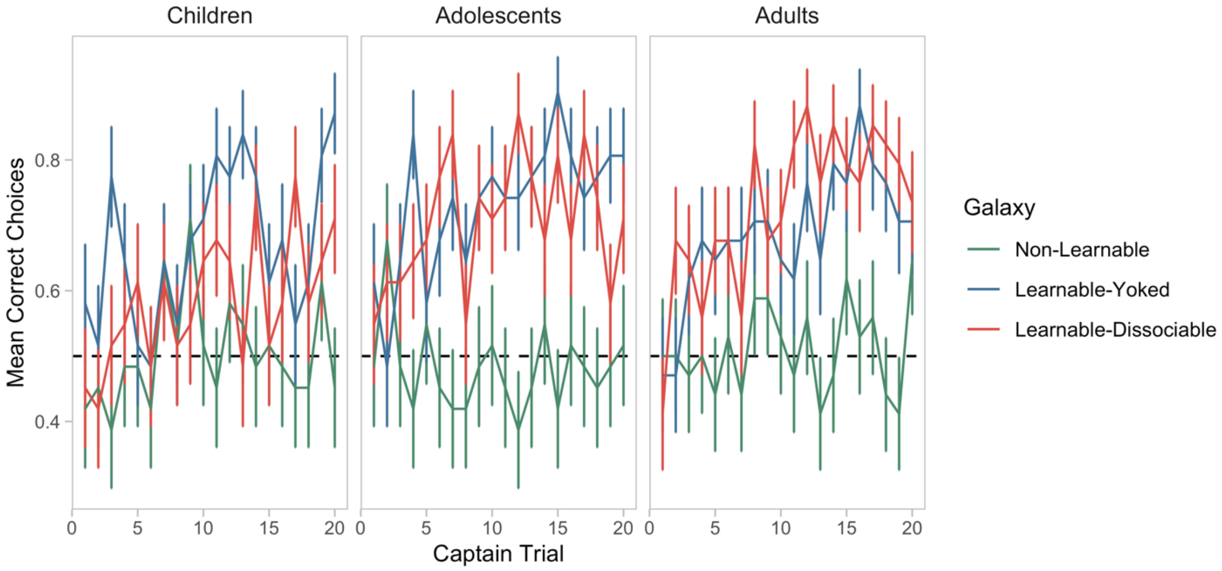
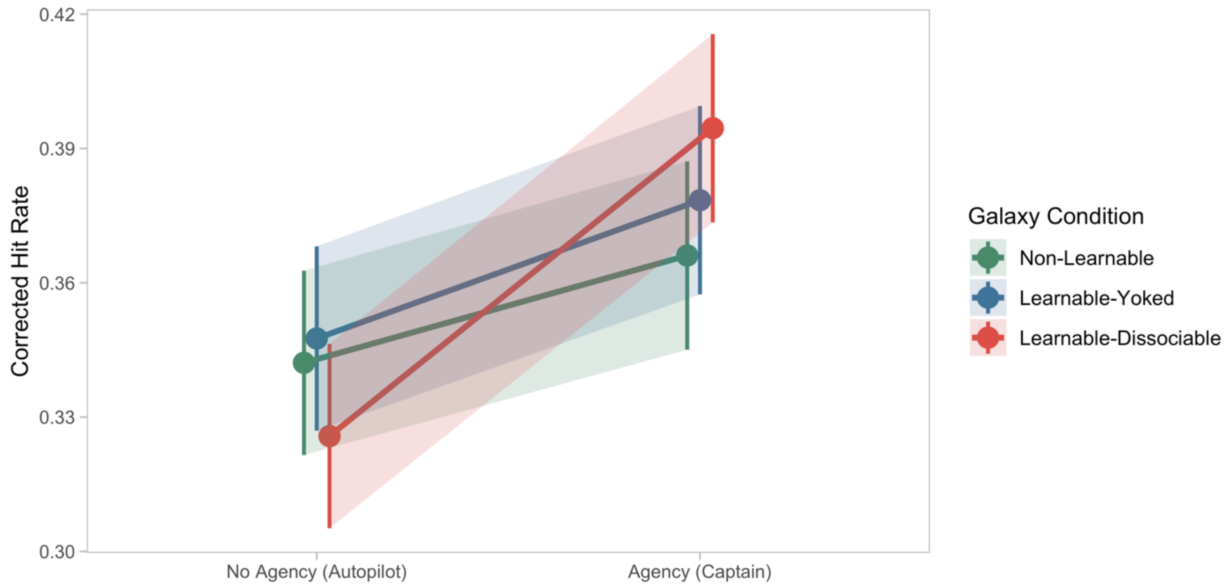
Katzman and Hartley, 2020
The value of choice facilitates subsequent memory across development

