Age-related Declines in the Fidelity of

Newly Acquired Category Representations

Supplementary Materials

Methods

Subjects

Thirty-Seven older adults (Average age = 69.65; Range = 59 - 82) and 32 younger adults (Average age = 20; Range = 18 - 26) participated in the experiment for monetary compensation. Older adults were given several standardized neuropsychological tests in a pre-screening session to ensure that they fell within normal ranges (see Procedure). All older adults were within normal ranges for memory and executive function (see Table 2). As subjects were given the rule that would categorize each of the rule-following items perfectly prior to the beginning of the experiment and reminded of the rule on each trial (see Procedure), failure to significantly exceed chance performance on the rule-following items was taken as evidence of non-compliance or poor comprehension of instructions, and used as an exclusion criterion. Five older adults and one younger adult were excluded from the study for failing to categorize the rule-following items at greater than chance levels.

Materials

The stimuli used for the category learning task consisted of schematic beetles that varied along four perceptual dimensions (see Figure 1A) and were assigned to categories (Hole A or Hole B based on their combinations of feature values. For each of the stimuli, four

of five possible dimensions (eyes, tail, legs, antennae, and fangs) were randomly selected to vary, and the unselected dimension was held fixed at a constant value. Six of the stimuli were rule-following items and could be categorized correctly based on the value of a single rule-relevant dimension. In the example in Figure 1A, the rule-relevant dimension was the legs; all but one of the beetles in Hole A had thick legs, and all but one of the beetles in Hole B had thin legs. The other two beetles (circles in Figure 1A) served as exceptions to the rule and appeared to belong to the opposing category based on their value on the rule-relevant dimension (legs). An abstract representation of the category structure is given in Table 1. In order to minimize the effects of feature salience, the mapping of each abstract dimension to a physical dimension was randomized for each subject.

Procedure

Neuropsychological Testing Session. Older adults were given a battery of standardized neuropsychological tests that spanned three functional realms: *memory* (Wechsler Memory Scale Third Edition (WMS-III) subtests: Weschler 1997; California Verbal Learning Test (CVLT): Delis et al. 1987), *mood* (Geriatric Depression Scale (GDS): Brink 1984), and *executive functioning and mental flexibility* (Stroop Color-Word Test: Stroop 1935; Trail Making Test A&B (TMT): Lezak 1995; Controlled Oral Word Association (COWA): Lezak 1995; Wisconsin Card Sorting Task (WCST): Heaton 1981). The tests were administered in a single two hour session in the following order: CVLT, GDS, WAIS-III Information subtest, WAIS-III Arithmetic subtest, WAIS-III Vocabulary subtest, CVLT delayed recall, WMS-III Logical Memory subtest, Stroop, TMT A&B, WAIS-III Similarities subtest, COWA, WAIS-III Digit Span subtest, WMS-

III Logical Memory delayed-recall, WMS-III Visual Reproduction subtest, WAIS-III Letter/Numbering Sequencing subtest, WCST computerized version, WMS-III Visual Reproduction delayed-recall. The delay period of tests requiring a delay was kept constant, and was comprised of other tests not requiring any long term memory storage. The standard, age appropriate, published norms were used to calculate normative scores for each subject. For all of the WAIS subtests, the percentile was calculated according to testing instructions, and this score was then converted to a standardized z-score. For the Stroop, CVLT, and WCST standardized T-scores were calculated according to testing directions, and this score was then converted to a standardized z-score. Finally, for the TMT and COWA standard z-scores were calculated according to the testing instructions. Older adults who had z-scores on two or more tests in the same functional realm that were two standard deviations below the mean were not asked to participate in the study. Table 2 shows the means, standard deviations and ranges of raw and standardized z-scores for each of the tests.

Both Older and Younger adults completed the WAIS vocabulary in order to control for basic verbal intelligence. Older adults had higher scores (**M** = 54.63) on the vocabulary sub-test than younger adults (**M** = 47.35), $t(61) = 3.17$, $p = 0.002$, $d = 0.81$. However, when included as a covariate, vocabulary did not have an appreciable impact on results presented below.

