumulator for the correct source is less likely to be close to the threshold, leading to higher confidence for errors on average. Critically, this leads to a higher proportion of high-confidence errors, as is shown in Figure 11.

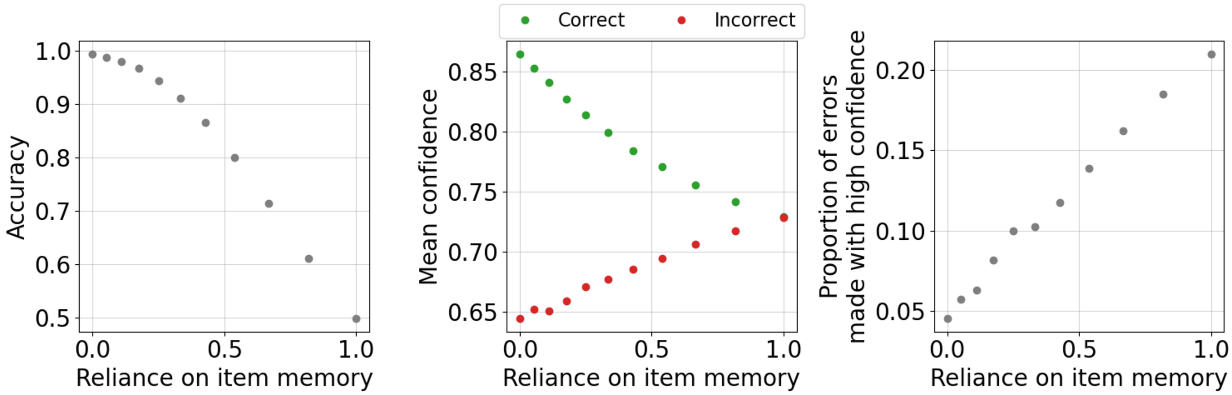

Figure 11. *Simulated effects of modulating reliance on item memory.* To investigate how reliance on item memory could impact performance, we used the sigmoid model to simulate choices and confidence values for a 2-choice task. We assumed that item strength (ρ_f) would contribute to the drift rate for both choices, whereas source strength (ρ_s) would only contribute to the drift rate for the correct choice. We kept the total amount of drift rate across the two choices constant while varying the magnitude of item strength between 0 and 0.5. Reliance on item memory was quantified as $\frac{\rho_s}{\rho_s+\rho_f}$. We simulated the model with each level of item reliance along with the following values of the other parameters, which were held constant for each simulation: $\rho_b = 0.5$, $\alpha = 10$, $\beta = 0.07$, $\kappa = 0.05$, $\tau = 0.5$, $\delta = 5$, $\phi = 0$. The plots in the figure show effects of item memory reliance on choice accuracy (left), mean confidence for correct and incorrect responses (middle), and the proportion of errors made with high confidence (right).

Importantly, this relationship is not restricted to source memory, or any kind of memory, per se; it simply falls out of the dynamics of SSMs and how confidence is calculated in the models we have presented here. When there is a proportionally high drift rate for the incorrect response alternative, errors are more likely to be made with high confidence. One implication of this is that more difficult tasks will lend themselves more to high-confidence errors. This relationship has been found in perceptual decision-making, where task difficulty can be straightforwardly manipulated (Baranski & Petrusic, 1994). This may have important consequences for developmental differences in memory tasks as well. Older adults, who are often thought to have a deficit in associative forms of memory such as source memory and other episodic memory tasks, may rely more on memory for items, such that memory strengths for choice alternatives are less distinguishable, leading to more high-confidence errors.

A recent example of higher high-confidence errors in older adults was reported by Greene, Chism, and Naveh-Benjamin (2022), who demonstrated that older adults are more prone to high-confidence errors when greater memory discriminability is needed in an associative memory task. In this task, faces were paired with scenes, some of which were very similar across pairs because they came from the same category (e.g. two kitchens). When faces and pairs were recombined in an associative recognition test, some faces were paired with a high similarity scene (e.g. a face that had been paired with a kitchen was now paired with a different kitchen), and some were paired with a scene that was quite different (e.g. instead of a kitchen, a face was now paired with a mountain). Critically, older adults had a much greater associative memory deficit when they needed to discriminate between high-similarity pairs, and were more prone to high-confidence errors, compared to young adults. The authors interpreted this finding as evidence for misrecollection, in that older adults may recall the scene category that a face had been paired with. This could lead them to incorrectly believe that a recombined but highly similar scene had been paired with a given face, resulting in low accuracy but high confidence.