Supplemental Information

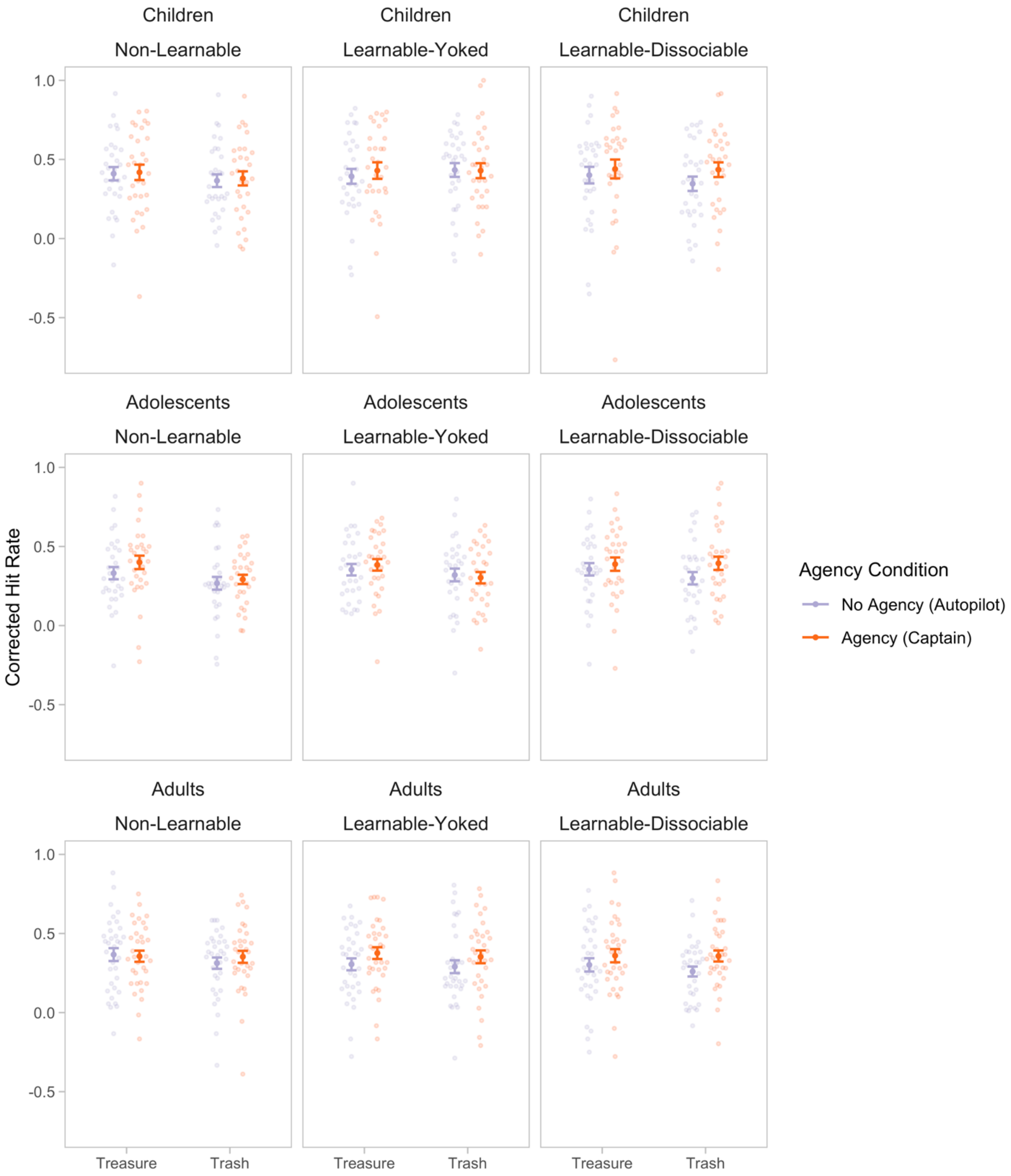
Participant data and analysis code can be found online at the Open Science Framework: https://osf.io/n4kam/



