



ISSN: (Print) (Online) Journal homepage: [www.tandfonline.com/journals/pvis20](http://www.tandfonline.com/journals/pvis20)

## What can 9 million trials tell us about memorability in a hybrid search task?

Dyllan D. Simpson, Jeremy M. Wolfe & Anna Kosovicheva

To cite this article: Dyllan D. Simpson, Jeremy M. Wolfe & Anna Kosovicheva (23 Apr 2025): What can 9 million trials tell us about memorability in a hybrid search task?, Visual Cognition, DOI: [10.1080/13506285.2025.2492667](https://doi.org/10.1080/13506285.2025.2492667)

To link to this article: <https://doi.org/10.1080/13506285.2025.2492667>



© 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



[View supplementary material](#)



Published online: 23 Apr 2025.



[Submit your article to this journal](#)



[View related articles](#)



[View Crossmark data](#)

## What can 9 million trials tell us about memorability in a hybrid search task?

Dyllan D. Simpson<sup>a,b</sup>, Jeremy M. Wolfe<sup>c,d</sup> and Anna Kosovicheva<sup>b</sup>

<sup>a</sup>Department of Psychology, University of California San Diego, La Jolla, CA, USA; <sup>b</sup>Department of Psychology, University of Toronto Mississauga, Mississauga, Canada; <sup>c</sup>Brigham and Women's Hospital, Boston, MA, USA; <sup>d</sup>Departments of Ophthalmology and Radiology, Harvard Medical School, Boston, MA, USA

### ABSTRACT

Hybrid visual search tasks involve searching for multiple targets held in memory, but some targets are more memorable than others. Furthermore, some items are readily identified as being in the memory set, while others are readily identified as *not* being in the memory set; these may be considered to vary in their “hittability” and “rejectability”, respectively. In principle, both factors should impact error rates and reaction times in hybrid search. Using a set of 9 million trials from an online hybrid search game, we analyze participants’ errors and show that hittability and rejectability are largely separable. It is possible for items to be rejectable without being particularly hittable, and to be hittable without being particularly rejectable. Both factors are consistent across participants and stable across age, training, and performance. Rejectability strongly predicted reaction times in the search for new items, while hittability was more weakly associated with reaction times.

### ARTICLE HISTORY

Received 19 December 2024  
Accepted 8 April 2025

### KEYWORDS



Memorability; hybrid search; visual search; signal detection theory; long-term memory


Throughout our lives, we encounter a wide variety of images that are not all remembered equally well. We all have objects that we find uniquely easier to recall than others. Conversely, some objects may be particularly difficult to recall. It is easy to see how this difficulty in recognizing items as previously encountered might impact our ability to perform daily tasks. One can imagine, for example, going clothing shopping and accidentally repurchasing the exact same dress that you already have at home. Previous work on long-term memory for sets of images has shown that some items are intrinsically more memorable than others (e.g., Bainbridge et al., 2013; Bylinskii et al., 2015). In other words, some items are more likely to be correctly identified as a member of the memorized set. Previous studies have consistently demonstrated that memorability is stable for stimuli such as scenes and faces across different observers (Bainbridge et al., 2013; Khosla et al., 2015), but the precise factors determining image memorability are still under investigation (Isola et al., 2014; Kramer et al., 2023).

Memorability is typically measured by having participants complete a visual recognition task, in which

they are shown a series of images, one on each trial, and asked to indicate when an image is repeated. Responses are then aggregated across a large number of participants. Memorability can be calculated from the proportion of correct responses out of the repeated images (i.e., hit rate; Isola et al., 2011, 2014), though many studies subtract the false alarm rate (proportion of new images reported as old) from the proportion of hits to calculate a single memorability score in an effort to minimize measurement variability (Khosla et al., 2015). Other studies report hit rate and false alarm rate separately (e.g., Bainbridge et al., 2013; Bainbridge & Rissman, 2018), in addition to reporting a single memorability score. Despite these different approaches to measuring memorability, consistent patterns emerge in which objects are more or less memorable across observers.

More recent work has separately analyzed the relationship between hit rates and correct rejection rates (i.e., proportion of correct responses on target-absent trials) to determine whether these capture distinct object properties. In other words, are objects that are readily identified as being *in* the memory

**CONTACT** Dyllan D. Simpson  dysimpson@ucsd.edu  Department of Psychology, University of California San Diego, 9500 Gilman Dr, La Jolla, CA 92093, USA

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/13506285.2025.2492667>.

© 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

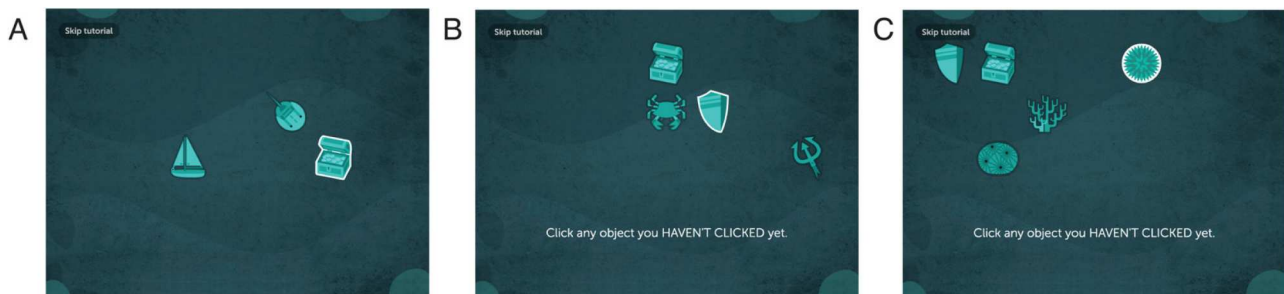
set (i.e., items with low miss rates) also the same objects that are easily identified as *not* being in the memory set (i.e., items with low false alarm rates)? Using a recognition memory task, Zhao et al. (2023) found that hit rates and correct rejections for individual items were not correlated with one another. In addition, while participants' ratings of how frequently they encountered items in daily life were significantly associated with correct rejection rates ("rejectability") for images, the same ratings did not correlate with hit rates ("hittability"). These findings provide evidence that these two properties are separable. Importantly, hittability is a property of old items, and rejectability is a property of new items. This distinction highlights the importance of examining both aspects of memory performance to fully understand the underlying factors.

In principle, this multifaceted nature of memorability should also impact how we interact with objects in a range of tasks that extend beyond simple recognition tasks. For instance, many daily tasks require us to use information about items stored in memory to search for an item (for example, searching grocery store shelves for the items needed for a recipe). In these sorts of hybrid search tasks, observers search for multiple types of targets held in memory, or for new items among distractors consisting of old items (Wolfe, 2012; Wolfe et al., 2015). If rejectability and hittability are indeed separable constructs, they should impact performance in hybrid search tasks in different ways, which would be seen in the types of errors that participants make. For example, if participants search for new items among a set of previously seen items, items with poor hittability would produce consistently high miss rates across participants (i.e., participants would consistently mistake an old item for a new item). Moreover, we expect that hittability and rejectability should have distinct effects on participants' reaction times in a hybrid search task, reflecting their ability to identify the search target. For example, if the task is to search for new items, items with poor rejectability (but not hittability) should elevate reaction times as participants would have difficulty identifying the target, while items with high rejectability would decrease reaction times. Conversely, if the task is to search for old items, the presence of items with poor hittability would elevate reaction times, while items with high hittability would produce faster responses. While

previous work has examined how reaction times vary with memory and visual set size (e.g., Drew et al., 2017; Gronau et al., 2024; Wolfe et al., 2016), the impact of individual item properties (hittability and rejectability) on reaction times is less well understood.