The current modeling framework could likely account for these results by assuming that older adults retain fewer specific details of the scenes that would allow them to discriminate between similar exemplars, forcing them to rely on more categorical cues to make their memory decisions, in the same way that we hypothesize that older adults relied more on item cues in the data we have modeled here (Dodson, Bawa, & Slotnick, 2007). In other words, older adults in the task of Greene and colleagues (2022) would be expected to have more similar drift rates between correct and incorrect responses in this condition than when pairs are less categorically similar. As a result, accuracy in the task would be reduced, while increasing high-confidence errors, without a misrecollection mechanism. Of course, future work would need to test this prediction by applying our model to the data of Greene and colleagues (2022), but we expect that the model we have presented would be able to capture the general trends observed.

Relatedly, the current modeling framework may be able to shed light on why older adults are more prone to high-confidence errors in some tasks than in others. In tasks in which older adults can rely on more item-based or semantic kinds of memory, there may be sufficient memory to provide separation in the magnitude of the drift rates governing evidence accumulation for older adults to both perform relatively well on the task and to avoid many high-confidence errors. For tasks that require more associative memory or mnemonic specificity (Greene & Naveh-Benjamin, 2023), however, older adults may be forced to rely on less discriminative information that result in lower accuracy and more high-confidence errors.

Misrecollection. Recall that one hypothesis that has been put forward to account for older adults' deficits in source memory performance, and susceptibility to high-confidence errors, is that older adults are more likely *misrecollect* erroneous source information (Dodson, Bawa, & Slotnick, 2007; Dodson, Bawa, & Krueger, 2007; Dodson & Krueger, 2006; Shing et al., 2009). This would suggest that the ability to form associations with source information may not decline with age – rather, older adults may simply be more likely to form the wrong associations. In the experiments modeled here (Dodson, Bawa, & Slotnick, 2007), older participants may have been more likely to bind statements to the incorrect source, especially as there were only two sources repeated across statements, and the sources were presented close together in time.

Dodson, Bawa, and Slotnick (2007) provided evidence for the misrecollection account with a SDT model of responses in the same dataset we have analyzed in the current work. These researchers found that an SDT model assuming misrecollections better fit the data from older adults, particularly the O_1 and O_2 groups, whereas this mechanism was not necessary to fit the data from young adults. However, we found no evidence that misrecollection was needed to account for these data within the current modeling framework for any of the datasets. This suggests that a misrecollection mechanism may not be necessary to account for age differences in source memory performance and subjective confidence.

We were somewhat surprised that our model results did not provide evidence for the misrecollection account, as we analyzed the same data that were used to support the same account in earlier work (Dodson, Bawa, & Slotnick, 2007). Why did the current results not align with Dodson et al.'s original analyses? One reason could be that the current models were fit to trial-level RTs as well as old/new choices and confidence values, whereas Dodson et al. fit their SDT models to summary ROC curves, which included no information about RTs.

It is also the case that the distribution of drift rates across accumulators modulates high-confidence errors, as described above, such that the additional process of misrecollection may not be needed to account for the data. Indeed, we conducted a second model simulation to investigate the effects of misrecollection. As may be seen in Figure 12, the effects of misrecollection (modulated by changing the parameter ϕ, are quite similar to the effects of increased reliance on item memory: as misrecollection increases, accuracy and confidence in correct responses decrease, whereas confidence in incorrect responses increases along with the proportion of high-confidence errors. At least when it comes to the data of Dodson, Bawa, and Slotnick (2007), then, a more general property of the decision-making process (i.e., differences in drift rates), along with a comparison of evidence for different choices such as the one we have proposed, may account for what appears to be a memory- and aging-specific process of misrecollection.