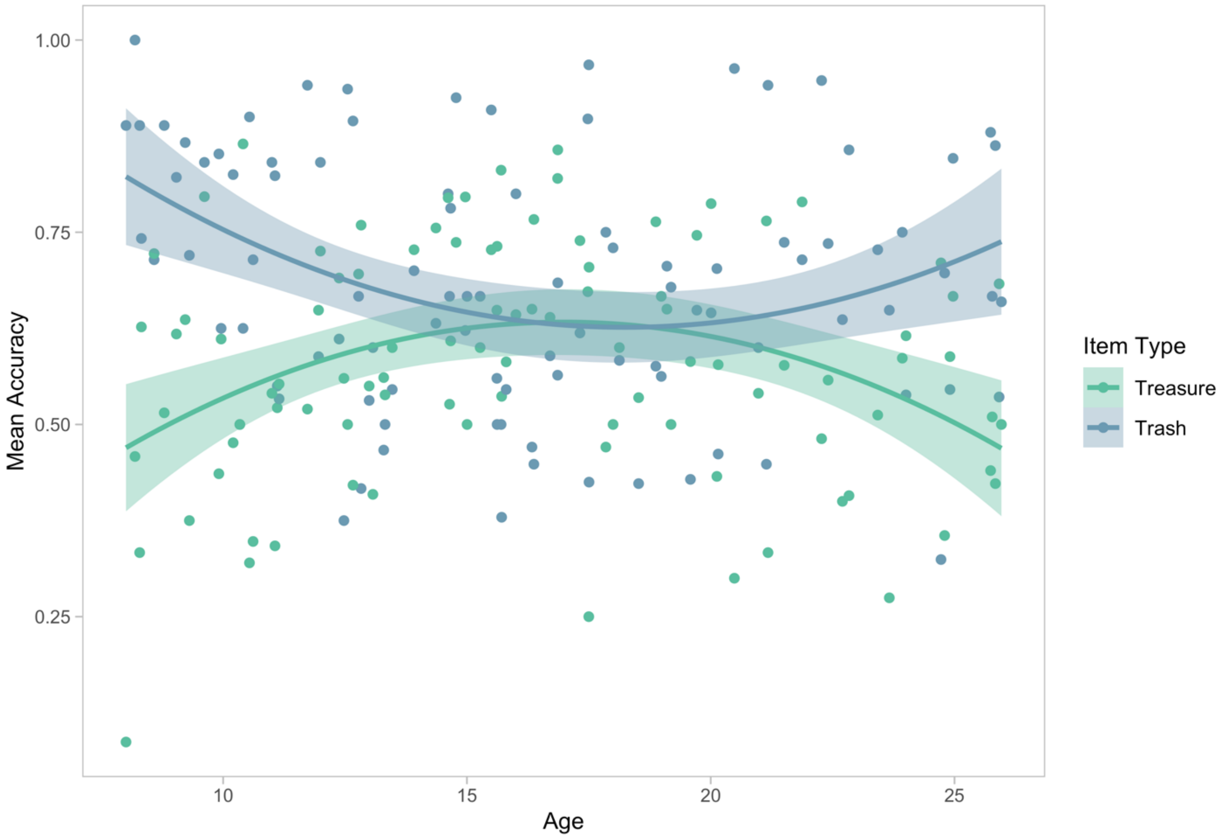
Supplementary Figure 1. Learning Phase performance, averaged across all participants in each age group (children: age 8-12, adolescents: age 13-17, adults: age 18-25) for each Captain trial for each galaxy. Chance performance denoted by dashed line. Error bars denote standard error.



