2024-04-25

Everyday Amnesia: Residual Memory for High Confidence Misses and Implications for Decision Models of Recognition

Berry, CJ

10.1037/xge0001599
Journal of Experimental Psychology: General
American Psychological Association

Date: 26/03/2024. Please note: This is the accepted manuscript, not the final published version. For the final version, see the publisher's website (*Journal of Experimental Psychology: General*). DOI: http://doi.org/10.1037/xge0001599.

Everyday Amnesia: Residual Memory for High Confidence Misses and Implications for Decision Models of Recognition

Christopher J. Berry[1] and David R. Shanks[2]

[1]School of Psychology, University of Plymouth

[2]Department of Experimental Psychology, University College London

Author Note

Christopher J. Berry https://orcid.org/0000-0002-3512-3604

David R. Shanks https://orcid.org/0000-0002-4600-6323

We have no known conflicts of interest to disclose.

Correspondence concerning this article should be addressed to Christopher J. Berry, School of Psychology, University of Plymouth, Drake Circus, PL4 8AA, UK. Email: christopher.berry@plymouth.ac.uk.

The data, analysis scripts, and materials for all experiments are available at the Open Science Framework at https://osf.io/2q5yw/

This research was partially funded by grants ES/S014616/1 and ES/Y002482/1 from the United Kingdom Economic and Social Research Council to D. R. Shanks.

The findings were presented at the 2023 meeting of the Experimental Psychology Society, Plymouth, UK. We thank Henry Roediger III and Ian Dobbins for comments on an earlier version of the article. We also thank Acquina Akbar, Thomas Biggs, Esme Doe, William Drover-Taylor, Leyre Honrubia Arribas, Oliver Mortimer, Gayathri Parackaparambil Subramanian, Aaron Ponting, Charlie Richards, Andre Santa Barbara Escarigo, and Nethmi Vithana Weerasinghe Arachchilage for their contribution to data collection.

Word count: 14,861

Abstract

Despite studying a list of items only minutes earlier, when reencountered in a recognition memory test, undergraduate participants often say with total confidence that they have not studied some of the items before. Such *high confidence miss* (HCM) responses have been taken as evidence of rapid and complete forgetting and of *everyday amnesia* (Roediger & Tekin, 2020). We investigated 1) if memory for HCMs is completely lost or whether a residual memory effect exists, and 2) whether dominant decision models predict the effect. Participants studied faces (Experiments 1a, 2, 3) or words (Experiment 1b), then completed a single-item recognition memory task, followed by either 1) a two-alternative forced-choice recognition task, in which the studied and non-studied alternatives on each trial were matched for their previous old/new decision and confidence rating (Experiments 1-2), or 2) a second single-item recognition task in which the targets and foils were HCMs and high confidence correct rejections (HCCRs), respectively (Experiment 3). In each experiment, participants reliably distinguished HCMs from HCCRs. The unequal variance signal detection (UVSD) and dual-process signal detection (DPSD) models were fit to the single-item recognition data, and the parameter estimates used to predict the memory effect for HCMs. The DPSD model predicted the residual memory effect (as did another popular model, the mixture-SDT model). However, the UVSD model incorrectly predicted a negative, or no, effect, invalidating this model. The residual memory effect for HCMs demonstrates that everyday amnesia is not associated with complete memory loss and distinguishes between decision models.

Keywords: Everyday amnesia; forgetting; UVSD model; DPSD model; signal detection theory; recognition memory.

Public Significance Statement

Participants in our experiments appeared to completely forget studying particular items (pictures of faces or words) over short intervals in tests of recognition memory—thereby showing everyday amnesia. However, memory for such items was evident in a follow-up memory test. This suggests that the memory loss that occurs in everyday amnesia is not complete and also has implications for formal decision models of recognition.

Everyday Amnesia: Residual Memory for High Confidence Misses and Implications for Decision Models of Recognition

In tests of recognition memory, participants will often claim with complete confidence that they did not study some of the items before, despite studying them only minutes earlier. As an example, in Experiment 2 of Tekin and Roediger (2017), undergraduate students first studied a list of 50 faces for 2s each. After a 10-minute retention interval they were shown the same faces intermixed with 50 new faces and asked to decide whether each one was previously studied or not by responding "old" or "new". Participants indicated how confident they were in each decision on a rating scale ranging from "not confident at all" to "totally confident". Although participants tended to correctly judge a greater proportion of actual old items to be "old" compared to new items—indicating that they had memory for old items—a substantial proportion received "new" decisions (i.e., were *misses*). Strikingly, 19.7% of these misses were made with total confidence. In a reanalysis of their study, Roediger and Tekin (2020) drew attention to the relatively high propensity with which undergraduate students made such *high confidence misses* (HCMs), and also reported a similar percentage (16%) in a reanalysis of other published studies (Tekin & Roediger, 2017, Experiment 1; DeSoto & Roediger, 2014).

From one perspective, these findings are relatively surprising: the participants were undergraduate students, presumably of sufficient capacity to learn and retain new information given their position in higher education, and yet were declaring with total confidence that they did not study something they had in fact studied only minutes earlier. Indeed, Roediger and Tekin (2020) referred to this as an example of *everyday amnesia* and took it to mean that "rapid and complete forgetting" (Roediger & Tekin, 2020, pp. 6, 8) of fully processed experiences can occur in all individuals, not only those with neurological disorders, for

example, as a result of damage to the hippocampus/medial temporal lobes (Squire, 1992; Squire et al., 2004), as is most commonly associated.

Signal Detection Theory and HCMs

An alternative view was proposed by Levi et al. (2022) and Goshen-Gottstein et al. (2022) from the perspective of signal detection theory (SDT), a widely used framework for conceptualising decision and memory signal components of recognition processes (Green & Swets, 1966; Hautus et al., 2022; see Wixted, 2020, for a historical review). In its standard form, SDT assumes that each item in a recognition task is associated with a continuous memory strength of evidence variable, most commonly assumed to be Gaussian, the mean of which is greater for studied items, owing to their presentation in the study phase. In order to decide whether a test item was studied or not, participants compare its strength against a decision criterion, C. If the strength exceeds the criterion, the item is judged old, otherwise it is judged new. To model N confidence ratings, $N - 1$ decision criteria can be used. For example, with six ratings (i.e., where 1 = "sure new"…6 = "sure old"), there are five criteria, C_1-C_5. If the strength of an item exceeds C_5, it receives a "6" rating. If it falls between C_5 and C_4, it receives a "5" rating, representing a medium confidence old decision; if it falls between C_4 and C_3 it receives a "4" rating, representing a low confidence old decision; if it falls between C_3 and C_2 it receives a low confidence new "3" rating; if it falls between C_2 and C_1 it receives a medium confidence new "2" rating, and if the strength value falls below C_1, it receives a sure new "1" rating.

Levi et al. (2022) and Goshen-Gottstein et al. (2022) advocated a popular version of SDT as applied to recognition tasks, the unequal variance signal detection (UVSD) model. In this model, the variance of the old item strength distribution can take on a different value from that of the new item strength distribution. The UVSD model is motivated by the properties of the empirical receiver operating characteristic (ROC), which is a plot of the hit

rate against the false alarm rate at different levels of bias (also known as an isosensitivity curve). The slope of the *z*-transformed empirical ROC is typically less than 1, which is inconsistent with the expectation of a slope value equal to one if the old and new item strength distributions have equal variances. The UVSD model, in contrast, produces a *z*-ROC slope less than one when the variance of the old item strength distribution is greater than that of new items (Egan, 1958; see Wixted, 2007, Yonelinas & Parks, 2007, for reviews). Levi et al. (2022) and Goshen-Gottstein et al. (2022) pointed out that, in the UVSD model (and SDT more generally), a miss is simply an item with a strength value that does not exceed the old-new decision criterion (i.e., C, or C_3), and the miss will be made with the highest level of confidence if its strength value falls below the lowest decision criterion C_1. Thus, a proportion of HCMs are to be expected given the way that the decision process is represented in SDT, and in this sense their occurrence is "predicted", even trivial. Levi et al. (2022) and Goshen-Gottstein et al. (2022) also used the UVSD model to reproduce the proportions reported in Roediger and Tekin (2020) in two Monte Carlo simulations and fitted the model to the data of Tekin and Roediger (2017, Experiment 1) using maximum likelihood estimation.

Roediger and Tekin (2022) and Dobbins (2022) in turn questioned the usefulness of the UVSD model as an explanation of HCMs and also its predictive value, pointing out that the model can reproduce any proportion of HCMs in an *ad hoc* manner by varying its parameters. To illustrate, we can express the proportion of HCMs that will occur in the UVSD model with the following equation:

$$P(\mathrm{HCM}) = \frac{\Phi\left(\frac{C_1 - d}{\sigma_\mathrm{o}}\right)}{\Phi\left(\frac{C_3 - d}{\sigma_\mathrm{o}}\right)} \tag{1}$$

where Φ is the cumulative distribution function of the standard normal distribution, C_1 is the criterion value (relative to the mean of the new item distribution, μ_n, which is fixed to $\mu_n = 0$ with no loss of generality), d is the mean difference in strength of old and new items, and σ_o is the standard deviation of the strength distribution of old items relative to that of new items, σ_n (which is fixed to $\sigma_n = 1$ to allow the other parameters to be identified, again with no loss of generality). To restate Equation 1 in words, the proportion of HCMs is the proportion of old items that fall below the C_1 criterion, divided by the proportion of old items judged new. The proportion of HCMs therefore depends on the values of C_1, C_3, d, and σ_o, and any proportion can potentially be reproduced exactly by varying these parameter values. For instance, all other parameter values being held constant, the proportion of HCMs 1) increases as the criterion C_1 takes on higher values (i.e., becomes more conservative), 2) decreases as d increases, and 3) interacts with C_1 and d as σ_o increases; for example, when d is relatively low and C_1 is relatively high, P(HCM) decreases as σ_o increases, but when d is relatively high, P(HCM) increases with σ_o. For comparison, the proportion of high confidence new judgments to new items (i.e., high confidence correct rejections, henceforth HCCRs), is

$$P(\text{HCCR}) = \frac{\Phi(C_1)}{\Phi(C_3)}. \tag{2}$$

Roediger and Tekin (2022), Dobbins (2022), and Roediger and Dobbins (2022), additionally questioned the theoretical informativeness of signal detection accounts of everyday amnesia, given their abstractness, favouring instead accounts that provide explanations in terms of psychological constructs, mechanisms, or neural processes. Roediger and Dobbins (2022) maintained that HCMs are worthy of further attention.

We agree that HCMs are worthy of further attention and, in this article, we investigate whether SDT models do in fact make any interesting predictions concerning them. Our

overall goal is to shed theoretical light on the important and novel concept of everyday amnesia, via two specific aims: First, we sought to establish whether memory for HCMs is completely lost, or whether some degree of "residual" memory for HCMs can be demonstrated. We did this using additional memory tasks in which participants must distinguish HCMs from HCCRs. If a residual memory effect for HCMs can be shown, this would inform the phenomenon of everyday amnesia in that it would have implications for the completeness of forgetting that can be said to have occurred. Second, we sought to determine whether the UVSD model successfully predicts any residual memory effect that can be demonstrated for HCMs.

Our approach was to first fit the model to the single-item recognition memory data and then use the parameter estimates to derive *ex ante* predictions in the additional memory tasks. Predictions derived in this way can be considered true "predictions" of the model, as opposed to mere fits or reproductions of the data, since the parameters are estimated on the basis of the single-item recognition task data, and the data from the subsequent memory task can therefore have no bearing whatsoever on these parameter estimates (Busemeyer & Wang, 2000). A similar approach to testing decision models was recently taken by Ma et al. (2022) and Dobbins (2023). We also compared the predictions of the UVSD model to those of another dominant model in the literature, the dual-process signal detection (DPSD) model (Yonelinas, 1994), which is commonly pitted against the UVSD model (see e.g., Wixted, 2007; Parks & Yonelinas, 2007; Rotello, 2017, for reviews). Finally, in further modelling, we broadened out the range of models by also considering the predictions of the two-high threshold (2HT) model (Bröder & Schütz, 2009; Egan, 1958; Moran, 2016; Snodgrass & Corwin, 1988), the mixture SDT model (DeCarlo, 2002), and versions of the UVSD model in which the distributions are not Gaussian.

Expected Strength of HCMs vs. HCCRs

The potential for the UVSD and DPSD models to make opposing predictions concerning the residual memory effects for HCMs can be identified by considering the expected strength values for HCMs and HCCRs in each model. Despite receiving the same recognition response, the expected strength value for HCMs can differ from that of HCCRs in both models. Interestingly, in the UVSD model, when σ_o is greater than σ_n, the expected strength of HCMs can be *lower* than that of HCCRs. This is because the old item strength curve intersects the new item curve at two points—the lower and upper tails of the distribution—meaning that the likelihoods of the lowest old item strengths are greater than those of new items with the same strength values. In other words, the ratio of the densities of new and old item strengths at each strength value does not vary monotonically along the strength axis. This feature is well-known and has been discussed by others (e.g., Green & Swets, 1966; see also DeCarlo, 2002; Dube, 2023; Glanzer et al., 2009; Kellen et al., 2021; Stretch & Wixted, 1998; Yonelinas & Parks, 2007). The characteristic is clearly counterintuitive from a psychological perspective since it means that, as a result of study, some old items will end up with a lower strength than the lowest strength of all new items. Despite this, it has not prevented widespread adoption of the model by many in the literature, presumably because of the model's successes in accounting for other aspects of recognition data.

To demonstrate the conditions under which the expected strength of HCMs will be lower than that of HCCRs in the UVSD model, we can state the expected strength (*S*) of a HCM as:

$$E(S|\,\mathrm{HCM}) = \; d - \sigma_{\mathrm{o}}\left(\frac{\phi\left(\frac{(C_1 - d)}{\sigma_{\mathrm{o}}}\right)}{\Phi\left(\frac{(C_1 - d)}{\sigma_{\mathrm{o}}}\right)}\right) \tag{3}$$

where ϕ is the normal density function. The expected strength of a HCCR is

$$E(S|\ \text{HCCR}) = -\left(\frac{\phi(C_1)}{\Phi(C_1)}\right). \tag{4}$$

The difference in expected strength of HCMs and HCCRs as the σ_o and d parameters are varied in Equations 3 and 4 is shown in Figure 1. It is evident that the expected strength of HCMs is more likely to be lower than that of HCCMs as σ_o increases, d is relatively low, and C_1 is relatively low (i.e., when the propensity to make high confidence new decisions is lower). Interestingly, HCM strength can be lower than HCCR strength even with relatively "standard" values of σ_o and d. For example, in the σ_o =1.25 panel of Figure 1, where the ratio of lure to target variance is 1/1.25 = 0.8 (Ratcliff et al., 1992), the difference in expected strength to HCMs and HCCRs is negative or close to zero when d is between 0-1.5 and C_1 is relatively low.

A visual representation of the UVSD model when fit to the data from Experiment 1 of Tekin and Roediger (2017) is shown in the top left panel of Figure 2 (where the MLE estimates are d = 1.08, σ_o = 1.29, C_1 = -1.04, C_2 = -0.39, C_3 = 0.38, C_4 = 0.96, C_5 = 1.48). With these estimates, the proportion of high confidence new responses to old items in the model is lower than that of new items (0.05 vs. 0.15), but, using Equations 3 and 4, the expected strength of HCMs is simultaneously lower than that of HCCRs (-1.58 vs. -1.56) (Figure 2, lower left panel). That is, despite being studied, a HCM is expected to have a lower strength than a HCCR in Experiment 1 of Tekin and Roediger (2017). Although the difference is small in this example, it can be greater (or indeed reverse), depending on the estimates of d, σ_o, and C_1, as shown in Figure 1.

Figure 1

Expected difference in strength of HCMs and HCCRs in the UVSD model. The solid lines indicate d = 0, 0.5, 1, 1.5, 2, and 2.5 (dark to light). The horizontal dashed line indicates zero difference.