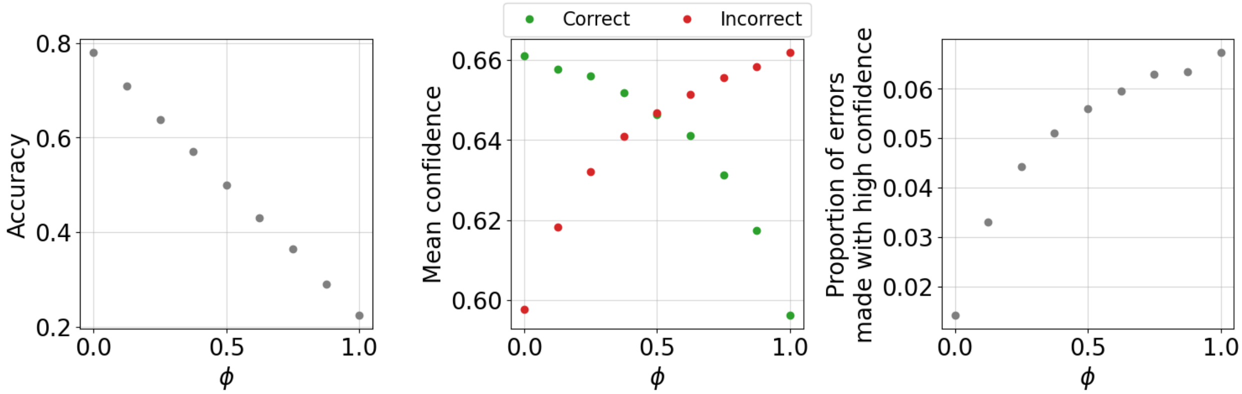

Figure 12. Simulated effects of misrecollection. To investigate how misrecollection could impact performance, we used the sigmoid model to simulate choices and confidence values for a 2-choice task. We simulated the model with different values of the misrecollection ϕ parameter along with the following values of the other parameters, which were held constant for each simulation: $\rho_f = 1.2$, $rho_s = 0.8$, $\rho_b = 0.5$, $\alpha = 10$, $\beta = 0.07$, $\kappa = 0.05$, $\tau = 0.5$, $\delta = 5$. The plots in the figure show effects of misrecollection on choice accuracy (left), mean confidence for correct and incorrect responses (middle), and the proportion of errors made with high confidence (right).

Although the misrecollection mechanism we implemented was not necessary to account for the data of Dodson, Bawa, and Slotnick (2007), we do not claim that the results of the current study disprove the misrecollection account. It is possible that misrecollection would be necessary to account for other datasets, or that misrecollection would improve the fit of other models with different mechanistic assumptions. We therefore emphasize the need for more research into misrecollection as well as alternative accounts.

Limitations

Although we believe this work provides a compelling account of the origins of confidence in memory-guided decisions, we acknowledge limitations of the work that provide opportunities for future research. First, the model somewhat struggled to fit some aspects of the data, particularly in young adults. For one thing, the model underestimated young adults' correct rejection rates. In addition, the model predicted somewhat higher confidence and faster RTs for young adults' correct rejections compared to correct source responses, even though such differences were not observed (they were observed, however, in older adults). We believe these issues were likely linked, in that the model could have produced higher correct rejection responses, for example by increasing the drift rate for "new" responses (estimated with parameter ρ_n). Increasing the correct rejection rates, however, would also result in higher confidence and faster RTs, exacerbating the predicted difference in these metrics between correct rejection and correct source responses. Future work is needed to assess whether these different patterns of performance between conditions for young and older adults are replicated in more experiments and the mechanisms by which they may arise.

A more general limitation of this modeling framework is that there is currently not a closed-form solution for the likelihood function. The current work circumvented this challenge by approximating likelihoods through simulating data. This approach has often been taken to fit LCA (Kirkpatrick et al., 2021; Ratcliff, 2006; Usher & McClelland, 2001; Weichart et al., 2020). Although simulation-based methods are popular, they are much more computationally intensive than analytical likelihood functions. Interestingly, an analytical likelihood has recently been proposed for LCA (Lo & Ip, 2021), although this approach does not capture the full RT distributions that arise from the first passage of time. A challenge for future work will be to integrate a sigmoidal confidence calculation, as we have applied in the current work, to an analytical likelihood solution of LCA and/or other SSM frameworks.

We also acknowledge that more research is needed to verify that this model generalizes to other paradigms. For example, a general model of memory-guided confidence should be able to account for simultaneous judgments of memory and confidence as well as secondary judgments of confidence. More generally, we expect that this framework would be well-suited for other kinds of decision-making tasks, such as perceptual tasks. Testing the model with a wide range of decision-making tasks is a goal for future research.

In addition, testing the model with other datasets will be needed to test more specific aspects of the modeling framework. For example, as noted in the "Comparing sigmoid-based models of confidence" section above, manipulating the number and characteristics of different response options will be needed to better test the functional form of a sigmoidal mapping between evidence accumulation and subjective confidence. In addition, although we have applied a sigmoid to the current modeling approach, and have compared several alternative approaches to assess what accumulators are guiding confidence judgments, we note that there are alternative functions that we have not tested in the current work, such as power functions or more complicated nonlinear functions that require more parameters. We also note that we have made the assumption that the relationship between distance and confidence is monotonic, such that greater amounts of distance of the losing accumulator(s) from the threshold always lead to higher levels of confidence, but it is possible that this is not the case. Future work is needed to further investigate the functional form of the mapping between evidence and confidence.

Finally, we note that participants in the experiments of Dodson et al. (2007) were instructed that they could take as long as they needed to respond in the test phase. This may have impacted the results, particularly because young and older adults often differ in terms of a speed-accuracy tradeoff (Starns & Ratcliff, 2010). In addition, it is possible that the absence of instructions to respond quickly may have increased the probability of other processes impacting RTs and the decision process itself, such as greater fluctuations of attention or different use of strategies. It is also the case that SSMs are typically applied to data in which participants have been encouraged to respond as quickly and accurately as possible, although this type of model has been applied to a wide variety of tasks and instructions, including those emphasizing accuracy over RTs (Ratcliff & McKoon, 2008).