Some evidence suggests that observers' prior experience with items impacts hybrid search performance only rather minimally. Wolfe et al. (2015) previously tested whether the familiarity of distractor items impacts performance in a hybrid search task. In principle, if participants encounter distractors repeatedly when searching for old items, they might falsely select one of the "lures", mistaking it for an item in the memory set. However, participants' performance was largely unaffected by this type of familiarity, based on repetition, regardless of whether the task was to search for an old item among repeatedly-seen distractors or to search for a new item among infrequently-seen distractors. Although participants were remarkably good at rejecting such lures within a hybrid search setting, one possibility is that hittability and rejectability represent more stable properties of objects, and that these could be reflected in the errors that participants make in a hybrid search task.

To examine hittability and rejectability in the context of a hybrid search task, we performed a retrospective analysis of a large-scale dataset from the cognitive training app, Lumosity, provided by the company, Lumos Labs. One of their online games, Tidal Treasures (Figure 1), was selected for this study due to its close resemblance to previous hybrid search tasks (e.g., Wolfe et al., 2015). In Tidal Treasures, participants are instructed to select a new item from a set of distractor items that they had previously selected (i.e., a search for a new item on each trial). As the observer progresses through the game, all of the previously chosen items are presented as distractor objects, while new (previously unselected) objects are presented on each trial as targets, with the memory set growing on each trial (Figure 1). This evolving visual set, where participants must remember the selected items, presents a novel opportunity to study object hittability, based on the set of remembered items. Items that participants mistakenly identify as "new" when they are in the memory set indicate poor hittability. Simultaneously, the new items that are not selected enable the study of object rejectability within the same task. That is, items that



**Figure 1.** Task design. **(A)** Participants are initially presented with three randomly selected objects and instructed to select one. In this example, they are shown a sailboat, a horseshoe crab, and a treasure chest. A white outline is shown around the selected object (the treasure chest). Note that an outline was only visible in the game if the participant hovered over a given item with their cursor (prior to clicking it) but was otherwise invisible. **(B)** Once selected, on the next display, the remaining two unselected objects (in this case, the sailboat and horseshoe crab) are removed and three new objects are added (Dungeness crab, trident, and seashell). The previously selected object (the treasure chest) remains on the next display in a randomly selected location. Participants are instructed to click on one of the newly presented objects. Selecting any new item would be correct (for example, the seashell), allowing the participant to continue. If the participant had incorrectly selected the old item (treasure chest), this would end the game. **(C)** This process is repeated on subsequent trials. The two unselected new items from the previous display (trident and Dungeness crab) have been removed and replaced with three new items: a sea urchin, and two different types of coral (the oval and spiny shape). Now both previously selected items (treasure chest and seashell) remain on screen but in new positions. Again, the participant must select one of the three new items to continue. In this example, the participant correctly selected the sea urchin, allowing gameplay to continue.

participants consistently fail to identify as “new” are those participants mistakenly believe to be “old” (i.e., in the memory set), indicating poor rejectability.

We note that classic signal detection models typically classify responses relative to the items in the memory set (i.e., an item correctly classified as “old” that is in the memory would be a “hit”). For consistency with this literature, we also analyze hittability and rejectability with respect to the remembered set, even though the task requires participants to search for new items. Figure 2A and B provide an illustration of how we classified participants’ responses with respect to the memory set.

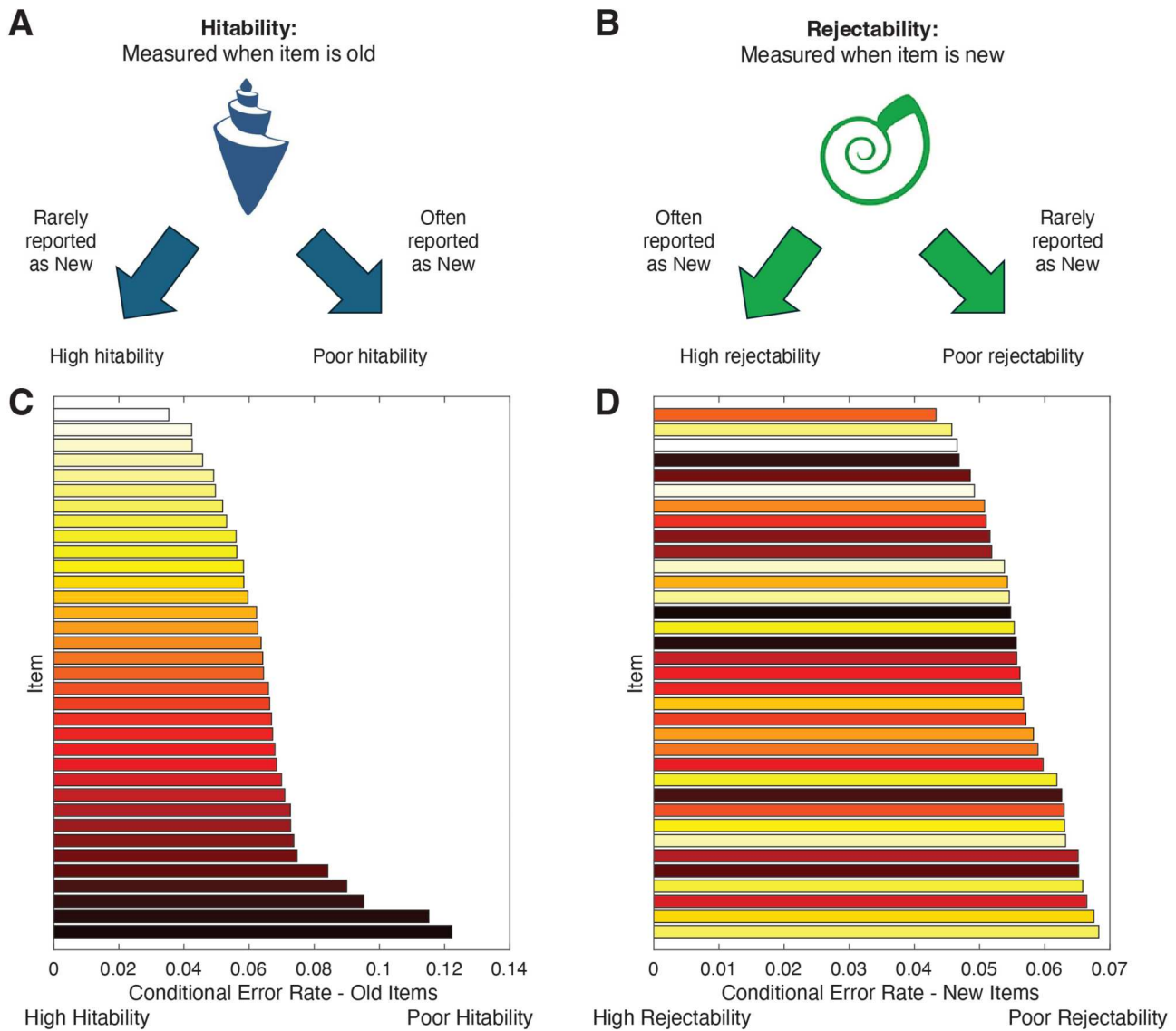
This dataset was collected over repeated sessions from nearly 10000 participants. The size and scope of the data allow us to examine how memorability and rejectability vary as a function of age, performance, and training. Previous studies have demonstrated a strong relationship between cognitive decline and age (e.g., Brockmole & Logie, 2013; Kirova et al., 2015) which would impact overall performance in memory-based tasks. To the extent that memorability is related to high-level object properties, such as typicality (e.g., Kramer et al., 2023; Vokey & Read, 1992) and meaningfulness (e.g., Brady & Störmer, 2022; Shoval et al., 2023a; Shoval et al., 2023b), memorability may vary between older and younger adults and differentially impact performance in a hybrid search task. Additionally, this

dataset allows us to examine the effect of training and practice. Do some items benefit more from repetitive exposure or practice? Do high-performing individuals make different errors than average-performing or low-performing individuals? Finally, as we practice and get better at memory tasks, do we become more resistant to the kinds of mistakes that we made initially when first exposed to objects we’ve never encountered?