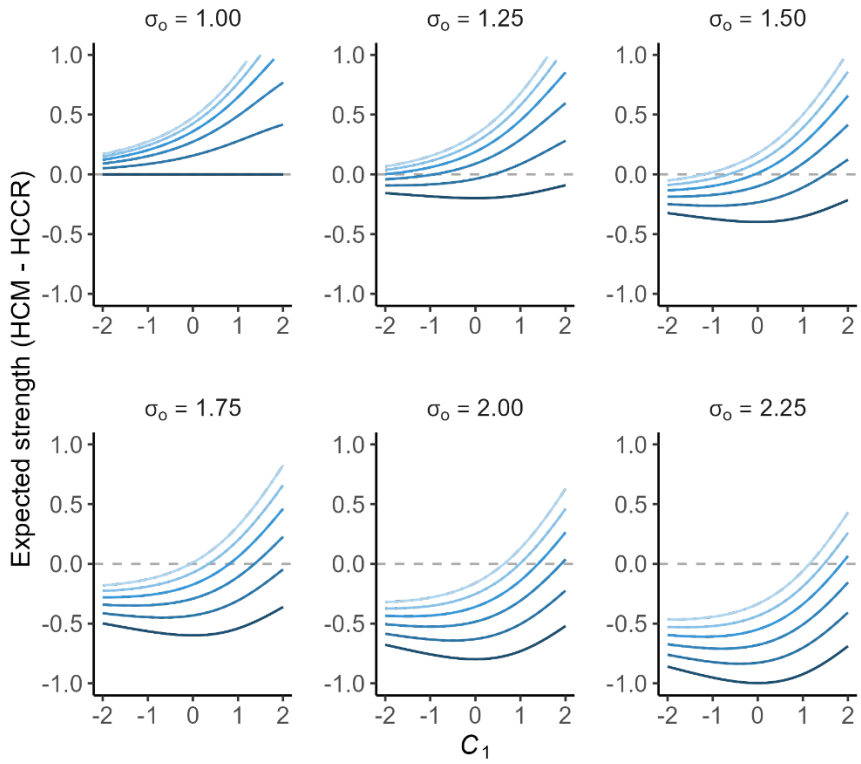
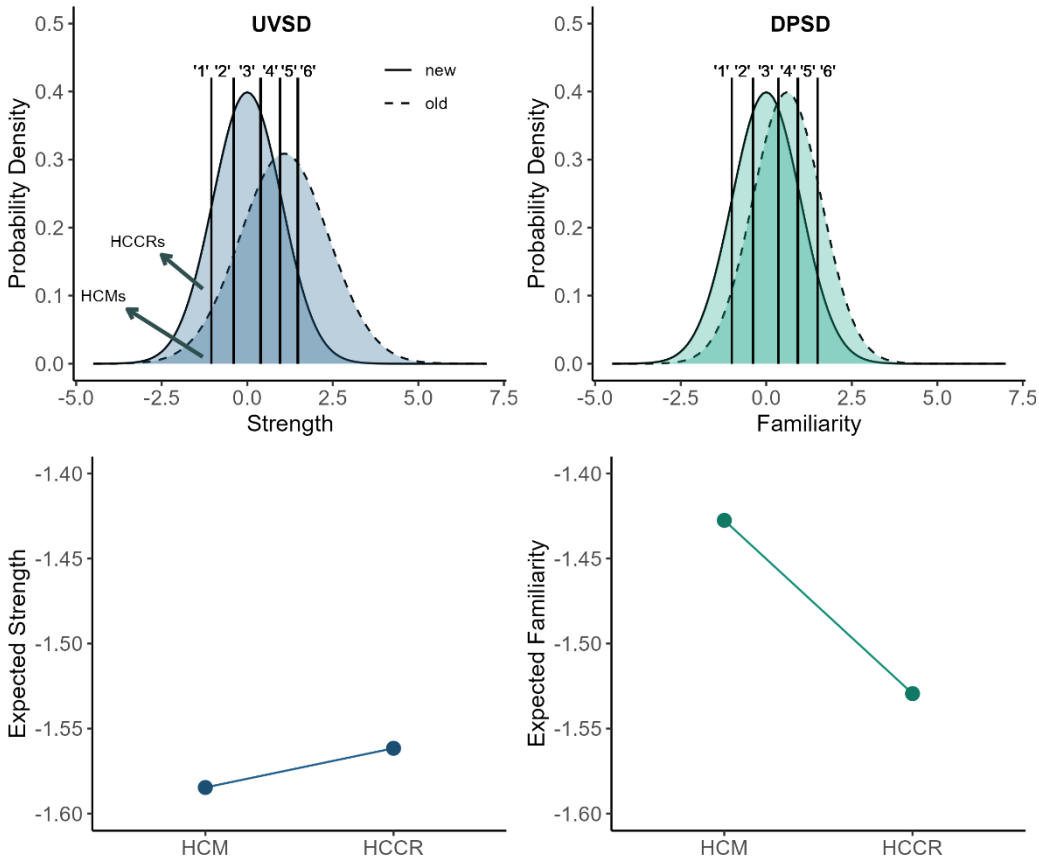


Figure 2

Visual representation of the UVSD and DPSD models when fit to Experiment 1 of Tekin and Roediger (2017) (top row), and the expected strengths of HCMs and HCCRs, given these fits (bottom row).



For comparison, consider the dual-process signal detection (DPSD) model (Yonelinas, 1994), according to which, recognition decisions are made on the basis of two distinct processes, recollection and familiarity. Recollection involves the retrieval of qualitative information associated with an item's previous presentation (e.g., where or when it was encountered), whereas familiarity is not associated with the retrieval of such contextual information and instead is strength-based, being modelled as an equal variance signal-detection process. If an old item is recollected, with probability R_o, it receives an "old" decision with the highest level of confidence; if it is not recollected, then the decision is

based on familiarity. HCMs are therefore described by an equal-variance SDT process, and their expected strength is given as:

$$E(S|\,\mathrm{HCM}) = d' - \left(\frac{\phi(C_1 - d')}{\Phi(C_1 - d')}\right) \tag{5}$$

HCCRs are also described by the same process, and the equation for their expected strength is the same as that of the UVSD model (Equation 4). Note that d is used to refer to the mean difference in strength of old and new items in the UVSD model while d' refers to the equivalent difference, in units of $\sigma = \sigma_o = \sigma_n$, in the DPSD model.

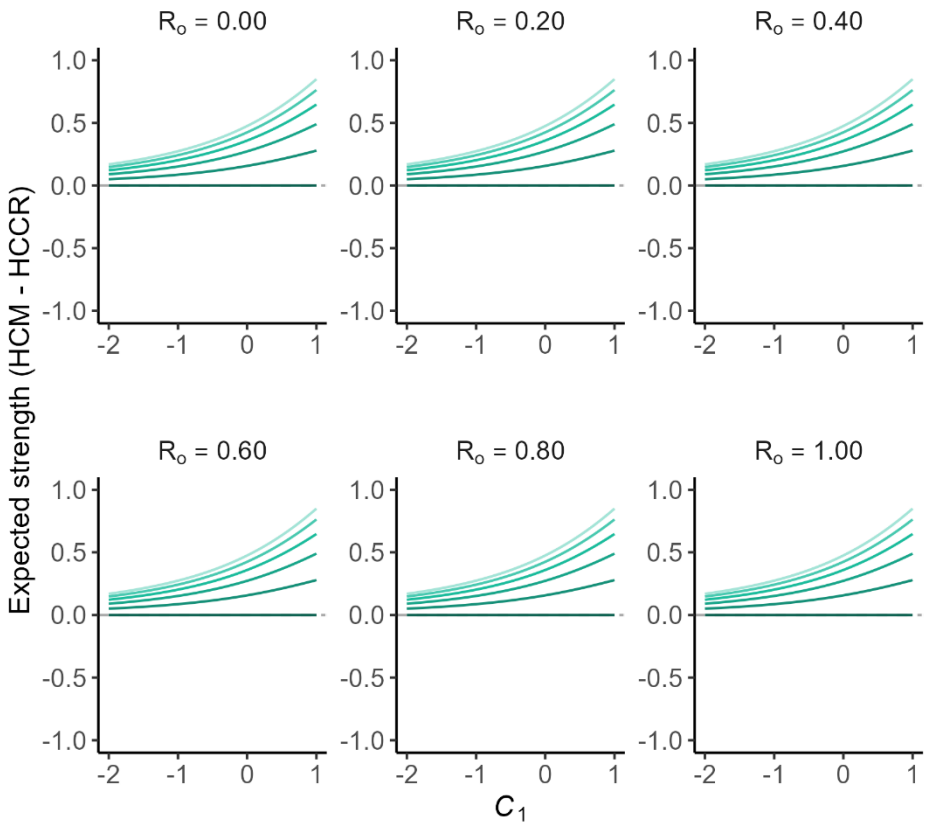
As shown in Figure 3, the expected strength of HCMs is greater than that of HCCRs in the DPSD model as d' and C_1 increases, and the difference is independent of R_o. A visual representation of the DPSD model when fit to Experiment 1 of Tekin and Roediger (2017) is shown in the top right panel of Figure 2 (where the MLE estimates are $d' = 0.61$, $R_o = 0.24$, $C_1 = -1.01$, $C_2 = -0.38$, $C_3 = 0.36$, $C_4 = 0.92$, $C_5 = 1.50$). Using these estimates, the probability of a high confidence new decision to an old item is lower than that of a high confidence new decision to a new item (0.04 vs. 0.16) and the expected strength of HCMs is greater than that of HCCRs (-1.42 vs. -1.53). Thus, HCMs are more familiar than HCCRs in the DPSD model.

In sum, the expected strength of a HCM can be *lower* than that of a HCCR in the UVSD model, but not in the DPSD model. Assuming that participants can be sensitive to differences in the strength of HCMs and HCCRs, and this translates to the capacity to discriminate between such items when presented in an additional memory test, there is a strong possibility that the models will make opposing predictions concerning participants' ability to positively discriminate HCMs from HCCRs once their parameters have first been fixed by fitting them to the recognition data. This constitutes a test of strong inference (Platt, 1964). Specifically, given suitable fits to the recognition data, the UVSD model predicts that

participants will either be unable to discriminate between HCMs and HCCRs, or even that participants will respond as if HCCRs are associated with greater strength than HCMs (i.e., a negative residual memory effect for HCMs). In contrast, the DPSD model predicts that HCMs and HCCRs can be discriminated, that is, there will be a residual memory effect for HCMs.

Figure 3

Expected difference in strength of HCMs and HCCRs in the DPSD model. The solid lines indicate d' = 0, 0.5, 1, 1.5, 2, and 2.5 (dark to light). The horizontal dashed line indicates zero difference.



Measuring residual memory for HCMs

To measure residual memory for HCMs, we gave participants one of two additional tasks after the standard single-item recognition task: a modified two-alternative forced choice (2AFC) task, or a second single-item recognition task in which the targets and foils were HCM and HCCR items and participants were instructed to decide whether each item was previously studied or not (Lee & Shanks, 2023, recently adopted a similar approach in the context of implicit learning). In the modified 2AFC task, studied and nonstudied alternatives were presented on each trial that were matched in terms of the previous single-item recognition decision and confidence rating they had received, and participants had to decide which one was presented in the study phase. The key 2AFC trials are those where a HCM is paired with a HCCM. From the participant's perspective, even though they previously indicated with total confidence that both items are new, they still have to select the one they think was studied. If participants reliably select the HCM on such trials (i.e., percentage of correct decisions is greater than 50%), then this can only be due to the presentation of the HCM but not the HCCR in the study phase and would therefore demonstrate that participants do in fact have some residual memory for HCMs.

Re-presenting items from the single-item recognition phase in 2AFC trials is similar to the *error correction* paradigm devised by Starns and colleagues (e.g., Ma et al., 2022; Starns et al., 2018), in which, after 12 trials of a recognition task where participants make old-new decisions to studied and nonstudied items, error items (i.e., misses or false alarms) are re-presented with ones for which responses were correct (correct rejections or hits) in two-alternative forced choice trials, and participants are required to select which of the alternatives they think was from the study list. In doing so, they are thereby given an opportunity to correct their previous errors. The 2AFC trials in this task are presented after every 12 trials to reduce the likelihood that items will be in a different state in the repeated test, due for example to forgetting. Given that our intention was to use similar procedures to

Tekin and Roediger (2017) to investigate HCMs, we presented the 2AFC trials after all the trials of the single-item recognition task, rather than after a subset of trials. Our 2AFC task – devised independently from the error correction paradigm – also differs from it in that the alternatives are matched according to the decision and confidence rating, whereas in the error correction paradigm, studied and non-studied items are matched only by the old/new decision. Instead of using confidence ratings, Ma et al. (2022) identified and tested competing predictions of the UVSD and 2HT models through manipulations of the response criterion and biased the tendency for participants to respond old or new using payoff manipulations.

Ma et al. (2022) provided equations to determine the probability with which an old item will be selected on a 2AFC trial consisting of a studied and non-studied item, where both items were judged as old or new. We adapted these functions for trials where the alternatives are matched according to confidence ratings. The probability that a studied item will be correctly selected from a forced-choice pair, *J*- *J*, comprising a studied and non-studied item that both received the same rating *J* in the preceding single-item recognition task (e.g., 1-1, when $J = 1$, where 1 = high confidence new), is given as:

$$P(correct|\, J - J) = \int_{C_{j-1}}^{C_j} \frac{\phi(f, d, \sigma_o)}{\Phi(C_j, d, \sigma_o) - \Phi(C_{j-1}, d, \sigma_o)} \frac{\Phi(f) - \Phi(C_{j-1})}{\Phi(C_j) - \Phi(C_{j-1})}\, dx \qquad (6)$$

where $j = J = 1,\ldots,6$, $C_0 = -\infty$, C_1-C_5 are the decision criteria, and $C_6 = \infty$; f is the strength value. In Figure 4, Equation 6 is used to plot accuracy as a function of the UVSD model parameters, and shows that, all else being equal in the UVSD model, accuracy on 1-1 trials (i.e., those where the alternatives are a HCM and HCCR) tends to increase as d and C_1 increase but decreases as σ_o increases. Accuracy can be at or below chance (50% correct) when σ_o is relatively high and d is relatively low. With fairly typical parameter values (i.e., d

~ 1, σ_o ~ 1.25), accuracy is around 50% correct for items falling below a C_1 of approximately -1.5. Thus, predicted accuracy follows approximately the same pattern as the expected differences in HCM and HCCR strength (Figure 1). The values in Figure 4 additionally serve to demonstrate that the strength differences in Figure 1 can translate to non-trivial quantitative differences in predicted accuracy.

In the DPSD model, accuracy on 1-1, 2-2, 3-3, 4-4, and 5-5 2AFC trials (i.e., when J = 1-5) is determined by the formula:

$$P(correct|\, J-J) = \int_{C_{j-1}}^{C_j} \frac{\phi(f, d', 1)}{\Phi(C_j, d', 1) - \Phi(C_{j-1}, d', 1)} \frac{\Phi(f) - \Phi(C_{j-1})}{\Phi(C_j) - \Phi(C_{j-1})}\, dx \qquad (7)$$

In the 6-6 forced-choice condition, because of the influence of recollection, an old item is either recollected with probability R_o, in which case it is correctly selected as the studied item, or else the decision is based on familiarity as in Equation 7. Thus, accuracy on 6-6 trials is given by:

$$P(correct, 6-6) = R_o + (1 - R_o) \int_{C_5}^{\mathrm{Inf}} \frac{\phi(f,\ d', 1)}{1 - \Phi(C_5, d', 1)} \frac{\Phi(f) - \Phi(C_5)}{1 - \Phi(C_5)}\, dx \qquad (8)$$

Accuracy of 1-1 trials across parameters in the DPSD model is shown in Figure 5. As was the case with the expected difference in strength to HCMs and HCCRs (Figure 3), accuracy tends to increase as d' and C_1 increases and is greater than chance when d' is greater than zero. Accuracy for these trials is unaffected by R_o, since recollection only occurs for the highest confidence old ratings, and 1-1 trials are based on familiarity.

Next, we describe four experiments designed to determine whether a residual memory effect can be demonstrated for HCMs before considering how well the UVSD and DPSD models predict the effect once fit to the single-item recognition data. To foreshadow our

behavioural findings, we found evidence of residual memory for HCMs using the aforementioned 2AFC task (in Experiments 1a, 1b, and 2), and also when a second single-item recognition task was given in which the targets and foils are HCMs and HCCRs, respectively (in Experiment 3).

Figure 4

Predicted accuracy on 1-1 2AFC trials (percentage of trials on which the HCM strength exceeds that of HCCRs) in the UVSD model. The solid lines indicate d = 0, 0.5, 1, 1.5, 2 and 2.5 (dark to light). The horizontal dashed line indicates 50%.

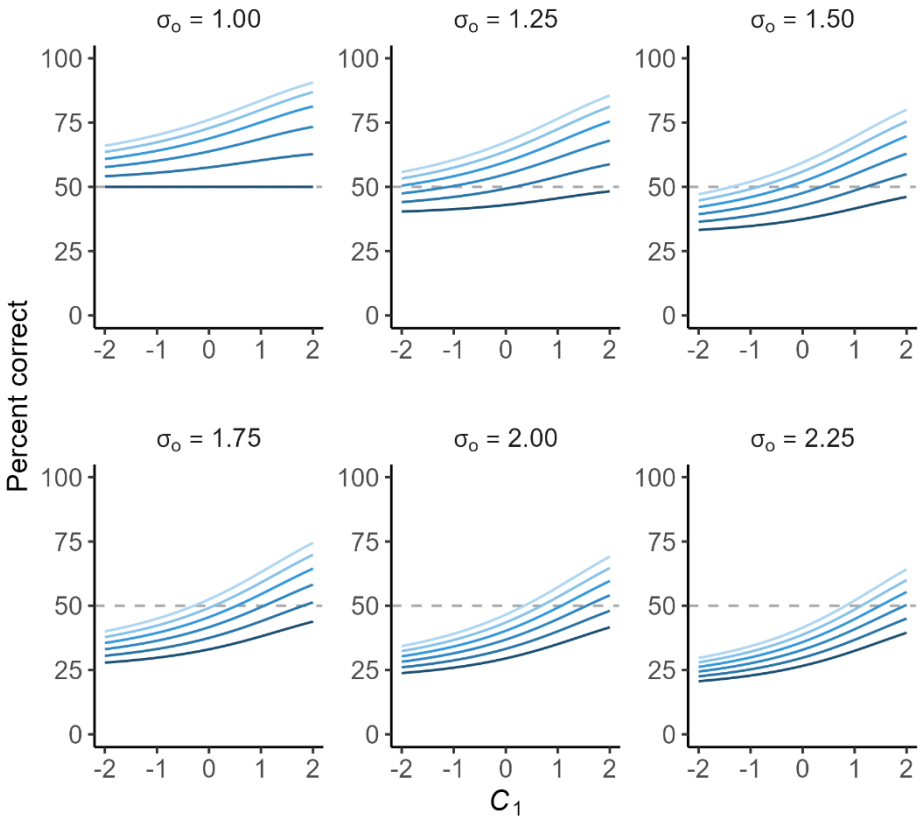
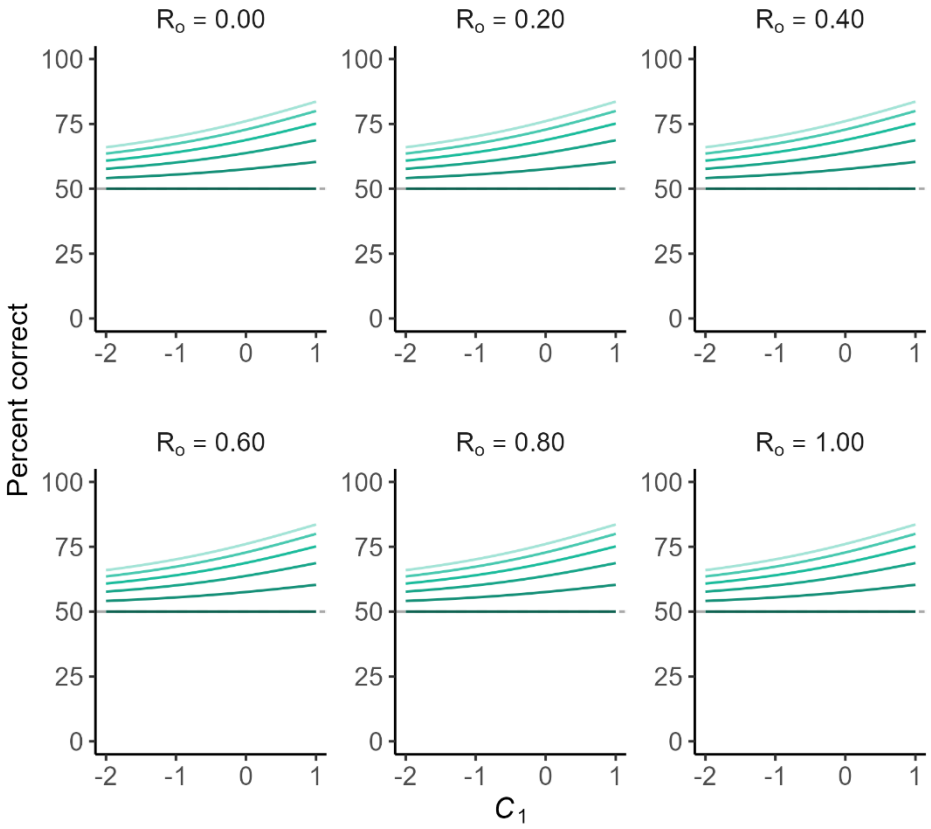


Figure 5

Predicted accuracy on 1-1 2AFC trials (percentage of trials on which the HCM strength exceeds that of HCCRs) in the DPSD model. The solid line indicates d' = 0, 0.5, 1, 1.5, 2 and 2.5 (dark to light). Horizontal dashed line indicates 50% correct.