Conclusions

In this work, we have presented a new account of subjective confidence in memory-based decisions, in which confidence results from processing evidence for different response options through a sigmoid function. This account was able to account substantially better for patterns in choices, RTs, and confidence values for source memory in young and older adults, compared to the simpler but less flexible RBOE framework (Merkle & Van Zandt, 2006; Vickers, 1979). This suggests that it is important to allow for individual differences in sensitivity and bias in the mapping between accumulated evidence and confidence judgments. In addition, we found evidence that older adults relied more on item memory than young adults, and exhibited evidence of weaker source-based memory when the experimental procedure was the same for both age groups. We also found evidence of greater sensitivity to differences in accumulated evidence in older adults who performed the same procedure as young adults, as well as greater bias in older adults to respond with higher confidence overall. We did not find any evidence to support the misrecollection account of aging deficits in source memory and confidence calibration in older adults, suggesting that a misrecollection mechanism may not be necessary to account for older adults' deficits in source memory performance and confidence-accuracy calibration.

Overall, the results of this study suggest that subjective confidence may be generated based on the state of evidence accumulated at the time a memory-based decision is made. However, the way evidence is processed likely differs between individuals and across the lifespan, as parameters controlling the shape of the sigmoid mapping evidence onto confidence, as well as bias toward higher or lower confidence overall, differed between age groups when equating the experimental task. This work, then, provides preliminary evidence that a sigmoid-based confidence calculation based on accumulated evidence for alternative choices may provide a useful account of subject confidence and how it changes across the lifespan.

Constraints on generality

The target populations for the reported findings are young adults (college age) and older adults (60-80 years). We chose to model data from these age groups (Dodson, Bawa, & Slotnick, 2007) to (1) identify the latent processes that could underlie subjective confidence in source memory decisions across the adult lifespan, and (2) identify processes that could be sources of behavioral changes due to aging. Future work will be needed to assess the model's ability to account for memory and metacognitive performance in other age groups, including children and middle-aged adults.

References

Amer, T., Wynn, J. S., & Hasher, L. (2022). Cluttered memory representations shape cognition in old age. *Trends in Cognitive Sciences*, *26*(3), 255–267. https://doi.org/10.1016/j.tics.2021.12.002

Ando, T. (2007). Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. *Biometrika*, *94*(2), 443–458. https://doi.org/10.1093/biomet/asm017

Baltes, P. B., Staudinger, U. M., & Lindenberger, U. (1999). LIFESPAN PSYCHOLOGY: Theory and Application to Intellectual Functioning. *Annual Review of Psychology*, *50*(1), 471–507. https://doi.org/10.1146/annurev.psych.50.1.471

Baranski, J. V., & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception & Psychophysics*, *55*(4), 412–428. https://doi.org/10.3758/BF03205299

Benjamin, A. S. (2010). Representational explanations of "process" dissociations in recognition: The DRYAD theory of aging and memory judgments. *Psychological Review*, *117*(4), 1055–1079. https://doi.org/10.1037/a0020810

Brown, S. D., Ratcliff, R., & Smith, P. L. (2006). Evaluating methods for approximating stochastic differential equations. *Journal of Mathematical Psychology*, *50*(4), 402–410. https://doi.org/10.1016/j.jmp.2006.03.004

Cansino, S., Torres-Trejo, F., Estrada-Manilla, C., Pérez-Loyda, M., Vargas-Martínez, C., Tapia-Jaimes, G., & Ruiz-Velasco, S. (2020). Contributions of Cognitive Aging

Models to the Explanation of Source Memory Decline across the Adult Lifespan. *Experimental Aging Research*, *46*(3), 194–213. https://doi.org/10.1080/0361073X.2020.1743920

Carlebach, N., & Yeung, N. (2020). Subjective confidence acts as an internal cost-benefit factor when choosing between tasks. *Journal of Experimental Psychology: Human Perception and Performance*, *46*(7), 729–748. https://doi.org/10.1037/xhp0000747

Castel, A. D., & Craik, F. I. M. (2003). The Effects of Aging and Divided Attention on Memory for Item and Associative Information. *Psychology and Aging*, *18*(4), 873–885.

Chalfonte, B. l., & Johnson, M. K. (1996). Feature memory and binding in young and older adults. *Memory & Cognition*, *24*(4), 403–416. https://doi.org/10.3758/BF03200930

Chernev, A., Böckenholt, U., & Goodman, J. (2015). Choice overload: A conceptual review and meta-analysis. *Journal of Consumer Psychology*, *25*(2), 333–358. https://doi.org/10.1016/j.jcps.2014.08.002

Craik, F. I. M. (1968). Two components in free recall. *Journal of Verbal Learning and Verbal Behavior*, *7*(6), 996–1004. https://doi.org/10.1016/S0022-5371(68)80058-1

Craik, F. I. M., 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. https://doi.org/10.1037/a0020276

Dar, R., Sarna, N., Yardeni, G., & Lazarov, A. (2022). Are people with obsessive-compulsive disorder under-confident in their memory and perception? A review and meta-analysis. *Psychological Medicine*, *52*(13), 2404–2412. https://doi.org/10.1017/S0033291722001908