## Method

### Lumosity dataset

The study was a retrospective analysis of a de-identified dataset from Lumosity.com and provided by Lumos Labs. The dataset consisted of gameplay data from the game, “Tidal Treasures”. All gameplays were completed on desktop devices between 2014-01-01 and 2018-12-30 and consisted of the first 25 gameplays of the game from all participants who had completed at least 25 gameplays. Analyzing these early gameplays allowed us to examine changes in error rates for individual items as participants learned to play the game. The final dataset consisted of 9748 participants. Demographic details about the participants can be found in the supplemental materials (Figure S1). Briefly, participants were between 21–80 years old ( $M=57.13$ ;  $SD=$



**Figure 2.** (A). Hittability is a property of old items, such that an item rarely reported as new has high hittability, while an item frequently reported as new has poor hittability. (B) Conversely, rejectability is a property of new items; an item often reported as new has high rejectability, while an item rarely reported as new has poor rejectability. (C) Distribution of conditional error rates for old items in the first round of the game, as a measure of hittability. Items are sorted from high hittability (at the top), to low hittability (bottom). (D) Distribution of conditional error rates for new items for the same set of items, as a measure of rejectability. Items are sorted by rejectability as in (C). Colours correspond to the same items shown in panel C.

12.28), with 69% identified as female and 31% male participants. The analysis of this data was approved by the Research Ethics Board at the University of Toronto (protocol #40461).

### **Stimuli and procedure**

Participants completed a beach-themed hybrid search game in which they searched for a new item on every trial embedded among previously seen distractors (see Figure 1). The object of the game was

to collect as many unique new items as possible, without selecting the same item twice. Each gameplay consisted of three consecutive attempts (which we will refer to as “rounds”) to reach this goal. A round ended when the participant collected 35 new items or when they made an error by selecting an old object they had already collected on that round. Items consisted of beach-themed objects (shells, seaweed, crabs, driftwood, etc.) and varied between the three rounds of the game (see Figure S2 in the Supplemental Materials for images of all the items

used). The items shown within each round were drawn from a possible set of up to 35 unique items. However, there was some overlap in items across the three rounds (rounds 1 & 2 had 13 items in common between them, rounds 2 & 3 had 9 items in common, and round 1 & 3 had 8 in common). Each successive round used more challenging items, such that the items in the first round consisted of the largest variety of items by category (28 unique categories: anchor, coral, driftwood, sailboat, shell, etc.), the second had 13 unique categories, and the third consisted of just four categories. Obviously, these were choices made in the original game design, and a prospective study might be structured somewhat differently.

Items in each search display were arranged in an invisible 5-row x 7-column grid (approximately 400 x 560 pixels for the whole grid, with 74 x 74 pixels for each item, and a horizontal and vertical separation of approximately 7 pixels on all sides). The full game display spanned 480 x 640 pixels, including a margin and additional game elements around the grid. The number of items on each display (i.e., trial) varied as described below. Each round of the game began with a display consisting of 3 randomly selected items (positioned in random locations on the grid, with the rest of the spots blank), and participants were asked to select one item by clicking on it. The display was then cleared, and the next trial consisted of three new randomly selected items and the one previously selected item, with all item locations randomly shuffled. The participant was then asked to select an item that they had not previously clicked on. The process repeated on each trial, such that each trial consisted of 3 new items (i.e., search targets) presented alongside all the previously selected items (distractors), with the memory set size growing on each trial. Participants had an unlimited amount of time to select a new item on every trial. This continuous round terminated when the participant either made an error by clicking on a previously selected item or otherwise after a correct response once the screen was filled with all 35 possible items.

Following each response, participants received feedback regarding accuracy in the form of a green checkmark or a red "x" overlaid on the item they had just selected (for correct and incorrect responses, respectively). An incorrect response would terminate

the round, at which point, participants were shown a display of all the items they had selected, in the order they were chosen, with the duplicate selection highlighted. In the event of a correct response, the game immediately proceeded to the next trial. At the end of each round, participants manually initiated the next round by clicking a button labelled "next". After round 3, participants were given the option to play the game again; this was counted as a separate gameplay.

To minimize repetition of the unselected search targets (i.e., unselected "new" items), the targets that were not selected on each trial were only repeated once all unselected items were shown once. Once all 35 items were presented as "new" items, the unselected items were then redisplayed following the same pattern of 3 new items each trial. To minimize repetition, they were displayed in the same order that they were originally shown; for example, the first two items that weren't selected on trial 1 were the first to be displayed as "new" items again.

Each gameplay consisted of three consecutive rounds; round length varied based on performance and consisted of an average of 21.92 trials (SD = 9.18) for round 1, 14.98 (SD = 7.74) for round 2 and 13.96 (SD = 7.4316) for round 3. Together, the three rounds took an average of 3.12 minutes to complete. Consecutive gameplays were separated by a median of three days (SD = 7.26). We included participants who had completed at least 25 gameplays, resulting in 9,302,690 trials across all participants.

## **Analysis**

### **Exclusions**

Prior to analysis, we removed 0.74% of gameplays (0.44% of all trials), as these gameplays were from an older version of the game. In addition, to filter out excessively long response times, we excluded trials with reaction times greater than 30 s (0.61% of trials).

### **Error rates**

Due to the structure of the game, most trials had three potential correct responses (i.e., participants could select any one of the three new items to advance to the next trial). Rarely, participants would advance to a set size of 34 or 35, in which there would be only two new items or only one new item



remaining in the set, respectively. Errors could occur when, instead of clicking on one of the new items on each trial, participants clicked on one of the old, previously selected items. The number of potential incorrect responses therefore increased with set size, varying from zero (on the first trial, in which they simply select an item) to 34 (when the set size reached 35).

Participants could make an error by selecting one of these old items, but errors can occur for different reasons. In some cases, an error could occur because participants forgot items that were held in memory (i.e., participants believed an item was new, when it was in fact old). In other cases, errors might occur when participants falsely remember a new item as being in the memory set (due to poor “rejectability”), resulting in them being more likely to click on an actual old item.

We separately analyzed these sources of error for each of the 35 items in a given round to calculate both their respective hittability and rejectability. To calculate error rates based on item hittability, we calculated the number of trials on which the item was incorrectly chosen as “new”, divided by the number of times it appeared in the memory set on these trials (this normalization is meant to account for the fact that, due to the nature of the game, some items appeared more frequently than others). In a signal detection framework, this is analogous to the miss rate (relative to the memory set) for the item; however, as there is more than one way to make a correct (or incorrect) response in the game, it is not possible to calculate miss rates (or other signal detection measures) directly. This provides a measure of item hittability; the items with largest proportion of these mistakes have the poorest hittability, and conversely, those with the fewest have the highest hittability.