Experiment 1a

Experiment 1a was based on Experiment 2 of Tekin and Roediger (2017), which was the first dataset that Roediger and Tekin (2020) referred to when describing the phenomenon of everyday amnesia. There were the following key differences to enable the 2AFC task to be added after the single-item recognition task. First, in Tekin and Roediger's (2017) experiment, after each "old" / "new" decision, participants indicated their confidence on either on a 4-, 5-, 20- or 100-point scale, whereas we only used a 4-point scale. We did this to ensure that the number of trials that could be presented in the 2AFC task was as high as

possible, since the number of studied and non-studied items that receive the same rating becomes less likely as the number of ratings increases. Another difference was that we had only one study-test phase cycle, rather than two, in order to reduce the likelihood of potential carry-over effects between 2AFC tasks. Given that we wanted to have the same number of studied faces (100) as Tekin and Roediger, and also that presenting them in a single study phase would effectively increase the study list length, relative to their experiment, we attempted to offset the poorer memory that would be expected from the longer list length by showing faces for a longer duration (4s rather than 2s as in Tekin & Roediger, 2017). The additional 2AFC task was presented on completion of the single-item recognition task. For completeness, in addition to the key 1-1 trials, we also included trials on which the items were matched for all possible combinations of old-new decision and confidence rating. The stages of the experiment are shown in Figure 6.

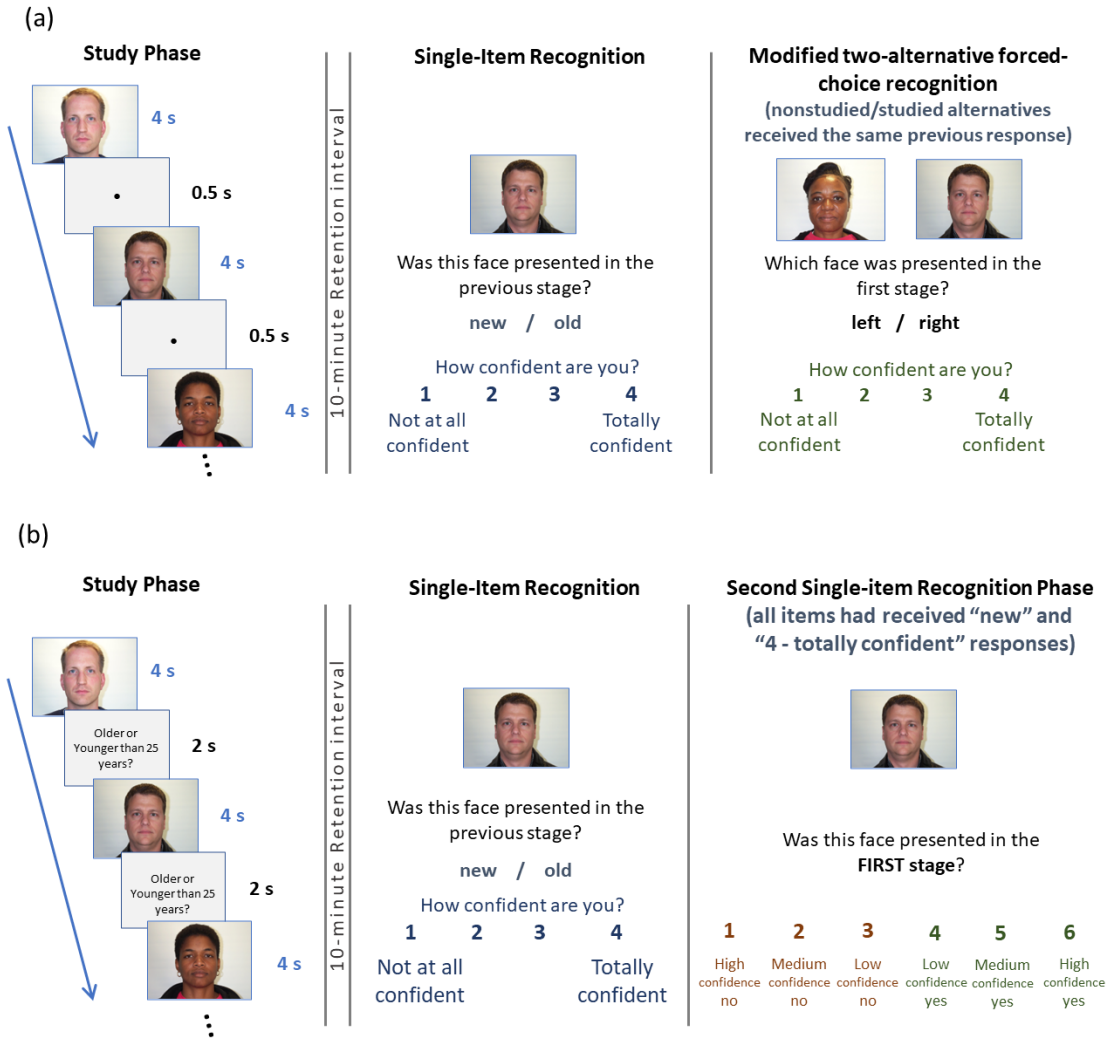
Method

Participants

We aimed to have an appropriate number of participants to match the statistics reported in Table 1 of Roediger and Tekin (2020) (i.e., 7200 study items / 100 study items per participant = 72 participants). Seventy-two participants were recruited but one was excluded from the analysis because their performance in the single-item recognition task was at floor (their hit-minus-false alarm rate was equal to 0.02), and so an additional participant was recruited to replace them. The 72 participants (60 female, 11 male, one non-binary/other) had a mean age of 19.56 years (*SD* = 1.73). All individuals in this and subsequent experiments were psychology students from the University of Plymouth, who participated in partial fulfilment of a course requirement, and were recruited through the University's online participant pool. Ethical approval for all experiments was obtained via the School of Psychology Ethics Committee, Faculty of Health, University of Plymouth.

Figure 6

(a) Experiment 1a task and (b) Experiment 3 task. Photos of faces are from the Minear and Park (2004) research database.



Materials

As in Tekin and Roediger (2017), the stimuli were 200 neutral faces of individuals between 19-50 years of age, taken from the Minear and Park (2004) database; 160 of the faces were white (80 male, 80 female), and 40 were black (20 male, 20 female). The faces were divided into two lists, matched in terms of the proportion of black/white x male/female

faces. One of the lists was randomly assigned to act as the studied stimuli for each participant, with the other list acting as the non-studied stimuli. The experiment was written in OpenSesame (Mathot et al., 2012) and the OSWeb functionality of the program was used to run it on a web browser on a desktop PC for each participant, via hosting on a JATOS server (Lange et al., 2015).

Procedure

Participants were tested individually in a quiet laboratory room. After reading a brief introduction and giving consent, participants read the instructions for the study phase, which told them that they would see faces presented one at a time, each for a few seconds, and that they should try to memorise each one for an upcoming but unspecified memory test.

On each trial of the study phase, a face was presented for 4 seconds, followed by a central fixation point for 500 ms. After the study phase, there was a 10-minute retention interval during which participants engaged in an unrelated task (word searches of countries of the world or counties in the UK). Next, the instructions for the single-item recognition memory test phase were presented. Participants were told that they would see a previously studied or non-studied face on each trial, and for each one they must decide whether it was presented in the first stage or not by responding “old” (if they thought it was studied) or “new” (if they thought it was not), after which they must indicate how confident they are in their decision on a 4-point scale, where 1 corresponds to “not at all confident” and 4 corresponds to “totally confident”.

The single-item recognition phase consisted of 200 trials, comprising 100 old and 100 new faces in a random order. On each trial, a face was shown in the centre of the screen, with the cue “Was this shown in the first phase? Z = New / M = Old” shown below it. All responses were self-paced, and once participants had pressed Z or M, the cue was replaced with the text “How confident are you in your decision?”, with the numbers 1-4 shown below

this, and the text “Not at all confident” below the number 1, and “Totally confident” shown below the number 4. After participants made their confidence rating, there was a 200 ms blank interval before the next trial was presented.

On completion of the single-item recognition memory phase, instructions for the 2AFC phase were presented. Participants were told that they would see a pair of faces side-by-side on each trial, and that one was presented in both the first stage (the study phase) and also the previous phase (the recognition task), whereas the other was not presented in the first phase and was only shown in the phase just completed. Their task was to select the face from the pair that they thought had been presented in the first phase. After making their selection they were to once again indicate how confident they were in their decision on a 4-point scale. On each 2AFC trial, a pair of faces was presented, side-by-side. Each pair consisted of one studied face and one non-studied face. Below the two faces, the question “Which one was presented in the first phase? Left / Right” was presented with the keypress response prompt “D = Left, J = Right” beneath. Crucially, the faces on each trial had been given identical ratings in the single-item recognition stage. An algorithm was built into the experimental program whereby, before the 2AFC phase commenced, studied and non-studied items that had received the same old/new decision and rating were randomly paired, for as many pairings as the responses made would allow. Following Roediger and Tekin (2020), responses at the lowest two confidence ratings (1 and 2) were binned. There were therefore six possible types of 2AFC trial, arising from 2 decisions (old, new) x 3 ratings (1 or 2, 3, 4), henceforth referred to as 1-1, 2-2, 3-3, 4-4, 5-5, and 6-6 trial types, where the number denotes the rating of the studied and non-studied alternatives on each 2AFC trial and the responses have been remapped to a 1-6 scale, with end points 1 = “totally confident new” and 6 = “totally confident old”. Thus, the items on 1-1 trials had both received a “new” decision and a “4 - totally confident” rating, those on 2-2 trials received a “new” decision and a medium

confidence “3” rating, those on 3-3 trials received a “new” decision and a “1 – not at all confident” or “2” rating, those on 4-4 trials received an “old” decision and a “1 – not at all confident” or “2” rating, those on 5-5 trials received an “old” decision and a “3” rating, and those on 6-6 trials received an “old” decision and a “4 – totally confident” rating. Hence the total number of 2AFC trials was variable across participants. 2AFC trials were randomly ordered for a given participant. Similarly, the left-right position of the studied item was randomly determined on each trial. Trials were self-paced. On completion of the final phase, participants were debriefed. The entire experiment took approximately 45 minutes, depending on the pace of the participant.

Transparency and Openness

An alpha level of .05 was used for all statistical tests. Data were analysed in R (R Core Team, 2023) and Bayes factors were obtained using the BayesFactor package (Morey & Rouder, 2022) using the default priors. The experiment was not preregistered (only Experiment 2 was preregistered, see https://osf.io/2q5yw/). The data, analysis scripts, and materials are available on the Open Science Framework at the repository for this article (https://osf.io/2q5yw/).

Results

In the single-item recognition task, the proportion of old items judged old (hits) was reliably greater than that of new items judged old (false alarms), indicating that participants could discriminate old from new items (Table 1), $t(71) = 18.06$, $p < .001$, $BF_{10} = 1.11 \times 10^{25}$, Cohen’s $d = 2.13$. The measure of discriminability ($d' = 0.98$) was comparable to that of Tekin and Roediger (2017, Experiment 2), who found that d' ranged from 0.89-1.17.

The proportion of low (1-2), medium (3) and high (4) confidence ratings made to hits, misses, false alarms, and correct rejections is shown in Table 2. 18.40% of misses received

high confidence ratings, which is comparable to that reported by Roediger and Tekin (2020) (19.74%).

Table 1

Single-item recognition phase hit rate, false alarm rate, and discriminability scores in Experiments 1a, 1b, 2, and 3.

Exp.	Study instruction	Hits		False alarms		*d′*	
		M	*SD*	*M*	*SD*	*M*	*SD*
1a	Memorise items	0.57	0.14	0.23	0.10	0.98	0.53
1b	Memorise items	0.63	0.13	0.29	0.12	0.97	0.63
2	Decide age	0.67	0.13	0.16	0.09	1.53	0.55
3	Decide age	0.67	0.12	0.20	0.09	1.34	0.47

Table 2

Number of observations and percentages of hits, misses, false alarms, and correct rejections in the single-item recognition task for Experiments 1a, 1b, 2, and 3. Percentages are within a response type (e.g., hit, miss)

	1-2		3		4		Total
	n	*%*	*n*	*%*	*n*	*%*	*n*
Experiment 1a							
Hit	872	21.32	986	24.10	2233	54.58	4091
Miss	1482	47.67	1055	33.93	572	18.40	3109
FA	781	47.25	551	33.33	321	19.42	1653
CR	2282	41.14	1871	33.73	1394	25.13	5547
Experiment 1b							
Hit	1019	22.31	1073	23.49	2476	54.20	4568
Miss	1564	59.42	817	31.04	251	9.54	2632
FA	1060	50.33	621	29.49	425	20.18	2106
CR	2518	49.43	1810	35.53	766	15.04	5094
Experiment 2							
Hit	931	19.31	1175	24.37	2716	56.33	4822
Miss	1147	48.23	802	33.73	429	18.04	2378
FA	625	54.44	345	30.05	178	15.51	1148
CR	2041	33.72	2309	38.15	1702	28.12	6052
Experiment 3							
Hit	1068	22.08	1307	27.02	2462	50.90	4837
Miss	1275	53.96	736	31.15	352	14.90	2363
FA	787	53.39	403	27.34	284	19.27	1474
CR	2445	42.70	2022	35.31	1259	21.99	5726

Having established a similar level of recognition discriminability and propensity for participants to make HCMs as Tekin and Roediger (2017, Experiment 2), we turned to performance in the 2AFC task. The mean number of trials in each 2AFC condition is shown in Table 3. Given that the number of trials in each condition differed across participants, and that some participants, due to their individual responses, had zero trials in a given 2AFC condition, we analysed the data from this phase using generalised linear mixed models, using the glmer() function in the lme4 package in R (Bates et al., 2015). The outcome variable was whether the decision on the 2AFC trial was correct or not, the fixed effect was the type of forced choice trial (i.e., 1-1, 2-2, 3-3, 4-4, 5-5, 6-6), and the random effect grouping factor was the participant. Item was not included as a random effect, since the items on each 2AFC trial were uniquely determined for each participant according to their previous responses. The model with binomial family and logit link function was fit using maximum likelihood estimation. Goodness of fit was assessed with AIC. To allow model convergence, the model contained random intercepts but not random slopes associated with the fixed factor. Overdispersion in the residuals was evaluated using the DHARMa package (Hartig, 2022) and was not detected.

Table 3

Mean number trials in each FC condition

Forced-choice condition	1-1		2-2		3-3		4-4		5-5		6-6	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Exp. 1a	9.88	9.39	14.44	10.96	20.11	12.98	9.72	6.31	7.68	6.28	6.29	7.20
Exp. 1b	4.76	4.44	11.23	6.78	21.62	12.51	12.56	6.71	8.29	5.03	6.54	6.95
Exp. 2	7.64	10.05	11.76	9.07	16.26	11.87	8.28	5.60	4.95	5.15	3.56	3.83

Accuracy differed reliably across 2AFC conditions, $\chi^2(5) = 33.10$, $p < .001$. The estimated marginal mean and 95% confidence interval for each condition are shown in Figure 7. Tests of each mean against 0.5 (i.e., proportion correct performance expected due to

chance) were performed, with *p*-values adjusted using the Holm method for six tests. Of key interest, accuracy in the 1-1 condition significantly exceeded chance ($M = 0.57$, $SE = 0.02$, 95% CI [0.51, 0.63], $z = 3.074$, $p = .0042$), suggesting that participants had some residual memory for HCMs. Accuracy did not exceed the level expected due to chance in the 2-2 condition ($M = 0.52$, $SE = 0.02$, 95% CI [0.48, 0.57], $z = 1.26$, $p = 0.21$), but did in the other conditions (Ms > 0.54, zs > 3.33, ps < .0027).

