Journal of Experimental Psychology:
Learning, Memory, and Cognition
2018, Vol. 44, No. 1, 1–23

0278-7393/18/$12.00 http://dx.doi.org/10.1037/xlm0000425

Aging and Confidence Judgments in Item Recognition

Chelsea Voskuilen, Roger Ratcliff, and Gail McKoon
The Ohio State University

We examined the effects of aging on performance in an item-recognition experiment with confidence judgments. A model for confidence judgments and response time (RTs; Ratcliff & Starns, 2013) was used to fit a large amount of data from a new sample of older adults and a previously reported sample of younger adults. This model of confidence judgments allows us to distinguish between changes evidence from memory and changes in decision-related components and it accounts for both RT distributions and response proportions. Older adults took longer to respond than younger adults, older adults exhibited a small decrease in the strength of evidence from memory compared with younger adults and a slight bias toward judging items as "old." The difference in RTs between the 2 age groups was primarily explained by the difference in the nondecision component. Although our small sample size makes the general conclusions about aging tentative, the results are consistent with other research examining the effects of aging in two-choice RT tasks and response-signal tasks, and the study demonstrates that confidence judgment choice proportion and RT distribution data from older adults can be fit with the response time and confidence 2 (RTCON2) model.

Keywords: aging, confidence judgments, RT models, RTCON2 model

For most cognitive tasks, older adults consistently make decisions more slowly but not necessarily less accurately than younger adults. This slowdown has often been interpreted as a generalized processing deficit in the central nervous system (e.g., Cerella, 1985; Deary, 2000; Salthouse, 1996). Previous modeling work, however, has demonstrated that this slow-down, at least in two-choice tasks, can be attributed to increased decision thresholds (older adults are more cautious) and an increase in the duration of processes outside the decision process, such as encoding, memory access, and motor response processes. However, this earlier research has consistently found only slight decreases in the quality of evidence extracted from the stimulus in item recognition memory tasks (Ratcliff, Thapar, & McKoon, 2004, 2006a, 2010, 2011). This experiment was designed to examine the effects of aging on performance in an item recognition memory task with confidence judgments using a model of confidence judgments that has not previously been applied to this population.

We used item recognition because the diffusion model has provided a complete explanation of accuracy and response time (RT) data (including RT distributions for both correct and error responses) from this type of memory task in previous experiments (McKoon & Ratcliff, 2012; Ratcliff et al., 2004, 2006a, 2007, 2010, 2011). In item recognition experiments, subjects study lists of items (items may be words, pictures, etc.) and then, during a later test, must distinguish between items that were on the previous study list ("old" items) and items that were not on the previous study list ("new" items). Other studies investigating aging and item recognition have found only small effects of age on accuracy in this task (e.g., Balota, Dolan, & Duchek, 2000; Bowles & Poon, 1982; Craik, 1994; Craik & Jennings, 1992; Erber, 1974; Gordon & Clark, 1974; Kausler, 1994; Naveh-Benjamin, 2000; Neath, 1998, chap. 16; Old & Naveh-Benjamin, 2008; Rabinowitz, 1984; Schonfield & Robertson, 1966). Older adults make similar patterns of responses as younger adults when making confidence judgments in item recognition tasks (Dodson, Bawa, & Krueger, 2007; Pacheco et al., 2012). This is in contrast to other memory tasks, such as associative recognition, cued and free recall, and source memory, where age has a larger effect on accuracy (Buchler & Reder, 2007; Craik, 1983, 1986; Craik & McDowd, 1987; Healy, Light, & Chung, 2005; Kausler, 1994; Naveh-Benjamin, 2000; Old & Naveh-Benjamin, 2008; Ratcliff et al., 2011; Schonfield & Robertson, 1966; Wahlin, Backman, & Winblad, 1995). Older adults also make different patterns of confidence responses than younger adults in associative recognition and source memory tasks: older adults make more high-confidence false alarms in these tasks (Chua, Schacter, & Sperling, 2009; Dodson, Bawa, & Krueger, 2007; Fandakova et al., 2013; Kelley & Sahakyan, 2003; Pacheco et al., 2012; Shing et al., 2009).

We examined confidence responses in item recognition because confidence judgments have been used extensively to investigate claims about the number and nature of the processes involved in recognition memory (Egan, 1958; Lockhart & Murdock, 1970; Malmberg & Xu, 2007; Ratcliff, McKoon, & Tindall, 1994; Ratcliff, Sheu, & Gronlund, 1992; Yonelinas, 1994). In confidence judgment procedures, subjects rate their confidence that an item is old or new using a response scale with levels ranging, for example, from "very sure old" to "very sure new." These ratings are then used to create receiver operating characteristic (ROC) functions,

This article was published Online First June 22, 2017.

Chelsea Voskuilen, Roger Ratcliff, and Gail McKoon, Department of Psychology, The Ohio State University.

This article was supported by National Institute on Aging Grant R01 AG041176 to Roger Ratcliff.

Correspondence concerning this article should be addressed to Roger Ratcliff, Department of Psychology, The Ohio State University, 291 Psychology Building, 1835 Neil Avenue, Columbus, OH 43210. E-mail: ratcliff.22@osu.edu

which are plots of the cumulative hit rate (old responses to old items) against the cumulative false alarm rate (old responses to new items). Hit and false alarm rates are calculated first for the highest confidence old responses, then for the two highest confidence old response categories (adding together the number of responses in those two categories), then for the three highest, and so on. These hit and false alarm rates are frequently converted to z-scores, resulting in a function called a z-ROC. The slopes of these z-ROC functions have been used to test the predictions of global memory models (Ratcliff & McKoon, 1991; Ratcliff et al., 1994, 1992) and the shapes of these z-ROC functions have been used to make claims about the number of processes and sources of information involved in recognition memory decisions (e.g., DeCarlo, 2002; Yonelinas, 1997, 1999). However, this type of analysis (and the use of ROCs to test the global memory models) typically ignores the RTs associated with these confidence judgments, it assumes that the only source of variability in the decision process is the variability in memory strength between items, and analyses are often conducted on data that have been averaged across subjects. All of these problems with standard z-ROC analyses can be addressed by using the response time and confidence 2 (RTCON2) model (Ratcliff & Starns, 2013). This model produces both accuracy and RT predictions, it includes several sources of variability related to the decision process, it can be fit to individual subjects, and it is able to fit a variety of z-ROC function shapes. This provides an alternative explanation for the shapes of z-ROC functions in item recognition that is based on how subjects utilize confidence response scales and the model and the explanation of shape is constrained by RT data (Ratcliff & Starns, 2013). Additionally, this model can distinguish between various causes of longer RTs such as differences in memory strength or differences in how much subjects emphasize accuracy over speed. This is especially important when modeling data from older adults.

There are many studies that have demonstrated that older adults perform cognitive tasks more slowly than younger adults. This slowing has often been interpreted as a general slowing of processing in the central nervous system because it occurs across a variety of tasks (e.g., Birren, 1965; Brinley, 1965; Cerella, 1985, 1990, 1991, 1994; Fisk & Warr, 1996; Salthouse, 1985, 1996; Salthouse, Kausler, & Saults, 1988). This slowdown is generally considered to be a deficit in that, as all cognitive processes slow down, certain operations may not be completed in time for later operations to use the results of earlier operations and overall processing will be impaired (e.g., Salthouse, 1996). However, this interpretation implies that a slowdown in RT for older adults should be accompanied by a decrease in accuracy and this is not always true (e.g., Ratcliff, Thapar, Gomez, & McKoon, 2004; Ratcliff, Thapar, & McKoon, 2001, 2003, 2004, 2006a, 2006b, 2007, 2010, 2011; Spieler, Balota, & Faust, 1996; Starns & Ratcliff, 2010; Thapar et al., 2003). To properly assess changes in RT and accuracy as a function of age, it is necessary to have a model for understanding both response proportions and RTs and the connection between them. There is a well-known relationship between the speed and accuracy with which people make decisions (Wickelgren, 1977), and any method of evaluating RTs and accuracy from older adults should be able to account for these speed/accuracy tradeoffs because older adults tend to emphasize accuracy more than younger adults (Salthouse, 1979).

Previous work by Ratcliff et al. (2004, 2006a, 2006b, 2007, 2010) has used Ratcliff's (1978) diffusion model to examine changes in performance in two-choice tasks as a function of age. These studies have demonstrated that, at least in these tasks, most of the RT slowdown observed in older adults can be attributed to a combination of different response strategies and longer nondecision times, but not to any deficit in the evidence used in the decision process. That is, older adults took longer to make responses because they were more cautious about making mistakes and so responded more slowly and carefully than younger adults, and the older adults took longer to encode a stimulus, extract decision-related information from memory, and make a motor response. All of these studies have found differences between older and younger adults in the amount of time needed to encode a stimulus, extract decision-related information from memory, and make a response (i.e., nondecision time), but results have been varied in terms of response caution. Most studies have found that older adults require more information to make a decision than younger adults (i.e., have higher decision thresholds; e.g., Ratcliff et al., 2001, 2006a), but some studies have found no difference between older and younger adults' decision thresholds for some tasks or conditions (e.g., Ratcliff, 2008; Ratcliff et al., 2003, 2006a). Specifically, no difference between older and younger adults in boundary separation was observed in a brightness discrimination task (Ratcliff et al., 2003), a response signal task (Ratcliff, 2008) and the difference is reduced or eliminated when subjects are instructed to emphasize accuracy over speed (Ratcliff et al., 2004, 2006a).

Our goal is to apply a similar analysis to results from a confidence judgment paradigm. As mentioned previously, confidence judgments have been used to test different models of memory and are used, inappropriately in our view, to determine how many memory processes are being used to make a decision. However, most previous work examining confidence judgments and aging have ignored RTs. This is especially problematic when comparing data from older and younger adults since younger adults may be more willing than older adults to make faster responses at the expense of accuracy (Basowitz & Korchin, 1957; Silverman, 1963; Starns & Ratcliff, 2010; Strayer & Kramer, 1994; Thorndike, Bregman, Tilton, & Woodyard, 1928). We are also generally interested in examining how performance on memory tasks changes with age and it is important to attempt to disentangle differences in memory strength from differences in how people use confidence response scales. To do this, we will apply the RTCON2 model to confidence judgments from older and younger adults in an item recognition task.

RTCON2 Model

The RTCON2 model has previously been applied to confidence judgments in item recognition and associative recognition as well as motion discrimination tasks and was shown to provide a better fit to the data than several competing decision models (Ratcliff & Starns, 2013; Voskuilen & Ratcliff, 2016). In the RTCON2 model, the evidence available to the decision process on a single trial (i.e., the memory strength for a particular item) is assumed to be a distribution across the evidence-strength dimension rather than a discrete value (cf. Beck et al., 2008; Gomez, Ratcliff, & Perea, 2008; Jazayeri & Movshon, 2006; Ratcliff, 1981; Ratcliff &

Starns, 2009). These evidence distributions have a *SD* of 1 and their mean location varies across trials, according to a distribution of the mean drift rate. The bottom portion of Figure 1 illustrates how the distribution of evidence for a single item feeds into the decision process. Confidence criteria are used to divide the evidence-strength dimension into multiple response regions corresponding to different levels of confidence. Each response region has its own accumulator and decision boundary, as shown in the top portion of Figure 1, and the accumulators race until one of them reaches its decision boundary and that response is made. In this kind of paradigm, each confidence category requires a different motor response which we argue requires a different accumulator and decision boundary for each possible response.

The mean of the evidence distribution is called the drift rate (v), and it is determined by the quality of information extracted from the stimulus. The quality of information from stimuli of the same type (e.g., high-frequency words) is allowed to vary across trials to reflect differences in the encoding and retrieval of information from memory. This between-trial variability in drift rate is assumed to be normally distributed with *SD* η. The average rate of accumulation for each response accumulator is determined by the proportion of the within-trial distribution of evidence in each of the response regions. This accumulation process is subject to moment-to-moment variability such that processes with the same drift rate will not always terminate at the same time or with the same confidence response.

RTCON2 uses a constant summed evidence algorithm to model the accumulation of evidence in each response accumulator. In this algorithm, the change in evidence on each time step is determined by its drift rate and noise. On each time step, one of the response accumulators is selected randomly and increased, and some of the other response accumulators are decreased such that the sum of the total decrease is equal to the increase in the selected accumulator.

Two versions of the constant summed evidence model have been examined, one with all accumulators decremented and one with only some of the accumulators decremented. (Ratcliff & Starns, 2013; Voskuilen & Ratcliff, 2016). For this application, we used a version of the model where an increase in evidence in one of the new accumulators would cause a decrease in evidence only in the old accumulators, but not the other new accumulators (and vice versa). This version of the algorithm (compared with the one in which all other accumulators are decremented) represents the assumption that evidence for one type of response (old or new) does not compete with other confidence levels of that same response. The expressions for the changes in evidence (Δx) for each accumulator at each time step (Δt) are given in Equations 1 and 2. Equation 1 describes the update in evidence for the selected accumulator and Equation 2 describes the corresponding change in activity for the nonselected accumulators.

$$\Delta x_i = a v_i \Delta t + \sigma \eta_i \sqrt{\Delta t} \tag{1}$$

$$\Delta x_j = -\left(\frac{1}{N - \frac{N}{2}}\right)\left(a v_i \Delta t + \sigma \eta_i \sqrt{\Delta t}\right) = -\left(\frac{1}{N - \frac{N}{2}}\right)\Delta x_i \tag{2}$$

If the selected accumulator was one of the new accumulators, then Equation 2 would be used to adjust the old accumulators, but the other new accumulators would be unchanged. In these equations, a is a scaling parameter that adjusts drift rate (v_i, the area under the distribution in a particular response region), σ is within-trial variability in the accumulation process, and η is a normally distributed random variable with mean 0 and *SD* 1.

RT predictions are obtained from the model by adding the decision time (the time taken for one of the evidence accumulators to reach a decision boundary) to a uniformly distributed nondecision time (the exact choice of the distribution shape is not critical so long as the *SD* in nondecision time is smaller than that of the decision time, Ratcliff, 2013). RT predictions are also dependent on the height of the decision boundaries, which vary from trial to trial over a uniform distribution with a range of s_b.