In addition, to examine how newly presented objects affect memory errors, we calculated each object’s error rate based on its “rejectability” – the likelihood that a new object would be falsely identified as old. For each item, we calculated its error rate based on its rejectability, or the likelihood of the item being mistakenly identified as old when it was, in fact, new. To measure these types of errors, we counted the number of trials in which each item appeared as a new item when an incorrect response was made. To account for the uneven distribution of

items due to the nature of the game, we normalized this by the total number of times the item appeared as new across all trials. If some objects are more difficult to reject than others, we would expect higher error rates when they are presented as new, leading to incorrect responses. In other words, this reflects the probability of making a mistake when identifying a new item, similar to its false alarm rate in relation to the memory set, providing a measure of “rejectability.” Objects frequently appearing as “new” during error trials had poor rejectability, meaning participants were hesitant to label them as new, even though they were not part of the memory set, resulting in errors. Therefore, items that rarely appeared as “new” in error trials were highly rejectable.

To summarize these error calculations, as shown in [Figure 2A](#) and [C](#), hittability is a property of old items, while rejectability is a property of new items; these are measured from the conditional error rates for old and new items, respectively. Importantly, any given item could be a “new” or “old” item depending on the trial and gameplay. For both analyses, error rates were calculated separately for each item and round (i.e., if an item was used in both round 1 and round 2, we calculated separate hittability and rejectability error rates for that item for each round). The three rounds of the game had different error rates, due to differences in the difficulty of the game (see Procedure). Therefore, to account for these differences in error rates between the three rounds and analyze each item’s relative rejectability or hittability relative to the items in that round, we calculated a z-scored error rate for each item, relative to all of the items in that round (see [Figure S3](#) for the raw error rates).

### *Item-based analysis of learning*

We quantified changes in learning rate for individual items by calculating hittability- and rejectability-based error rates for each of the 25 gameplays across participants and fitted an exponential decay function to each set of error rates. We used a least-squares fitting procedure to fit a three-parameter function of the form:

$$Y = (C - A) \times e^{-\lambda \times (x-1)} + A \quad (1)$$

where  $Y$  represents the error rate on gameplay  $x$ ,  $C$  represents the intercept on the first gameplay,  $\lambda$

corresponds to the decay rate, and  $A$  represents the asymptotic error rate.

### Reaction time

To analyze the impact of rejectability and hittability on reaction time, we measured, for each item, the impact of its presence or absence on reaction time for a given trial. To do this, we calculated, for each item, the median reaction time when the item was present as a new item and the median reaction time when the item was absent as a new item, separately for each unique combination of set size (i.e., the number of items present in the screen), and accuracy (whether the response on that trial was correct or incorrect). We then subtracted the median reaction times on the item-absent trials from the reaction times on the item-present trials. To account for differences in reaction times between different set sizes and accuracy levels (correct or incorrect trials), the resulting values were then z-scored within each set size and averaged. We correlated the resulting z-scores with the conditional error rates for new items. This analysis was repeated for the “old” items on each trial; we took the reaction time when the item was present as an old item minus the median reaction time when the item was absent as an old item, and we correlated the resulting z-scores with the conditional error rates for old items.

## Results

### Error rates

The task was designed such that as participants progressed through the rounds, the game becomes more difficult by reducing the number of categories and increasing the number of exemplars per category. This effect can be seen in the average conditional error rates for old items, which increased monotonically across the three rounds (mean for rounds 1-3: 0.066, 0.097, 0.106; SD: 0.018, 0.019, 0.022). The average conditional error rates for new items also increased similarly across rounds (M: 0.057, 0.088, 0.096; SD: 0.007, 0.012, 0.015).

Figure 2C and 2D show the distribution of conditional error rates for individual items shown in the first round (see Figure S3 in the supplemental materials for data from all rounds with labels), colour coded based on their conditional error rate

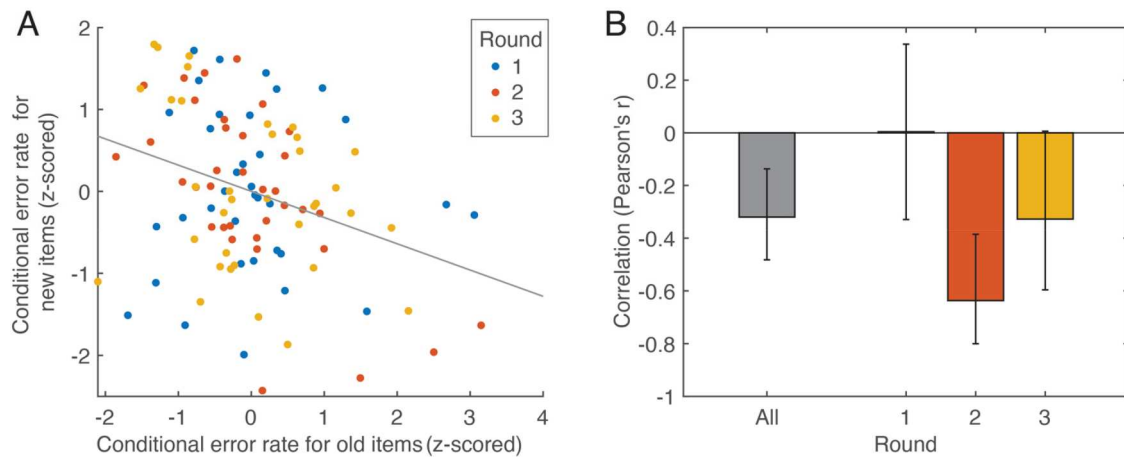
when “old”. While individual items vary in their respective error rates, the items most prone to false memories (i.e., poor rejectability) were not the same as the most forgettable items (i.e., poor hittability). To quantify the relationship between these measures, we first z-scored the error rates for each item respect to the other items in the round to account for the differences in error rates between rounds. As shown in Figure 3a, across all three rounds of the game, correlating the z-scored conditional error rates for old vs. new items between matched pairs of items showed a negative relationship between the two measures ( $r(103) = -0.32$ ,  $p = 0.0009$ ). In other words, items that were more forgettable (i.e., prone to misses) were not also more prone to false alarms; they were in fact somewhat *less* prone to them. As shown in Figure 3b, separately correlating the items within each round revealed that this relationship was largely driven by the items in round 2 ( $r(33) = -0.64$ ,  $p < .001$ ), with a weaker association for the other two rounds ( $r(33) < -0.33$ ,  $p > .05$ ).

### Consistency across age, practice, performance

Conditional error rates for old vs. new items were also well-correlated between participants, and stable between different age groups, levels of performance, and experience with the game. As shown in Figure 4, conditional error rates were highly consistent between the oldest (ages 68 - 80; top 20th percentile) and youngest (ages 21-49; bottom 20th percentile) participants ( $r(103) = 0.80$ ,  $p < .001$  for old items,  $r(103) = 0.91$ ,  $p < .001$  for new items).