Darby, K. P., & Sederberg, P. B. (2022). Transparency, replicability, and discovery in cognitive aging research: A computational modeling approach. *Psychology and Aging*, *37*, 10–29. https://doi.org/10.1037/pag0000665

Darby, K. P., Sederberg, P. B., & Sloutsky, V. M. (2022). Intraobject and extraobject memory binding across early development. *Developmental Psychology*, *58*, 1237–1253. https://doi.org/10.1037/dev0001355

Dautriche, I., Rabagliati, H., & Smith, K. (2021). Subjective confidence influences word learning in a cross-situational statistical learning task. *Journal of Memory and Language*, *121*, 104277. https://doi.org/10.1016/j.jml.2021.104277

Davidson, P. S. R., Vidjen, P., Trincao-Batra, S., & Collin, C. A. (2019). Older Adults' Lure Discrimination Difficulties on the Mnemonic Similarity Task Are Significantly Correlated With Their Visual Perception. *The Journals of Gerontology: Series B*, *74*(8), 1298–1307. https://doi.org/10.1093/geronb/gby130

DeCarlo, L. T. (2021). On Joining a Signal Detection Choice Model with Response Time Models. *Journal of Educational Measurement*, *58*(4), 438–464. https://doi.org/10.1111/jedm.12300

Delhaye, E., Tibon, R., Gronau, N., Levy, D. A., & Bastin, C. (2018). Misrecollection prevents older adults from benefitting from semantic relatedness of the memoranda in associative memory. *Aging, Neuropsychology, and Cognition*, *25*(5), 634–654. https://doi.org/10.1080/13825585.2017.1358351

Desender, K., Donner, T. H., & Verguts, T. (2021). Dynamic expressions of confidence within an evidence accumulation framework. *Cognition*, *207*, 104522. https://doi.org/10.1016/j.cognition.2020.104522

Destan, N., & Roebers, C. M. (2015). What are the metacognitive costs of young children's overconfidence? *Metacognition and Learning*, *10*(3), 347–374. https://doi.org/10.1007/s11409-014-9133-z

Dodson, C. S. (2017). Aging and Memory73. In *Learning and Memory: A Comprehensive Reference* (pp. 403–421). Elsevier. https://doi.org/10.1016/B978-0-12-809324-5.21053-5

Dodson, C. S., Bawa, S., & Krueger, L. E. (2007). Aging, metamemory, and high-confidence errors: A misrecollection account. *Psychology and Aging*, *22*(1), 122–133.

Dodson, C. S., Bawa, S., & Slotnick, S. D. (2007). Aging, source memory, and misrecollections. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(1), 169–181. https://doi.org/10.1037/0278-7393.33.1.169

Dodson, C. S., & Krueger, L. E. (2006). I misremember it well: Why older adults are unreliable eyewitnesses. *Psychonomic Bulletin & Review*, *13*(5), 770–775. https://doi.org/10.3758/BF03193995

Eifler, S., Rausch, F., Schirmbeck, F., Veckenstedt, R., Mier, D., Esslinger, C., Englisch, S., Meyer-Lindenberg, A., Kirsch, P., & Zink, M. (2015). Metamemory in schizophrenia: Retrospective confidence ratings interact with neurocognitive deficits. *Psychiatry Research*, *225*(3), 596–603. https://doi.org/10.1016/j.psychres.2014.11.040

Falbén, J. K., Golubickis, M., Tamulaitis, S., Caughey, S., Tsamadi, D., Persson, L.

M., Svensson, S. L., Sahraie, A., & Macrae, C. N. (2020). Self-relevance enhances evidence gathering during decision-making. *Acta Psychologica*, *209*, 103122. https://doi.org/10.1016/j.actpsy.2020.103122

Fandakova, Y., Shing, Y. L., & Lindenberger, U. (2013). Differences in binding and monitoring mechanisms contribute to lifespan age differences in false memory. *Developmental Psychology*, *49*(10), 1822–1832. https://doi.org/10.1037/a0031361

Fraundorf, S. H., Hourihan, K. L., Peters, R. A., & Benjamin, A. S. (2019). Aging and recognition memory: A meta-analysis. *Psychological Bulletin*, *145*(4), 339–371. https://doi.org/10.1037/bul0000185

Gettleman, J. N., Grabman, J. H., Dobolyi, D. G., & Dodson, C. S. (2021). A decision processes account of the differences in the eyewitness confidence-accuracy relationship between strong and weak face recognizers under suboptimal exposure and delay conditions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(3), 402–421. https://doi.org/10.1037/xlm0000922

Gold, J. I., & Ding, L. (2013). How mechanisms of perceptual decision-making affect the psychometric function. *Progress in Neurobiology*, *103*, 98–114. https://doi.org/10.1016/j.pneurobio.2012.05.008

Golomb, J. D., Peelle, J. E., Addis, K. M., Kahana, M. J., & 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. https://doi.org/10.3758/MC.36.5.947

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (pp. xi, 455). John Wiley.