The confidence rating made after each 2AFC decision (1 = not at all confident…4 = total confidence) was also analysed using linear mixed models with condition (1-1…6-6) and decision (correct vs. incorrect) as fixed factors and participant as a random factor. Model terms were tested with the Satterthwaite method. Both the effects of rating and decision were statistically significant: $F(5, 4536.3) = 30.34$, $p < .001$, and $F(1, 4501.3) = 50.12$, $p < .001$, respectively. Moreover, a condition x decision interaction was found, $F(5, 4498.2) = 5.25$, $p < .001$, indicating that confidence generally increased across conditions 1-1 to 6-6 and tended to be greater for correct decisions than incorrect ones, but with the difference in the 2-2 and 3-3 conditions tending to be smaller than the others (Figure 8). Thus, the differences in confidence ratings to correct and incorrect decisions generally followed the same pattern as the accuracy data.

Discussion

In Experiment 1a, participants tended to select a HCM as the previously studied item rather than a HCCR in the 2AFC task. The finding that accuracy exceeded 50% in the 1-1 condition demonstrates that HCMs were encoded to some degree, and that some residual memory exists for these items. Interestingly, accuracy on 2-2 trials did not exceed 50%, indicating an absence of memory for misses made with medium confidence. On the surface, this result could be taken to reflect everyday amnesia, yet Roediger and Tekin (2020) did not make any claims with regard to these items, and there is no justification a priori for why

memory should be absent for these responses but not HCMs. Furthermore, to foreshadow the findings of our other experiments, the absence of memory for misses made with a medium level of confidence is not a robust finding. Finally, although not the focus of our investigation, an interesting aspect of our results is that accuracy was significantly greater than 50% correct on 6-6 trials, which shows that highly confident "false memories" (high-confidence false alarms) could be distinguished from true ones (high confidence hits).

Figure 7

2AFC task accuracy according to 2AFC condition (1-1, 2-2, 3-3, 4-4, 5-5, 6-6). In the left column, the estimated marginal means (controlling for participant; transformed to percentages) are plotted as dark circles, and the error bars represent the 95% confidence intervals of these estimated means from the model; light circles denote data from individual participants. Predicted accuracies according to the UVSD and DPSD models are shown in the middle and right columns, where each dark circle represents the mean expected accuracy across participants and the light circles denote the expected value for each participant.

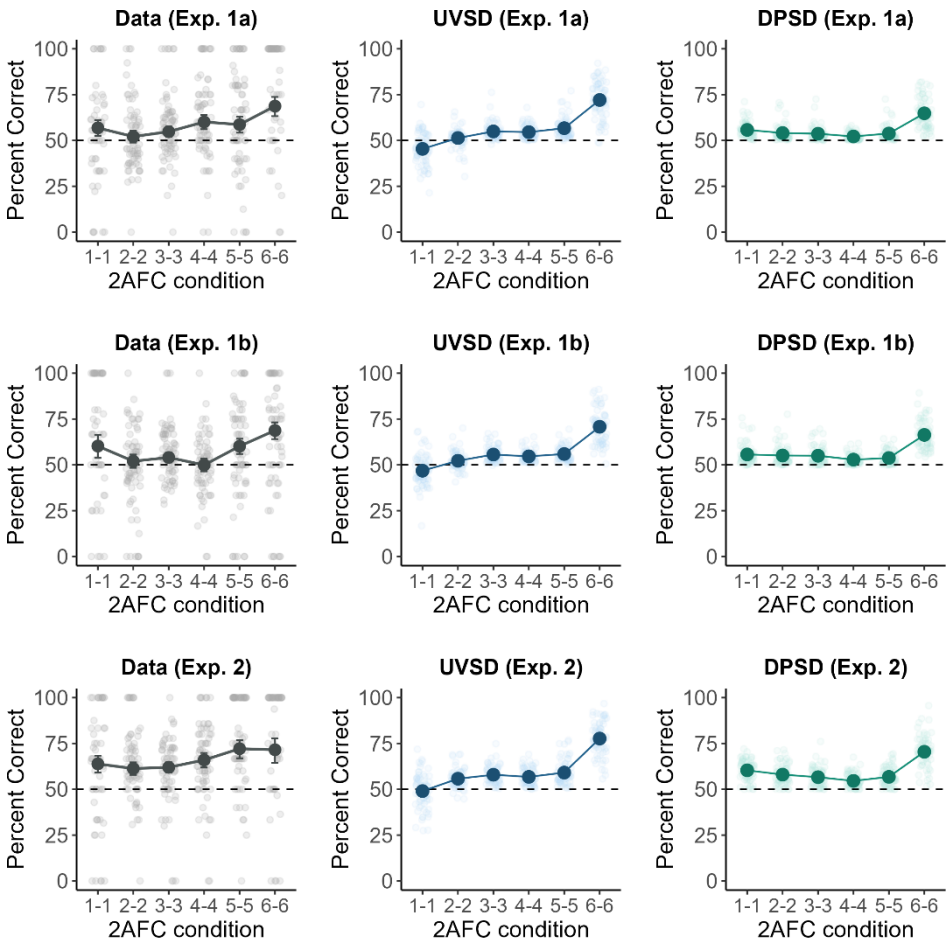
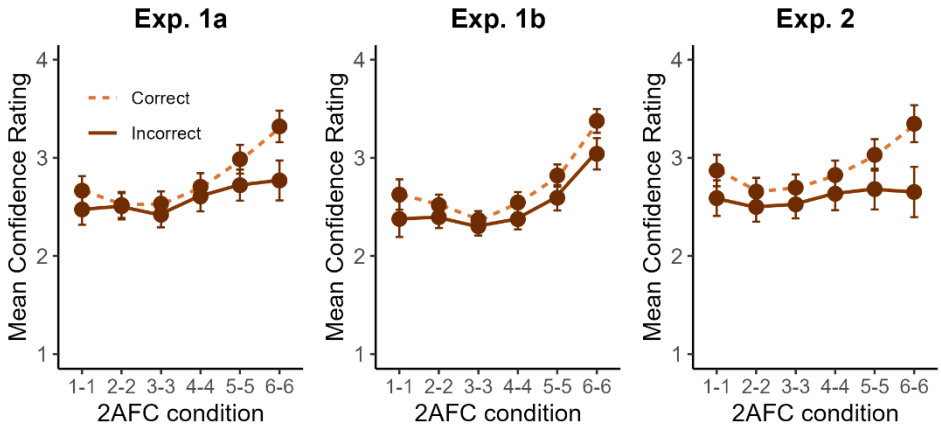


Figure 8

Confidence ratings to correct and incorrect decisions in each 2AFC condition in Experiments 1a, 1b, and 2



Experiment 1b

The aim of Experiment 1b was to determine if the residual memory effect for HCMs would also be demonstrated with word stimuli. The experiment was therefore similar in design to Tekin and Roediger (2017, Experiment 1) except for the following key differences: As in our Experiment 1a, there was a single study-test phase cycle, rather than the two cycles used by Tekin and Roediger. Once again, the reason for this was to avoid potential carry-over effects from the inclusion of the additional 2AFC task. We also used half the total number of study items as Tekin and Roediger (2017, Exp. 1) (i.e., 100 words in a single study phase) to facilitate comparison with Experiment 1a. Given that overall levels of memory performance were lower in Tekin and Roediger's Experiment 1 (with words) compared to their Experiment 2 (with faces), we attempted to counteract the anticipated lower levels of memory by doubling the study exposure duration of each word (from 2s to 4s). Thus, Experiment 1b was in fact identical to our Experiment 1a, except that the stimuli were words.

Method

Participants

We recruited 72 participants for parity with Experiment 1a. One participant performed at floor in the single-item recognition task (their false alarm rate was greater than their hit rate) and was replaced. The 72 participants (69 female, nine male, one non-binary/other) had a mean age of 19.62 years (*SD* = 2.15).

Materials and Procedure

Words were selected from Nelson et al. (2004) with similar constraints to Tekin and Roediger (2017, Exp 1). One hundred words comprised one list, and a further 100 corresponding associates comprised another. Each associate was one of the top three words associated to the other word according to Nelson et al. (2004). For example, if 'table' was on the first list, the associate 'chair' would be on the other list. All words had concreteness scores between 3.5 and 7 and log HAL frequencies (Balota et al., 2007) of between 5.99 and

13.55. Words were between 5-6 letters in length and each list had the same number of 5- and 6-letter words (66 and 34, respectively). There were no duplicates of words across lists. As in Experiment 1a, one of the lists was randomly assigned to act as the studied stimuli for each participant, with the other list acting as the non-studied stimuli. The procedure used in Experiment 1b was identical to that of Experiment 1a except that the stimuli were words instead of faces.

Results and Discussion

Participants could reliably discriminate old from new items (Table 1), $t(71) = 16.11$, $p < .001$, $BF_{10} = 1.82 \times 10^{22}$, Cohen's $d = 1.90$. Mean d' was virtually identical to Experiment 1a ($d' = 0.97$, Table 1). Interestingly though, the percentage of misses made with high confidence (9.54%) was approximately half the level found in Experiment 1a (18.40%, Table 2) and in Roediger and Tekin (2020).

The mean number of 2AFC trials in each condition is shown in Table 3. As in Experiment 1a, we analysed the data from this phase in the same manner using generalised linear mixed models, and derived expected marginal mean accuracy with tests of each mean against expected chance levels of performance that were adjusted for multiple comparisons using a Holm correction. Accuracy significantly differed across conditions, $\chi^2(5) = 51.63$, $p < .001$. As shown in Figure 7, accuracy in the 1-1 condition exceeded chance performance ($M = 0.60$, $SE = 0.03$, 95% CI [0.52, 0.68], $z = 3.09$, $p = .008$), suggesting once again that participants had some residual memory for HCMs. As in Experiment 1a, accuracy did not exceed the level expected due to chance in the 2-2 condition ($M = 0.52$, $SE = 0.02$, 95% CI [0.47, 0.57], $z = 1.08$, $p = .56$). Unlike Experiment 1a, accuracy in the 4-4 condition did not exceed the level expected due to chance ($M = 0.50$, $SE = 0.02$, 95% CI [0.45, 0.55], $z = -0.04$, $p = 0.97$), but accuracy in the 3-3, 5-5- and 6-6 conditions did (Ms > 0.54, zs > 2.91, ps < .011).

As in Experiment 1a, the confidence rating made after each 2AFC decision (1 = not at all confident…4 = total confidence) was analysed using linear mixed models with the condition (1-1 to 6-6) and decision (correct vs. incorrect) as fixed factors and participant as a random factor. Model terms were tested with the Satterthwaite method. Both the effects of rating and decision were statistically significant, $F(5, 4449.7) = 64.56$, $p < .001$, and $F(1, 4416.1) = 40.54$, $p < .001$, respectively. A marginal condition x decision interaction was also found, $F(5, 4409.1) = 2.14$, $p = .06$, indicating that confidence generally increased from conditions 1-1 to 6-6 and tended to be greater for correct decisions than incorrect ones, except in conditions 2-2 and 3-3 where the difference tended to be smaller (Figure 8). Thus, like Experiment 1a, confidence and accuracy were generally related in the 2AFC task.

Once again, accuracy in the 1-1 condition exceeded the level of performance expected due to chance, demonstrating a residual memory effect for HCMs, this time with word stimuli. Interestingly, the effect was shown even though the percentage of HCMs made was roughly half that of Experiment 1a.

Experiment 2

In Experiment 2, we aimed to replicate the residual memory effect for HCMs, but under more demanding encoding conditions where a given study item was more likely to have been processed. Although participants were instructed to memorise the set of study images/words in Experiments 1a and 1b, it is possible that not all items were attended to, and so these items may be functionally equivalent to new items at the time of the test phase, which could have given rise to HCM responses. This alternative explanation for HCMs was considered by Roediger and Tekin (2020), and if it occurred, may have diluted the residual memory effect we found for HCMs. Thus, to help ensure that participants processed each item during the study phase, they were required to make a decision to each one, rather than simply memorise the study list. To achieve this, the design was identical to Experiment 1a,

except that in the study phase, participants decided whether they thought each face was of a person older or younger than 25 years of age. We chose 25 years for the decision, since approximately half of the faces in each counterbalance condition were older than 25 and approximately half were younger. We based this experiment on Experiment 1a rather than Experiment 1b, since the proportion of HCMs was higher, and more comparable to the level reported by Roediger and Tekin (2020). Experiment 2 was preregistered (at https://osf.io/2q5yw/).

Method

Participants

Seventy-two participants were recruited from the same participant pool database as Experiment 1a but had not taken part in that experiment. One participant was replaced due to their performance in the single-item recognition task being at floor (their hit minus false alarm rate was less than the 0.05 criterion we preregistered, and was equal to zero). Participants received either course credit or £7.50 in exchange for their participation. Fifty-six participants identified as male, 14 as female, one as non-binary/other, and one did not provide this information. Their mean age was 20.15 years (*SD* = 2.69).

Design and Procedure

The design and procedure were identical to Experiment 1a, except that after each face was shown in the study phase, participants had to decide whether the face was of a person who was older or younger than 25 years of age. The study instructions informed participants that they would see faces presented one at a time, each for a few seconds, and that they would have to decide whether the person looked older or younger than 25 by pressing one of two keys. The decision would sometimes be difficult to make but they should try their best to do so within 2 seconds after they are presented. On each trial of the study phase, a central fixation point was presented for 500 ms, then a face was presented for 4 seconds. After the

face disappeared the question "Did the face look older or younger than 25 years?" was presented and the keypress prompts "Q = older" and "P = younger" appeared below the question. Participants had up to 2 seconds to make their keypress before the program automatically advanced to the next trial (if no keypress was made). If a keypress was made in under two seconds, the program advanced to the next trial.

Results

In the study phase, the mean proportion of items receiving an "older" or "younger" judgment within the trial duration was 0.98 (*SD* = 0.11). There were no participants who made no responses, and no participants exclusively responded "older" or "younger".

In the single-item recognition task, the hit rate was reliably greater than the false alarm rate (Table 1), indicating that participants could successfully discriminate old from new items, $t(71) = 26.53$, $p < .001$, $BF_{10} = 1.53 \times 10^{35}$, Cohen's $d = 3.13$. Mean d' (1.53) was numerically greater than those of Experiments 1a and 1b and Tekin and Roediger (2017, Experiment 2) (where d' ranged from 0.89-1.17), in line with the deeper encoding task performed by participants.

The proportions of low (1-2), medium (3) and high (4) confidence ratings made to hits, misses, false alarms, and correct rejections are shown in Table 2. 18.04% of misses received high confidence ratings, which is comparable to the level in Experiment 1a and Roediger and Tekin (2020). This percentage remained similar (18.20%) even when old items for which no key press decision was made during the study phase were excluded. This bolsters our confidence that HCMs were extensively attended and processed during the study phase.

We turn next to the central issue of whether residual memory could be detected for HCMs in the 2AFC task. The mean numbers of 2AFC trials in each condition are shown in Table 3. As in Experiments 1a and 1b, we analysed the data from this phase in the same

manner, using generalised linear mixed models and conducted tests of accuracy versus levels expected due to chance with *p*-values adjusted for multiple comparisons using the Holm correction. Accuracy significantly differed across conditions, $\chi^2(5) = 19.41$, $p < .001$. As shown in Figure 7, accuracy in the 1-1 condition exceeded the level expected due to chance (0.5) ($M = 0.64$, $SE = 0.024$, 95% CI [0.57, 0.70], $z = 5.56$, $p < .0001$). Once again, this suggests that some residual memory for HCMs could be detected in this task. Accuracy also exceeded the level expected due to chance in the other 2AFC conditions (*M*s > 0.61, *SE*s < 0.035, 95% CIs [> 0.57, < 0.80], *z*s > 5.47, *p*s < .0001).

As in the previous experiments, confidence ratings made after each 2AFC decision (1 = not at all confident…4 = total confidence) were analysed using linear mixed models with condition (1-1…6-6) and decision (correct vs. incorrect) as fixed factors and participant as a random factor. This analysis was exploratory and was not pre-registered. Model terms were tested with the Satterthwaite method. Both the effects of rating and decision were statistically significant, $F(5, 3386.3) = 10.62$, $p < .001$, and $F(1, 3353.9) = 72.89$, $p < .001$, respectively. A condition x decision interaction was also found, $F(5, 3354.5) = 3.34$, $p = .005$, indicating that confidence generally increased across conditions 1-1 to 6-6 and tended to be greater for correct decisions than incorrect ones, with a smaller difference apparent in the 2-2, 3-3, and 4-4 conditions (Figure 8). Thus, confidence tended to follow accuracy in the 2AFC task.

Discussion

Once again, a residual memory effect for HCMs was demonstrated in Experiment 2, this time following a deeper encoding task, where we can have greater confidence that HCMs were attended to and processed during encoding. For almost every HCM, the participant had processed the face stimulus sufficiently in the study phase to make a decision about the person's age. Indeed, in line with there being deeper processing, performance in both

recognition tasks was numerically greater than in Experiments 1a and 1b, and accuracy in all 2AFC conditions was greater than chance.