These item-based effects were also resistant to practice across the 25 game sessions. Although participants’ performance improved over the course of repeated practice (measured as an improvement in run length; see Figure S4), items that were highly memorable (or rejectable) in the first five sessions were also memorable (or rejectable) in the last five sessions (Figure 5:  $r(103) = 0.66$ ,  $p < .001$  for old items,  $r(103) = 0.85$ ,  $p < .001$  for new items). Consistent with this, conditional error rates were also stable between high-performing and lower-performing participants (top 20% vs bottom 20%; Figure 6:  $r(103) = 0.49$ ,  $p < .001$  for old items,  $r(103) = 0.74$ ,  $p < .001$  for new items); similar correlations were also observed when comparing the top and bottom 20th percentiles of the younger age group ( $r(103) > 0.36$





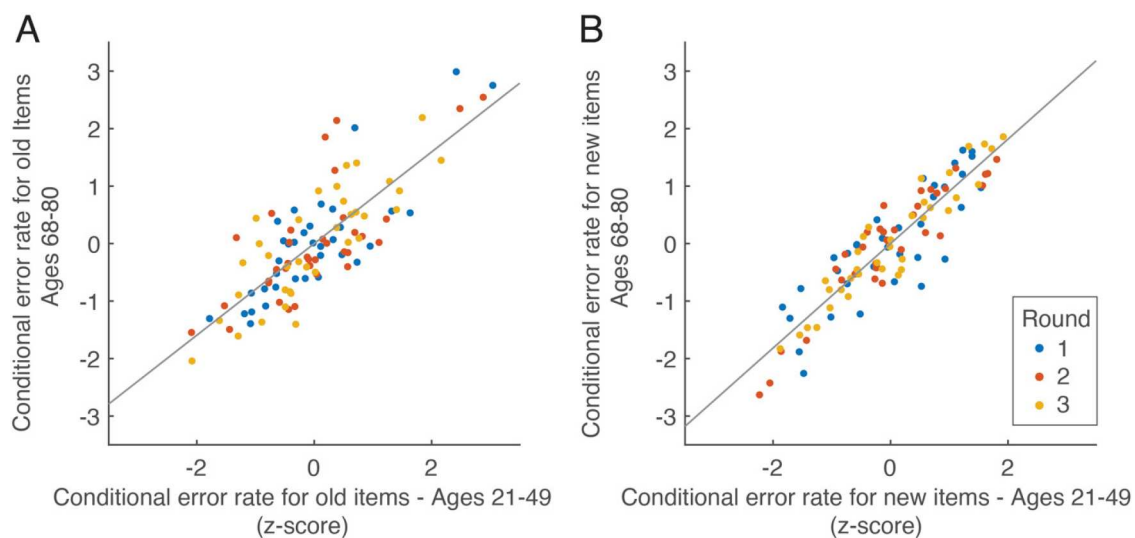
**Figure 3.** (A) Scatter plot showing the relationship between conditional error rates for old and new items averaged across all gameplays for all users. Each point represents each of the 105 items, and the colours represent the three rounds (blue, red, and yellow, for rounds 1, 2, and 3). (B) Correlation coefficient for conditional error rates for old vs. new items, calculated across all three rounds, and for each round separately. Error bars represent 95% confidence intervals.

$p < .001$ ), and the older age group ( $r(103) > 0.59$ ,  $p < .001$ ) separately.

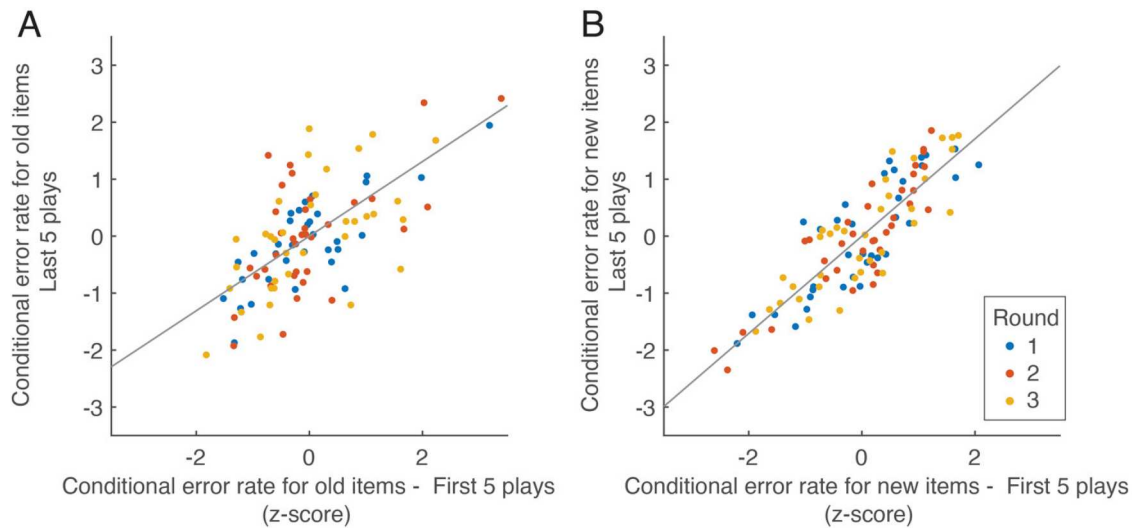
### Item-based analysis of learning

Although we observed consistent patterns of error for early gameplays compared to later gameplays, the rate at which error rates changed for particular items varied considerably. Comparing error rates between the first five and last five gameplays (Figure 5) suggested that conditional error rates for old items were less stable than those for new items.

We quantified this by calculating separate conditional error rates for old and new items for each of the 25 gameplays across participants and fitted an exponential decay function to each (see Figure S5 for individual fits). Figure 7A shows the conditional error rates for old items for each of the 105 items across the three rounds as a function of gameplay number, expressed as a proportion of the error rate on the first gameplay for that item (i.e., a value of 0.4 indicates that the error rate is equal to 40% of the rate observed on the first gameplay). Items varied considerably in the change in error rate; final error rates



**Figure 4.** (A) Scatter plot showing the relationship between z-scored error rates for old items averaged across the youngest 20th percentile of users plotted against the oldest 20th percentile of users. Each point represents the average z-scored error rate of one of 105 items when appearing as “old”, and the colours represent the three rounds. (B) Scatter plot showing the relationship between error rates for new items across age groups, following the same conventions as in (A)



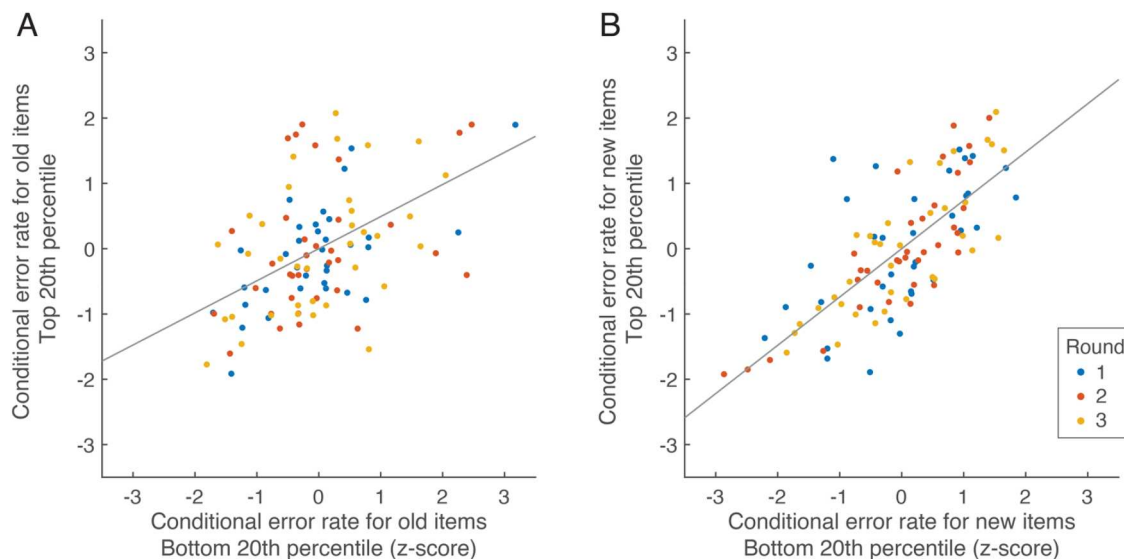
**Figure 5.** (A) Scatter plot showing the relationship between z-scored error rates for old items averaged across the first 5 gameplays and last 5 gameplays. Each point represents the average z-scored error rate of one of 105 items when appearing as “old”, and the colours represent the three rounds. (B) Scatter plot showing the relationship between error rates for new items across gameplays, following the same conventions as in (A)