Category Learning Task. Subjects were trained on rule-plus-exception task for 16 blocks in which each of the beetles stimuli were presented once in a random order. Prior to beginning the task, subjects were given the rule that would allow them to

successfully categorize the rule-following items and were told that they would need to memorize exceptions to the rule (Davis et al. 2011; Love and Gureckis 2007).

On each trial of the category learning task, a single beetle was presented in the center of the screen, and subjects were asked "Does This One Belong In Hole A or Hole B?" (Figure 1B). A hint was presented on each trial below the prompt cuing subjects to the rule to reduce the attentional demands of the task and focus on pattern separation and pattern completion mechanisms. Subjects were allowed to respond at their own pace by hitting the A or B button. After categorizing a beetle, corrective feedback that showed the beetle next to the correct category was immediately presented for 2.5 seconds, followed by a blank screen for 0.25 seconds.

Model Formalism and Procedure

The RMC models category learning behavior by estimating, on each trial, the probability that a stimulus will belong to a given category based on its observed features and the features associated with cluster representations stored in memory. A cluster is a grouping of stimuli that code the features and category labels associated with stimuli in the task, and can contain anywhere from a single stimulus to all of the stimuli observed in a task. The RMC assumes that one cluster was responsible for generating the observed stimulus, and generates an overall probability of the category label by aggregating the category label predictions over clusters.

Formally, the probability that a stimulus i belongs to category j given its observed feature structure F and the features of k stored clusters is given by:

$$P_i(j \mid F) = \sum_k P_i(k \mid F) P_i(j \mid k) \tag{1}$$

where $P_i(k|F)$ is the probability the stimulus i coming from cluster k given the observed features and $P_i(j|k)$ is the probability of i having the category label j given the cluster. Each $P_i(k|F)$ is a posterior probability that combines the prior probability of cluster k and the likelihood of the stimulus features F given cluster k:

$$P(k|F) = \frac{P(k)P(F|k)}{\sum_k P(k)P(F|k)} \tag{2}$$

where $P(k)$ is the prior probability of the cluster over possible partitionings of the stimulus space. The denominator in equation 2 normalizes the cluster probabilities to sum to one.

$P(k)$ is given by:

$$P(k) = \begin{cases} \frac{cn_k}{(1-c)+cn} & \text{if } n_k > 0 \ (k \text{ is old}) \\ \frac{(1-c)}{(1-c)+cn} & \text{if } n_k = 0 \ (k \text{ is new}) \end{cases} \tag{3}$$

where c is a coupling parameter that controls how likely objects are to be stored in the same cluster, n is the total number of stimuli presented so far, and n_k is the number of stimuli assigned to cluster k. When c is low, objects tend to be stored in separate clusters, whereas when c is high, objects tend to be stored in separate clusters. Here we use a separate coupling parameter for each group c_{older} and c_{young} to evaluate the hypothesis that the difference between older and younger adults is in their ability to pattern separate objects into separate clusters. $P(k)$ has a modulating effect on the posterior probabilities such that as the number of previous items stored in cluster k increases, the RMC judges the cluster as more probable to have generated a stimulus.

$P(F|k)$ (in equation 2) is the likelihood of the stimulus features F given cluster k given by:

$$P(F \mid k) = \prod_i P_i(j \mid k) \tag{4}$$

where j's are the values on each i stimulus dimension in the feature set F. Equation 4 assumes that features within a cluster are independent of one another, and so the individual contribution of each feature dimension can be multiplied to obtain the likelihood. In discrete dimension cases, like the present application, the contribution of matches/mismatches on an individual feature dimension i is given by:

$$P_i(j \mid k) = \frac{n_j + \beta_i}{n_k + 2\beta_i} \tag{5}$$

where n_j is the number of stimuli in cluster k with the same value as the to be classified stimulus on dimension i, n_k is the number of stimuli in cluster k (assuming all stimuli have a value on dimension i), and the β_i's are symmetric beta priors that controls the weight of matches (/mismatches) on dimension i in the likelihood computation (i.e., dimensional salience). The β_i's are constrained to be in the range (0,1) with lower values indicating higher salience. In the present application, we fit a separate beta prior for the rule-relevant dimension β_r and the last three other dimensions β_o to reflect the instructions given to subjects to focus on the rule-relevant dimension. The beta prior for the category label dimension was fixed at 0.01, which reflects the assumption that stimuli with different category labels will belong to different clusters. All priors are symmetric such that, for a given stimulus dimension/category label, matches and mismatches are weighted the same regardless of the particular feature instantiation (e.g., whether the category label is A or B).

In the present application, the probability with which the RMC responds with category *m* given stimulus *i* is scaled using a probabilistic choice rule (Nosofsky et al. 1994):

$$P(resp = m) = \frac{P_i(j = m \mid F)^{\gamma}}{\sum\limits_{j=1}^{J} P_i(j \mid F)^{\gamma}} \tag{6}$$

where γ is a decision parameter that scales the probabilities such that for $\gamma = 1$ (see also, Ashby and Maddox 1993) the model probability matches and responds proportional to the probability of the category label given *i* calculated in (1) and as γ approaches infinity the responses become more deterministic such that the model always chooses the most probable category. After a response is made and feedback is delivered, the cluster probabilities are recalculated taking into account the now observed category label, and the stimulus is stored in is the most likely given the stimulus' conjunction of features and category label, or a new cluster if a candidate new cluster is judged more likely to have generated the stimulus.

Values for β_r, β_o, c_{older}, $c_{younger}$, and γ were estimated by fitting the model, using a a grid search, to subjects' average performance for each trial type in the final 5 blocks of learning, minimizing the sum of squared error. The final 5 blocks of learning (blocks 12-16) have been used in previous rule-plus-exception tasks as a basis for computing learning criteria for determining whether or not subjects learned (Davis et al. 2009; Davis et al. 2011). Importantly, in the present case older adults exhibited exception performance that was not significantly different from the previous 5 blocks (blocks 6-11; $t(31) = 1.11$); indicating that their performance was relatively stable by blocks 12-16.

All parameters except the coupling parameters were constrained to be equal across older and younger adults. The best fitting parameter values were: $\beta_r = 0.83$, $\gamma = 1.56$, $c_{older} = 0.60$, $c_{younger} = 0.50$ ($SSE = 0.003$), and the model approximated each group's final learning performance well (RMC_{older}: exceptions $= 0.35$; rule-following $= 0.86$; $RMC_{younger}$: exceptions $= 0.64$; rule-following $= 0.88$). To assess the RMC's predicted clustering for each group at the fitted parameter values, we simulated the RMC for older and younger adults over 10,000 runs of the model and computed the modal number of clusters over runs ($RMC_{older} = 2$; $RMC_{younger} = 4$). For individual subject analysis, we used the parameter values from the group subject fits for all parameters except for the coupling parameters, which we fit to each subject separately.

Supplementary Table. Response time means and standard deviations for each item type, accuracy (correct or incorrect), and age group. Outlier reaction times greater than 2.5 standard deviations above each subjects' median reaction time were removed from calculations.