This model can produce longer RTs in several ways. Smaller values of drift rate for a particular response region will produce longer RTs for that confidence response than larger values of drift rate (note that the drift rate associated with a response is determined by both the location of the drift distribution for a particular condition and the positions of the confidence criteria). Larger boundary values, indicating a more conservative threshold for making a response, will produce longer RTs than smaller boundary values. Larger values of nondecision time will also produce longer RTs than smaller values of nondecision time. However, while all of these parameter changes can produce longer mean RTs, they also all produce different predictions regarding the shape and location of the RT distributions and are therefore distinguishable. Changes in drift rate primarily affect the tails of the RT distributions and have only a slight effect on the location of the leading edges of the distributions. Smaller drift rates will also be associated with lower accuracy. Changes in the boundary settings affect both the leading edge and the tails of the RT distribution and will have a small effect on accuracy (larger boundary values are associated with higher accuracy). Changes in nondecision time will change the position of the RT distribution but will not affect its shape or the accuracy of the responses. Therefore, this model can disentangle differences in memory performance (as measured by drift rates) from differences in decision process settings (such as changes in boundary separation) and differences in nondecision

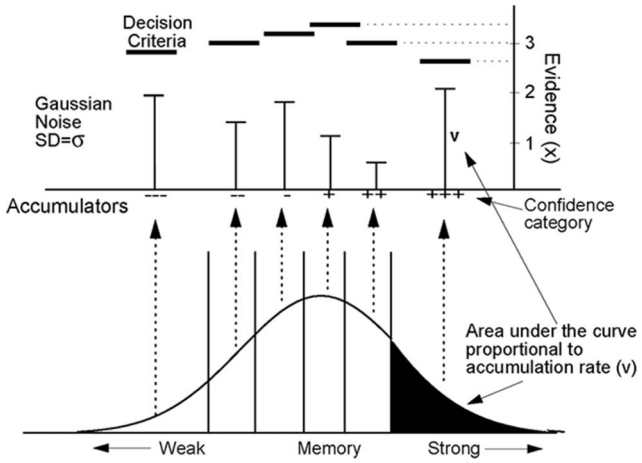

Figure 1. Response time and confidence 2 (RTCON2). The distribution of evidence for an item on a given trial drives six mutually inhibitory accumulators (one for each confidence category). The proportion of the distribution between the confidence criteria on match dimension drives the drift rate for each confidence category. When one of the accumulators reaches its decision boundary, the corresponding response is made.

times and so can be used to examine which of these processes are responsible for slowing in older adults.

Experiment

The goal of the experiment was to collect data from an item-recognition task with confidence judgments that we will fit with the RTCON2 model to use the resulting model parameters to investigate the effects of aging on performance in this task. Data from the younger subjects were originally collected and reported in Ratcliff et al. (1994, Experiment 5) and also modeled in Ratcliff and Starns (2013) while the data from the older subjects were collected more recently. Data from the younger subjects are refit here identically to the way the data from the older subjects are fit. Because this model is fit to both response proportions and RT quantiles, to obtain high quality model fits, we need a large amount of data per subject to provide reliable estimates of RTs and choice proportions across conditions and confidence levels. To that end, we have collected a data from many sessions from a relatively small number of subjects. While this limits the conclusions we can draw about aging effects in general, the patterns of results and effect sizes are consistent with results from previous aging studies with two-choice tasks. This experiment was approved by The Ohio State University Social and Behavioral Institutional Review Board (IRB).

Method

Subjects. Eleven Northwestern undergraduates completed a total of 97 1-hr sessions, after one practice session per subject was eliminated (resulting in 7–11 sessions per subject). These data were previously reported in Ratcliff et al. (1994). Twelve older adults (age 60–80) completed a total of 96 1-hr sessions (eight sessions per subject). We did not eliminate the first session of data from the older adults (i.e., treat it as a practice session) both because the older adults had participated in other experiments from this lab (and so were practiced at this type of task) and because we did not want to eliminate any data if it was not necessary (because it was more difficult to get multiple sessions of data with this population). To ensure this did not affect our results, we also fit the model to data from just Sessions 2–7 from the older adults and the results were unchanged (see Appendix). Older adults were recruited from senior citizen centers in the Columbus area and paid $15 for each session they completed. The older subjects had to meet the following inclusion criteria to participate in the study: a score of 26 or above on the Mini-Mental State Examination (Folstein, Folstein, & McHugh, 1975) and no evidence of disturbances in consciousness, medical or neurological disease causing cognitive impairment, head injury with loss of consciousness, or current psychiatric disorder. One of the older adult subject's data were excluded because that person's performance was at chance for most of the conditions. Older subject characteristics are presented in Table 1. Subject characteristics for the younger adults are not available, but the characteristics of the older subjects match those of the Northwestern undergraduate population used in similar studies (see Table 1 with characteristics from Ratcliff et al., 2001, 2003, 2004).

Materials. The stimuli were drawn from two pools of words formed from the Kucera and Francis (1967) word frequency lists. Words in the low-frequency pool had frequencies of either 4 or 5, and words in the high-frequency pool had frequencies between 78 and 10,601. The words varied from 4 to 10 letters in length. Words derived from other common words by adding suffixes (e.g., -ing, -ed, or -tion), plurals, and proper names were eliminated. This resulted in a high-frequency pool of 815 words and a low-frequency pool of 871 words.

Procedure. Study lists were composed of pairs of words to minimize the possibility of rehearsal trading strategies (see Ratcliff et al., 1990). Study lists consisted of either pure or mixed lists. In a pure list, each of 16 pairs was presented for the same amount of time, 1.5 s for weak or 5 s for strong items. In a mixed list, sequential blocks of pairs in the study list had different study times: the first 2 pairs at 1.5 s, the next 6 pairs at 5 s, the next 6 pairs at 1.5 s, and the last 2 pairs at 5 s, or the reverse ordering of presentation times. For both pure and mixed lists, within each middle block of 6 pairs, 3 pairs for which both words were high frequency and 3 pairs for which both words were low frequency were placed in random positions. The first and last 2 pairs in a list were buffer items, and one word of each buffer pair was high frequency and one low frequency. Subjects were instructed to learn the pairs for later cued-recall tests. The cued recall task was included to encourage subjects to focus on the pairs of words and minimize rehearsal, but performance on this task was not scored. In the 16 lists for a session, there were four of each type of list: pure weak, pure strong, and the two kinds of mixed lists. There

Table 1
Subject Characteristics

		Younger adults				
Measure	Older adults	Ratcliff, Thapar, Gomez, et al., 2004; Experiment 1	Ratcliff, Thapar, Gomez, et al., 2004 Experiment 2	Ratcliff, Thapar, & McKoon, 2004	Ratcliff, Thapar, & McKoon, 2003	Thapar, Ratcliff, & McKoon, 2003
Years education	16.14 (3.12)	13.12 (1.11)	13.58 (1.55)	12.6 (.9)	12.67 (1.03)	12.36 (1.04)
MMSE	29.27 (.65)	29.00 (.80)	29.04 (1.21)	29.0 (1.1)	29.11 (.94)	29.13 (1.06)
WAIS-III vocabulary (scaled score)	12.64 (2.58)	—	—	14.4 (1.9)	14.49 (2.26)	14.24 (2.12)
WAIS-III matrix reasoning (scaled score)	12.91 (3.45)	—	—	10.8 (2.5)	11.24 (2.79)	10.71 (2.32)
WAIS-III IQ	115.82 (12.43)	117.31 (8.76)	116.69 (11.91)	—	116.76 (12.11)	114.46 (9.14)

Note. MMSE = Mini-Mental State Examination; WAIS-III = Wechsler Adult Intelligence Scale-3rd edition. The young adult background characteristics are for a group of subjects from the same pool as those tested here. These data are from: Ratcliff, Thapar, Gomez, & McKoon, 2004; Ratcliff, Thapar, & McKoon, 2003, 2004; Thapar, Ratcliff, & McKoon, 2003.

were 64 test items associated with each study list (the 32 items from the study list along with 32 new items) and these items were presented in random order. Test lists immediately followed each study list with no retention interval. Subjects responded using a PC keyboard with the *X*, *C*, *V*, *B*, *N*, and *M* keys. Subjects were instructed to place their left-hand ring, middle, and index fingers on the *X*, *C*, and *V* keys and their right-hand index, middle, and ring fingers on the *B*, *N*, and *M* keys. Subjects were instructed that the *M* key stood for "very sure old," the *N* key stood for "somewhat sure old" and so on. After each response, there was a 250-ms blank interval followed by the next test item. For two randomly chosen study lists, the recognition test list was followed by a cued-recall test (the left member of the study pair was presented and the subject was required to recall the right member). For our analyses, we collapsed across list type leaving us with six conditions: Word Frequency (high or low) × Strength (new, weak, or strong). These manipulations were originally designed to investigate the slope of the z-ROC function across various conditions, as discussed in Ratcliff et al. (1994), and are repeated here to produce an equivalent data set from older subjects.

Model Fitting

The RTCON2 model was fit to each individual subject's response proportion and RT quantiles (.1, .3, .5, .7, and .9) for each of the six confidence response for each of the six conditions (as in Ratcliff & Starns, 2013). The RT quantiles segment the response proportion data into six bins of response proportions for each confidence category. Initial parameter values were chosen that produced predictions similar to the empirical data and then a simplex function (Nelder & Mead, 1965) was used to adjust the parameters of the model until the predictions matched the data as closely as possible. The match between the empirical data and the model predictions was quantified by a χ^2 statistic, which was minimized by the simplex function (see Ratcliff & Tuerlinckx, 2002 for more detail). Because there are no exact solutions for this model, simulations are used to generate predicted values from the model. To simulate the process of accumulation given by Equations 1 and 2, we used the simple Euler's method with 1-ms steps (cf. Brown, Ratcliff, & Smith, 2006; Usher & McClelland, 2001). For each millisecond step, one accumulator was chosen randomly, and the evidence in it was incremented or decremented according to Equation 1 and opposite accumulators were incremented or decremented according to Equation 2 (e.g., if the selected accumulator was for one of the new responses, then the evidence in the old accumulators would be adjusted according to Equation 2 and the other new accumulators would be unchanged). For each condition, 20,000 simulations of the decision process were used to generate the response proportions and RT quantiles for each confidence category.

There are six RT bins for each confidence response, which gives 36 degrees of freedom for the six-choice task. However, these response proportions have to add to one, which reduces the degrees of freedom to 35 for each condition. With six conditions, this gives a total of 210 degrees of freedom in the data. To enable comparisons across the two age groups, some of the parameter values were fixed across the two age groups. These parameters were the scale on the drift rate (a), the within-trial variability of the diffusion process (σ), and the between-trial variability in the decision boundaries (s_b). These parameters were chosen because they can potentially complicate the comparison of other parameters of interest (e.g., the two age groups could have the same drift rates but different scaling parameters such that the rate of evidence accumulation would be different for the two groups). To fix these parameters, we first fit all of the data with these parameters allowed to freely vary and then refit all of the data with these parameters fixed to their mean values (across old and young the subjects) from the initial fit.

Results

Data from this experiment consisted of response proportions and RT quantiles for each subject from each condition and each confidence response. For the older subjects, RTs less than 400 ms or greater than 7,000 ms were excluded from our analyses. This excluded 0.2% of the data. For the younger subjects, the cutoffs were 300 ms and 3,000 ms, excluding 1.7% of the data (as in Ratcliff & Starns, 2013). Different cutoff values are required for the different age groups because the older adults' RTs are considerably longer than the younger adults'. Lower cutoff values were chosen based on performance (i.e., at what time point does performance rise above chance) and upper cutoff values were chosen to exclude extreme values while minimizing loss of data.

We conducted several analysis of variances (ANOVAs) to investigate the effects of age and the experimental manipulations on our various dependent variables. First, we investigated differences in median RTs as a function of age, experimental condition, and confidence response. Figure 2 plots the average median RT as a function of the confidence response for each condition and age group. Across all of the conditions and confidence responses, the older adults' median RTs (M = 1,209 ms) were longer than the younger adults' median RTs (M = 988 ms) and this difference was significant ($F(1, 20) = 11.69$, $p < .05$, $\eta^2 = 0.13$). Across all of the conditions and age groups, there were significant differences in median RTs across confidence level ($F(5, 100) = 28.87$, $p < .05$, $\eta^2 = 0.24$) with higher confidence responses having smaller median RTs than lower confidence responses. However, the pattern of median RTs across confidence responses was not identical for older and younger adults ($F(5, 100) = 2.80$, $p < .05$, $\eta^2 = 0.02$). The changes in RTs across confidence levels are more extreme for the younger adults than for the older adults such that there is a larger difference between the two groups for high-confidence responses than for low-confidence responses. Second, we investigated differences in response proportions as a function of age, experimental condition, and confidence response. For this analysis, the ordering of the confidence scale for the new conditions has been reversed such that for all conditions the confidence scale ranges from a high-confidence incorrect response to a high-confidence correct response (i.e., we're collapsing across old and new responses). Some confidence responses were made more frequently than others ($F(5, 100) = 49.33$, $p < .05$, $\eta^2 = 0.56$) that is unsurprising given that performance was above chance (such that correct responses should be made more often than incorrect responses) and the pattern of confidence responses varied across conditions ($F(25, 500) = 16.43$, $p < .05$, $\eta^2 = 0.08$). There was no main effect of age ($F(1, 20) = 4.00$, $\eta^2 < 0.001$) and no significant interactions between age and the other factors (all F values <1.7). Figure 3 shows the response proportions for older

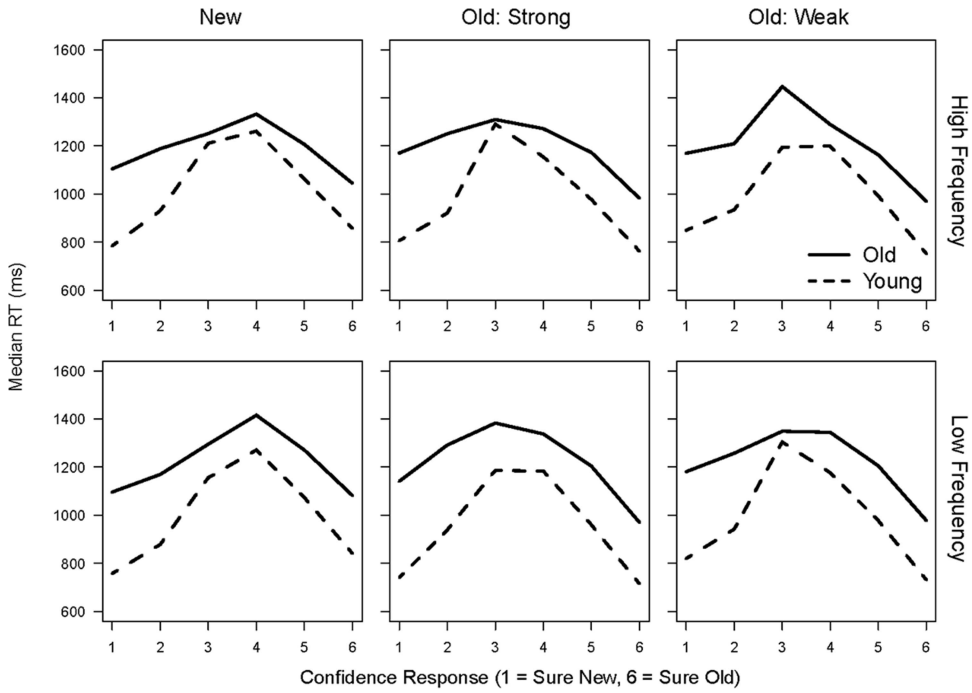

Figure 2. Median responses time (RTs) as a function of the confidence response, experimental condition, and age group. The dashed lines represent the averaged data from the younger adults and the solid lines represent the averaged data from the older adults.