ranged from 34% of the rate on the first gameplay, to an increase of 16% above the error rate on the first gameplay. Figure 7b shows the same analysis, based on conditional error rates for new items, where there was considerably less variation; final error rates ranged from 48% – 79% of the error rate on the first gameplay. As shown in Figure 7c, the percentage change in error rates for new items were not correlated with the percentage change in error rates for old items (z-scored within-round:  $r(103) = -0.15$ ,  $p =$

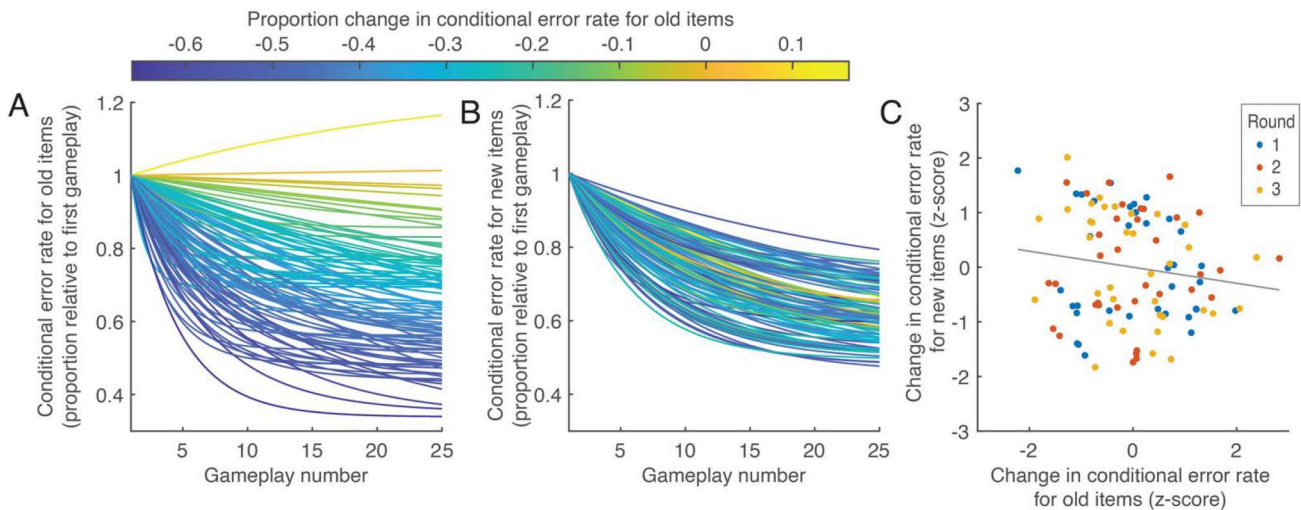
0.14). The rate of change parameter,  $\lambda$ , was also not correlated between conditional error rates for old vs. new items ( $r(103) = 0.06$ ,  $p = 0.52$ ).

#### Effects of rejectability and hittability on reaction time

When searching for new items, the tendency to mistake a new item for an old one should elevate RTs; in these instances, participants failed to



**Figure 6.** (A) Scatter plot showing the relationship between z-scored error rates for old items averaged across the bottom 20th percentile of user by performance (average run length) plotted against the upper 20th percentile of users. Each point represents the average z-scored conditional error rate of one of 105 items when appearing as “old”, averaged across all games, and the colours represent the three rounds. (B) Scatter plot showing the relationship between error rates for new items across performance, following the same conventions as in (A)



**Figure 7.** (A) Conditional error rates for old items as a function of gameplay number, expressed as a proportion of the error on the first gameplay. Lines represent the best-fitting exponential decay function for each of the 105 items across all three rounds, and are coloured based on the change in the error rate for old items. (B) Conditional error rates for new items across gameplays, following the same conventions and colours as (A). (C) Relationship between the change in conditional error rates for old vs. new items. Points represent individual items and rounds 1, 2, and 3 are represented by blue, red, and yellow dots, respectively.

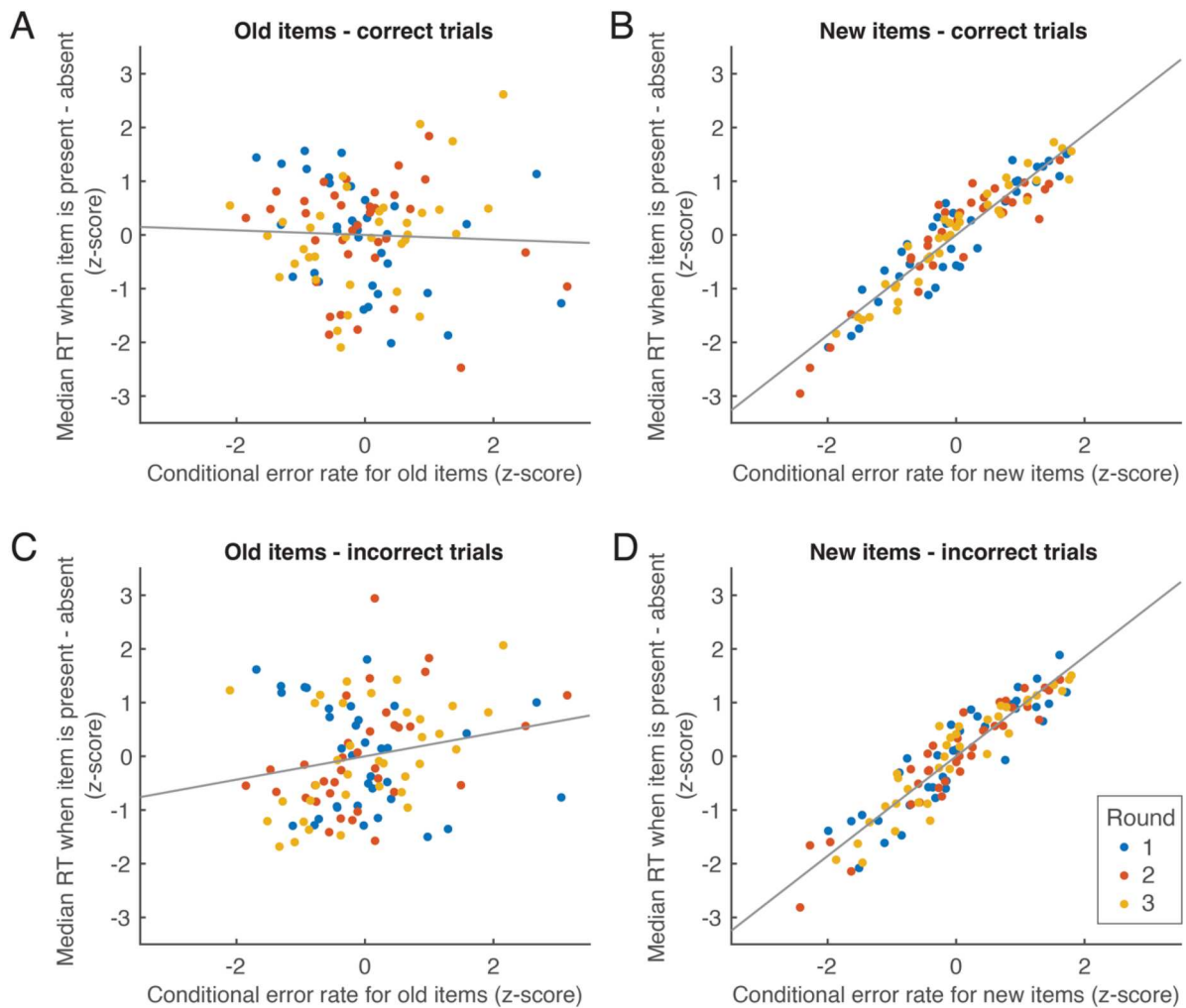
correctly identify one or more of the search targets and therefore should spend longer searching. Conversely, the tendency to forget that an item is in the memory set might shorten RTs if participants mistake an old item for a new one (i.e., they mistake a distractor for a search target). Therefore, when searching for new items, the presence of an item with poor rejectability within the “new” set should increase reaction times, while the presence of an item with poor hittability in the “old” set should decrease reaction times. We tested these predictions by calculating the impact of the presence of an item on reaction time (e.g., median reaction time when item present as a new item – median reaction time when item is not present as a new item). The resulting values were z-scored separately for each set size and separately for correct and incorrect trials. Figure 8 shows these results separated by accuracy and presence of the item in the “old” vs the “new” set.