Supplementary Figure 2. The interaction plot depicts the difference in memory performance for Captain trials compared to Autopilot trials by galaxy condition. The agency benefit (indicated by the slope of each line) is greater for the Learnable-Dissociable galaxy (in red) relative to the Non-Learnable galaxy (in green).



Supplementary Figure 3. Corrected hit rates for trash and treasure items in each galaxy for each age group (children: age 8-12, adolescents: age 13-17, adults: age 18-25) . A corrected hit rate of 0 indicates an equal proportion of hits and false alarms for that condition. Positive values indicate better memory performance. Smaller dots represent individual subjects' corrected hit rate for the corresponding trial types, with the means represented as larger dots, and error bars reflecting standard error.



Supplementary Figure 4. Source memory accuracy for trash and treasure items, plotted by age.

			Children	Adolescents	Adults
Learnable-Dissociable	Captain	Treasure	.574 (.251)	.692 (.201)	.642 (.203)
		Trash	.569 (.227)	.697 (.253)	.640 (.211)
	Autopilot	Treasure	.535 (.226)	.659 (.188)	.584 (.223)
		Trash	.480 (.237)	.602 (.211)	.542 (.226)
Learnable-Yoked	Captain	Treasure	.564 (.230)	.687 (.189)	.658 (.186)
		Trash	.563 (.230)	.606 (.238)	.635 (.232)
	Autopilot	Treasure	.527 (.209)	.656 (.198)	.587 (.211)
		Trash	.567 (.186)	.623 (.224)	.572 (.222)
Non-Learnable	Captain	Treasure	.553 (.219)	.703 (.150)	.639 (.195)
		Trash	.514 (.232)	.595 (.209)	.635 (.178)
	Autopilot	Treasure	.544 (.200)	.635 (.222)	.649 (.243)
		Trash	.500 (.229)	.571 (.206)	.595 (.183)
False Alarm Rate			.134 (.155)	.303 (.185)	.282 (.190)

Supplementary Table 1. Mean (SD) hit rates and false alarm rates per trial type by age group.

Supplementary Analyses

Excluding Participants for Low Performance
We identified participants with an overall corrected hit rate (cHR) at or below zero (i.e., those for whom the proportion of remembered items was equal to or less than the proportion of correctly rejected lures). This lowered our number of participants from 96 to 92. We then re-fit a linear mixed-effects model with agency condition, galaxy, age, and reward (treasure or trash) as regressors to predict cHR, and included a random intercept and random slopes for agency and reward for each participant. Our effects still held with those participants excluded. We found a main effect of reward such that participants had greater cHR for treasure than trash items overall *(*X^2 = 5.4, df =1, *p* = .02). Additionally, the interaction between galaxy and agency remained with greater cHR for Captain than Autopilot trials in the Learnable-Dissociable galaxy compared to the Non-Learnable galaxy (X^2 = 4.6, df = 1, *p* = .031). No other main effects or interactions reached significance.

Excluding Trials with Short/Long Reaction Times
We assessed whether the inclusion of trials with exceedingly short or long reaction times (RTs) affected our results. For the learning task, we decided not to impose any minimum reaction time because participants are able to see the options and plan their choice prior to making a response. Even so, most responses were made within 10 seconds (median RT = 0.84s), and the longest RT was 34s. To test whether inclusion of outlier RTs influenced our finding, we defined an upper limit by first log-transforming the RTs and setting a cutoff at 2.5 standard deviations above the mean. Transformed back from log space, the upper limit was 13.18s. We then re-fit the learning phase accuracy model excluding trials above the RT threshold and all effects remained significant. Accuracy was greater for the Learnable-Dissociable (X^2 = 27.0, df = 1, *p* < .001) and Learnable-Yoked (X^2 = 37.9, df = 1, *p* < .001) galaxies relative to the Non-Learnable galaxy. Accuracy also increased over the course of the task in the Learnable-Dissociable (X^2 = 19.4, df = 1, *p* < .001) and Learnable-Yoked (X^2 = 21.4, df = 1, *p* < .001) galaxies relative to the Non-Learnable galaxy.