and younger subjects collapsed across all conditions (note that the response scale in this figure ranges from high-confident incorrect to high-confident correct). Although older adults made slightly more medium-confidence error responses (Response 2), this difference was not significant. This is consistent with previous studies finding no difference between older and younger adults in patterns of confidence responses in item recognition tasks (Dodson et al., 2007; Pacheco et al., 2012). Figure 4 plots the average proportion of responses for each confidence response for each condition and age group (with the regular ordering of the confidence scale). Across all of the conditions the pattern of responses for older and younger subjects is remarkably similar. Third, we investigated differences in accuracy as a function of age and experimental condition. For this analysis, response proportions were combined across confidence levels to yield a single accuracy level for each condition for each subject (e.g., response proportions for the three new confidence categories were combined to yield an accuracy value for the new conditions). The younger subjects were slightly more accurate (about 5%) than the older subjects (see Table 2 for mean accuracy values across conditions, collapsed across confidence levels), and this difference was significant ($F(1, 20) = 4.71$, $p < .05$, $\eta^2 = 0.05$). There were also significant differences in accuracy across conditions ($F(5, 100) = 14.22$, $p < .05$, $\eta^2 = 0.29$) but these differences were consistent for older and younger subjects ($F(5, 100) = 0.96$, $\eta^2 = 0.02$). Overall, the behavioral effects of age on RT and accuracy were consistent with previous work (cf. Ratcliff et al., 2004, 2007).

The RTCON2 model was applied to data from each individual subject. Averages of the fits are presented here and individual fits are in the Appendix. Mean parameter estimates and *SD*s for each age group are shown in Table 3 along with average χ^2 values. To enable comparisons across the two age groups, the scale on the drift rate (a), the within-trial variability of the diffusion process (σ), and the between-trial variability in the decision boundaries (s_b) were set to fixed values (shown in Table 3). These values were chosen by initially allowing these parameters to vary freely and then using the mean estimates across all subjects as the fixed values. When these parameters were allowed to vary freely, none of the results described below were changed and there were no significant differences between the two groups for these parameters (mean parameter values from fits with parameters varying freely are presented in the Appendix). To make the model identifiable, the mean drift rate for the first condition was also fixed to zero (with all other mean drift rates and all of the confidence criteria freely varying). In perceptual tasks, older adults have been shown to have larger practice effects than younger adults (Ratcliff et al., 2006b). To control for a possible practice confound, we repeated all of our modeling analyses using only data from Sessions 2–8 for each subject and the pattern of results was unchanged. Mean parameter values from fits to just Sessions 2–8 are also presented in the Appendix.

There were several significant differences in model parameters across the two age groups. The older subjects had significantly larger nondecision time components ($M = 559$ ms) than the younger subjects ($M = 370$ ms; $F(1, 20) = 29.10$, $p < .05$, $\eta^2 = 0.59$) that is consistent with previous research (e.g., Ratcliff et al., 2004, 2007). To compare drift rates across the two age groups, we subtracted the middle confidence criterion from all of the mean drift rates (so that all of the values would be centered around the middle of each subjects' confidence scale) and reversed the sign of

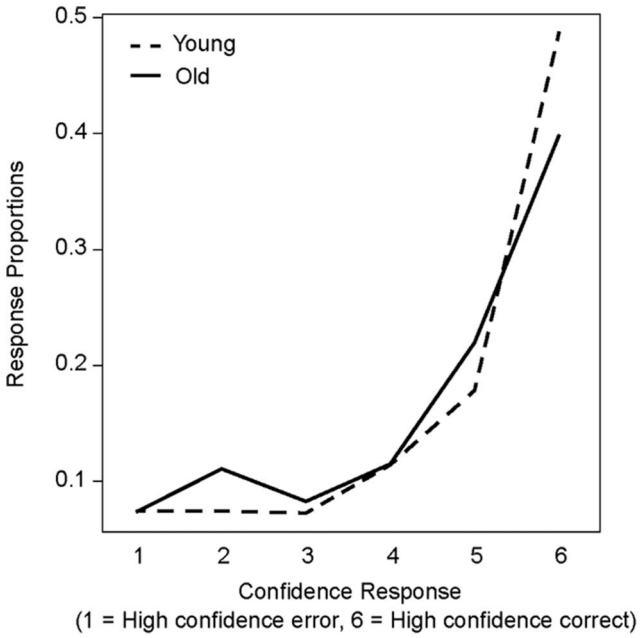

Figure 3. Response proportions for each confidence response averaged across conditions. The ordering of the confidence scale for the "new" conditions has been reversed such that for all conditions the confidence scale ranges from a *high-confidence error response* (1) to a *high-confidence correct response* (6). The dashed line represents the averaged data from the younger adults and the solid line represents the averaged data from the older adults.

the mean drift rates for new conditions (so that for all conditions a larger drift rate indicates more evidence in favor of the correct response). We conducted an ANOVA with experimental condition and age as factors and found that the younger adults had more extreme drift rates ($M = 1.76$) than the older adults ($M = 1.35$; $F(1, 20) = 6.60$, $p < .05$, $\eta^2 = 0.07$) and there were significant differences in drift rates across conditions ($F(5, 100) = 52.32$, $p < .05$, $\eta^2 = 0.50$). Across conditions, mean drift rates ranged from 0.71 to 2.50. We conducted a similar analysis to compare decision boundaries across age groups and conditions. Unlike some previous research (e.g., Ratcliff et al., 2004, 2007), we found no significant difference between the old ($M = 2.15$) and young ($M = 2.12$) subjects in terms of the height of their decision boundaries ($F(1, 20) = 0.10$, $\eta^2 = 0.001$). There were significant differences across conditions ($F(5, 100) = 5.14$, $p < .05$, $\eta^2 = 0.14$) with high-confidence response categories having lower bounds ($M = 2.08, 1.82$) than low-confidence response categories ($M = 2.33, 2.45$) that is the same pattern that was observed for RTs. The changes in bounds across conditions were not the same for older and younger adults ($F(5, 100) = 2.82$, $p < .05$, $\eta^2 = 0.08$), which will be discussed below and is consistent with the behavioral finding that the pattern of RTs across confidence levels was not the same for older and younger adults.

The best-fitting parameter estimates for each subject were used to generate predicted data from the model and the predicted data were then compared with the empirical data. The averaged empirical RT data are plotted along with the averaged predicted RT data in Figure 5. The six confidence categories are plotted along the *x*-axis and the five RT quantiles for each confidence level are plotted vertically. Overall, the model's predictions appear to match the data quite well. Figure 6 shows all of the empirical data (both the RT quantiles and the response proportions) from all of the subjects and conditions plotted against the predicted data along with a reference line (with slope of one and intercept of zero). The red (gray) points represent conditions with fewer than 25 observations. Again, the model's predictions appear to match the data quite well.

The ROC and z-ROC functions for the averaged data and model predictions are shown in Figure 7 along with the average decision boundaries. The ROC functions for the older adults, shown in Panel A, are more symmetric over the negative diagonal than the ROC functions for younger adults. The z-ROC functions for both groups of subjects, shown in Panel B, are approximately linear. Both the ROC and z-ROC functions for the younger adults are slightly higher than the functions for the younger adults as a result of the small difference in accuracy between the two groups. The decision boundaries for each confidence category and each age group are shown in Panel C. As mentioned previously, there was no significant effect of age on boundary settings (i.e., younger and older adults had similar average boundary values). However, the changes in bounds across conditions were not the same for older and younger adults. On average the older adults adopted slightly lower boundaries for the old response categories and slightly higher boundaries for the new response categories (see the bottom row of Figure 7). Averaged across the different confidence levels, for older adults the mean decision boundary for old responses was 2.06, the mean boundary for new responses was 2.24, and this difference was significant, $t(10) = 2.45$, $p < .05$. In other words, the older adults required a slightly greater amount of evidence to decide an item was new than they did to decide an item was old. Behaviorally, this pattern of decision boundaries would result in slight increase in both hit and false alarm rates relative to a symmetric pattern of decision boundaries. In contrast to the older adults, the younger subjects' decision boundaries were relatively symmetric across the old and new response categories. For younger adults the mean decision boundary for old responses was 2.19, the mean boundary for new responses was 2.04, and this difference was not significant, $t(10) = -1.80$, $p > .05$. The changes in the decision boundaries across confidence levels were also less extreme for the older adults than for the younger adults. The younger adults show a more pronounced inverted u-shape across confidence levels where the older adults show a slight u-shape across confidence levels for only the old responses. This difference captures the observed RT pattern where the difference between older and younger subjects' RTs was greater for the high-confidence responses than for the low-confidence responses.

Overall, the older adults made responses more slowly and slightly less accurately than the younger adults and the model was able to account for these effects. The change in RTs for the older adults relative to younger adults was fit by the model mainly as a change in nondecision time (see Ratcliff, 2008). A change in nondecision time produces shifts in the RT distribution, but does not affect the overall shape of the distribution. This is consistent with the observed RT distributions for the older and younger adults. If we plot the RT quantiles of the older subjects against the RT quantiles of the younger subjects, the resulting line is close to linear and is well described by a line with a slope of one and an

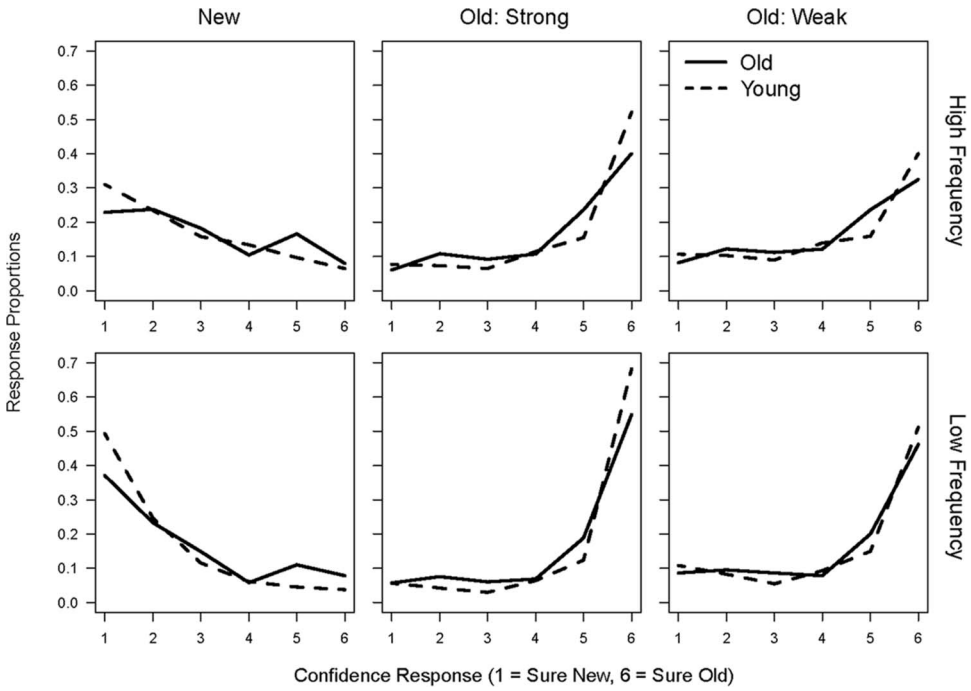

Figure 4. Response proportions for each confidence response, experimental condition, and age group. Dashed lines represent the averaged data from the younger adults and solid lines represent the averaged data from the older adults.

intercept equal to the average T_{er} difference between the old and young subjects (as shown in Figure 8). The points fall quite close to this line indicating that the older adults' RT distribution is of the same shape as the younger adults' RT distribution and is merely shifted by an increase in nondecision time. For comparison, the best-fitting line for these data points is also included in Figure 8 (in red). To examine (and eliminate) the possibility that these results reflect tradeoffs between parameter values rather than valid individual differences, we simulated 40 sets of data using the mean parameter values for the older adults and the same number of observations per condition as in our behavioral data, and then fit RTCON2 to the simulated data. The best-fitting estimates for nondecision time, variability in nondecision time, and the average of the six decision boundaries for each fit are plotted against each other in Figure 9A along with correlations between these parameters. There is a correlation between the nondecision estimates and the decision boundary estimates showing that these parameters may tradeoff. However, the range of recovered nondecision estimates is about 30 ms, which is much smaller than the difference between older and younger subjects observed in our experiments (around 190 ms). We also generated simulated data with either large or small nondecision times (600 or 300 ms) or large or small decision boundaries (average height of 2.35 or 1.95) and then attempted to fit the data forcing the wrong parameter to account for the effects. That is, if the data were simulated using the smaller nondecision time, we fixed the nondecision time to the larger value, fixed all other parameters except the boundaries to the true generating values, and only allowed the decision boundaries to vary. The resulting fits are shown in Figure 9B. The first row shows the data simulated with a large or small value of nondecision time (the numbers) along with the best-fitting predictions from the model fit with only decision boundaries changing (the dashed lines). The model can produce reasonable predictions for the median RT values across conditions, but produces RT distributions with too much or too little spread. The second row shows the data simulated with large or small average decision boundaries

Table 2
Mean Probability of Correct Responses ("Old" Responses in Old Conditions and "New" Responses in New Conditions) and SDs Across Conditions and Age Groups, Collapsed Across Confidence Levels

Age group	New, HF words	Old strong, HF words	Old weak, HF words	New, LF words	Old strong, LF words	Old weak, LF words
Old	.65 (.13)	.74 (.06)	.68 (.08)	.75 (.14)	.81 (.07)	.74 (.08)
Young	.70 (.12)	.79 (.08)	.70 (.10)	.86 (.05)	.87 (.04)	.75 (.06)

Note. HF = high-frequency; LF = low-frequency.