As shown in Figure 8B and D, the presence of items with poor rejectability in the “new” set (the presence of items with high conditional error rates for new items) was associated with increased reaction time when the item was present, for both correct and incorrect trials ( $r(103) = 0.93$ ,  $p < .001$  and  $r(103) = 0.93$ ,  $p < .001$ ). However, as shown in Figure 8A and C, the presence of items with poor hittability in the “old” set (i.e., high conditional error rates for old items) was not associated with decreased reaction

times correct trials ( $r(103) = -0.04$ ,  $p = 0.67$ ), and was associated with a slight increase in reaction time on incorrect trials ( $r(103) = 0.22$ ,  $p = 0.026$ ).

## Discussion

Previous work has shown that some items are more easily remembered than others (e.g., Bainbridge et al., 2013; Bylinskii et al., 2015) and that this may capture distinct aspects of memory performance – the ability to correctly identify an item as being in the memory set, and the ability to identify an item as *not* being in the memory set (Zhao et al., 2023). In this study, we evaluated these properties, hittability and rejectability, respectively, in the context of a hybrid search game. Our findings suggest a nuanced relationship between the likelihood of items being falsely identified as “new” or “old”. Notably, as observed in the error rates for individual items, items prone to false memories (poor rejectability) were not the same ones that participants forgot were in the memory set (poor hittability). This was also observed in the effects of learning on participants’ errors; the rate of change in participants’ errors across gameplays was distinct for each of these two error types. Furthermore, we observed distinct effects of the hittability and rejectability of each item on participants’ reaction times. The results of this study contribute to the broader discourse on memory and learning by demonstrating that hittability and



**Figure 8.** (A) Relationship between the conditional error rate for individual items when “old” and the impact of the presence of the corresponding item on RTs. Positive values on the y-axis indicate slower RTs on trials when the item is present as an “old” item compared to trials when it is absent. RTs are z-scored within each set size, and shown for correct trials only. Points indicate individual items, and colours represent the three rounds. (B) Relationship between the conditional error rate for new items and the impact of item presence on RTs (correct trials only), following the same conventions as (A). (C and D). Same as A & B, showing the relationship for trials in which the participant responded incorrectly.

rejectability are not simply two ends of the same spectrum but may reflect distinct cognitive processes.

Importantly, error rates based on the hittability and rejectability of items were highly consistent between individuals, and were well-correlated across age groups, training and practice. The consistency of hittability and rejectability across different age groups adds an interesting layer to our understanding of cognitive aging. Despite a decrease in performance in memory-based tasks with age (Brockmole & Logie, 2013; Kirova et al., 2015), both hittability and rejectability scores remained highly correlated across our younger and older observers. While high-level factors such as typicality (e.g., Kramer et al., 2023; Vokey & Read, 1992), distinctiveness (Nosofsky & Osth,

2024), and meaningfulness (e.g., Brady & Störmer, 2022; Shoval et al., 2023a; Shoval et al., 2023b) can predict memorability for specific items, these may not change substantially with age to produce substantial differences between younger and older adults, for either the conditional error rates for old or new items. In addition, hittability and rejectability were consistent across multiple sessions, as well as between high – and low-performing observers, suggesting that these may be highly stable properties of objects within the context that they were tested.

Despite this degree of consistency, we observed variation in the learning rates for different items. Specifically, conditional error rates declined faster for some items compared to others, and this variation

was larger for the error rates for old items compared to new items. Importantly, the changes in these two types of error rates were not correlated with each other, pointing to separate mechanisms underlying learning-based improvements in memory performance. Understanding the sources of these different learning rates and what results in these differential changes in error rates would be an important avenue for future work.

Furthermore, in addition to demonstrating that hittability and rejectability can be separable, we show that these properties have distinct effects on participants' reaction time. Notably, items with poor rejectability produce a false memory effect, in that participants mistakenly identify them as one of the items in the memory set. This failure to recognize them as newly presented objects results in reliable increases in reaction time (whether participants are correct or incorrect), as they fail to identify their search target. In contrast, the presence of items with poor hittability did not affect reaction times on trials when participants responded correctly, and slightly increased reaction times when participants responded incorrectly. These results contrast with previous work indicating that participants are remarkably good at rejecting familiar "lures" in a hybrid search task, which minimally impact participants' reaction time (Wolfe et al., 2015). It seems that while familiarity may not strongly impact participants' performance in the search for new (or old) items, participants' performance may be affected for certain classes of targets that are prone to being mistaken for distractors.

Despite evidence that hittability and rejectability are separable, we did observe a negative correlation between rejectability and conditional error rates for old items. This would suggest selection biases to choose certain objects and not others (i.e., a tendency to believe some items are "new" while others are "old", regardless of whether they are in the memory set or not). In signal detection terms, this would translate into variations in the decision criterion ( $c$ ) for each item – i.e., willingness to say that an item is in the memory set. Interestingly, this relationship seems to depend on the context, such that the negative association was largely driven by round 2 and to some extent round 3, where items were more similar to each other. With a diverse set of objects, as shown in round 1, there was no association between these error rates. Global matching models like Retrieving Effectively from

Memory (REM; Shiffrin & Steyvers, 1997) provide a good starting point to explain why these relationships might be dependent to the homogeneity of the image set. This class of models predict that as memory sets become more homogenous (i.e., similar to rounds 2 or 3), they elicit stronger familiarity, which could increase the likelihood of participants reporting items as "old". This highlights a complex and potentially relevant interaction between the homogeneity of the dataset and the response tendencies of participants (Osth & Dennis, 2015).