Experiment 3

In Experiment 3, we sought to determine whether the residual memory effect for HCMs would be also found when the 2AFC task was replaced with a second single-item recognition task in which the targets and foils were HCMs and HCCRs, respectively. If so, this would suggest that the conditions necessary to demonstrate the residual memory effect for HCMs are not restricted to those imposed by a 2AFC task, for example, the requirement to make a relative assessment of the strength of two items. Experiment 3 was identical to Experiment 2 except that, after the first single-item recognition task, participants completed a second single-item recognition task comprising solely previous HCMs and HCCRs (see Figure 6b). Participants were told that half of the items in this phase were in fact from the study phase, and that they must decide for each item whether it was presented in this phase or not. The design was otherwise equivalent to Experiment 2.

Method

Participants

As in the previous experiments, we recruited 72 participants on the basis of acquiring a set of data of roughly the same size (i.e., 7200 study items) as Roediger and Tekin (2020). None had taken part in the other experiments with face stimuli (Experiments 1a or 2). Demographic data for 5 participants were lost due to technical failure. The remaining 67 individuals (51 female, 15 male, 1 non-binary/other) had a mean age of 19.51 years (*SD* = 1.85).

Materials and Procedure

The materials and procedure of Experiment 3 were identical to those of Experiment 2 except that the 2AFC task was replaced with a second single-item recognition task. The

instructions for this phase told participants that they would see a mixture of faces that they had just seen, each presented one at a time. For each face, they were to decide whether they thought it was one that was presented in the first phase or not, indicating their decision using a 6-point rating scale, where 1 = high confidence no, 2 = medium confidence no, 3 = low confidence no, 4 = low confidence yes, 5 = medium confidence yes, and 6 = high confidence yes. This one-step rating scale was used, rather than the two-step procedure used in Experiment 1a, in order to help distinguish the two single-item recognition phases, and also to help reduce the likelihood that participants would attempt to simply reproduce their response from the first test phase. That is, we wanted to avoid a situation where, for a particular face, a participant could adopt a strategy whereby they remembered that they responded "new" followed by "4" to it, and then attempt to reproduce these exact same responses in order to be consistent with their previous responding. In addition, participants were told that half of the faces they would see were in fact from the first phase and half were not. We reasoned that by giving participants this information, they may be less likely to adopt a strategy in which they attempt to be consistent in their responses across phases.

An algorithm was built into the experimental program that ensured that the number of HCMs and HCCRs presented in the second single-item recognition phase was the same. If the number of HCCRs was greater than that of HCMs, a random sample of HCCRs was selected, with *N* equal to the number of HCMs, and vice versa if the number of HCMs was greater than the number of HCCRs. The order of presentation of HCMs and HCCRs was randomised for each participant. In the event that a participant made no HCMs or HCCRs in the first single-item recognition phase, no items could be presented in the second single-item recognition phase, so the experiment ended, and the participant was debriefed.

Results

In the study phase, the mean proportion of items receiving an “older” or “younger” judgment within the trial duration was 0.94 (*SD* = 0.19). There were no participants who made no responses or who exclusively responded “older” or “younger”.

As in the previous experiments, the hit rate was reliably greater than the false alarm rate, indicating that participants could successfully discriminate old from new items (Table 1), $t(71) = 27.45$, $p < .001$, $BF_{10} = 1.33 \times 10^{36}$, Cohen’s $d = 3.23$. Mean d' (1.34) was numerically greater than those of Experiments 1a and 1b and Tekin and Roediger (2017, Experiment 2) (where d' ranged from 0.89-1.17), in line with the deeper encoding task performed by participants.

The proportion of low (1-2), medium (3) and high (4) confidence ratings made to hits, misses, false alarms and correct rejections is shown in Table 2. 14.90% of misses received high confidence ratings, which is slightly lower than that found in Experiment 1a and reported by Roediger and Tekin (2020). As in Experiment 2, this value remained similar even when old items for which no key press decision was made during the study phase were excluded (13.87%).

We turn next to the question of whether any residual memory could be detected for HCMs in the second single-item recognition task. Fifty out of 72 participants made at least one HCM and one HCCR response in the first single-item recognition task and were therefore presented with the second single-item recognition task. The mean number of trials in this phase was 14 across participants (*SE* = 2.35, range 2-104 trials). Participants were able to discriminate between old and new items in this phase as indicated by the mean confidence rating (from 1 = high confidence new to 6 = high confidence old) being greater for old items ($M = 2.98$, $SE = 0.15$) than new items ($M = 2.47$, $SE = 0.15$), $t(49) = 3.52$, $p < .001$, $BF_{10} = 30.11$, $d = 0.48$. Likewise, the hit rate (i.e., classifying HCMs as old; $M = 0.37$, $SE = 0.04$) was significantly greater than the false alarm rate (i.e., classifying HCCRs as old; $M = 0.23$,

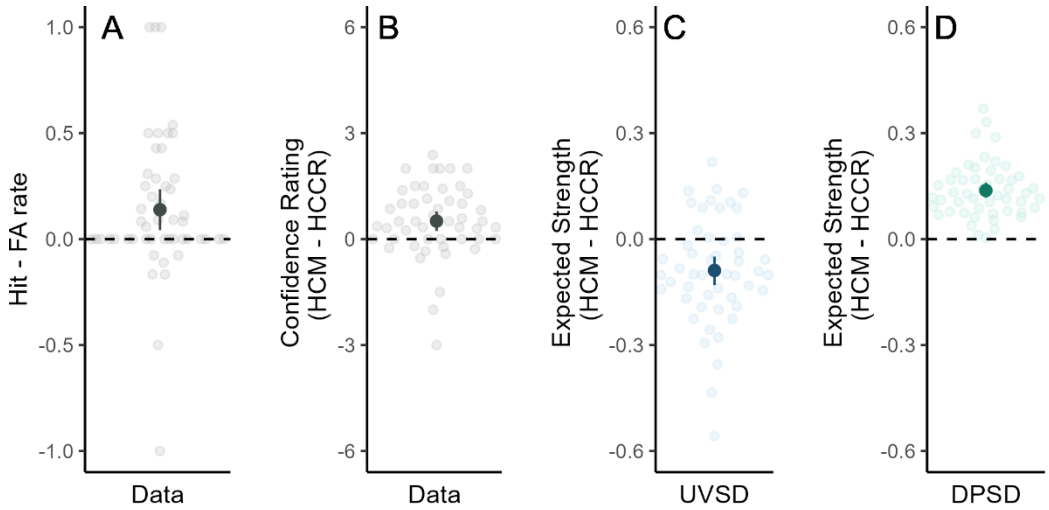
SE = 0.04), $t(49) = 2.90$, $p = .006$, $BF_{10} = 6.29$, $d = 0.45$. The mean differences in confidence ratings and proportions of old judgments are shown in panels A and B of Figure 9.

Discussion

In Experiment 3, when re-presented with HCMs and HCCRs in a second single-item recognition task, participants were more likely to judge HCMs to be old and assign them higher confidence ratings, compared to HCCRs. As with the previous experiments, this demonstrates a residual memory effect for HCMs and that the effect generalises to another type of recognition task other than 2AFC.

Figure 9

Mean difference in hit and false alarm rate (panel A) and six-point confidence ratings (panel B) for HCMs and HCCRs when presented in the second single-item recognition memory phase of Experiment 3. Panels C and D show the expected difference in strength of HCMs and HCCRs, given fits of the UVSD and DPSD models to the single-item recognition data. Twenty-two individuals did not make any HCMs so are omitted from the experimental data panels and model data. One individual had an extreme negative outlying difference in expected strength to HCMs and HCCRs in the UVSD model (difference less than -3) and is not shown in either the UVSD or DPSD panels. Black circles indicate mean; error bars denote 95% confidence interval of the mean. Grey circles denote individual participant data (panels A and B) or expected values under each model (panels C and D).



Modelling

UVSD and DPSD models

Having established a residual memory effect for HCMs, we turn next to our second main aim, which was to determine the extent to which the UVSD and DPSD models could predict this effect once their parameters were first fixed by fitting them to the data from the initial single-item recognition phase.

Fits to the single-item recognition data

The parameters of the UVSD and DPSD models were obtained from the single-item recognition data for every participant in each experiment using maximum likelihood estimation. This involved obtaining the likelihood of every response given particular parameter values and using the optim function in R (R Core Team, 2023) to obtain the values that maximised the summed log likelihood across trials. In Experiment 2, the data of two participants could not be fit by the models due to there being no responses in some of the stimulus × ratings cells; data from one other participant could not be fit in Experiment 3 for the same reason. A number of participants had extreme positive C_5 parameter estimates (i.e., $C_5 > 100$) when fit by the DPSD model (nine in Experiment 1a, one in Experiment 1b, four in Experiment 2, seven in Experiment 3). These participants made no "totally confident old" decisions to new items and were not included in the calculation of the mean parameter estimates shown in Table 4. In each experiment, we also fit the data aggregated across participants and the same pattern of predictions for HCMs that we report below was found in each model.

We assessed the fit of each model by obtaining G^2 values for each participant, comparing the observed frequencies of each response, and the expected frequencies given the parameter estimates. Each model yielded a satisfactory fit to the majority of participants across experiments (79-90%), as indicated by G^2 values with associated p-values greater than 0.05 (see Table 5). If anything, the UVSD model tended to fit a greater proportion than the DPSD model.

We also calculated ΔAIC for each model, where the AIC for a participant for a given model is AIC = $-2\ln(L) + 2p$, where L is the maximum likelihood value, p is the number of free parameters (seven in each model: σ_o, d, C_1-C_5 in the UVSD model, and d', R_o, C_1-C_5 in the DPSD model), and ΔAIC is the AIC value minus the AIC for the best fitting model for

that participant (Akaike, 1973). Note that both DPSD and UVSD models have an equal number of parameters, so comparisons of AIC are equivalent to comparisons of the log likelihood. ΔAIC values less than 2 do not distinguish the models, offering little support for the best fitting one (Burnham & Anderson, 2002). Although the UVSD model tended to fit the majority of participants best by this criterion, the mean ΔAIC values were less than 2 in each experiment. Overall, the goodness-of-fit statistics confirm that the models fit the single-item recognition data well, as might be expected from the literature (e.g., Wixted, 2007; Yonelinas & Parks, 2007). Both models also closely reproduced the percentage of HCMs found in each experiment (see Table 6).

Table 4

Mean parameter estimates of the UVSD and DPSD models in each experiment

	Experiment							
	1a		1b		2		3	
Parameter	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
UVSD								
d	1.16	0.08	1.18	0.08	1.83	0.10	1.62	0.08
σ_o	1.46	0.04	1.38	0.05	1.58	0.06	1.48	0.04
C_1	-1.42	0.21	-1.98	0.27	-1.02	0.15	-1.69	0.30
C_2	-0.12	0.07	-0.40	0.06	0.18	0.07	-0.13	0.06
C_3	0.82	0.05	0.62	0.05	1.06	0.04	0.87	0.04
C_4	1.31	0.06	1.16	0.05	1.59	0.05	1.48	0.07
C_5	1.93	0.08	1.73	0.06	2.25	0.08	2.27	0.13
DPSD								
d'	0.50	0.04	0.64	0.07	0.99	0.06	0.88	0.06
R_o	0.23	0.02	0.23	0.02	0.30	0.02	0.26	0.02
C_1	-1.24	0.18	-1.76	0.22	-0.86	0.12	-1.29	0.16
C_2	-0.11	0.07	-0.36	0.06	0.21	0.06	-0.07	0.06
C_3	0.70	0.04	0.57	0.04	0.99	0.04	0.79	0.04
C_4	1.17	0.05	1.10	0.05	1.50	0.05	1.32	0.06
C_5	5.74	1.54	2.56	0.44	7.42	1.89	5.75	1.18

Table 5

Goodness of fit of the UVSD and DPSD models in Experiments 1-3.

Experiment	Percentage of participants with non-significant G^2		ΔAIC		Percentage of participants best fit by AIC		*N*
	UVSD	DPSD	UVSD	DPSD	UVSD	DPSD	
1a	86.11	81.94	1.38	1.53	44.44	55.56	72
1b	84.72	79.17	0.97	1.16	55.56	44.44	72
2	90.00	80.00	0.66	1.78	68.57	31.43	70
3	88.73	81.69	0.47	1.90	61.97	38.03	71

Table 6

Mean percentage (and SE) of each response type across participants according to the parameter estimates of the UVSD and DPSD models for each experiment

	UVSD						DPSD					
	1-2		3		4		1-2		3		4	
	M	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
Experiment 1a												
Hit	22.85	1.82	26.33	1.81	50.81	2.48	22.77	1.82	23.59	1.75	53.65	2.44
Miss	48.82	2.74	30.42	1.82	20.76	2.33	51.47	2.95	30.89	1.94	17.64	2.32
FA	49.08	2.86	30.94	2.03	19.98	2.22	47.95	2.80	36.06	2.12	15.98	2.34
CR	40.88	2.86	34.77	1.94	24.35	2.72	39.54	2.71	34.59	1.88	25.86	2.69
Experiment 1b												
Hit	24.44	1.58	24.47	1.18	51.10	2.23	24.52	1.57	22.37	1.13	53.11	2.18
Miss	59.57	2.37	29.40	1.65	11.03	1.63	62.05	2.45	29.31	1.74	8.63	1.39
FA	51.16	2.13	29.05	1.22	19.79	1.88	50.19	2.08	32.73	1.28	17.08	1.90
CR	49.30	2.37	36.51	1.67	14.19	1.73	48.21	2.33	36.38	1.65	15.42	1.81
Experiment 2												
Hit	19.48	1.52	23.88	1.93	56.64	2.75	19.66	1.57	22.30	1.91	58.05	2.80
Miss	51.21	2.78	30.58	2.06	18.20	2.43	53.98	2.91	31.66	2.18	14.35	2.35
FA	55.72	2.73	27.50	1.87	16.79	2.28	53.83	2.59	32.39	2.00	13.79	2.18
CR	32.99	2.55	38.90	2.24	28.12	2.75	32.09	2.48	38.24	2.23	29.67	2.77
Experiment 3												
Hit	22.79	1.82	27.32	1.59	49.89	2.77	23.22	1.86	25.43	1.56	51.35	2.82
Miss	57.17	2.82	28.25	1.81	14.58	2.04	60.61	3.04	27.85	2.02	11.54	1.92
FA	55.39	3.41	27.27	1.92	17.35	2.39	52.67	3.14	32.25	1.87	15.08	2.34
CR	42.86	2.67	36.73	1.90	20.40	2.35	41.54	2.57	36.63	1.85	21.83	2.38

Forced-choice accuracy predictions

For Experiments 1a, 1b, and 2, the parameter estimates were then used to obtain predicted accuracy for each 2AFC condition. In Experiment 2, it was not possible to derive the expected accuracy in the 6-6 condition of the 2AFC task for those individuals with extreme positive C_5 estimates in the DPSD model, since derivation of the cumulative normal probability failed, so a simulation-based approach was used to derive their predicted accuracy in this condition instead. For parity across conditions and models, we obtained predicted 2AFC accuracy in each condition using this simulation-based approach (which yielded the same results as Equations 6-8). For each participant, 200,000 2AFC trials were simulated per condition, and the item with the greater strength value on each trial was assumed to be selected as the alternative that had been studied. On 6-6 trials in DPSD model, if recollection occurred for the old item, it was assumed to be selected, otherwise the item with the greater strength value was selected. Even with this simulation approach, derivation of the expected accuracy of 6-6 trials in the DPSD model failed for three participants with the most extreme C_5 estimates (two participants in Experiment 1a, and one in Experiment 1b) since familiarity based 2AFC decisions could not be determined. These participants were not included in the analysis for this model below.

Predicted accuracy in each condition is shown in Figure 7. The UVSD model did not predict a residual memory effect for HCMs, as indicated by the mean predicted accuracy across participants in the 1-1 condition being below or no different from 50%. In Experiment 1a, predicted accuracy was significantly below 50% (M = 45.39%), $t(71) = -4.54$, $p < .001$; likewise in Experiment 1b (M = 46.83%), $t(71) = -3.22$, $p = .002$. In Experiment 2, predicted accuracy did not differ from 50%, (M = 49.02%), $t(69) = -0.87$, $p = .39$, and is far below the lower limit of the 95% CI on the observed data. The DPSD model did, however, predict the effect, as indicated by the predicted mean accuracy being greater than 50% in Experiment 1a

(M = 55.76%), $t(69) = 11.82$, $p < .001$, Experiment 1b (M = 55.69%), $t(70) = 9.38$, $p < .001$, and Experiment 2 (M = 60.35%), $t(69) = 16.10$, $p < .001$. In each case the predicted effect falls inside the observed 95% CI.

It is also evident from Figure 7 that the UVSD and DPSD models tended to numerically underestimate accuracy in the 2-2, 3-3, 4-4, and 5-5 conditions. In this sense, there is room for improvement in the quantitative predictions made by both models. Predicted accuracy in the 6-6 condition was closer to levels observed in the data. Most importantly though, the main qualitative patterns were predicted in the remaining conditions: both models predicted that accuracy is greater than chance, tends to be greatest in the 6-6 condition, and is generally higher in Experiment 2, where memory was stronger following the encoding manipulation employed.

Second single-item recognition phase predictions

In Experiment 3, where a second single-item recognition phase was used as the additional memory test, Equations 3-5 were used to obtain the expected difference in strength to HCMs and HCCRs from the model parameter estimates. We did not derive predictions for the hit and false alarm rates or confidence ratings to HCMs and HCCRs in this phase, since doing so would require estimating further parameters for this stage (e.g., decision criteria), which would require fitting the data from this phase rather than deriving *ex ante* predictions for it. Figure 9 shows that the mean expected strength of HCMs was lower than that of HCCRs in the UVSD model, $t(70) = -3.88$, $p < .001$, contrary to the observed pattern, but the DPSD model predicted the opposite, $t(70) = 14.57$, $p < .001$. Assuming that differences in strength translate to subsequent levels of discriminability, the UVSD model did not predict the residual memory effect for HCMs, and instead predicted a negative effect, whereas the DPSD model successfully predicted the effect.