Greene, N. R., Chism, S., & Naveh-Benjamin, M. (2022). Levels of specificity in episodic memory: Insights from response accuracy and subjective confidence ratings in older adults and in younger adults under full or divided attention. *Journal of Experimental Psychology: General*, *151*(4), 804–819. https://doi.org/10.1037/xge0001113

Greene, N. R., Forsberg, A., Guitard, D., Naveh-Benjamin, M., & Cowan, N. (2024). A lifespan study of the confidence–accuracy relation in working memory and episodic long-term memory. *Journal of Experimental Psychology: General*, No Pagination Specified–No Pagination Specified. https://doi.org/10.1037/xge0001551

Greene, N. R., & Naveh-Benjamin, M. (2023). Adult age-related changes in the specificity of episodic memory representations: A review and theoretical framework. *Psychology and Aging*.

Hellmann, S., Zehetleitner, M., & Rausch, M. (2023). Simultaneous modeling of choice, confidence, and response time in visual perception. *Psychological Review*, No Pagination Specified–No Pagination Specified. https://doi.org/10.1037/rev0000411

Hoven, M., Lebreton, M., Engelmann, J. B., Denys, D., Luigjes, J., & van Holst, R. J. (2019). Abnormalities of confidence in psychiatry: An overview and future perspectives. *Translational Psychiatry*, *9*(1), 1–18. https://doi.org/10.1038/s41398-019-0602-7

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford University Press.

Kiani, R., Corthell, L., & Shadlen, M. N. (2014). Choice Certainty Is Informed by

Both Evidence and Decision Time. *Neuron*, *84*(6), 1329–1342. https://doi.org/10.1016/j.neuron.2014.12.015

Kirkpatrick, R. P., Turner, B. M., & Sederberg, P. B. (2021). Equal evidence perceptual tasks suggest a key role for interactive competition in decision-making. *Psychological Review*, No Pagination Specified–No Pagination Specified. https://doi.org/10.1037/rev0000284

Klopukh, I. R. S., & Darby, K. P. (2024). Timing Is Everything: Effects Of Temporal Delay Of Confidence Judgments In Memory Decision-Making. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*(0).

Korkki, S. M., Richter, F. R., Jeyarathnarajah, P., & Simons, J. S. (2020). Healthy ageing reduces the precision of episodic memory retrieval. *Psychology and Aging*, *35*(1), 124–142. https://doi.org/10.1037/pag0000432

Lo, C.-F., & Ip, H.-Y. (2021). Modified leaky competing accumulator model of decision making with multiple alternatives: The Lie-algebraic approach. *Scientific Reports*, *11*(1), 10923. https://doi.org/10.1038/s41598-021-90356-7

Lohnas, L. J., Polyn, S. M., & Kahana, M. J. (2015). Expanding the scope of memory search: Modeling intralist and interlist effects in free recall. *Psychological Review*, *122*(2), 337–363. https://doi.org/10.1037/a0039036

McCarley, J. S., & Yamani, Y. (2021). Psychometric Curves Reveal Three Mechanisms of Vigilance Decrement. *Psychological Science*, *32*(10), 1675–1683. https://doi.org/10.1177/09567976211007559

Merkle, E. C., & Van Zandt, T. (2006). An Application of the Poisson Race Model to Confidence Calibration. *Journal of Experimental Psychology: General*, *135*(3), 391–408.

Moreno-Bote, R. (2010). Decision Confidence and Uncertainty in Diffusion Models with Partially Correlated Neuronal Integrators. *Neural Computation*, *22*(7), 1786–1811. https://doi.org/10.1162/neco.2010.12-08-930

Morgan, M., Dillenburger, B., Raphael, S., & Solomon, J. A. (2012). Observers can voluntarily shift their psychometric functions without losing sensitivity. *Attention, Perception, & Psychophysics*, *74*(1), 185–193. https://doi.org/10.3758/s13414-011-0222-7

Moritz, S., & Woodward, T. S. (2006). The contribution of metamemory deficits to schizophrenia. *Journal of Abnormal Psychology*, *115*(1), 15–25. https://doi.org/10.1037/0021-843X.15.1.15

Murnane, K., & Bayen, U. J. (1996). An evaluation of empirical measures of source identification. *Memory & Cognition*, *24*(4), 417–428. https://doi.org/10.3758/BF03200931

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. https://doi.org/10.1037/0278-7393.26.5.1170

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. https://doi.org/10.1037/0882-7974.19.3.541

Naveh-Benjamin, M., & Kilb, A. (2014). Age-related differences in associative memory: The role of sensory decline. *Psychology and Aging*, *29*(3), 672–683. https://doi.org/10.1037/a0037138