One important consideration is that, due to the nature of the game, the items were all from a narrow set of categories (even for beach 1, which had the largest range of objects). While the data consists of items from a narrow set of categories, the hittability and rejectability estimates for each item were based on many thousands of observations, resulting in a high degree of precision for any individual item. In contrast to this approach, previous work (Zhao et al., 2023) used fewer participants, but a wider range of objects to provide similar evidence that hits and correction rejections are supported by different mechanisms. Together, these results provide converging evidence for the separability of these two processes. Nevertheless, further work would be needed to establish the degree to which our results generalize to larger and more diverse sets of objects.

Together, these findings underscore the complexity of memorability and highlight the need for a nuanced approach in its study. Throughout this paper, we have shown the distinct impact rejectability has on reaction time, as well as the distinct influence of practice on hittability compared to rejectability. These results highlight the importance of separating different classes of errors to understand memory-based performance in a range of tasks that extend beyond single-item recognition.

## Acknowledgments

The authors thank Lumos Labs for providing the dataset, and Keisuke Fukuda for helpful discussions.

## Data availability statement

The data that support the findings of this study were provided by Lumos Labs. Restrictions apply to the availability of these data, which were used under license for this study. Data are available to researchers with the permission of Lumos Labs



via the Human Cognition Project (<https://www.lumoslabs.com/hcp-apply>).

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

This work was supported by an NSERC Undergraduate Student Research Award to DS and National Eye Institute: [grant number EY017001]; National Science Foundation: [grant number 2146617]

## ORCID

Jeremy M. Wolfe  <http://orcid.org/0000-0002-6475-1984>  
Anna Kosovicheva  <http://orcid.org/0000-0002-5219-3006>

## References

- Bainbridge, W. A., Isola, P., & Oliva, A. (2013). The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, 142(4), 1323–1334. <https://doi.org/10.1037/a0033872>
- Bainbridge, W. A., & Rissman, J. (2018). Dissociating neural markers of stimulus memorability and subjective recognition during episodic retrieval. *Scientific Reports*, 8(1), 1–11. <https://doi.org/10.1038/s41598-018-26467-5>
- Brady, T. F., & Störmer, V. S. (2022). The role of meaning in visual working memory: Real-world objects, but not simple features, benefit from deeper processing. *Journal of Experimental Psychology: Learning Memory and Cognition*, 48(7), 942–958. <https://doi.org/10.1037/xlm0001014>
- Brockmole, J. R., & Logie, R. H. (2013). Age-related change in visual working memory: A study of 55,753 participants aged 8–75. *Frontiers in Psychology*, 4(JAN), 1–5. <https://doi.org/10.3389/fpsyg.2013.00012>
- Bylinskii, Z., Isola, P., Bainbridge, C., Torralba, A., & Oliva, A. (2015). Intrinsic and extrinsic effects on image memorability. *Vision Research*, 116, 165–178. <https://doi.org/10.1016/j.visres.2015.03.005>
- Drew, T., Boettcher, S. E. P., & Wolfe, J. M. (2017). One visual search, many memory searches: An eye-tracking investigation of hybrid search. *Journal of Vision*, 17(11), 1–10. <https://doi.org/10.1167/17.11.5>
- Gronau, N., Nartker, M., Yakim, S., Utochkin, I., & Wolfe, J. (2024). Categorically distinct subsets allow flexible memory selection in hybrid search. *Journal of Experimental Psychology: Learning Memory and Cognition*, 15, 1–45. <https://doi.org/10.1037/xlm0001377>
- Isola, P., Parikh, D., Torralba, A., & Oliva, A. (2011). *Understanding the intrinsic memorability of images*. Advances in neural information processing systems 24: 25th annual conference on neural information processing systems 2011, NIPS 2011, 1–9. <https://doi.org/10.1167/12.9.1082>
- Isola, P., Xiao, J., Parikh, D., Torralba, A., & Oliva, A. (2014). What makes a photograph memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7), 1469–1482. <https://doi.org/10.1109/TPAMI.2013.200>
- Khosla, A., Raju, A. S., Torralba, A., & Oliva, A. (2015). Understanding and predicting image memorability at a large scale. *Proceedings of the IEEE International Conference on Computer Vision, 2015 Inter*, 2390–2398. <https://doi.org/10.1109/ICCV.2015.275>
- Kirova, A. M., Bays, R. B., & Lagalwar, S. (2015). Working memory and executive function decline across normal aging, mild cognitive impairment, and Alzheimer's disease. *BioMed Research International*, 2015, <https://doi.org/10.1155/2015/748212>
- Kramer, M. A., Hebart, M. N., Baker, C. I., & Bainbridge, W. A. (2023). The features underlying the memorability of objects. *Science Advances*, 9(17), 1–14. <https://doi.org/10.1126/SCIADV.ADD2981>
- Nosofsky, R., & Osth, A. (2024). Hybrid-Similarity exemplar model of context-dependent memorability. In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th annual conference of the cognitive science society* (pp. 996–1002). Cognitive Science Society.
- Osth, A. F., & Dennis, S. (2015). Sources of interference in item and associative recognition memory. *Psychological Review*, 122(2), 260–311. <https://doi.org/10.1037/a0038692>
- Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM-retrieving efficiently from memory. *Psychonomic Bulletin & Review*, 4(2), 145–166.
- Shoval, R., Gronau, N., & Makovski, T. (2023a). Massive visual long-term memory is largely dependent on meaning. *Psychonomic Bulletin and Review*, 30(2), 666–675. <https://doi.org/10.3758/s13423-022-02193-y>
- Shoval, R., Gronau, N., Sidi, Y., & Makovski, T. (2023b). Objects' perceived meaningfulness predicts both subjective memorability judgments and actual memory performance. *Visual Cognition*, 31(6), 472–484. <https://doi.org/10.1080/13506285.2023.2288433>
- Vokey, J. R., & Read, J. D. (1992). Familiarity, memorability, and the effect of typicality on the recognition of faces. *Memory & Cognition*, 20(3), 291–302. <https://doi.org/10.3758/BF03199666>
- Wolfe, J. M. (2012). Saved by a log: How do humans perform hybrid visual and memory search? *Psychological Science*, 23(7), 698–703. <https://doi.org/10.1177/0956797612443968>
- Wolfe, J. M., Aizenman, A. M., Boettcher, S. E. P., & Cain, M. S. (2016). Hybrid foraging search: Searching for multiple instances of multiple types of target. *Vision Research*, 119, 50–59. <https://doi.org/10.1016/j.visres.2015.12.006>
- Wolfe, J. M., Boettcher, S. E. P., Josephs, E. L., Cunningham, C. A., & Drew, T. (2015). You look familiar, but I don't care: Lure rejection in familiarity. *Journal of Experimental Psychology: Human Perception and Performance*, 41(6), 1576–1587. <https://doi.org/10.1037/xhp0000096>
- Zhao, C., Fukuda, K., & Woodman, G. F. (2023). Target recognition and lure rejection: Two sides of the same memorability coin? *Visual Cognition*, 31(9), 633–641. <https://doi.org/10.1080/13506285.2024.2315803>