Table 3
Mean Parameter Values and SDs Across Subject Groups

Age group	T_{er}	s_t	a	σ	s_b	χ^2
Old	559 (104)	124.0 (23.1)	0.028	0.100	0.400	861 (347)
Young	370 (52.0)	83.6 (33.0)	0.028	0.100	0.400	910 (226)
	b_1	b_2	b_3	b_4	b_5	b_6
Old	2.33 (.81)	2.20 (.60)	2.20 (.40)	2.29 (.35)	2.09 (.34)	1.80 (.45)
Young	1.83 (.47)	1.84 (.27)	2.45 (.62)	2.61 (.62)	2.13 (.28)	1.84 (.50)
	c_1	c_2	c_3	c_4	c_5	
Old	−1.24 (.47)	−.06 (.21)	0.69 (.15)	1.47 (.33)	2.84 (.37)	
Young	−1.07 (.55)	−.07 (.37)	0.74 (.24)	1.76 (.37)	2.83 (.48)	
	$v_{N\text{-}HF}$	$v_{OS\text{-}HF}$	$v_{OW\text{-}HF}$	$v_{N\text{-}LF}$	$v_{OS\text{-}LF}$	$v_{OW\text{-}LF}$
Old	0.00	2.03 (.63)	1.68 (.52)	−0.69 (.74)	2.76 (.87)	2.29 (.73)
Young	0.00	2.61 (.49)	2.03 (.50)	−1.05 (.41)	3.67 (.34)	2.68 (.52)
	$s_{N\text{-}HF}$	$s_{OS\text{-}HF}$	$s_{OW\text{-}HF}$	$s_{N\text{-}LF}$	$s_{OS\text{-}LF}$	$s_{OW\text{-}LF}$
Old	1.34 (.20)	1.67 (.07)	1.65 (.12)	1.57 (.22)	1.75 (.11)	1.82 (.09)
Young	1.06 (.22)	1.68 (.10)	1.68 (.14)	1.37 (.29)	1.74 (.21)	1.89 (.20)

Note. T_{er} is the mean nondecision time, s_t is the range in nondecision time, σ is the SD in within trial variability, a is the scaling factor that multiplies drift rate, s_b is the range in variability in the decision boundaries, b_1–b_6 Are the decision boundaries, c_1–c_5 are the confidence criteria, the *v* values are the mean values of the drift rate distributions for each experimental condition, and the s values are the between-trial variability values for each experimental condition. χ^2 is the goodness-of-fit value for the model fits. N-HF = New, high-frequency; OS-HF = Old strong, high-frequency; OW-HF = Old weak, high-frequency; N-LF = New, low-frequency; OS-LF = Old strong, low-frequency; OW-LF = Old weak, low-frequency.

(the numbers) along with the best-fitting predictions from the model fit with only nondecision time changing (the dashed lines). The model can produce reasonable predictions for the .1 quantiles of the RT distributions, but produces distributions with the wrong amount of skew in the tails. These simulations demonstrate that, while there may be small tradeoffs between nondecision and bound height parameters when fitting RTCON2, these parameters affect RT distributions in distinguishable ways such that the two parameters are not interchangeable.

The slight decrease in accuracy for the older adults was fit by the model as a change in drift rate. If we compare a measure of d' generated from the data with a model-based measure of d', we see that the two measures are in close agreement (see Figure 10). The behavioral d' is calculated by subtracting the normalized (z transformed) false alarm rate from the normalized hit rate for each condition and subject (collapsing across confidence levels). The model-based d' is calculated from the parameter estimates for the evidence strength distributions by plugging the appropriate values into the formula below (where μ_o and σ_o are the mean and *SD* of the old items and μ_N and σ_N are the mean and *SD* of the new items):

$$d' = \frac{\mu_o - \mu_N}{\sqrt{\frac{1}{2}(\sigma_o^2 + \sigma_N^2)}}$$

The v_i parameters may be used for μ_o and σ_N as these parameters give the means of the evidence strength distributions across trials. On each trial, the evidence strength distribution has a *SD* of one and the location of the distribution varies across-trials according to the parameters such that the *SD* of the evidence distribution across trials is $\sqrt{s_i^2+1^2}$.

While the younger adults had larger d' values on average than the older adults, it's also worth noting that there is a large degree of overlap between the two groups illustrated in Figure 10 (each point represents one subject and one condition). Although each subject is contributing four points to this figure (one per condition), the results are not dependent on any one subject. For example, for the older adults the six largest d' values come from five different subjects. The model is thus able to account for the differences in both RT and accuracy observed between the two age groups.

Discussion

We examined the effects of aging on performance in an item-recognition experiment with confidence judgments. Because our sample size is relatively small, our general conclusions about aging are tentative. However, the results and effect sizes are consistent with previous research investigating decision-making and aging. Consistent with previous research, the older adults in this experiment responded more slowly than the younger adults and slightly less accurately. We fit these data with the RTCON2 model and the model was able to capture these effects in an appropriate way. The change in RT was reproduced in the model as a change in nondecision time. The slight change in accuracy was reproduced in the model as a slight decrease in drift parameters. The effect of age on nondecision time was moderately large ($\eta^2 = 0.59$). The effect of age on drift rates was small ($\eta^2 = 0.07$) and by contrast, the effect

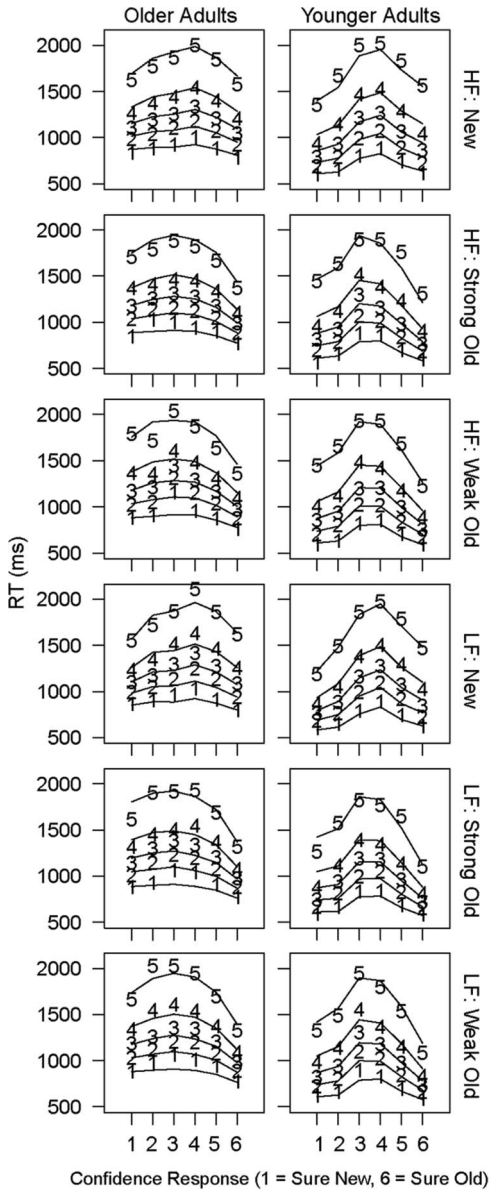

Figure 5. Averaged empirical response time (RT) data and averaged predicted RT data for each condition and age group. The six confidence categories are plotted along the *x*-axis and the five RT quantiles for each confidence category are plotted vertically. The numbers 1–5 represent the average RT quantiles from the behavioral data and the lines represent the average predictions from RTCON2.

of condition on drift rates was large ($\eta^2 = 0.50$). This shows that the drift rate result across age groups is not the result of large amounts of noise or insufficient sample size. Both of these findings are consistent with previous modeling work investigating changes in decision-making as a function of age (Ratcliff et al., 2001, 2003, 2004, 2006a, 2006b, 2007, 2010, 2011). Earlier research has found only slight decreases in drift rates in item recognition memory tasks (Ratcliff et al., 2004, 2006a, 2010, 2011) and larger decreases in other tasks (e.g., letter discrimination: Thapar, Ratcliff, & McKoon, 2003; associative recognition: Ratcliff et al., 2011; Ratcliff & McKoon, 2008; McKoon & Ratcliff, 2012). Although differences in drift rate will also affect RTs, the difference in RTs between the two age groups

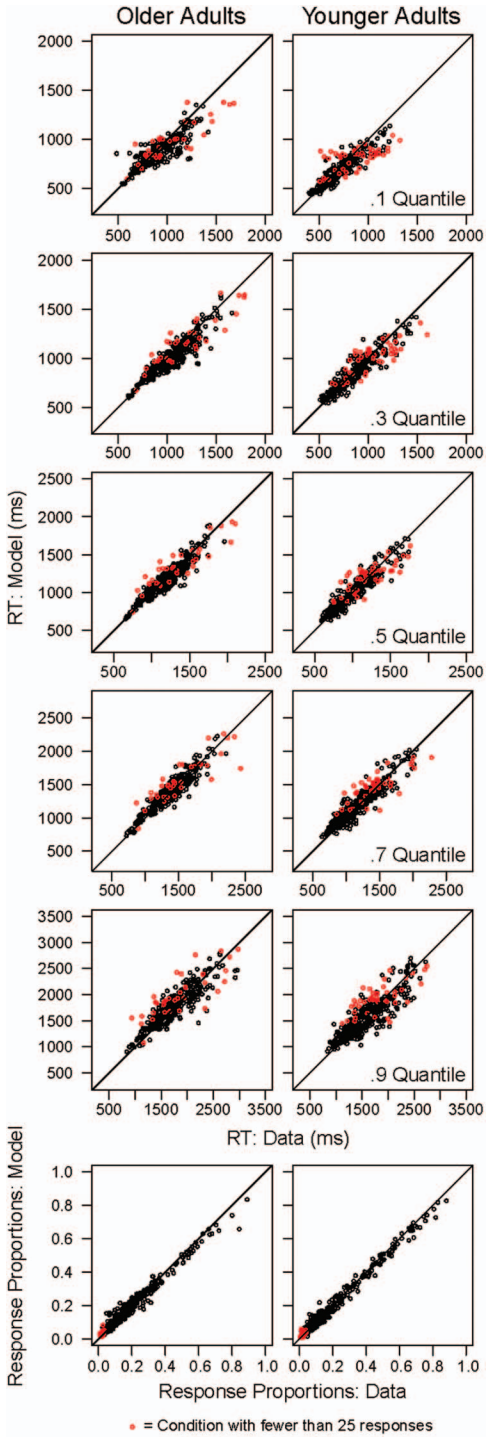

Figure 6. Empirical data (quantile response time (RTs) and response proportions) plotted against model predictions from all subjects and all experimental conditions with reference lines with intercept of 0 and slope of 1. Conditions with fewer than 25 responses are plotted in red. See the online article for the color version of this figure.

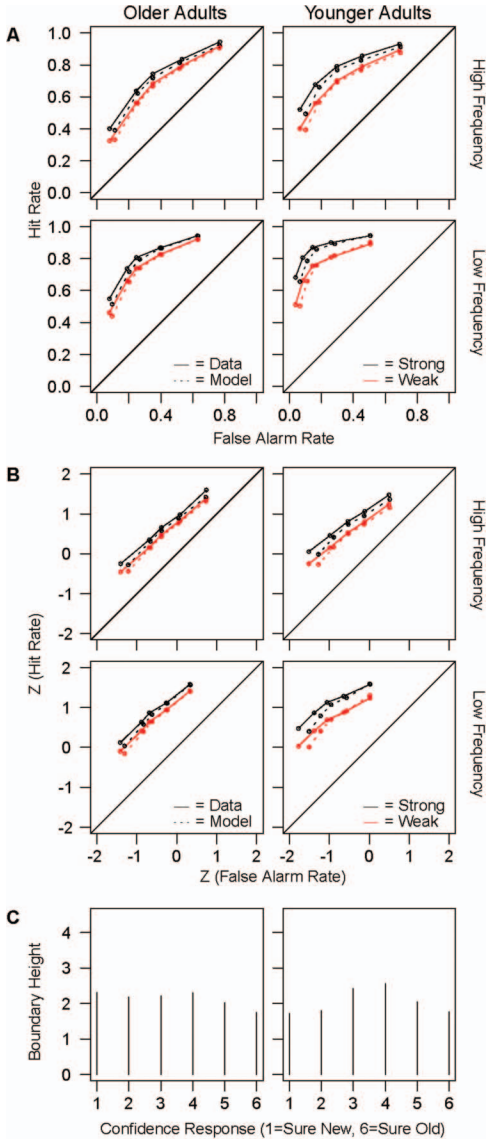

Figure 7. Average receiver operating characteristics (ROCs), z-ROCs, and decision boundaries. (A) Average empirical ROC functions (solid lines) and average predicted ROC functions (dashed lines) for old and young subjects. (B) Average empirical z-ROC functions (solid lines) and average predicted z-ROC functions (dashed lines) for old and young subjects. (C) Average decision boundary height for each confidence response for old and young subjects. See the online article for the color version of this figure.

in this experiment is primarily explained by the difference in nondecision time. The approximate linearity of the points in Figure 8 indicates that, relative to the RT distribution of the younger subjects, the RT distribution of the older subjects is shifted but does not appear to differ in terms of its shape or skew. Changes in nondecision time can explain such a shift in RT distributions whereas changes in mean drift rates or boundary separation will affect the shape of the RT distribution (as demonstrated in Figure 9B).

Previous applications of the diffusion model to two-choice data from older adults have often found that older adults have larger values of boundary separation than younger adults when no special instructions are given or when both groups are instructed to emphasize speed (Ratcliff et al., 2004). However, when both groups are instructed to emphasize accuracy then sometimes the difference in boundary separation is reduced or eliminated (Ratcliff et al., 2004, 2006a). Furthermore, no difference in boundary separation was observed in a response-signal task with older and younger adults (Ratcliff, 2008) and in a brightness discrimination task with older and younger adults (Ratcliff et al., 2003).

Ratcliff (2008) fit data from older and younger adults from both a response-signal and a standard RT version of a numerosity discrimination task. In modeling the response signal task, the only significant difference between the two groups was that the older adults had longer nondecision times. That is, the older adults were able to extract the same quality of information from the stimulus as younger adults and set similar decision thresholds for making their responses, but took longer to extract the relevant information from the stimulus and make a response. The size of the boundary separation in the response signal task was also comparable to the size of the boundary separation in the standard RT task when subjects were instructed to emphasize accuracy. Ratcliff et al. (2003) fit data from older and younger adults from a brightness discrimination task and found only differences in nondecision times across the two age groups. This confidence judgment paradigm is thus not the first to show no difference between old and young adults in terms of decision boundaries, although it is not yet clear why some tasks and response paradigms produce a difference and others do not.