Oberauer, K., & Lewandowsky, S. (2019). Simple measurement models for complex working-memory tasks. *Psychological Review*, *126*(6), 880–932. https://doi.org/10.1037/rev0000159

Old, S. R., & 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. https://doi.org/10.1037/0882-7974.23.1.104

Oskamp, S. (1962). The relationship of clinical experience and training methods to several criteria of clinical prediction. *Psychological Monographs: General and Applied*, *76*(28), 1–27. https://doi.org/10.1037/h0093849

Pereira, M., Faivre, N., Iturrate, I., Wirthlin, M., Serafini, L., Martin, S., Desvachez, A., Blanke, O., Van De Ville, D., & Millán, J. del R. (2020). Disentangling the origins of confidence in speeded perceptual judgments through multimodal imaging. *Proceedings of the National Academy of Sciences*, *117*(15), 8382–8390. https://doi.org/10.1073/pnas.1918335117

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-Stage Dynamic Signal Detection: A Theory of Choice, Decision Time, and Confidence. *Psychological Review*, *117*(3), 864–901.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. https://doi.org/10.1037/0033-295X.85.2.59

Ratcliff, R. (2006). Modeling response signal and response time data. *Cognitive Psychology*, *53*(3), 195–237. https://doi.org/10.1016/j.cogpsych.2005.10.002

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

Ratcliff, R., & McKoon, G. (2015). Aging effects in item and associative recognition memory for pictures and words. *Psychology and Aging*, *30*(3), 669–674. https://doi.org/10.1037/pag0000030

Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological Review*, *116*, 59–83. https://doi.org/10.1037/a0014086

Ratcliff, R., & Starns, J. J. (2013). Modeling Confidence Judgments, Response Times, and Multiple Choices in Decision Making: Recognition Memory and Motion Discrimination. *PSYCHOLOGICAL REVIEW*, *120*(3), 697–719. https://doi.org/10.1037/a0033152

Rhodes, S., Greene, N. R., & Naveh-Benjamin, M. (2019). Age-related differences in recall and recognition: A meta-analysis. *Psychonomic Bulletin & Review*, *26*(5), 1529–1547. https://doi.org/10.3758/s13423-019-01649-y

Robey, A. M., Dougherty, M. R., & Buttaccio, D. R. (2017). Making Retrospective Confidence Judgments Improves Learners' Ability to Decide What Not to Study. *Psychological Science*, *28*(11), 1683–1693. https://doi.org/10.1177/0956797617718800

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review*, *103*(3), 403–428. https://doi.org/10.1037/0033-295X.103.3.403

Samanez-Larkin, G., Mottola, G., Heflin, D., Yu, L., & Boyle, P. (2023). *Overconfidence in financial knowledge associated with financial risk tolerance in older adults*. https://doi.org/10.31234/osf.io/p5gec

Schacter, D. L., Kaszniak, A. W., Kihlstrom, J. F., & Valdiserri, M. (1991). The relation between source memory and aging. *Psychology and Aging*, *6*(4), 559–568. https://doi.org/10.1037/0882-7974.6.4.559

Scheibehenne, B., Greifeneder, R., & Todd, P. M. (2010). Can There Ever Be Too Many Options? A Meta-Analytic Review of Choice Overload. *Journal of Consumer Research*, *37*(3), 409–425. https://doi.org/10.1086/651235

Sederberg, P. B., Howard, M. W., & Kahana, M. J. (2008). A context-based theory of recency and contiguity in free recall. *Psychological Review*, *115*(4), 893–912. https://doi.org/10.1037/a0013396

Shekhar, M., & Rahnev, D. (2024). How do humans give confidence? A comprehensive comparison of process models of perceptual metacognition. *Journal of Experimental Psychology: General*, *153*(3), 656–688. https://doi.org/10.1037/xge0001524

Shing, Y. L., Werkle-Bergner, M., Li, S.-C., & Lindenberger, U. (2009). Committing memory errors with high confidence: Older adults do but children don't. *Memory*, *17*(2), 169–179. https://doi.org/10.1080/09658210802190596

Slane, C. R., & Dodson, C. S. (2022). Eyewitness confidence and mock juror decisions of guilt: A meta-analytic review. *Law and Human Behavior*, *46*(1), 45–66. https://doi.org/10.1037/lhb0000481

Slotnick, S. D., Klein, S. A., Dodson, C. S., & Shimamura, A. P. (2000). An analysis of signal detection and threshold models of source memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(6), 1499–1517. https://doi.org/10.1037/0278-7393.26.6.1499

Smith, A. D. (1977). Adult age differences in cued recall. *Developmental Psychology*, *13*(4), 326–331. https://doi.org/10.1037/0012-1649.13.4.326

Smith, P. L., Saber, S., Corbett, E. A., & Lilburn, S. D. (2020). Modeling continuous outcome color decisions with the circular diffusion model: Metric and categorical properties. *Psychological Review*, *127*, 562–590. https://doi.org/10.1037/rev0000185