Older Adult		Rule-following (milliseconds)	Exceptions (milliseconds)
	Correct	M = 3277; SD = 1816	M = 3977; SD = 1603
	Incorrect	M = 4661; SD = 2671	M = 3561; SD = 2106
Younger Adult			
	Correct	M = 1530; SD = 636	M = 1990; SD = 976
	Incorrect	M = 1661; SD = 978	M = 1468; SD = 749

Figure S1. Rule-following and exception item performance as a function of learning block for older and younger adults. Error-bars depict +/- standard errors of the mean. Because our focus is on differences in final representations between groups, and performance is stable in both groups during the last 5 learning blocks, all model-based and neuropsychological analyses focus on performance at the end of learning.
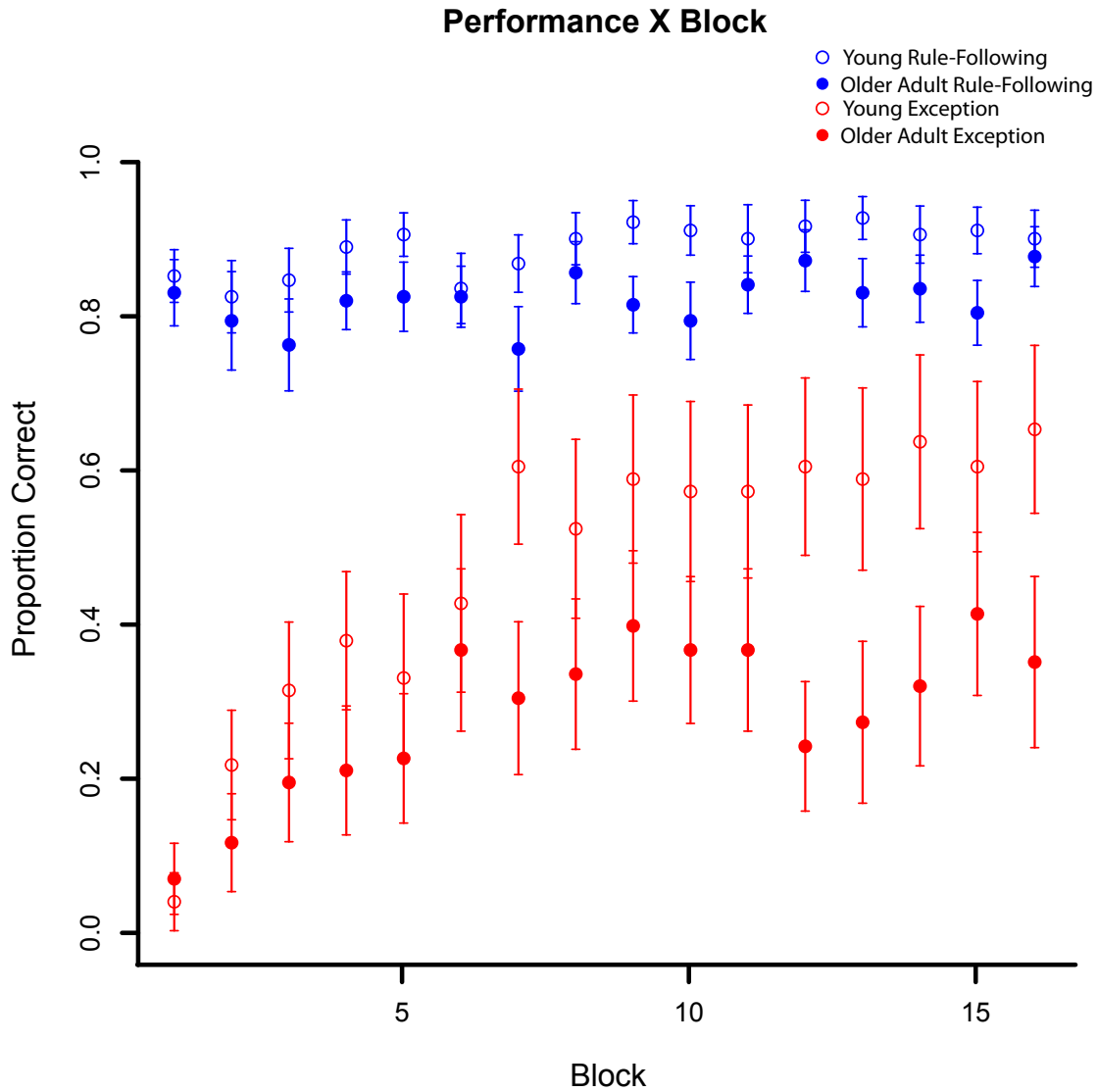
Figure S2. Histogram detailing distribution of proportion correct for exception items during last 5 blocks of learning for older and younger adults.
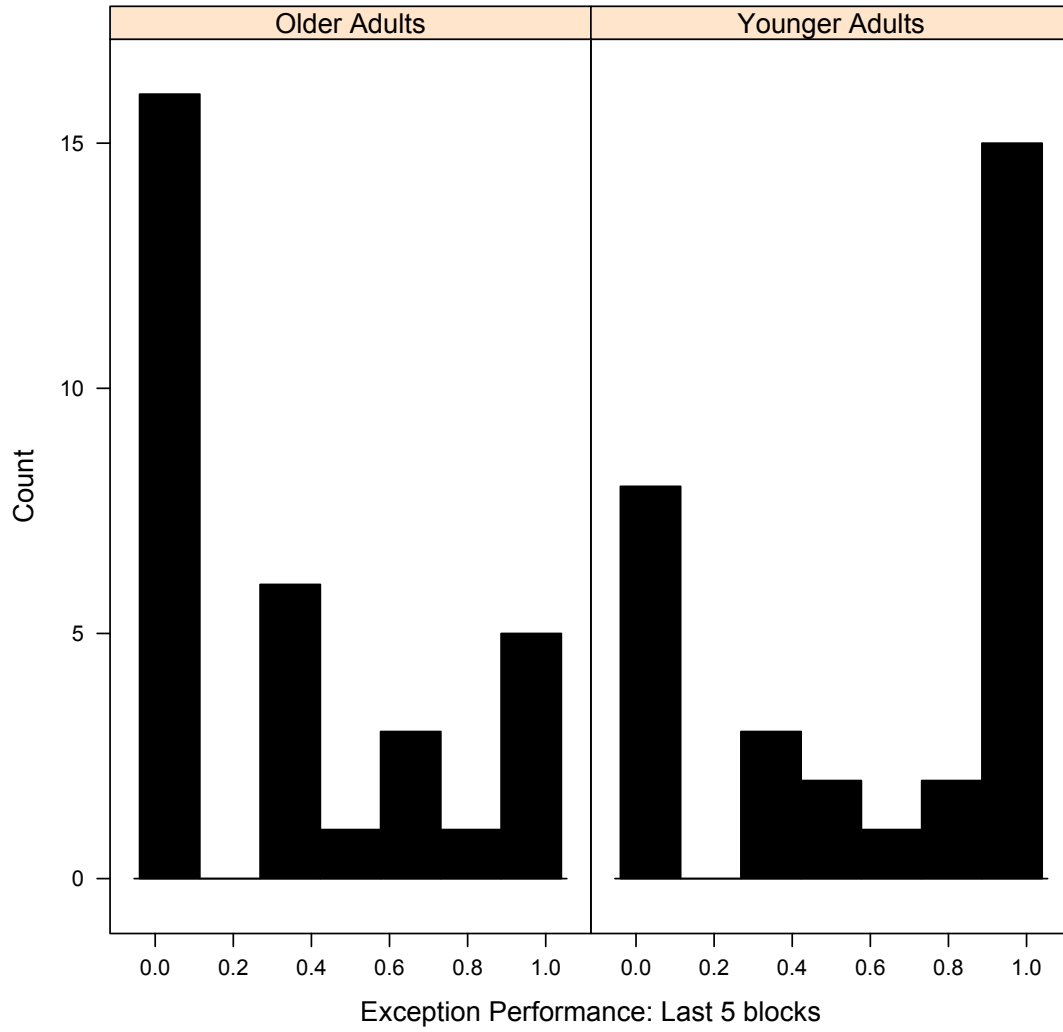
Figure S3. Histogram detailing distribution of coupling parameter for older and younger adults.