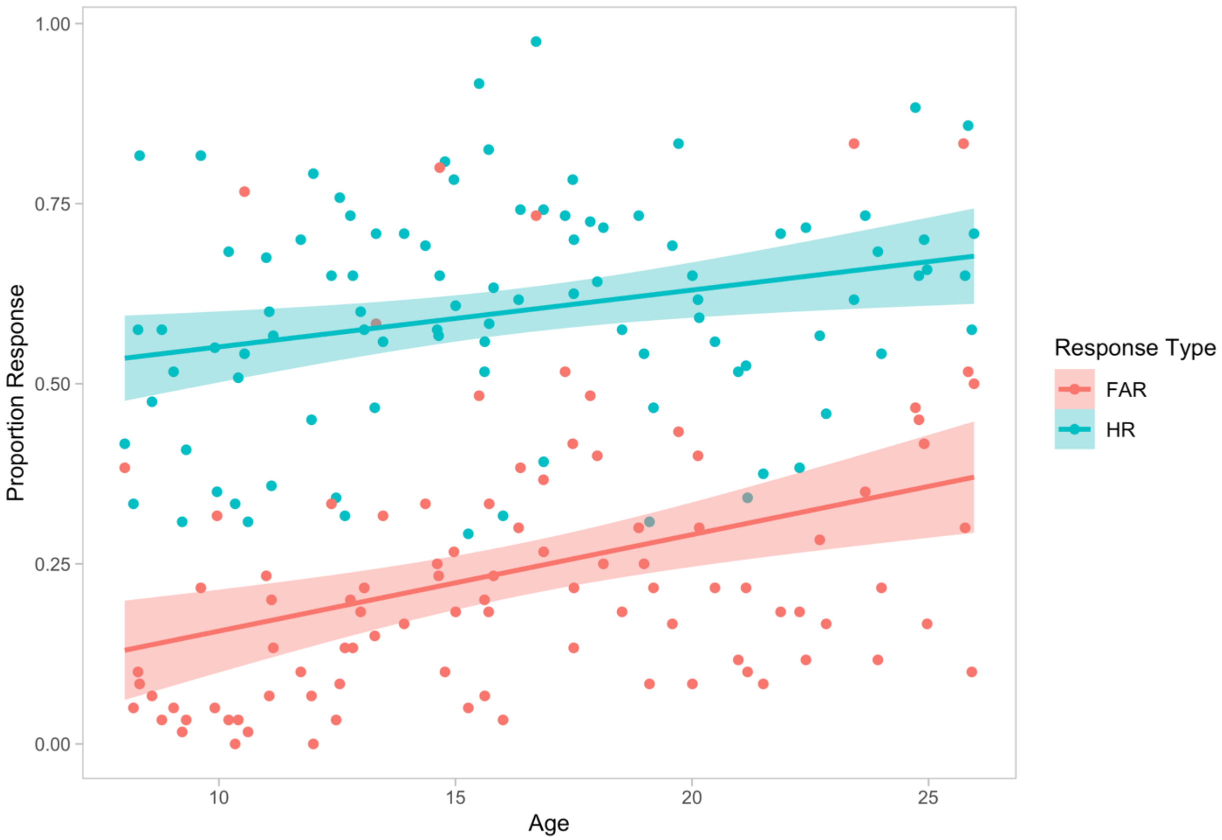
For the recognition memory trials, we set a lower bound for RTs less than 250ms. To determine an upper bound we again calculated 2.5 standard deviations above the mean of the log RT distribution. When transformed back from log space, this put the upper bound at 10.3s. We next identified all memory phase trials in which the recognition memory RT was above the upper or below the lower threshold and trials for which the learning phase RT for that item was above its respective threshold, and removed them from analysis. We then re-calculated hit rates and re-fit the recognition memory accuracy model. Our results still held, with a main effect of reward (X^2 =5.1, df = 1, *p* = .023) and an interaction of agency and the Learnable-Dissociable vs. Non-Learnable galaxy (X^2 = 4.3, df = 1, *p* = .037).