Median RTs in this confidence-judgment task were considerably longer than those observed in two-choice versions of this task. In

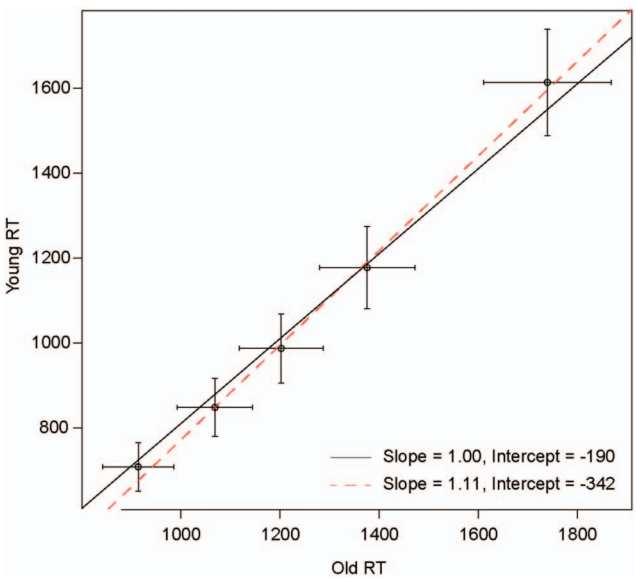

Figure 8. Comparison of older and younger subject response time (RT) quantiles averaged across subjects and conditions. The quantiles increase from the 0.1 quantile in the bottom left to the 0.9 quantile in the top right (in steps of 0.2). *SE* bars for each quantile are shown as well as the best-fitting linear regression line (the red dashed line) and reference line with a slope of 1 and intercept equal to the mean Ter difference between old and young subjects (the black line). See the online article for the color version of this figure.

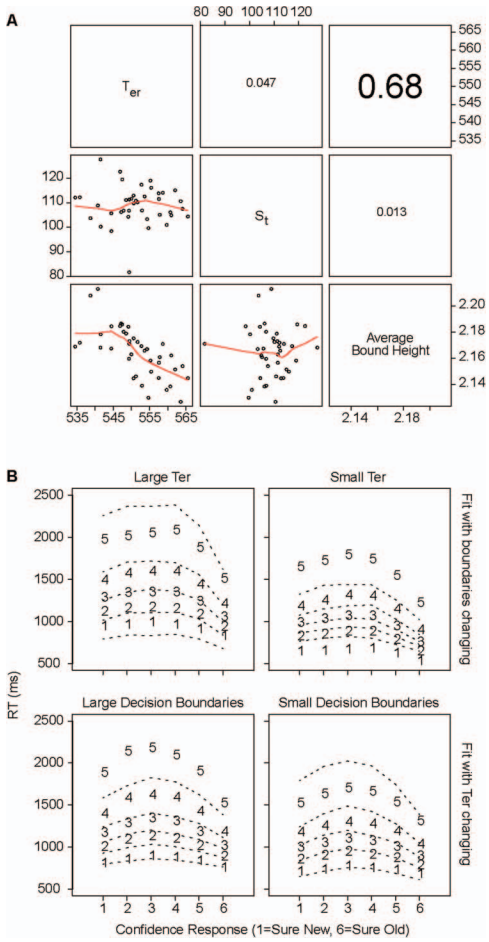

Figure 9. Simulations of tradeoffs between parameters that affect response time (RTs). (A) Parameter estimates for nondecision time (T_{er}), variability in nondecision time (s_t) and the average of the six decision boundaries plotted against each other along with correlations for each pair of parameters. (B) Fits of the model to simulated data sets. The six confidence categories are plotted along the *x*-axis and the five RT quantiles for each confidence category are plotted vertically. The numbers 1–5 represent the RT quantiles from the simulated data and the lines represent the predictions from the model fits. The first row shows the data simulated using either a small or large value of Ter and the second row shows the data simulated using either small or large decision boundaries. See the online article for the color version of this figure.

a two-choice recognition memory experiment in Ratcliff et al. (2011), the average median RTs for college-age adults were around 600 ms across all conditions. In our confidence judgment experiment, the average median RTs for the college-age adults ranged from 800 to 1,200 ms, depending on the confidence level. These RTs are more similar to those observed from college-age adults when they have been instructed to emphasize accuracy over speed (cf. Ratcliff et al., 2006a). It is possible that making confidence judgments implicitly encourages subjects to emphasize accuracy over speed and so induces behavior that is more similar across the two age groups. Alternatively, it is possible that making confidence judgments, with the increased number of response options, seems more difficult than making two-choice judgments so that even young subjects set more conservative boundaries than they would in a two-choice task (cf. Starns & Ratcliff, 2010).

There are numerous methods for assessing subjects' confidence in their responses. Confidence judgments may be collected at the same time as a decision (this is the more traditional approach, especially in memory tasks; Banks, 1970; Egan, 1958; Lockhart & Murdock, 1970; Ratcliff et al., 1992, 1994; Wickelgren & Norman, 1966) or following a two-choice decision (that is used more often in perceptual tasks, Baranski & Petrusic, 1998; Merkle & Van Zandt, 2006; Pleskac & Busemeyer, 2010; Van Zandt, 2000; Van Zandt & Maldonado-Molina, 2004; Vickers, 1979; Vickers & Lee, 1998, 2000). Confidence may be reported using an ordinal scale (as in more of the studies references above) or on a continuous scale (Kvam et al., 2015; Province & Rouder, 2012). Confidence judgments have also been associated with a variety of metamemory concepts such as prospective confidence judgments (i.e., predictions made at the time of learning about future ability to recall an item), judgments of learning, feeling-of-knowing, and other associated concepts.

Our study is novel in examining both RTs and response proportions from older adults in a memory task with confidence judgments. Other studies, however, have examined just response proportions from older adults making confidence judgments and have analyzed these results according to a dual-process description of memory (Healy, Light, & Chung, 2005; Howard, Bessette-Symons, Zhang, & Hoyer, 2006; Toth & Parks, 2006). According to a general dual-process account, recognition memory relies on two components: recollection and familiarity. Familiarity is thought to consist of a general sense of oldness whereas recollection is thought to include qualitative information about the remembered item or its context. This account is often implemented as an equal-variance signal-detection process for

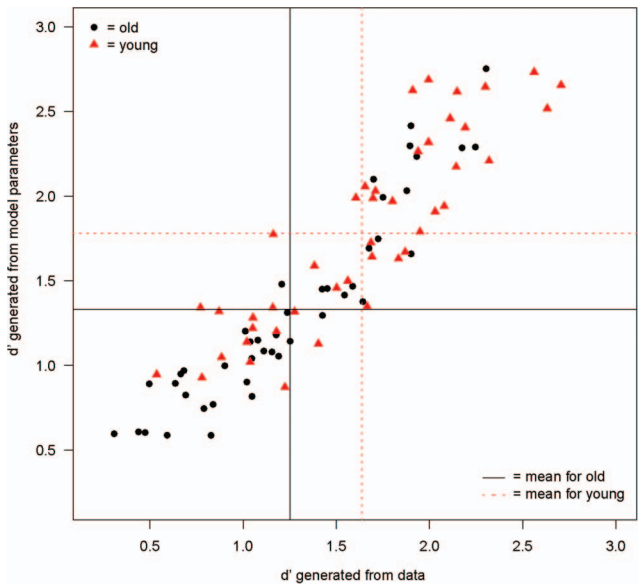

Figure 10. Comparison of performance measures derived from the data and from the model parameters. A measure of d' based on hit and false alarm rates (averaged across confidence levels) is plotted against a model-based measure of d' are based on the means and *SD*s of the drift rate distributions from the model fits. Each point represents the d' value from a single condition for a single subject. The horizontal and vertical lines represent the mean d' values for each age group and each type of d'. See the online article for the color version of this figure.

familiarity plus a discrete threshold process for recollection (Yonelinas, 1994; Yonelinas & Parks, 2007). When responding is based entirely on familiarity, this model predicts asymmetrical curvilinear ROC functions and linear z-ROC functions with a slope equal to one. When responding is based on recollection for some proportion of the word pairs, the model predicts linear ROC functions and slightly nonlinear (i.e., slightly U-shaped) z-ROC functions with slopes less than one.

Researchers invested in a dual-process account of memory have collected data from older adults using a variety of procedures aimed at identifying which components of memory decline with age (e.g., remember or know judgments: Bastin & Van Der Linden, 2003; Toth & Parks, 2006; process-dissociation: Jennings & Jacoby, 1997; structural equation modeling: Quamme, Yonelinas, Widaman, Kroll, & Sauvé, 2004; confidence judgments: Howard et al., 2006; Toth & Parks, 2006; Healy, Light, & Chung, 2005). Most of the studies have claimed that older adults show a decrease in recollection (see Yonelinas, 2002), but results about familiarity have been mixed with some studies finding a decrease in familiarity for older adults (Mark & Rugg, 1998; Schacter, Koutstaal, Johnson, Gross, & Angell, 1997; Toth & Parks, 2006) and other studies finding no difference or a nonsignificant increase in familiarity (Bastin & Van Der Linden, 2003; Howard et al., 2006; Quamme et al., 2004). It has been argued that these differences may be a result of the procedures used to estimate recollection and familiarity components (Light, Prull, LaVoie, & Healy, 2000; Prull, Dawes, Martin, Rosenberg, & Light, 2006) and the overall performance levels (Yonelinas, 2002). It is worth noting that more symmetric ROC functions can also be produced in a standard signal detection model by including criterion variability (Benjamin, Diaz, & Wee, 2009). Older adults could have more variability in criterion placements and this would produce weaker and more symmetric ROC functions. However, none of these analyses consider RTs or any kind of process model for making decisions.

The RTCON2 model distinguishes between the evidence used to make a decision (i.e., the information from memory) and the actual process of making a decision. This makes the model well suited not only for examining differences across age groups, but also for disentangling changes in evidence from memory from changes in response strategies. The ROC functions observed in our experiment are consistent with a dual-process account in that a decrease in recollection would produce ROC functions that are more symmetric. This is what we observed for the older adults. However, we were able to fit these results with a single-process model that was also able to account for the RTs across all of the subjects and confidence levels. In contrast there is no dual process model that is capable of dealing with RTs and so such models are incomplete. The RTCON2 model is able to capture the effects in the experimental data including the behavior of RT distributions with a simple shift in the drift distribution, small changes in decision thresholds, and a large change in nondecision time.

Other experiments have examined aging and confidence judgments in associative recognition and source memory tasks. These paradigms have produced some results that differ from item recognition results. In associative and source-memory studies, older adults have been found to make more high-confidence false alarms and are less sensitive to differences in difficulty than younger adults (Chua, Schacter, & Sperling, 2009; Dodson, Bawa, & Krueger, 2007; Dodson, Bawa, & Slotnick, 2007; Dodson & Krueger, 2006; Fandakova et al., 2013; Kelley & Sahakyan, 2003; Norman & Schacter, 1997; Pacheco et al., 2012; Shing et al., 2009). Dodson et al. (2007) and Pacheco et al. (2012) found that older adults were less accurate than younger adults at assessing the accuracy of source judgments and cued-recall responses (even when matched on performance), but not less accurate when assessing item-recognition responses. Fandakova et al. (2013) and Shing et al. (2009) found that older adults made more high-confidence errors in an associative recognition task. Kelley and Sahakyan (2003) found that older and younger adults were equally able to judge the accuracy of their cued-recall responses in a control condition, but older adults showed a greater decrease in metamemory accuracy than younger adults in a deceptive condition involving associatively related lures.

Consistent with the item recognition findings, older adults in our study did not make more high confidence errors than younger adults overall, although they did make slightly more medium and high confidence errors to new words (i.e., they were slightly more likely than younger subjects to incorrectly claim they remembered studying a new word with medium or high confidence). This can be explained in the model by looking at the pattern of boundary heights for old and young subjects across the old and new response categories. On average, the older adults adopted slightly lower boundaries for the old response categories and slightly higher boundaries for the new response categories (see the bottom row of Figure 6). This is equivalent to a response bias in a two-choice task. As mentioned previously, studies of two-choice decision-making have typically found that older adults adopt more conservative decision-boundaries than younger adults (Ratcliff, Thapar, & McKoon, 2004). In other words, older adults tend to use a decision-making approach that reduces the number of incorrect responses. The asymmetric decision boundaries observed in this experiment may illustrate a similar strategy, albeit one that reduces one type of incorrect response (misses) while increasing another (false alarms). It would be interesting to see in future work if a similar bias could account for some of the high-confidence errors observed in associative and source memory tasks for older adults.

Fitting RT distributions is what enables RTCON2 to distinguish between changes in response proportions that are the result of changes in evidence and changes in response proportions that are the result of changes in decision settings. Changes in the strength of the evidence being used to make the decision (i.e., changes in drift rates) can change the proportion of responses made at each level of the confidence scale and this would correspond to an overall change in performance. However, changes in response proportions can also occur because of differences in how subjects set decision boundaries across the confidence scale. This type of change can reflect individual preferences (e.g., being more or less cautious about making errors in general, being more or less willing to make high confidence responses, and so on) or it can be the result of instructions (e.g., to respond more quickly or more accurately). These changes in response patterns are distinguishable because they have different effects on the RT distributions. Shifts in the evidence distribution will change the amount of evidence within each response region. These changes in evidence primarily affect the tail of the RT distribution and only minimally affect the leading edge of the RT distribution. In contrast, changing the height of the decision boundaries affects both the leading edge and the tail of the RT distribution. Fitting RT distributions thus enables us to distinguish between changes in evidence and changes in the decision-making process.

Conclusions

We modeled both response proportions and RT quantiles from older and younger adults in an item-recognition task with confidence judgments using RTCON2. This modeling approach provides a more complete account of the experimental data by accounting for both choice proportions and RT distributions for each of the choice categories. This allows us to distinguish between changes in memory performance and differences in decision-making preferences. As in previous modeling work with two-choice tasks, the longer RTs of older adults were well explained by changes in nondecision processing and there were only small changes in drift rate with age (and these changes were consistent with the observed changes in accuracy). Older adults did not use more conservative decision thresholds than younger adults in this task as in some earlier studies (Ratcliff, 2008; Ratcliff et al., 2003, 2004, 2006a), but did demonstrate a slight bias toward calling items old. The RTCON2 model was able to fit all of the behavioral results with a single memory process and a different pattern of decision boundaries for older adults than younger adults.