Other models

Alternative distributional assumptions

Although the dominant version of the UVSD model is one in which the distributions are Gaussian, versions in which the distributions are not Gaussian have occasionally been fit in the literature (see e.g., Wixted & Mickes, 2010), and, more broadly, the Gaussian assumption in SDT is technically an auxiliary assumption (Kellen et al., 2021). To explore the extent to which the mis-prediction of the UVSD model is due to this Gaussian assumption, we fit versions of the model where the underlying distributions were either Gumbel, logistic, Weibull, lognormal, exponential, or gamma, and then obtained their predictions in a similar manner. The parameter estimates are shown in Table 7. These models all closely reproduced the percentage of HCMs in each experiment (Table 8) and fits were comparable to the Gaussian-UVSD model (Table 9).

Notably, the Gumbel-, lognormal-, and logistic-UVSD models all still failed to predict the residual memory effect for HCMs in Experiments 1a, 1b, and 2 (see Figures 9-12), and the expected value for HCMs was lower than that of HCCRs in Experiment 3 (as shown by the solid-black points below the zero-difference line in Figure 13). In the Weibull-UVSD model, predicted accuracy did not differ from chance in Experiment 1a (M = 49.54%), $t(71)$ = -0.66, p = .51, or Experiment 1b (M = 51.00%), $t(71)$ = 1.22 p = .23. However, it was greater than chance in Experiment 2 (M = 53.52%), $t(69)$ = 3.81, p < .001, although below the lower limit of the 95% CI on the observed effect. The expected strength of HCMs was greater than that of HCCRs in Experiment 3, $t(70)$ = 3.37, p < .001, but the predicted difference was very small (M = 0.009). In the gamma-UVSD model, predicted accuracy was below chance in Experiment 1a (M = 47.60%), $t(71)$ = -2.62, p = .01, and did not differ from chance in Experiment 1b (M = 49.43%), $t(71)$ = -0.59, p = .56, or Experiment 2 (M = 51.90%), $t(69)$ = 1.72, p = .09. Again, these predicted effects fell outside the observed 95% CIs. Expected

strength of HCMs and HCCRs also did not differ in Experiment 3, $t(70) = 1.06$, $p = .29$ (*M* difference = 0.006).

Interestingly, the Exponential-UVSD model predicted that 1-1 accuracy was greater than chance in Experiment 1a ($M = 51.34\%$), $t(71) = 6.38$, $p < .001$, Experiment 1b ($M = 50.68\%$), $t(71) = 5.48$, $p < .001$, and also Experiment 2 ($M = 52.07\%$), $t(69) = 7.72$, $p < .001$, but, crucially, predicted accuracy was generally much lower than observed empirically. Although the predicted effect for Experiment 1a was inside the observed 95% CI, those of Experiments 1b and 2 fell outside the observed 95% CIs. The expected strength of HCMs was also greater than that of HCCRs in Experiment 3, $t(70) = 3.02$, $p = .004$, but again, the predicted effect was very small (*M* difference in strength = 0.007).

Our explorations of versions of a UVSD model with non-Gaussian distributions show that the failure of the UVSD model to predict the residual memory effect for HCMs is not due to the Gaussian assumption. Neither the Gumbel, logistic, lognormal, or gamma versions of the UVSD model predicted the residual memory effect. The Weibull-UVSD model did predict the presence of the effects in Experiments 2 and 3, but not in Experiments 1a and 1b, so is unsatisfactory overall. Interestingly, the exponential-UVSD model predicts a positive residual memory effect for HCMs and can do so because the likelihood ratio is monotonic with strength in this version, but the predicted effects were generally far smaller than we observed in our experiments and so in this sense the model is also unsatisfactory.

We could have extended this exploration by considering yet more distributions (e.g., exponentially modified Gaussian distribution), or by using different fixed values for the new item distributions, but such an exercise is clearly *post hoc*. Furthermore, any non-Gaussian implementation of the UVSD model would also need to explain the recognition literature at least as well as the Gaussian version, and this would require additional investigation. What this exploration of non-Gaussian distributions highlights is that, even when the distributions

are not Gaussian, it is possible for the likelihood of extreme low strength old items to be greater, or practically indistinguishable from the likelihood of new items, which can lead the UVSD model to mis-predict the residual memory effect for HCMs. This effect therefore represents a serious problem for the model.

Table 7

Parameter estimates of the 2HT, MSD and non-Gaussian-UVSD models for each experiment

	Experiment							
	1a		1b		2		3	
	M	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
2HT								
d_o	0.27	0.02	0.30	0.02	0.37	0.02	0.31	0.02
d_n	0.10	0.01	0.06	0.01	0.16	0.02	0.10	0.01
g_1	0.11	0.02	0.05	0.01	0.11	0.02	0.08	0.01
g_2	0.25	0.02	0.23	0.01	0.30	0.02	0.25	0.01
g_3	0.31	0.02	0.34	0.02	0.29	0.02	0.32	0.02
g_4	0.14	0.01	0.17	0.01	0.15	0.01	0.16	0.01
g_5	0.13	0.01	0.14	0.01	0.13	0.01	0.15	0.01
MSD								
d'	2.74	0.52	2.82	0.37	3.31	0.61	2.17	0.11
λ	0.58	0.03	0.62	0.03	0.76	0.02	0.73	0.02
C_1	-1.52	0.29	-1.90	0.24	-1.14	0.22	-1.70	0.33
C_2	-0.13	0.07	-0.41	0.06	0.16	0.07	-0.14	0.07
C_3	0.80	0.05	0.61	0.05	1.07	0.04	0.87	0.04
C_4	1.32	0.06	1.25	0.11	1.58	0.05	1.47	0.06
C_5	2.22	0.16	1.94	0.13	2.19	0.07	2.16	0.09
Gumbel-UVSD								
μ_o (location)	1.34	0.12	1.44	0.14	2.39	0.17	2.05	0.13
β_o (scale)	2.01	0.08	1.91	0.09	2.59	0.12	2.28	0.10
C_1	-0.86	0.16	-1.33	0.23	-0.59	0.16	-1.31	0.37
C_2	0.31	0.08	0.00	0.07	0.67	0.08	0.29	0.07
C_3	1.50	0.08	1.21	0.08	1.89	0.07	1.59	0.06
C_4	2.30	0.10	2.05	0.10	2.82	0.09	2.62	0.12
C_5	3.46	0.16	3.09	0.13	4.18	0.17	4.30	0.31
Logistic-UVSD								
μ_o (location)	1.95	0.14	1.99	0.15	3.15	0.17	2.75	0.15
s_o (scale)	1.50	0.04	1.39	0.05	1.62	0.07	1.49	0.04
C_1	-2.96	0.58	-5.32	1.44	-2.32	0.60	-3.80	0.90
C_2	-0.22	0.12	-0.68	0.11	0.30	0.12	-0.22	0.11
C_3	1.37	0.09	1.04	0.09	1.82	0.07	1.48	0.07
C_4	2.22	0.11	1.96	0.10	2.75	0.09	2.53	0.12
C_5	3.35	0.15	2.97	0.12	3.92	0.14	3.90	0.22

	Experiment							
	1a		1b		2		3	
	M	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
Weibull-UVSD								
k_o (shape)	2.87	0.07	3.15	0.10	3.34	0.12	3.23	0.08
λ_o (scale)	1.47	0.03	1.46	0.03	1.72	0.04	1.63	0.04
C_1	0.53	0.03	0.41	0.02	0.60	0.03	0.50	0.03
C_2	0.84	0.02	0.75	0.02	0.95	0.02	0.84	0.02
C_3	1.16	0.02	1.09	0.02	1.24	0.01	1.18	0.01
C_4	1.33	0.02	1.28	0.02	1.43	0.02	1.39	0.02
C_5	1.57	0.03	1.48	0.02	1.66	0.03	1.68	0.05
Lognormal-UVSD								
μ_o	0.29	0.02	0.29	0.02	0.46	0.02	0.40	0.02
σ_o	0.37	0.01	0.34	0.01	0.39	0.02	0.37	0.01
C_1	0.75	0.02	0.66	0.02	0.80	0.02	0.72	0.02
C_2	0.98	0.02	0.91	0.01	1.06	0.02	0.98	0.02
C_3	1.23	0.02	1.17	0.01	1.31	0.01	1.25	0.01
C_4	1.40	0.02	1.34	0.02	1.49	0.02	1.46	0.03
C_5	1.65	0.04	1.56	0.02	1.78	0.04	1.84	0.08
Exponential-UVSD								
λ_o (rate)	0.39	0.02	0.38	0.02	0.24	0.02	0.27	0.02
C_1	0.25	0.04	0.12	0.02	0.33	0.04	0.21	0.03
C_2	0.71	0.05	0.49	0.04	0.94	0.06	0.67	0.05
C_3	1.62	0.07	1.37	0.07	1.95	0.06	1.67	0.06
C_4	2.44	0.11	2.20	0.10	3.05	0.12	2.89	0.15
C_5	3.83	0.21	3.50	0.17	5.11	0.29	5.70	0.61
Gamma-UVSD								
k_o (shape)	1.70	0.07	1.91	0.07	1.96	0.09	1.91	0.07
θ_o (scale)	3.32	0.27	3.18	0.55	4.74	0.52	3.74	0.34
C_1	0.76	0.07	0.49	0.04	0.94	0.08	0.69	0.06
C_2	1.64	0.09	1.30	0.07	2.03	0.09	1.61	0.08
C_3	3.00	0.09	2.65	0.09	3.45	0.08	3.09	0.07
C_4	4.07	0.15	3.71	0.12	4.69	0.12	4.50	0.17
C_5	5.81	0.26	5.15	0.17	6.72	0.27	7.34	0.65

Note. The new item distribution parameters were fixed to the following values in each model. Gumbel-UVSD, μ_n (location) = 0, β_n (scale) = 1; logistic-UVSD, μ_n (location) = 0, s_n (scale) = 1; Weibull-UVSD, k_n (shape) = 3, λ_n (scale) = 1; lognormal-UVSD, $\mu_n = 0$, $\sigma_n = 0.25$; exponential-UVSD, λ_n (rate) = 1; gamma-UVSD, k_n (shape) = 2, θ_n (scale) = 1. Four participants with an extreme positive value of d' in the MSD model were not included in the calculation of the means and *SE* for the parameters of that model (one participant in Experiment 1b, three in Experiment 2).

Table 8

Percentage of HCMs produced by the 2HT, MSD and non-Gaussian-UVSD models

Exp.	2HT		MSD		Gumbel-UVSD		Logistic-UVSD		Weibull-UVSD		Lognormal-UVSD		Exponential-UVSD		Gamma-UVSD	
	M	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
1a	17.05	2.29	20.23	2.38	21.32	2.34	22.38	2.34	20.33	2.33	20.76	2.33	19.22	2.32	20.48	2.33
1b	8.29	1.38	10.52	1.45	11.72	1.65	12.67	1.65	10.59	1.60	11.03	1.63	10.70	1.38	10.71	1.62
2	14.44	2.34	19.02	2.45	19.31	2.43	20.85	2.41	17.55	2.43	18.20	2.43	18.64	2.24	17.78	2.44
3	11.42	1.89	14.77	2.08	15.46	2.05	16.77	2.04	14.03	2.02	14.58	2.04	14.46	1.87	14.22	2.04

Table 9

Goodness of fit of the 2HT, MSD and non-Gaussian-UVSD models

		Percentage of participants with non-significant G^2							
Exp.	*n*	2HT	MSD	Gumbel-UVSD	Logistic-UVSD	Weibull-UVSD	Log-normal-UVSD	Exponential-UVSD	Gamma-UVSD
1a	72	51.39	90.28	76.39	69.44	88.89	86.11	88.89	88.89
1b	72	48.61	86.11	76.39	77.78	84.72	84.72	81.94	86.11
2	70	17.14	90.00	84.29	82.86	92.86	90.00	80.00	90.00
3	71	26.76	90.14	85.92	81.69	92.96	88.73	87.32	90.14

2HT

Although our main interest was in comparing the UVSD and DPSD models, we also explored two other popular models. First, we considered the 2HT model (Snodgrass & Corwin, 1988), which is a discrete state model of recognition. The model assumes that old items are detected as "old" with probability d_o, whereas new items are detected as "new" with probability d_n. If an item is detected, it receives a high confidence decision ("sure old" if in the d_o state, and "sure new" if in the d_n state). If an item is not detected, then the probability

with which it will receive a given confidence rating is based on a guessing parameter, g_j, where j = 1…6 (e.g., with probability g_1 for the 1 "sure new" rating), and g_1-g_6 sum to 1.

For new items, the probability of a particular rating j is therefore given as

$$P(j|\, new) = d_{\mathrm{n}} r_{\mathrm{n}} + (1 - d_{\mathrm{n}}) g_j$$

where r_{n} = 1 when j = 1, and r_{n} = 0 when j = 2-6. For old items, the probability of a given rating is given as:

$$P(j|\, old) = d_{\mathrm{o}} r_{\mathrm{o}} + (1 - d_{\mathrm{o}}) g_j$$

where r_{o} = 1 when j = 6, and 0 when j = 1-5.

More complex versions of the model allow for the d_{o} and d_{n} states to give rise to intermediate confidence levels, but we did not implement these versions because they are not identifiable when there is only a single old item and a single new item condition, as is the case in our experiments (see Moran, 2016, for further details), and their parameters therefore cannot be estimated (but see Bröder et al., 2013, who fit such a model when the parameters were sufficiently constrained).

The percentage of HCMs reproduced by the model for each experiment is shown in Table 8. The model generally did not fit the data from individual participants as well as the other models (see Table 9), and did particularly poorly in Experiments 2 and 3, where, unlike the other models, it did not fit the majority of participants. Predicted 2AFC accuracy in each condition was obtained by adapting the equations described in Ma et al. (2022) but for a 1-6 ratings scale rather than binary "old" / "new" judgments. 2AFC accuracy is given for 1-3 ratings (i.e., "sure-", "medium-" and "high-confidence new") as:

$$P(correct|\,J\ =\ 1,2,3) = \frac{d_{\mathrm{n}}r_{\mathrm{n}} + 0.5(1-d_{\mathrm{n}})g_j}{d_{\mathrm{n}}r_{\mathrm{n}} + (1-d_{\mathrm{n}})g_j}$$

where $j = J$, and $r_{\mathrm{n}} = 1$ when $j = 1$, and $r_{\mathrm{n}} = 0$ when $j = 2$ or $j = 3$. 2AFC accuracy for 4-6 ratings (i.e., "sure-", "medium-" and "high-confidence old") is given as:

$$P(correct|\,J\ =\ 4,5,6) = \frac{d_{\mathrm{o}}r_{\mathrm{o}} + 0.5(1-d_{\mathrm{o}})g_j}{d_{\mathrm{o}}r_{\mathrm{o}} + (1-d_{\mathrm{o}})g_j}$$

where $r_{\mathrm{o}} = 1$ when $j = 6$, and $r_{\mathrm{o}} = 0$ when $j = 4$ or $j = 5$.

The model correctly predicted a residual memory effect for HCMs in each experiment (sees Figures 10-12), but the prediction was clearly greater than was observed empirically and outside the observed 95% CIs. It also predicted a memory effect on 6-6 trials (which was greater than that in the 1-1 condition), but again, this was clearly greater than observed empirically. The model incorrectly predicted accuracy would be at chance in the remaining 2-2, 3-3, 4-4, 5-5 conditions, due to the items appearing in such conditions arising only from guessing states. Since the model is a discrete-state model, we did not derive expected strength values for HCMs and HCCRs for Experiment 3, as we did for the other models.

Mixture Signal Detection Model

Lastly, we considered the mixture signal detection (MSD) model (DeCarlo, 2002). In this model, the studied item distribution is actually a mixture-distribution, made up of an attended (or more strongly encoded) item distribution and a non-attended (or more weakly encoded) distribution, representing the possibility that items may have been encoded in different states. Studied items are represented by the attended distribution with probability λ.

The remaining studied items are represented by a strength distribution with a lower mean value, with probability (1 - λ). The distributions are assumed to be Gaussian with equal variances. This means that the model does not have the same capacity to mis-predict the residual memory effect for HCMs as the UVSD model because, unlike the UVSD model, and as pointed out by DeCarlo (2002), the likelihood ratio is monotonic with the strength axis.

In the MSD model, when the mean of the non-attended distribution is greater than that of the new item distribution, but less than that of the attended old distribution, the model is not identifiable when there is only a single old-new item condition, as is the case in our experiments. A simplifying assumption is often made where the mean of the non-attended distribution is set equal to that of the new item distribution (e.g., DeCarlo, 2002; Spanton & Berry, 2020), and this model is identifiable with the design. The probability of a given rating *j* to an old item is given as:

$$P(j|\text{ old}) = \lambda\left(\Phi\left(C_j, d'\right) - \Phi\left(C_{j-1}, d'\right)\right) + (1-\lambda)\left(\Phi\left(C_j\right) - \Phi\left(C_{j-1}\right)\right)$$

where $j = 1\ldots6$, and $C_6 = \infty$, $C_0 = -\infty$. The probability of a given rating *j* to a new item is given as:

$$P(j|\text{ new}) = \Phi\left(C_j\right) - \Phi\left(C_{j-1}\right)$$

The parameter estimates of the model when fit to the single-item recognition data are shown in Table 7. Like the other models, it closely reproduced the percentage of HCMs in each experiment (Table 8). It also tended to fit the majority of participants (Table 9). As is shown in Figures 10-12, the model predicted the residual memory effect for HCMs in the 1-1 condition in Experiment 1a (M = 59.54%), $t(71) = 16.18$, $p < .001$, Experiment 1b (M =

59.37%), $t(71) = 14.62$, $p < .001$, and Experiment 2 ($M = 64.84\%$), $t(69) = 22.08$, $p < .001$, with the predicted effects falling within the observed 95% CIs. It also predicted the other trends in the accuracy data reasonably well. Similarly, it predicted the residual memory effect in Experiment 3 (Figure 13), $t(70) = 13.37$, $p < .001$, and the mean predicted difference in strength to HCMs and HCCRs was nonnegligible ($M = 0.17$).