Exploratory High- and Low-Confidence Memory Analyses
We calculated high-confidence (HC) and low-confidence (LC) hit rates as the proportion of HC/LC "old" responses out of all previously studied items, and calculated HC/LC false alarms as the proportion of HC/LC "old" responses out of all unstudied items. We used HC/LC corrected hit rate as our dependent variable, calculated as HC/LC hit rate minus HC/LC false alarm rate, respectively. Our models included galaxy, agency, reward, and age as predictors, with a random intercept for each participant. The LC model would not converge with the addition of any random slopes, so we removed the random slopes from both models for comparison. In our HC model, we found a main effect of reward (X^2 = 10.6, df = 1, *p* = 0.001), with better HC memory performance for treasure items than trash items, and a main effect of agency (X^2 = 5.0, df = 1, *p* = .025), with better HC memory performance for Captain trials than Autopilot trials. No other effects or interactions reached significance. In our LC model there were many significant effects. LC memory performance was worse in the Learnable-Dissociable galaxy compared to the Non-Learnable galaxy (X^2 = 6.3, df = 1, *p* = .012), better for treasure items than trash items in the Learnable-Dissociable galaxy compared to the Non-Learnable galaxy (X^2 = 4.5, df = 1, *p* = .034), and better for treasure items than trash item for Captain trials compared to Autopilot trials (X^2 = 4.7, df = 1, *p* = .03). There was a triple interaction of galaxy, agency, and reward such that the degree to which LC memory performance was better for treasure vs. trash items in Captain trials compared to

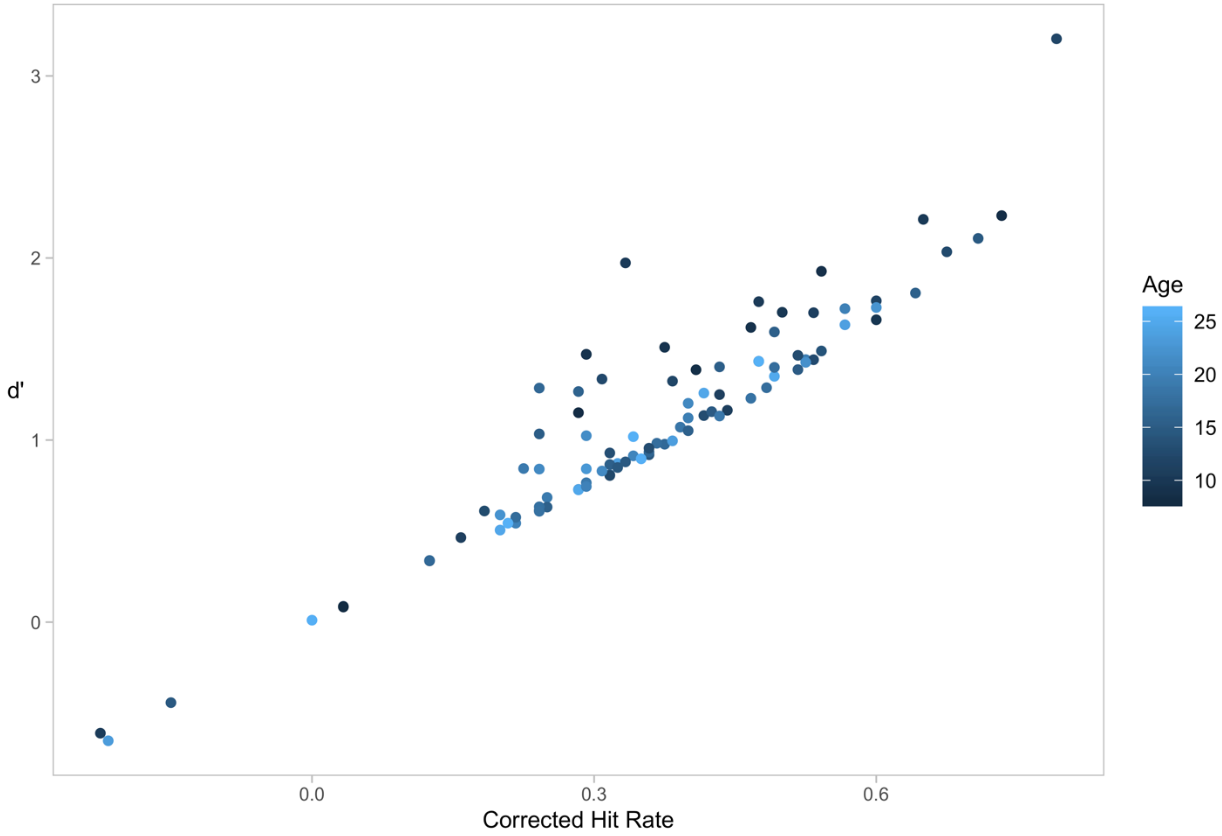
Autopilot trials was larger in the Non-Learnable galaxy relative to the Learnable-Dissociable galaxy (X^2 = 4.4, df = 1, *p* = 0.036). There was a triple interaction of galaxy, agency, and age such that the degree to which LC memory performance was better for Captain vs. Autopilot trials in the Learnable-Dissociable galaxy compared to the Non-Learnable galaxy decreased with age (X^2 = 4.6, df = 1, *p* = .031). Finally, a four-way interaction between galaxy, agency, reward, and age also reached significance, where the previously described three-way interaction between galaxy, agency, and age was greater for trash items compared to treasure items (X^2 = 4.7, df = 1, *p* = .029). Participants responded “old” with high confidence on roughly 64% of memory trials. This proportion did not vary by age, galaxy, agency, value, or any interaction. The fact that the majority of “old” responses were made with high confidence may explain why our overall results are more closely aligned with those of the HC model.