References

Balota, D. A., Dolan, P. O., & Duchek, J. M. (2000). Memory changes in healthy young and older adults. In E. Tulving & F. I. M. Craik (Eds.), *Handbook of memory* (pp. 395–410). New York, NY: Oxford University Press.

Banks, W. P. (1970). Signal detection theory and human memory. *Psychological Bulletin, 74,* 81–99. http://dx.doi.org/10.1037/h0029531

Baranski, J. V., & Petrusic, W. M. (1998). Probing the locus of confidence judgments: Experiments on the time to determine confidence. *Journal of Experimental Psychology: Human Perception and Performance, 24,* 929–945. http://dx.doi.org/10.1037/0096-1523.24.3.929

Basowitz, H., & Korchin, S. J. (1957). Age differences in the perception of closure. *The Journal of Abnormal and Social Psychology, 54,* 93–97. http://dx.doi.org/10.1037/h0040733

Bastin, C., & Van der Linden, M. (2003). The contribution of recollection and familiarity to recognition memory: A study of the effects of test format and aging. *Neuropsychology, 17,* 14–24. http://dx.doi.org/10.1037/0894-4105.17.1.14

Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J., . . . Pouget, A. (2008). Probabilistic population codes for Bayesian decision making. *Neuron, 60,* 1142–1152. http://dx.doi.org/10.1016/j.neuron.2008.09.021

Benjamin, A. S., Diaz, M., & Wee, S. (2009). Signal detection with criterion noise: Applications to recognition memory. *Psychological Review, 116,* 84–115. http://dx.doi.org/10.1037/a0014351

Birren, J. E. (1965). Age-changes in speed of behavior: Its central nature and physiological correlates. In A. T. Welford & J. E. Birren (Eds.), *Behavior, aging, and the nervous system* (pp. 191–216). Springfield, IL: Charles C Thomas.

Bowles, N. L., & Poon, L. W. (1982). An analysis of the effect of aging on recognition memory. *Journal of Gerontology, 37,* 212–219. http://dx.doi.org/10.1093/geronj/37.2.212

Brinley, J. F. (1965). Cognitive sets, speed and accuracy of performance in the elderly. In A. T. Welford & J. E. Birren (Eds.), *Behavior, aging and the nervous system* (pp. 114–149). Springfield, IL: Thomas.

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

Buchler, N. E. G., & Reder, L. M. (2007). Modeling age-related memory deficits: A two-parameter solution. *Psychology and Aging, 22,* 104–121. http://dx.doi.org/10.1037/0882-7974.22.1.104

Cerella, J. (1985). Information processing rates in the elderly. *Psychological Bulletin, 98,* 67–83. http://dx.doi.org/10.1037/0033-2909.98.1.67

Cerella, J. (1990). Aging and information-processing rate. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (3rd ed., pp. 201–221). San Diego, CA: Academic Press. http://dx.doi.org/10.1016/B978-0-12-101280-9.50018-8

Cerella, J. (1991). Age effects may be global, not local: Comment on Fisk and Rogers (1991). *Journal of Experimental Psychology: General, 120,* 215–223. http://dx.doi.org/10.1037/0096-3445.120.2.215

Cerella, J. (1994). Generalized slowing in Brinley plots. *The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences, 49,* P65–P71. http://dx.doi.org/10.1093/geronj/49.2.P65

Chua, E. F., Schacter, D. L., & Sperling, R. A. (2009). Neural basis for recognition confidence in younger and older adults. *Psychology and Aging, 24,* 139–153. http://dx.doi.org/10.1037/a0014029

Craik, F. I. M. (1983). On the transfer of information from temporary to permanent memory. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences, 302,* 341–359. http://dx.doi.org/10.1098/rstb.1983.0059

Craik, F. I. M. (1986). A functional account of age differences in memory. In F. Klix & H. Hagendorf (Eds.), *Human memory and cognitive capabilities: Mechanisms and performances* (pp. 409–422). Amsterdam, the Netherlands: Elsevier.

Craik, F. I. M. (1994). Memory changes in normal aging. *Current Directions in Psychological Science, 3,* 155–158. http://dx.doi.org/10.1111/1467-8721.ep10770653

Craik, F. I. M., & Jennings, J. (1992). Human memory. In F. I. M. Craik & T. A. Salthouse (Eds.), *The handbook of aging and cognition* (pp. 51–110). Hillsdale, NJ: Erlbaum.

Craik, F. I. M., & McDowd, J. M. (1987). Age differences in recall and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 474–479. http://dx.doi.org/10.1037/0278-7393.13.3.474

Deary, I. J. (2000). *Looking down on human intelligence: From psychometrics to the brain*. New York, NY: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780198524175.001.0001

DeCarlo, L. T. (2002). Signal detection theory with finite mixture distributions: Theoretical developments with applications to recognition memory. *Psychological Review, 109,* 710–721. http://dx.doi.org/10.1037/0033-295X.109.4.710

Dodson, C. S., Bawa, S., & Krueger, L. E. (2007). Aging, metamemory, and high-confidence errors: A misrecollection account. *Psychology and Aging, 22,* 122–133. http://dx.doi.org/10.1037/0882-7974.22.1.122

Dodson, C. S., Bawa, S., & Slotnick, S. D. (2007). Aging, source memory, and misrecollections. *Journal Of Experimental Psychology: Learning, Memory, And Cognition, 33,* 169–181.

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

Egan, J. P. (1958). Recognition memory and the operating characteristic. *USAF Operational Applications Laboratory Technical Note, 58-, 51,* 32.

Erber, J. T. (1974). Age differences in recognition memory. *Journal of Gerontology, 29,* 177–181. http://dx.doi.org/10.1093/geronj/29.2.177

Fandakova, Y., Shing, Y. L., & Lindenberger, U. (2013). High-confidence memory errors in old age: The roles of monitoring and binding processes. *Memory, 21,* 732–750. http://dx.doi.org/10.1080/09658211.2012.756038

Fisk, J. E., & Warr, P. (1996). Age and working memory: The role of perceptual speed, the central executive, and the phonological loop. *Psychology and Aging, 11,* 316–323. http://dx.doi.org/10.1037/0882-7974.11.2.316

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state." A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 12,* 189–198. http://dx.doi.org/10.1016/0022-3956(75)90026-6

Gomez, P., Ratcliff, R., & Perea, M. (2008). The overlap model: A model of letter position coding. *Psychological Review, 115,* 577–601. http://dx.doi.org/10.1037/a0012667

Gordon, S. K., & Clark, W. C. (1974). Adult age differences in word and nonsense syllable recognition memory and response criterion. *Journal of Gerontology, 29,* 659–665. http://dx.doi.org/10.1093/geronj/29.6.659

Healy, M. R., Light, L. L., & Chung, C. (2005). Dual-process models of associative recognition in young and older adults: Evidence from receiver operating characteristics. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 768–788. http://dx.doi.org/10.1037/0278-7393.31.4.768

Howard, M. W., Bessette-Symons, B., Zhang, Y., & Hoyer, W. J. (2006). Aging selectively impairs recollection in recognition memory for pictures: Evidence from modeling and receiver operating characteristic curves. *Psychology and Aging, 21,* 96–106. http://dx.doi.org/10.1037/0882-7974.21.1.96

Jazayeri, M., & Movshon, J. A. (2006). Optimal representation of sensory information by neural populations. *Nature Neuroscience, 9,* 690–696. http://dx.doi.org/10.1038/nn1691

Jennings, J. M., & Jacoby, L. L. (1997). An opposition procedure for detecting age-related deficits in recollection: Telling effects of repetition. *Psychology and Aging, 12,* 352–361. http://dx.doi.org/10.1037/0882-7974.12.2.352

Kausler, D. H. (1994). *Learning and memory in normal aging*. San Diego, CA: Academic Press.

Kelley, C. M., & Sahakyan, L. (2003). Memory, monitoring and control in attainment of memory accuracy. *Journal of Memory and Language, 48,* 704–721.

Kučera, H., & Francis, W. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Kvam, P. D., Pleskac, T. J., Yu, S., & Busemeyer, J. R. (2015). Interference effects of choice on confidence: Quantum characteristics of evidence accumulation. *Proceedings of the National Academy of Sciences of the United States of America, 112,* 10645–10650.

Light, L. L., Prull, M. W., LaVoie, D. J., & Healy, M. R. (2000). Dual process theories of memory in older age. In T. J. Perfect & E. A. Maylor (Eds.), *Models of cognitive aging* (pp. 238–300). Oxford, United Kingdom: Oxford University Press. 2001–00072-009.

Lockhart, R. S., & Murdock, B. B. (1970). Memory and the theory of signal detection. *Psychological Bulletin, 74,* 100–109. http://dx.doi.org/10.1037/h0029536

Malmberg, K. J., & Xu, J. (2007). On the flexibility and the fallibility of associative memory. *Memory & Cognition, 35,* 545–556. http://dx.doi.org/10.3758/BF03193293

Mark, R. E., & Rugg, M. D. (1998). Age effects on brain activity associated with episodic memory retrieval. An electrophysiological study. *Brain: A Journal of Neurology, 121,* 861–873. http://dx.doi.org/10.1093/brain/121.5.861

McKoon, G., & Ratcliff, R. (2012). Aging and IQ effects on associative recognition and priming in item recognition. *Journal of Memory and Language, 66,* 416–437. http://dx.doi.org/10.1016/j.jml.2011.12.001

Merkle, E. C., & Van Zandt, T. (2006). An application of the poisson race model to confidence calibration. *Journal of Experimental Psychology: General, 135,* 391–408. http://dx.doi.org/10.1037/0096-3445.135.3.391

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,* 1170–1187. http://dx.doi.org/10.1037/0278-7393.26.5.1170

Neath, I. (1998). *Human memory: An introduction to research, data, and theory*. Pacific Grove, CA: Brooks/Cole.

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal, 7,* 308–313. http://dx.doi.org/10.1093/comjnl/7.4.308

Norman, K. A., & Schacter, D. L. (1997). False recognition in younger and older adults: Exploring the characteristics of illusory memories. *Memory & Cognition, 25,* 838–848.

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,* 104–118. http://dx.doi.org/10.1037/0882-7974.23.1.104

Pacheco, J., Beevers, C. G., McGeary, J. E., & Schnyer, D. M. (2012). Memory monitoring performance and PFC activity are associated with 5-HTTLPR genotype in older adults. *Neuropsychologia, 50,* 2257–2270. http://dx.doi.org/10.1016/j.neuropsychologia.2012.05.030

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review, 117,* 864–901. http://dx.doi.org/10.1037/a0019737

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

Prull, M. W., Dawes, L. L. C., Martin, A. M., III, Rosenberg, H. F., & Light, L. L. (2006). Recollection and familiarity in recognition memory: Adult age differences and neuropsychological test correlates. *Psychology and Aging, 21,* 107–118. http://dx.doi.org/10.1037/0882-7974.21.1.107

Quamme, J. R., Yonelinas, A. P., Widaman, K. F., Kroll, N. E., & Sauvé, M. J. (2004). Recall and recognition in mild hypoxia: Using covariance structural modeling to test competing theories of explicit memory. *Neuropsychologia, 42,* 672–691. http://dx.doi.org/10.1016/j.neuropsychologia.2003.09.008

Rabinowitz, J. C. (1984). Aging and recognition failure. *Journal of Gerontology, 39,* 65–71. http://dx.doi.org/10.1093/geronj/39.1.65

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

Ratcliff, R. (1981). A theory of order relations in perceptual matching. *Psychological Review, 88,* 552–572. http://dx.doi.org/10.1037/0033-295X.88.6.552

Ratcliff, R. (2008). Modeling aging effects on two-choice tasks: Response signal and response time data. *Psychology and Aging, 23,* 900–916. http://dx.doi.org/10.1037/a0013930

Ratcliff, R. (2013). Parameter variability and distributional assumptions in the diffusion model. *Psychological Review, 120,* 281–292.

Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). List-strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 163–178. http://dx.doi.org/10.1037/0278-7393.16.2.163

Ratcliff, R., & McKoon, G. (1991). Using ROC data and priming results to test global memory models. In S. Lewandowsky & W. E. Hockley (Eds.), *Relating Theory and Data: Essays on Human Memory in Honor of Bennet B. Murdock, Jr* (pp. 279–296). Hillsdale, NJ: Erlbaum.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20,* 873–922. http://dx.doi.org/10.1162/neco.2008.12-06-420

Ratcliff, R., McKoon, G., & Tindall, M. (1994). Empirical generality of data from recognition memory receiver-operating characteristic functions and implications for the global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 763–785. http://dx.doi.org/10.1037/0278-7393.20.4.763

Ratcliff, R., Sheu, C. F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review, 99,* 518–535. http://dx.doi.org/10.1037/0033-295X.99.3.518

Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological Review, 116,* 59–83. http://dx.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,* 697–719. http://dx.doi.org/10.1037/a0033152

Ratcliff, R., Thapar, A., Gomez, P., & McKoon, G. (2004). A diffusion model analysis of the effects of aging in the lexical-decision task. *Psychology and Aging, 19,* 278–289. http://dx.doi.org/10.1037/0882-7974.19.2.278

Ratcliff, R., Thapar, A., & McKoon, G. (2001). The effects of aging on reaction time in a signal detection task. *Psychology and Aging, 16,* 323–341. http://dx.doi.org/10.1037/0882-7974.16.2.323

Ratcliff, R., Thapar, A., & McKoon, G. (2003). A diffusion model analysis of the effects of aging on brightness discrimination. *Perception & Psychophysics, 65,* 523–535. http://dx.doi.org/10.3758/BF03194580

Ratcliff, R., Thapar, A., & McKoon, G. (2004). A diffusion model analysis of the effects of aging on recognition memory. *Journal of Memory and Language, 50,* 408–424. http://dx.doi.org/10.1016/j.jml.2003.11.002

Ratcliff, R., Thapar, A., & McKoon, G. (2006a). Aging and individual differences in rapid two-choice decisions. *Psychonomic Bulletin & Review, 13,* 626–635. http://dx.doi.org/10.3758/BF03193973

Ratcliff, R., Thapar, A., & McKoon, G. (2006b). Aging, practice, and perceptual tasks: A diffusion model analysis. *Psychology and Aging, 21,* 353–371. http://dx.doi.org/10.1037/0882-7974.21.2.353

Ratcliff, R., Thapar, A., & McKoon, G. (2007). Application of the diffusion model to two-choice tasks for adults 75–90 years old. *Psychology and Aging, 22,* 56–66. http://dx.doi.org/10.1037/0882-7974.22.1.56

Ratcliff, R., Thapar, A., & McKoon, G. (2010). Individual differences, aging, and IQ in two-choice tasks. *Cognitive Psychology, 60,* 127–157. http://dx.doi.org/10.1016/j.cogpsych.2009.09.001

Ratcliff, R., Thapar, A., & McKoon, G. (2011). Effects of aging and IQ on item and associative memory. *Journal of Experimental Psychology: General, 140,* 464–487. http://dx.doi.org/10.1037/a0023810

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review, 9,* 438–481. http://dx.doi.org/10.3758/BF03196302

Salthouse, T. A. (1979). Adult age and the speed-accuracy trade-off. *Ergonomics, 22,* 811–821. http://dx.doi.org/10.1080/00140137908924659

Salthouse, T. A. (1985). Speed of behavior and its implications for cognition. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (2nd ed., pp. 400–426). New York, NY: Van Nostrand Reinhold Co.