Figure 10

Predicted 2AFC accuracy according to the 2HT, MSD, and non-Gaussian-UVSD models when fit to the data from Experiment 1a shown in Figure 7.

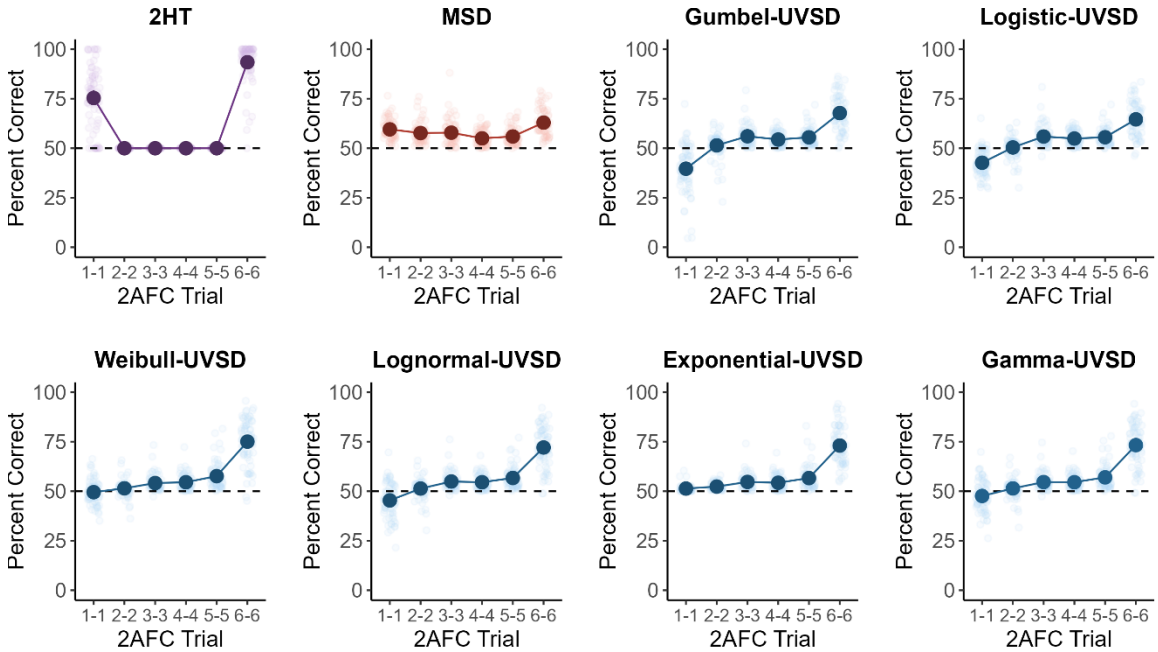


Figure 11

Predicted 2AFC accuracy of the 2HT, MSD, and non-Gaussian-UVSD models when fit to the data from Experiment 1b shown in Figure 7.

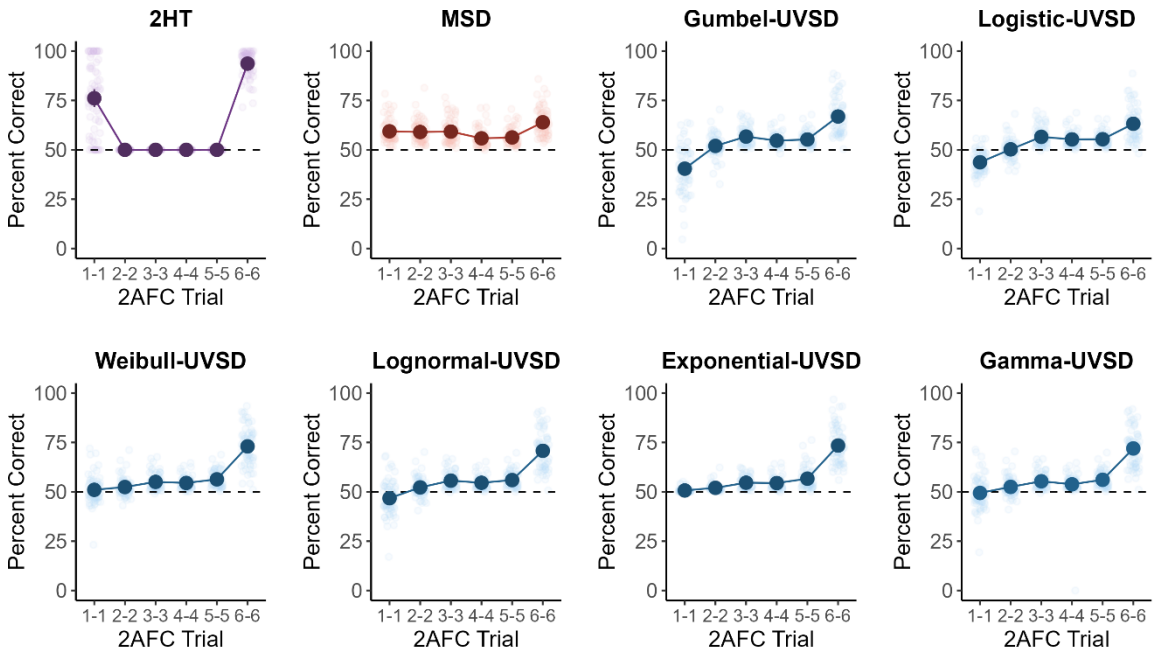


Figure 12

Predicted 2AFC accuracy of the 2HT, MSD, and non-Gaussian-UVSD models when fit to the data from Experiment 2 shown in Figure 7.

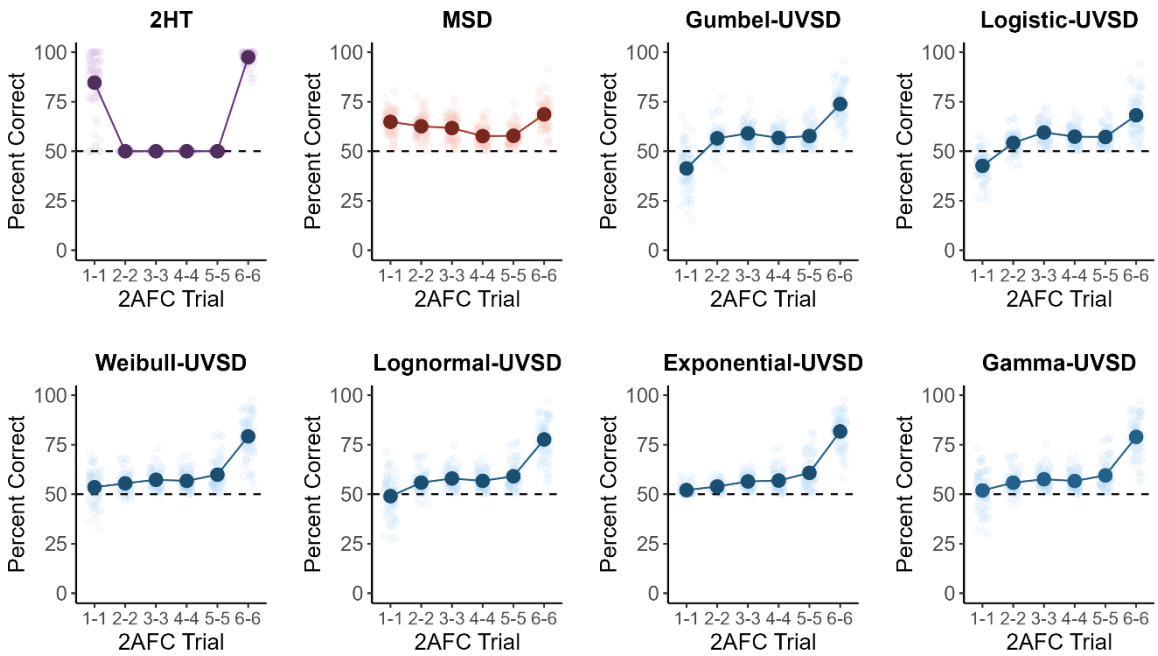
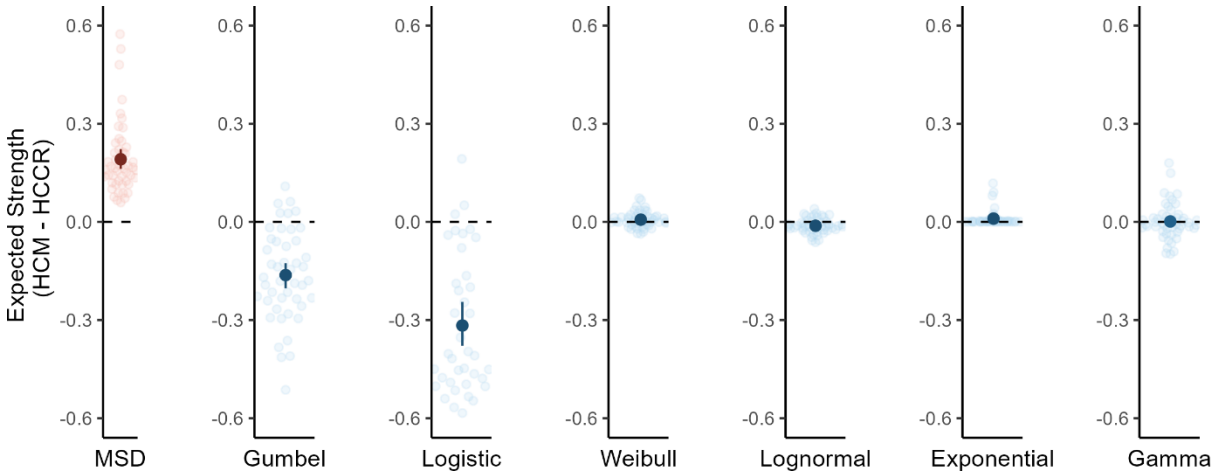


Figure 13

Expected difference in strength values of HCMs and HCCRs in the MSD and non-Gaussian-UVSD models when fit to the data of Experiment 3 shown in Figure 9.



General Discussion

We investigated whether a residual memory effect for HCMs exists and the implications of such an effect for the phenomenon of everyday amnesia and decision models of recognition. Residual memory for HCMs was found in 1) a modified 2AFC task in which the alternatives are matched in terms of their previous single-item recognition response (Experiment 1a, 1b, 2), and 2) a second single-item recognition task in which HCMs and HCCRs must be discriminated (Experiment 3). The effect was demonstrated with face stimuli (Experiments 1a, 2, 3) and also word stimuli (Experiment 2), and under study instructions to memorise the study stimuli (Experiments 1a and 1b) and under more demanding encoding conditions requiring a semantic decision to each studied stimulus (Experiments 2 and 3).

Implications for Everyday Amnesia

Although participants may say with total confidence that they do not remember previously studying an item, and in this sense complete and rapid forgetting of the item's

presentation during the study phase (and hence everyday amnesia) can be said to have occurred (Roediger & Tekin, 2020), participants are still able to reliably discriminate such items from HCCRs, which are matched in terms of the previous response received. This demonstrates that HCMs are not permanently lost from memory or inaccessible, and that the forgetting that occurs for HCMs in the first single-item recognition phase is due to retrieval failure and not necessarily the loss of a fully processed item from memory (Miller, 2021).

Implications for Decision Models

The residual memory effect for HCMs discriminated between dominant decision models of recognition. Importantly, the experiments provided strong tests (Platt, 1964) of the models as parameter-free predictions were derivable. When fit to the single-item recognition data, the UVSD model closely reproduced the proportion of HCMs, but when using these same parameter estimates to derive predictions for residual memory effects for HCMs, it incorrectly predicted either a sub-chance effect (Experiments 1a and 1b) or no effect (Experiment 2). The misprediction arises because, when the value of σ_o (the variance of the old item distribution) is greater than that of σ_n (the variance of the new item distribution), the expected strength of HCMs can actually be lower than that of HCCRs. This prediction is counterintuitive from a psychological perspective; it is not clear why studying an item would endow it with a lower strength value, relative to a non-studied item. Despite this feature being often acknowledged, this has not prevented the widespread adoption of the model over the past few decades (Egan, 1958; see Rotello, 2017; Wixted, 2007, for reviews), presumably due to its successes in accounting for other aspects of recognition data. To our knowledge though, evidence relevant to this specific feature of the model does not exist, and our findings therefore constitute the first direct evidence against it: our findings imply that the mean strength of HCMs is greater than that of HCCRs, not lower (or equal). Moreover, the UVSD model was still unable to sufficiently predict the effect in each experiment when non-

Gaussian distributions (Gumbel, logistic, lognormal, Weibull, exponential, gamma) were assumed.

Other models, however, did not suffer from the same problem as the UVSD model. The DPSD and MSD models successfully predicted the residual memory effect for HCMs in each experiment. Unlike the UVSD model, HCMs (and HCCRs) in these models are the product of an equal variance signal detection process, meaning that the negative effect for HCMs will not occur because the likelihood ratio is monotonic with the strength axis in these models. The expected strength of HCMs will always be greater than that of HCCRs whenever the mean strength of old items is greater than that of new items.

The 2HT model was also fit to the data and was able to predict the memory effect for HCMs, but substantially overestimated it (likewise for accuracy in the 6-6 condition). It also incorrectly predicted chance-levels of accuracy in the other 2AFC conditions. Future research should explore whether more complex versions of the 2HT model make more accurate predictions in extended experimental designs (e.g., with multiple old item conditions) that allow the parameters of more complex versions of the model to be identified (Moran, 2016).

Related recent studies

Others have recently used a similar approach to the one we have taken here. Ma et al. (2022) identified competing predictions of the UVSD and 2HT models in a single-item recognition paradigm with old/new ratings and where a payoff manipulation was used to manipulate response bias. They found that neither model satisfactorily predicted the relative bias effects on forced-choice accuracy. The DPSD model was found to produce a better quantitative prediction for the effect, but their data did not distinguish between competing qualitative predictions of the models in the way that ours do.

Dobbins (2023) recently found evidence against the UVSD model using a three-alternative forced choice (3AFC) test, in which participants must select the new item when

presented with two alternatives that are old items. Once fit to single-item recognition data, estimates of σ_o in the UVSD model were found to be positively correlated with 3AFC accuracy across participants, whereas the model actually predicted a negative correlation. Accuracy was, however, positively correlated with estimates of R_o in the DPSD model, as predicted by this model. Furthermore, estimates of σ_o and d in the UVSD model were positively correlated across participants.[1] This correlation would occur if, as σ_o increases, estimates of d become greater to compensate for the depression in the upper portion of the ROC that results. Interestingly, σ_o and d were also positively correlated when the UVSD model was fit to data generated from the DPSD model. In this sense, the DPSD model predicted the UVSD model's mispredictions. As in our study, the UVSD model's mispredictions were made despite providing slightly better fits to the data than the DPSD model. This again demonstrates the value of deriving and testing the predictions of models in additional tasks as we have done here, rather than relying solely on the goodness of fit of the models.

These two studies, together with ours, converge on highlighting problems for the UVSD model in predicting performance in additional tasks once its parameters are fixed by first fitting it to single-item recognition data. Furthermore, recent attempts to validate a psychological explanation offered for the unequal-variance assumption in the UVSD model in terms of encoding variability have proven unsuccessful (Spanton & Berry, 2020, 2022). Thus, although the UVSD model has been popular for several decades, these recent studies along with our findings here highlight major issues for the model.

Everyday amnesia and organic amnesia

Roediger and Tekin (2020, p. 6) defined amnesia as rapid and complete forgetting and argued that the occurrence of HCMs in memory-intact individuals implies that amnesia "occurs in all people (not just amnesia patients)" (p. 1). In light of the findings reported here,

to what extent is it reasonable to maintain this degree of continuity between everyday and organic amnesia?

If forgetting in explicit memory must be complete in order to qualify as amnesia, then the present findings clearly challenge Roediger and Tekin's (2020) claim. Residual memory exists for forgotten items, even ones classified as new with complete confidence. Although left largely implicit, this was in effect the whole point of the critiques of Levi et al. (2022) and Goshen-Gottstein et al. (2022). To the extent that SDT can explain HCMs, it does so by regarding these items simply as ones which fall below the relevant decision criterion; with a different criterion placement, these items would receive a different rating. Levi et al. (2022) and Goshen-Gottstein et al. (2022) were incorrect in suggesting that the UVSD model could explain the properties of HCMs, but the more general point that SDT (particularly as instantiated in the DPSD model) explains HCMs in terms of residual memory is correct, as we have shown here.

But just because the occurrence of HCMs in memory-intact individuals fails to meet Roediger and Tekin's definition of amnesia does not mean that it is unrelated to organic amnesia. Indeed a considerable body of work attempts to conceptualize organic amnesia precisely in terms of the DPSD model (see e.g., Yonelinas et al., 2010; Yonelinas et al., 2022, for reviews). Moreover, as we have argued elsewhere, complete forgetting in individuals with amnesia (that is, recognition at chance) occurs only in rare cases and can often be attributed to insufficient power to detect small residual memory effects (see Berry et al., 2014, for discussion; see also Wixted & Squire, 2004).