Exploratory d’ Analyses

We calculated d’ using the R package *psycho* (*v0.4.9;* Makowski, 2018). We fit a mixed-effects linear model with predictors of galaxy, agency, reward, and age, and included a random intercept and random slopes for agency and reward for each participant. The interaction between the random slopes was removed in order to get the model to converge. There was still a positive effect of reward (X^2 = 7.3, df = 1, *p* = 0.007), where d’ was greater for treasure items than trash items. Our interaction of interest (agency x (Learnable-Dissociable vs. Non-Learnable)) also remained significant (X^2 = 4.5, df = 1, *p* = .035). However there was also a significant negative effect of age (X^2 = 4.5, df = 1, *p* = 0.035), where d’ was lower for the older participants relative to younger participants. While cHR and d’ were highly correlated (t(94)=24.4, *p* < .001, r = .93), the discrepancy in results for these two measures appeared to stem from systematic age differences in the tendency to endorse an item as old (either correctly or incorrectly). Both hit rates and false alarm rates increase with age at a similar rate (Supplementary Fig. 5). As a result, cHR stays relatively consistent over age. Because the transformation to d’ is nonlinear, a constant difference between hit rate and false alarm rate can still result in a range of d’ values. As seen in Supplementary Figure 6, the discrepancy between cHR and d’ varies with age, with the majority of participants with strongly deviating d’ scores being younger participants (indicated by darker colors). Because these deviations inflate the d’ score relative to cHR, that can explain why the children have a higher d’ than adults.



Supplementary Figure 5. Overall hit rate (HR) and false alarm rate (FAR) plotted for each individual by age.



Supplementary Figure 6. Plotted is the relationship between each individual's overall corrected hit rate and d'. Though the two measures are highly correlated, the greatest deviations between cHR and d' are primarily found in younger participants (indicated by darker colors).

Reference

Makowski, D. (2018). The Psycho Package: An Efficient and Publishing-Oriented Workflow for Psychological Science. Journal of Open Source Software, 3(22), 470. Available from https://github.com/neuropsychology/psycho.R