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

Salthouse, T. A., Kausler, D., & Saults, J. S. (1988). Investigation of student status, background variables, and feasibility of standard tasks in cognitive aging research. *Psychology and Aging, 3,* 29–37. http://dx.doi.org/10.1037/0882-7974.3.1.29

Schacter, D. L., Koutstaal, W., Johnson, M. K., Gross, M. S., & Angell, K. E. (1997). False recollection induced by photographs: A comparison of older and younger adults. *Psychology and Aging, 12,* 203–215. http://dx.doi.org/10.1037/0882-7974.12.2.203

Schonfield, D., & Robertson, B. A. (1966). Memory storage and aging. *Canadian Journal of Psychology, 20,* 228–236. http://dx.doi.org/10.1037/h0082941

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,* 169–179. http://dx.doi.org/10.1080/09658210802190596

Silverman, I. (1963). Age and the tendency to withhold responses. *Journal of Gerontology, 18,* 372–375. http://dx.doi.org/10.1093/geronj/18.4.372

Spieler, D. H., Balota, D. A., & Faust, M. E. (1996). Stroop performance in healthy younger and older adults and in individuals with dementia of the Alzheimer's type. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 461–479. http://dx.doi.org/10.1037/0096-1523.22.2.461

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,* 377–390. http://dx.doi.org/10.1037/a0018022

Strayer, D. L., & Kramer, A. F. (1994). Aging and skill acquisition: Learning-performance distinctions. *Psychology and Aging, 9,* 589–605. http://dx.doi.org/10.1037/0882-7974.9.4.589

Thapar, A., Ratcliff, R., & McKoon, G. (2003). A diffusion model analysis of the effects of aging on letter discrimination. *Psychology and Aging, 18,* 415–429. http://dx.doi.org/10.1037/0882-7974.18.3.415

Thorndike, E. L., Bregman, E. O., Tilton, J. W., & Woodyard, E. (1928). *Adult learning*. New York, NY: Macmillan.

Toth, J. P., & Parks, C. M. (2006). Effects of age on estimated familiarity in the process dissociation procedure: The role of noncriterial recollection. *Memory & Cognition, 34,* 527–537. http://dx.doi.org/10.3758/BF03193576

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

Van Zandt, T. (2000). ROC curves and confidence judgements in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 582–600. http://dx.doi.org/10.1037/0278-7393.26.3.582

Van Zandt, T., & Maldonado-Molina, M. M. (2004). Response reversals in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 1147–1166. http://dx.doi.org/10.1037/0278-7393.30.6.1147

Vickers, D. (1979). *Decision processes in visual perception*. New York, NY: Academic Press.

Vickers, D., & Lee, M. D. (1998). Dynamic models of simple judgments: I. properties of a self-regulating accumulator module. *Nonlinear Dynamics, Psychology, and Life Sciences, 2,* 169–194. http://dx.doi.org/10.1023/A:1022371901259

Vickers, D., & Lee, M. D. (2000). Dynamic models of simple judgments: II. properties of a self-organizing PAGAN (parallel, adaptive, generalized accumulator network) model for multi-choice tasks. *Nonlinear Dynamics, Psychology, and Life Sciences, 4,* 1–31. http://dx.doi.org/10.1023/A:1009571011764

Voskuilen, C., & Ratcliff, R. (2016). Modeling confidence and response time in associative recognition. *Journal of Memory and Language, 86,* 60–96. http://dx.doi.org/10.1016/j.jml.2015.09.006

Wahlin, A., Bäckman, L., & Winblad, B. (1995). Free recall and recognition of slowly and rapidly presented words in very old age: A community-based study. *Experimental Aging Research, 21,* 251–271. http://dx.doi.org/10.1080/03610739508253984

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica, 41,* 67–85. http://dx.doi.org/10.1016/0001-6918(77)90012-9

Wickelgren, W. A., & Norman, D. A. (1966). Strength models and serial position in short-term recognition memory. *Journal of Mathematical Psychology, 3,* 316–347. http://dx.doi.org/10.1016/0022-2496(66)90018-6

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1341–1354. http://dx.doi.org/10.1037/0278-7393.20.6.1341

Yonelinas, A. P. (1997). Recognition memory ROCs for item and associative information: The contribution of recollection and familiarity. *Memory & Cognition, 25,* 747–763. http://dx.doi.org/10.3758/BF03211318

Yonelinas, A. P. (1999). The contribution of recollection and familiarity to recognition and source-memory judgments: A formal dual-process model and an analysis of receiver operating characteristics. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 1415–1434. http://dx.doi.org/10.1037/0278-7393.25.6.1415

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46,* 441–517. http://dx.doi.org/10.1006/jmla.2002.2864

Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. *Psychological Bulletin, 133,* 800–832. http://dx.doi.org/10.1037/0033-2909.133.5.800

Appendix

Individual Fits and Additional Analyses

See Figures A1–A4 for individual data and model fits for older and younger adults. RT quantiles are only included for high-frequency (HF) conditions but the patterns were consistent across LF conditions.

Table A1 contains mean parameter values from the model fits when the scale on drift, within-trial noise, and the variability in the decision boundaries are allowed to vary across age groups. Note that the best-fitting values for these parameters are very similar across the two age groups. Table A2 contains mean parameter values from when the model was fit to data from only seven sessions per subject. To account for practice effects, we fit data from Sessions 2–8 from the older subjects and Sessions 1–7 from the younger subjects (because the younger subjects initially completed one practice session). For both of these fits, all of the same patterns of results were observed.

(*Appendix continues*)

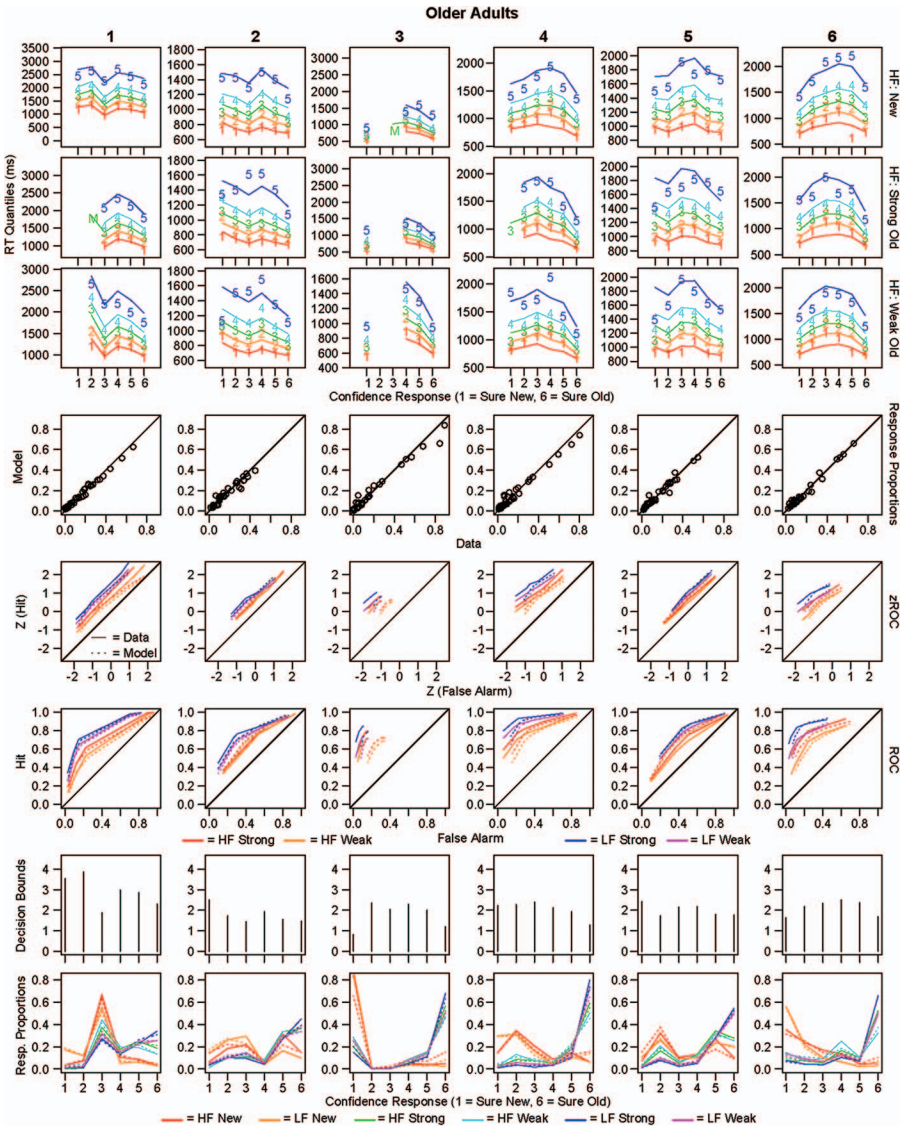

Figure A1. Older subjects' data and model fits. Each column contains data and model predictions for one subject. The first three rows plot the response time (RT) quantiles for each confidence response with the six response keys plotted on the *x*-axis (the "sure new" category is labeled 1 and the "sure old" category is labeled 6) and the RT quantiles plotted vertically with each line representing a RT quantile. Only the high-frequency conditions are shown to save space. The numbers plotted represent the empirical data and the lines represent predicted data from the model. In conditions where subjects made between 5 and 10 responses the median RT is plotted as an 'M' and the other quantiles are not included. Conditions where subjects made fewer than five responses are omitted from the figure. In conditions where the model predicted between 5 and 10 responses only the median RT is plotted and the other quantiles are not included. Conditions where the model predicted fewer than five responses are omitted from the figure. The fourth row plots the response proportions from the data against the model predictions for all conditions along with a reference line (with intercept of 0 and slope of 1). The fifth and sixth row in each column plot the empirical and predicted z-receiver operating characteristic (ROC) and ROC curves for each subject. The solid lines depict the empirical data and the dashed lines depict the model predictions. The blue, magenta, red, and orange lines depict the low-frequency (LF): Strong, LF: Weak, high-frequency (HF): Strong, and HF: Weak conditions, respectively. The seventh row plots the decision boundaries for each confidence response and the eighth row plots the response proportions (both empirical data and model predictions) for each confidence response and condition. The solid lines depict the empirical data and the dashed lines depict the model predictions. The red, orange, green, cyan, blue, and magenta lines depict the HF: New, LF: New, HF: Strong Old, HF: Weak Old, LF: Strong Old, and LF: Weak Old conditions, respectively. See the online article for the color version of this figure.

(*Appendix continues*)

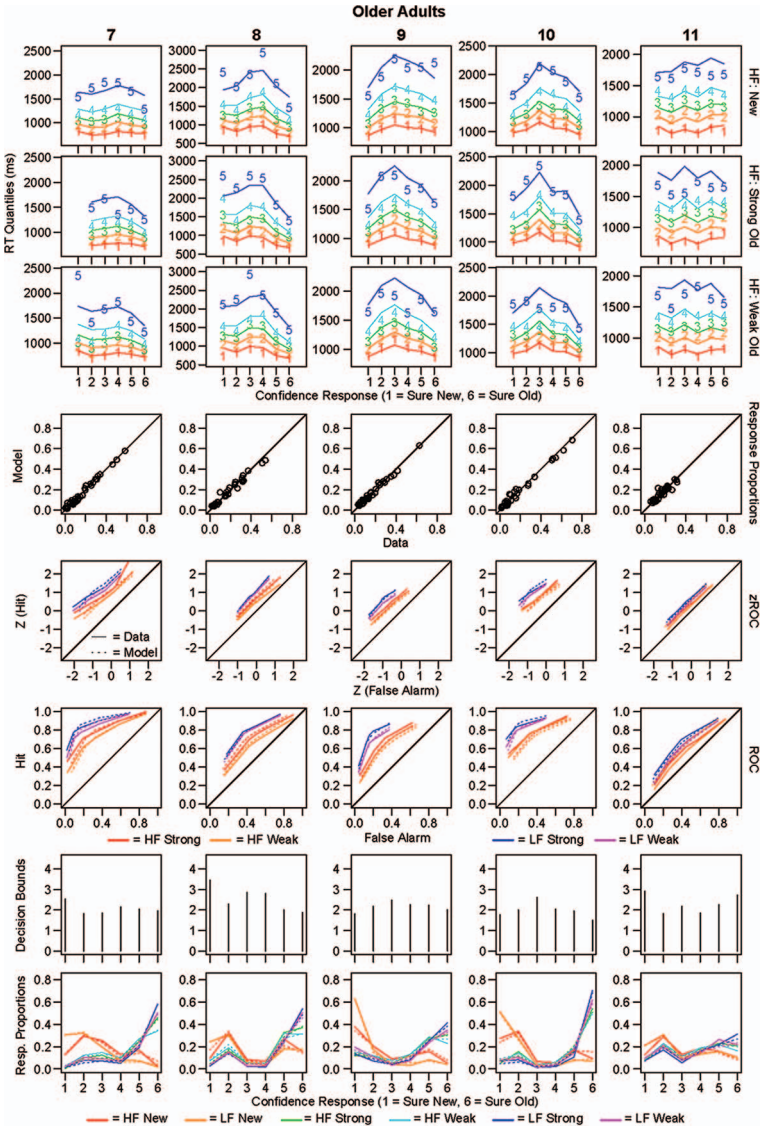

Figure A2. Older subjects' data and model fits. Each column contains data and model predictions for one subject. The first three rows plot the response time (RT) quantiles for each confidence response with the six response keys plotted on the *x*-axis (the "sure new" category is labeled 1 and the "sure old" category is labeled 6) and the RT quantiles plotted vertically with each line representing a RT quantile. Only the high-frequency conditions are shown to save space. The numbers plotted represent the empirical data and the lines represent predicted data from the model. In conditions where subjects made between 5 and 10 responses the median RT is plotted as an 'M' and the other quantiles are not included. Conditions where subjects made fewer than five responses are omitted from the figure. In conditions where the model predicted between 5 and 10 responses only the median RT is plotted and the other quantiles are not included. Conditions where the model predicted fewer than five responses are omitted from the figure. The fourth row plots the response proportions from the data against the model predictions for all conditions along with a reference line (with intercept of 0 and slope of 1). The fifth and sixth row in each column plot the empirical and predicted z-receiver operating characteristic (ROC) and ROC curves for each subject. The solid lines depict the empirical data and the dashed lines depict the model predictions. The blue, magenta, red, and orange lines depict the low-frequency (LF): Strong, LF: Weak, high-frequency (HF): Strong, and HF: Weak conditions, respectively. The seventh row plots the decision boundaries for each confidence response and the eighth row plots the response proportions (both empirical data and model predictions) for each confidence response and condition. The solid lines depict the empirical data and the dashed lines depict the model predictions. The red, orange, green, cyan, blue, and magenta lines depict the HF: New, LF: New, HF: Strong Old, HF: Weak Old, LF: Strong Old, and LF: Weak Old conditions, respectively. See the online article for the color version of this figure.