Starns, J. J., & 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. https://doi.org/10.1037/a0018022

Stephens, J. D. W., & Overman, A. A. (2018). Modeling age differences in effects of pair repetition and proactive interference using a single parameter. *Psychology and Aging*, *33*(1), 182. https://doi.org/10.1037/pag0000195

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

Szu-Ting Fu, T., Koutstaal, W., Poon, L., & Cleare, A. J. (2012). Confidence judgment in depression and dysphoria: The depressive realism vs. Negativity hypotheses. *Journal of Behavior Therapy and Experimental Psychiatry*, *43*(2), 699–704. https://doi.org/10.1016/j.jbtep.2011.09.014

Theisen, M., Lerche, V., von Krause, M., & Voss, A. (2021). Age differences in diffusion model parameters: A meta-analysis. *Psychological Research*, *85*(5), 2012–2021. https://doi.org/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. https://doi.org/10.1016/j.arr.2015.08.006

Turner, B. M., & Sederberg, P. B. (2014). A generalized, likelihood-free method for posterior estimation. *Psychonomic Bulletin & Review*, *21*(2), 227–250. https://doi.org/10.3758/s13423-013-0530-0

Turner, B. M., Sederberg, P. B., Brown, S. D., & Steyvers, M. (2013). A method for efficiently sampling from distributions with correlated dimensions. *Psychological Methods*, *18*(3), 368–384. https://doi.org/10.1037/a0032222

Turner, B. M., & Van Zandt, T. (2014). Hierarchical approximate Bayesian computation. *Psychometrika*, *79*(2), 185–209. https://doi.org/10.1007/s11336-013-9381-x

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, *108*(3), 550–592. https://doi.org/10.1037/0033-295X.108.3.550

Vallat, R. (2018). Pingouin: Statistics in Python. *Journal of Open Source Software*, *3*(31), 1026. https://doi.org/10.21105/joss.01026

van Ravenzwaaij, D., van der Maas, H. L. J., & Wagenmakers, E.-J. (2012). Optimal decision making in neural inhibition models. *Psychological Review*, *119*(1), 201–215. https://doi.org/10.1037/a0026275

Verhaeghen, P. (2003). Aging and vocabulary scores: A meta-analysis. *PSYCHOLOGY AND AGING*, *18*(2), 332–339. https://doi.org/10.1037/0882-7974.18.2.332

Vickers, D. (1978). An Adaptive Module for Simple Judgment. In *Attention and Performance VII*. Routledge.

Vickers, D. (1979). *Decision processes in visual perception*. Academic Press.

Voskuilen, C., Ratcliff, R., & McKoon, G. (2018). Aging and confidence judgments in item recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(1), 1–23. https://doi.org/10.1037/xlm0000425

Weichart, E. R., Turner, B. M., & Sederberg, P. B. (2020). A model of dynamic, within-trial conflict resolution for decision making. *Psychological Review*, *127*, 749–777. https://doi.org/10.1037/rev0000191

Weigard, A. S., Sathian, K., & Hampstead, B. M. (2020). Model-based assessment and neural correlates of spatial memory deficits in mild cognitive impairment. *Neuropsychologia*, *136*, 107251. https://doi.org/10.1016/j.neuropsychologia.2019.107251

Winkel, J., Hawkins, G. E., Ivry, R. B., Brown, S. D., Cools, R., & Forstmann, B. U. (2016). Focal striatum lesions impair cautiousness in humans. *Cortex*, *85*, 37–45. https://doi.org/10.1016/j.cortex.2016.09.023

Winsor, A. A., Flowe, H. D., Seale-Carlisle, T. M., Killeen, I. M., Hett, D., Jores, T., Ingham, M., Lee, B. P., Stevens, L. M., & Colloff, M. F. (2021). Child witness expressions of certainty are informative. *Journal of Experimental Psychology: General*, *150*(11), 2387–2407. https://doi.org/10.1037/xge0001049

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review*, *114*(1), 152–176. https://doi.org/10.1037/0033-295X.114.1.152

Wixted, J. T., & Wells, G. L. (2017). The Relationship Between Eyewitness Confidence and Identification Accuracy: A New Synthesis. *Psychological Science in the Public Interest*, *18*(1), 10–65. https://doi.org/10.1177/1529100616686966

Yu, S., Pleskac, T. J., & Zeigenfuse, M. D. (2015). Dynamics of postdecisional processing of confidence. *Journal of Experimental Psychology: General*, *144*(2), 489. https://doi.org/10.1037/xge0000062

Zhou, J., Osth, A. F., Lilburn, S. D., & Smith, P. L. (2021). A circular diffusion model of continuous-outcome source memory retrieval: Contrasting continuous and threshold accounts. *Psychonomic Bulletin & Review*, *28*(4), 1112–1130. https://doi.org/10.3758/s13423-020-01862-0