After outlining an SDT-based account of organic amnesia, in which the old and new item distributions are largely overlapping, Roediger and Tekin asked: "Will the scientific world accept the SDT-based explanations of anterograde amnesia… proposed here? We suspect not. SDT provides a useful conceptualization of the underlying memory signals and

the decision criteria, not a theoretical explanation in terms of psychological constructs or neural processes of why they are depicted as they are. Likewise, we do not find the SDT interpretation of everyday amnesia to be an explanation, for the same reasons." We disagree. SDT explains recognition memory phenomena including HCMs in term of theoretical constructs (signal and noise distributions, decision criteria) and the DPSD model explains the occurrence and detailed nature of residual memory, whereas the UVSD model does not. These explanations are equally applicable to both organic and everyday amnesia.

Potential limitations

A potential limitation of our findings is that, given that we used conventional procedures for testing recognition memory—namely, a single study-test phase separated by a retention interval—it is possible that the underlying memory state or strength of an item may have changed from its presentation in the first recognition test to the second (see also Ma et al., 2022, for a discussion of this issue). Memory for HCMs could, for example, have been weaker with the longer retention interval, or items could move from a detect state to a not detected state (from the perspective of the 2HT model), or from the attended item distribution to the unattended item distribution (from the perspective of the MSD model). We adopted a conventional single-item recognition design because we wanted to replicate the percentage of HCMs under the same conditions reported by Roediger and Tekin (2020). In the future, however, the predictions of other instantiations of these models that allow for the memory of individual items to change from one phase to the next could be explored. If, for example, memory weakens over time, and the variance of the old relative to the new item distribution is linked to overall strength as some have observed (e.g., Spanton & Berry, 2020, 2022), then the UVSD model predictions may be more similar to those that would be made under equal variance assumptions, and the model may therefore not mis-predict the residual memory effect for HCMs.

An issue with this account, however, is that, aside from being *post hoc*, it would need to be subjected to empirical test by devising a method to confirm that the value of σ_o is equivalent to that of σ_n in the additional memory test, despite the fact that studied and non-studied items can be discriminated reasonably well in this test, as our experiments showed. For example, overall accuracy across 2AFC conditions in Experiment 2 was 64.11% correct, the mean difference in the hit and false alarm rate in the second single-item recognition phase of Experiment 3 was 0.14, and these values would likely have been much greater had the alternatives not been matched for their previous single-item recognition response. Moreover, this type of *post hoc* explanation is not necessary to explain the ability of the DPSD and MSD models to successfully predict the residual memory effect for HCMs.

Finally, it is also possible that the first recognition test influenced performance on the second recognition test by acting as an additional learning episode. The UVSD model might be able to account for the residual memory effect for HCMs once the increment in strength that occurs for items as a result of their presentation in the first test is taken into account. If the increment is inversely related to an item's strength (Bjork & Bjork, 1992; Storm et al., 2008), then the increment for HCMs might be expected to be greater than that for HCCRs, and their strength for the second test will then be more likely to be greater than that of HCCRs, giving rise to the residual memory effect for HCMs. A problem with this account is that, for the other single-item recognition rating categories (e.g., items receiving a "2 – medium confidence new" rating), the expected strength of non-studied items is *lower* than that of studied items under the UVSD model (discussed further in Lee et al., 2024). If the increment these items receive from the first test is also inversely related to strength, the non-studied items would be expected to receive a greater strength increment than the studied items receiving the same rating. Expected 2AFC accuracy in all but the 1-1 condition ought then to be *below* 50%, yet this is not what we observed in our experiments. Indeed, in

Experiment 2, accuracy was reliably above 50% in all 2AFC conditions. Thus, this explanation of the residual memory effect for HCMs in terms of the UVSD model comes at the expense of mispredicting accuracy in the other 2AFC conditions and is therefore implausible.

Conclusions

Our conclusions are twofold: First, the residual memory effect for HCMs demonstrates that even though a studied item receives a "new" decision with total confidence in a recognition test, memory of the item is not completely lost. If given another opportunity, be it in a 2AFC task or additional single-item recognition task, participants can reliably distinguish these items from HCCRs. Second, once the parameters of the UVSD and DPSD models were fixed by fitting them to the single-item recognition data, they made opposing predictions. Specifically, the residual memory effect for HCMs was not predicted by the UVSD model, which instead tended to predict a sub-chance effect or an absence of an effect, providing evidence against this model. In contrast, the DPSD model did predict the effect, and this is due to the equal variance signal detection process assumed to give rise to HCMs and HCCRs. For the same reason, in additional modelling, the MSD model was found to correctly predict a residual memory effect for HCMs. The 2HT model also predicted the effect, but tended to overpredict it and incorrectly predicted an absence of memory on all other 2AFC trial types except 6-6 trials. The residual memory effect for HCMs therefore distinguished between decision models of recognition and provides a new benchmark for testing such models.

Constraints on Generality

We observed the residual memory effect for HCMs using a variety of stimuli (words, faces), and recognition tasks (2AFC and single-item recognition tasks). The stimuli were selected from the same sources used by Tekin and Roediger (2017) (words from Nelson et

al., 2004; faces from Minear & Park, 2004), whose data, like ours, demonstrate the phenomenon of everyday amnesia (HCMs). Although we expect our findings to generalize to other types of stimuli typically used in recognition tasks (e.g., pictures of objects or scenes), it will be important to demonstrate this in future research. As in Tekin and Roediger (2017), our participants were adults in higher education—all were students at the University of Plymouth, and almost all were studying psychology—and tended to be younger (18-32 years). Although we expect our findings to generalise to other age groups and similar participant pools, it will be important to demonstrate this in future research. We have no reason to believe that the results depend on other characteristics of the participants, materials, or context.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Caski (Eds.), *Second International Symposium on Information Theory* (pp. 267–281). Budapest, Hungary: Academiai Kiado.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon project. *Behavior Research Methods, 39*(3), 445-459. https://doi.org/10.3758/BF03193014

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1-48. https://doi:10.18637/jss.v067.i01

Berry, C. J., Kessels, R. P. C., Wester, A. J., & Shanks, D. R. (2014). A single-system model predicts recognition memory and repetition priming in amnesia. *Journal of Neuroscience, 34*, 10963-10974. https://doi.org/10.1523/jneurosci.0764-14.2014

Berry, C. J., & Shanks, D. R. (2023). *Memory for high confidence misses*. Retrieved from https://osf.io/2q5yw/

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), From Learning Processes to Cognitive Processes: Essays in Honor of William K. Estes (Vol. 2, pp. 35-67). Hillsdale, NJ: Erlbaum.

Bröder, A., Kellen, D., Schütz, J., & Rohrmeier, C. (2013). Validating a two-high-threshold measurement model for confidence rating data in recognition. *Memory, 21*(8), 916–944. https://doi.org/10.1080/09658211.2013.767348

Bröder, A., & Schütz, J. (2009). Recognition ROCs are curvilinear-or are they? On premature arguments against the two-high-threshold model of recognition. *Journal of*

Experimental Psychology: Learning, Memory, and Cognition, *35*(3), 587–606. https://doi.org/10.1037/a0015279

Busemeyer, J. R., & Wang, Y.-M. (2000). Model comparisons and model selections based on generalization criterion methodology. *Journal of Mathematical Psychology*, *44*(1), 171–189. https://doi.org/10.1006/jmps.1999.1282

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multi-model inference*. (2nd ed.). Springer-Verlag. https://doi.org/10.1007/b97636

DeCarlo, L. T. (2002). Signal detection theory with finite mixture distributions: Theoretical developments with applications to recognition memory. *Psychological Review*, *109*(4), 710-721. https://doi.org/10.1037/0033-295X.109.4.710

DeSoto, K. A., & Roediger III, H. L. (2014). Positive and negative correlations between confidence and accuracy for the same events in recognition of categorized lists. *Psychological Science*, *25*(3), 781-788. https://doi.org/10.1177/0956797613516149

Dobbins, I. (2022). Hindsight and the theories of signal detection: Commentary on Levi, Mickes and Goshen-Gottstein (2022). *Neuropsychologia*, *166*, 1-3. https://doi.org/10.1016/j.neuropsychologia.2021.108121

Dobbins, I. G. (2023). Recognition receiver operating characteristic asymmetry: Increased noise or information? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 49*(2), 216-229. https://doi.org/10.1037/xlm0001224

Dubé, C. (2023). ROC measures of memory accessibility. *Quarterly Journal of Experimental Psychology*, *76*(4) 881-887. https://doi.org/10.1177/17470218221113559

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

Glanzer, M., Hilford, A., & Maloney, L. T. (2009). Likelihood ratio decisions in memory: Three implied regularities. *Psychonomic Bulletin & Review*, *16*(3), 431-455. https://doi.org/10.3758/PBR.16.3.431

Goshen-Gottstein, Y., Levi, A., & Mickes, L. (2022). Signal-detection theory separates the chaff of bias from the wheat of memory: Illuminating the triviality of high-confidence judgments. *Neuropsychologia*, *166*, 1–4. https://doi.org/10.1016/j.neuropsychologia.2021.108116

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

Hartig, F. (2022). *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models*. R package version 0.4.6, https://CRAN.R-project.org/package=DHARMa

Hautus, M. J., MacMillan, N. A., & Creelman, C. D. (2022). *Detection theory: A user's guide*. Routledge.

Kellen, D., Winiger, S., Dunn, J. C., & Singmann, H. (2021). Testing the foundations of signal detection theory in recognition memory. *Psychological Review*, *128*(6), 1022–1050. https://doi.org/10.1037/rev0000288

Lange, K., Kühn S., & Filevich E. (2015). "Just Another Tool for Online Studies" (JATOS): An Easy Solution for Setup and Management of Web Servers Supporting Online Studies. *PLoS ONE*, *10*(6): e0130834. https://doi.org/10.1371/journal.pone.0130834

Lee, D. Y. H., Berry, C. J., & Shanks, D. R. (2024). *Kelley's Paradox and strength skewness in research on unconscious mental processes* [Manuscript submitted for publication].

Lee, D. Y. H., & Shanks, D. R. (2023). Conscious and unconscious memory and eye movements in context-guided visual search: A computational and experimental reassessment of Ramey, Yonelinas, and Henderson (2019). *Cognition*, *240*, 105539. https://doi.org/10.1016/j.cognition.2023.105539Levi, A., Mickes, L., & Goshen-

Gottstein, Y. (2022). The new hypothesis of everyday amnesia: An effect of criterion placement, not memory. *Neuropsychologia*, *166*, 1-3. https://doi.org/10.1016/j.neuropsychologia.2021.108114

Ma, Q., Starns, J. J., & Kellen, D. (2022). Bias effects in a two-stage recognition paradigm: A challenge for "pure" threshold and signal detection models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 48*(10), 1484-1506. https://doi.org/10.1037/xlm0001107

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314–324. https://doi.org/10.3758/s13428-011-0168-7

Miller, R. R. (2021). Failures of memory and the fate of forgotten memories. *Neurobiology of Learning and Memory*, *181*, 16. https://doi.org/10.1016/j.nlm.2021.107426

Minear, M., & Park, D. C. (2004). A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments & Computers, 36*(4), 630-633. https://doi.org/10.3758/BF03206543

Moran, R. (2016). Thou shalt identify! The identifiability of two high-threshold models in confidence-rating recognition (and super-recognition) paradigms. *Journal of Mathematical Psychology*, *73*, 1-11. https://doi.org/10.1016/j.jmp.2016.03.002

Morey, R., & Rouder, J. (2022). *BayesFactor: Computation of Bayes Factors for Common Designs*. R package version 0.9.12-4.4, https://CRAN.R-project.org/package=BayesFactor

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments & Computers, 36*(3), 402-407. https://doi.org/10.3758/BF03195588

Parks, C. M., & Yonelinas, A. P. (2007). Moving beyond pure signal-detection models: Comment on Wixted (2007). *Psychological Review*, *114*(1), 188–201. https://doi.org/10.1037/0033-295X.114.1.188

Platt, J. R. (1964). Strong Inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others. *Science*, *146*(3642), 347-353. https://doi.org/10.1126/science.146.3642.347

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Ratcliff, R., Sheu, C.-F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review*, *99*(3), 518–535. https://doi.org/10.1037/0033-295X.99.3.518

Roediger III, H. L., & Tekin, E. (2020). Recognition memory: Tulving's contributions and some new findings. *Neuropsychologia*, *139*, 9. https://doi.org/10.1016/j.neuropsychologia.2020.107350

Roediger III, H. L., & Dobbins, I. (2022). Predicting and "predicting" high confidence misses. *Neuropsychologia*, *166*, 1-4. https://doi.org/10.1016/j.neuropsychologia.2021.108117

Roediger III, H. L., & Tekin, E. (2022). Can signal detection theory explain everyday amnesia (high confident misses)? *Neuropsychologia*, *166*, 3. https://doi.org/10.1016/j.neuropsychologia.2021.108115

Rotello, C. M. (2017). Signal detection theories of recognition memory. In J. H. Byrne & J. T. Wixted (Eds.), *Learning and memory: A comprehensive reference*, Vol. 2: *Cognitive Psychology of Memory* (2nd ed., pp. 201–226). Academic Press. https://doi.org/10.1016/b978-0-12-809324-5.21044-4

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, 117(1), 34–50. https://doi.org/10.1037/0096-3445.117.1.34

Spanton, R. W., & Berry, C. J. (2020). The unequal variance signal-detection model of recognition memory: Investigating the encoding variability hypothesis. *Quarterly Journal of Experimental Psychology*, *73*(8), 1242-1260. https://doi.org/10.1177/1747021820906117

Spanton, R. W., & Berry, C. J. (2022). Does variability in recognition memory scale with mean memory strength or encoding variability in the UVSD model? *Quarterly Journal of Experimental Psychology*. https://doi.org/10.1177/17470218221136498

Squire, L. R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review*, *99*(2), 195-231. https://doi.org/10.1037/0033-295X.99.2.195

Squire, L. R., Stark, C. E. L., & Clark, R. E. (2004). The medial temporal lobe. *Annual Review of Neuroscience*, *27*, 279-306. https://doi.org/10.1146/annurev.neuro.27.070203.144130

Starns, J. J., Dubé, C., & Frelinger, M. E. (2018). The speed of memory errors shows the influence of misleading information: Testing the diffusion model and discrete-state models. *Cognitive Psychology*, *102*, 21–40. https://doi.org/10.1016/j.cogpsych.2018.01.001

Stretch, V., & Wixted, J. T. (1998). Decision rules for recognition memory confidence judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1397-1410. https://doi.org/10.1037/0278-7393.24.6.1397

Storm, B. C., Bjork, E. L., & Bjork, R. A. (2008). Accelerated relearning after retrieval-induced forgetting: The benefit of being forgotten. *Journal of Experimental*

Psychology: Learning, Memory, and Cognition, 34(1), 230–236. https://doi.org/10.1037/0278-7393.34.1.230

Tekin, E., & Roediger III, H. L. (2017). The range of confidence scales does not affect the relationship between confidence and accuracy in recognition memory. *Cognitive Research: Principles and Implications*, *2*, 1-13. https://doi.org/10.1186/s41235-017-0086-z

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

Wixted, J. T. (2020). The forgotten history of signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(2), 201-233. https://doi.org/10.1037/xlm0000732

Wixted, J. T., & Mickes, L. (2010). Useful scientific theories are useful: A reply to Rouder, Pratte, and Morey (2010). *Psychonomic Bulletin & Review*, *17*(3), 436–442. https://doi.org/10.3758/PBR.17.3.436

Wixted, J. T., & Squire, L. R. (2004). Recall and recognition are equally impaired in patients with selective hippocampal damage. *Cognitive, Affective & Behavioral Neuroscience*, *4*(1), 58–66. https://doi.org/10.3758/CABN.4.1.58

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*(6), 1341–1354. https://doi.org/10.1037/0278-7393.20.6.1341

Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. *Psychological Bulletin*, *133*(5), 800–832. https://doi.org/10.1037/0033-2909.133.5.800

Yonelinas, A. P., Aly, M., Wang, W.-C., & Koen, J. D. (2010). Recollection and familiarity: Examining controversial assumptions and new directions. *Hippocampus*, *20*(11), 1178–1194. https://doi.org/10.1002/hipo.20864

Yonelinas, A. P., Ramey, M. M., & Riddell, C. (2022). Recognition memory: The role of recollection and familiarity. In M. J. Kahana & A. D. Wagner (Eds.), *The Oxford handbook of human memory*. Oxford University Press.

Footnotes

[1] Estimates of σ_o and *d* were similarly positively correlated in each of our experiments ($0.42 < r\text{s} < 0.73$). Following Dobbins (2023), we also explored whether estimates of the memory evidence parameters in the UVSD and DPSD models were correlated with performance in the additional task, specifically 1-1 accuracy. This analysis did not shed further light on our model results though. In the UVSD model, although estimates of σ_o were strongly negatively correlated with predicted 1-1 accuracy across individuals ($r\text{s} < -0.73$ in Experiments 1a, 1b, and 2), they were not reliably correlated with observed 1-1 accuracy ($0.08 < r\text{s} < 0.16$). Estimates of *d* were also not correlated with predicted 1-1 accuracy ($-0.14 < r\text{s} < 0.02$). In the DPSD model, estimates of *d′* were strongly positively correlated with predicted 1-1 accuracy ($r\text{s} > 0.92$), but not with observed 1-1 accuracy ($-0.14 < r\text{s} < 0.28$). Estimates of R_o were also not consistently correlated with 1-1 accuracy across experiments ($-0.06 < r\text{s} < 0.3$).