(*Appendix continues*)

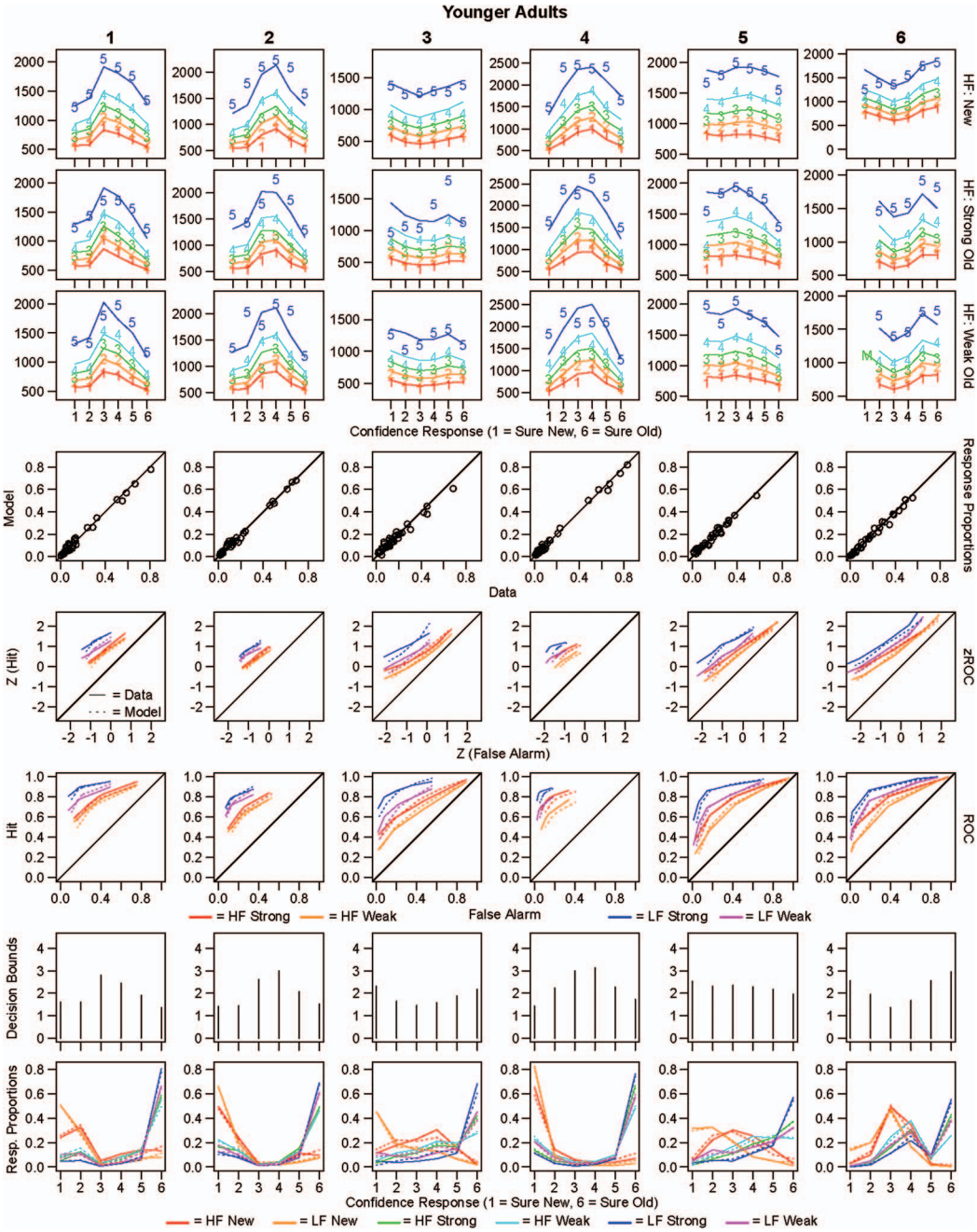

Figure A3. Younger subjects' data and model fits. Same plotting conventions as A1–A2. See the online article for the color version of this figure.

(*Appendix continues*)

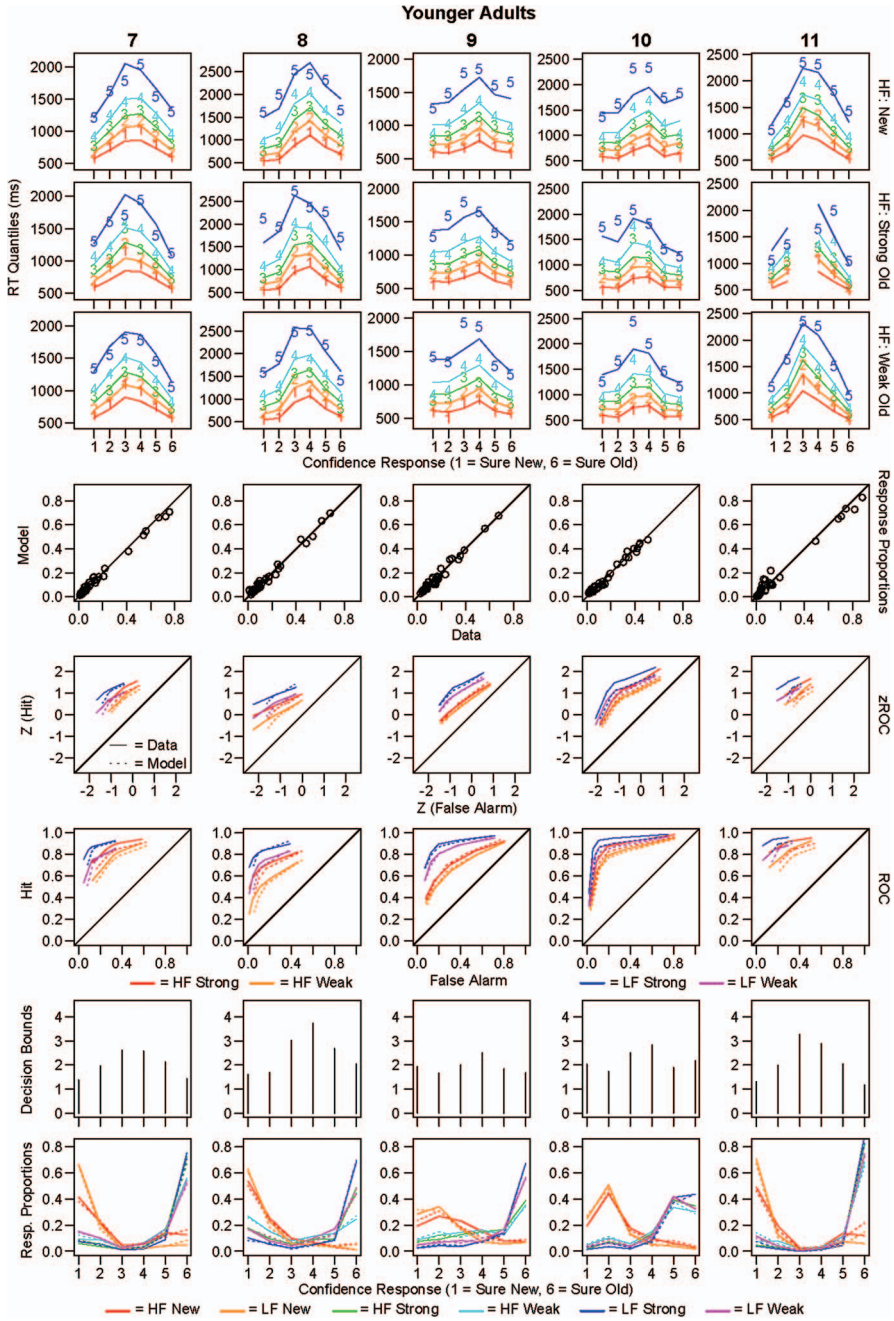

Figure A4. Younger subjects' data and model fits. Same plotting conventions as A1–A2. See the online article for the color version of this figure.

(*Appendix continues*)

Table A1
Mean Parameter Values and SDs Across Subject Groups With All Parameters Allowed to Vary Across Age Groups

Age group	T_{er}	s_t	a	σ	s_b	χ^2
Old	555 (108)	129.0 (28.1)	0.031 (.004)	0.093 (.006)	0.404 (.049)	682 (247)
Young	368 (51.2)	76.9 (36.1)	0.032 (.004)	0.090 (.005)	0.430 (.031)	726 (237)
	b_1	b_2	b_3	b_4	b_5	b_6
Old	2.31 (.81)	2.10 (.63)	2.15 (.42)	2.28 (.37)	2.06 (.33)	1.75 (.44)
Young	1.80 (.46)	1.75 (.28)	2.32 (.58)	2.48 (.58)	2.04 (.29)	1.79 (.47)
	c_1	c_2	c_3	c_4	c_5	
Old	−1.23 (.46)	−0.08 (.22)	0.70 (.15)	1.53 (.38)	2.93 (.43)	
Young	−1.09 (.59)	−0.09 (.40)	0.75 (.25)	1.73 (.35)	2.79 (.47)	
	$v_{N\text{-}HF}$	$v_{OS\text{-}HF}$	$v_{OW\text{-}HF}$	$v_{N\text{-}LF}$	$v_{OS\text{-}LF}$	$v_{OW\text{-}LF}$
Old	0.00	2.04 (.62)	1.65 (.55)	−0.57 (.66)	2.76 (.83)	2.25 (.69)
Young	0.00	2.47 (.51)	1.95 (.49)	−0.97 (.37)	3.45 (.31)	2.52 (.48)
	$s_{N\text{-}HF}$	$s_{OS\text{-}HF}$	$s_{OW\text{-}HF}$	$s_{N\text{-}LF}$	$s_{OS\text{-}LF}$	$s_{OW\text{-}LF}$
Old	1.39 (.18)	1.73 (.12)	1.69 (.13)	1.59 (.20)	1.86 (.17)	1.94 (.20)
Young	1.17 (.25)	1.73 (.11)	1.75 (.13)	1.43 (.25)	1.91 (.24)	1.98 (.22)

Note. T_{er} is the mean nondecision time, s_t is the range in nondecision time, σ is the *SD* in within trial variability, a is the scaling factor that multiplies drift rate, s_b is the range in variability in the decision boundaries, b_1–b_6 are the decision boundaries, c_1–c_5 are the confidence criteria, the *v* values are the mean values of the drift rate distributions for each experimental condition, and the *s* values are the between-trial variability values for each experimental condition. χ^2 is the goodness-of-fit value for the model fits. N-HF = New, high-frequency; OS-HF = Old strong, high-frequency; OW-HF = Old weak, high-frequency; N-LF = New, low-frequency; OS-LF = Old strong, low-frequency; OW-LF = Old weak, low-frequency.

(*Appendix continues*)

Table A2
Mean Parameter Values and SDs Across Subject Groups From Fits to Six Sessions Per Subject

Age group	T_{er}	s_t	a	σ	s_b	χ^2
Old	561 (108)	131.0 (36.7)	0.028	0.100	0.400	814 (344)
Young	375 (53.6)	87.0 (37.6)	0.028	0.100	0.400	722 (213)
	b_1	b_2	b_3	b_4	b_5	b_6
Old	2.27 (.82)	2.15 (.57)	2.17 (.38)	2.27 (.37)	2.07 (.34)	1.75 (.44)
Young	1.86 (.48)	1.90 (.32)	2.54 (.64)	2.65 (.64)	2.21 (.28)	1.83 (.46)
	c_1	c_2	c_3	c_4	c_5	
Old	−1.29 (.49)	−0.11 (.26)	0.69 (.16)	1.54 (.40)	2.98 (.44)	
Young	−1.17 (.60)	−0.07 (.38)	0.79 (.27)	1.84 (.43)	2.95 (.54)	
	$v_{N\text{-}HF}$	$v_{OS\text{-}HF}$	$v_{OW\text{-}HF}$	$v_{N\text{-}LF}$	$v_{OS\text{-}LF}$	$v_{OW\text{-}LF}$
Old	0.00	2.08 (.68)	1.70 (.57)	−0.68 (.72)	2.84 (.88)	2.30 (.72)
Young	0.00	2.69 (.60)	2.05 (.57)	−1.11 (.40)	3.73 (.41)	2.70 (.61)
	$s_{N\text{-}HF}$	$s_{OS\text{-}HF}$	$s_{OW\text{-}HF}$	$s_{N\text{-}LF}$	$s_{OS\text{-}LF}$	$s_{OW\text{-}LF}$
Old	1.39 (.20)	1.66 (.16)	1.64 (.11)	1.61 (.19)	1.80 (.16)	1.85 (.13)
Young	1.19 (.21)	1.71 (.08)	1.74 (.09)	1.49 (.18)	1.78 (.15)	1.91 (.18)

Note. T_{er} is the mean nondecision time, s_t is the range in nondecision time, σ is the *SD* in within trial variability, a is the scaling factor that multiplies drift rate, s_b is the range in variability in the decision boundaries, b_1–b_6 are the decision boundaries, c_1–c_5 are the confidence criteria, the v values are the mean values of the drift rate distributions for each experimental condition, and the s values are the between-trial variability values for each experimental condition. χ^2 is the goodness-of-fit value for the model fits. N-HF = New, high-frequency; OS-HF = Old strong, high-frequency; OW-HF = Old weak, high-frequency; N-LF = New, low-frequency; OS-LF = Old strong, low-frequency; OW-LF = Old weak, low-frequency.

Received January 12, 2017
Revision received March 27, 2017
Accepted April 6, 2017 ■