AMERICAN PSYCHOLOGICAL ASSOCIATION

Neuropsychology


ISSN: 0894-4105

2024, Vol. 38, No. 1, 58–68
https://doi.org/10.1037/neu0000910

# Multitrial Free Recall for Evaluating Memory

R. T. Adrogue[1], N. Herz[1], D. J. Halpern[1], J. Tracy[2], and M. J. Kahana[1]
[1] Department of Psychology, University of Pennsylvania
[2] Department of Neuropsychology, Thomas Jefferson University

***Objective:*** Much of our knowledge concerning the neural basis of human memory derives from lab-based verbal recall tasks. Outside of the lab, clinicians use validated and normed neuropsychological tests to assess patients' memory function and to evaluate clinical interventions. Here we sought to establish the clinical validity of examining memory through multitrial free recall of semantically organized and unrelated word lists. ***Method:*** We compare memory performance in multitrial free recall tasks with the Rey Auditory Verbal Learning Test and the California Verbal Learning Test, two common neuropsychological tests aimed at evaluating memory function in clinical settings. We compare predictive validity between the tasks by evaluating deficits in a patient sample and examining age-related declines in memory. We additionally compare test–retest reliability, establish convergent validity, and show the emergence of common recall dynamics between the tasks. ***Results:*** We demonstrate that both laboratory free recall tasks have better predictive validity and test–retest reliability than the established neuropsychological tests. We further show that all tasks have good convergent validity and reveal core memory processes, including temporal and semantic organization. However, we also demonstrate the benefits of repeated trials for evaluating the dynamics of memory search and their neuropsychological sequelae. ***Conclusions:*** These results provide evidence for the clinical validity of lab-based multitrial free recall tasks and highlight their psychometric benefits over neuropsychological measures. Based on these results, we discuss the need to bridge the gap between clinical understanding of putative mechanisms underlying memory disorders and neuroscientific findings obtained using lab-based free recall tasks.

***Key Points***
***Question:*** How do the reliability and validity of laboratory-based recall paradigms compare with traditional neuropsychological recall measures? ***Findings:*** Laboratory-based recall paradigms have strong internal validity, test–retest reliability, and convergent validity with established neuropsychological tests. ***Importance:*** These data support the clinical validity of multitrial free recall tests with unique word lists. ***Next Steps:*** Identifying the neural correlates of behavioral deficits in recall tasks can offer novel insights into clinical disorders resulting in impaired memory performance.

*Keywords:* free recall, categorized free recall, Rey Auditory Verbal Learning Test, California Verbal Learning Test

*Supplemental materials:* https://doi.org/10.1037/neu0000910.supp



This article was published Online First October 23, 2023.
R. T. Adrogue https://orcid.org/0000-0002-8851-756X
M. J. Kahana https://orcid.org/0000-0001-8122-9525
The authors express their gratitude to the patients who selflessly volunteered to participate in this experiment. This work was supported by the Foundation for the National Institutes of Health (Grant R01 NS 106611) awarded to M. J. Kahana and also by the Defense Advanced Research Projects Agency Restoring Active Memory program (Cooperative Agreement N66001-14-2-4032).

The views, opinions, and/or findings contained in this material are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government. M. J. Kahana holds a greater than 5% equity interest in Nia Therapeutics Inc., a company intended to develop and commercialize brain stimulation therapies for memory restoration.

R. T. Adrogue played a lead role in data curation, formal analysis, investigation, software, and visualization and an equal role in methodology, project administration, writing–original draft, and writing–review and editing. N. Herz played a supporting role in formal analysis and visualization and an equal role in conceptualization, data curation, investigation, methodology, supervision, validation, writing–original draft, and writing–review and editing. D. J. Halpern played a supporting role in formal analysis and writing–original draft and an equal role in data curation, methodology, project administration, software, supervision, validation, visualization, and writing–review and editing. J. Tracy played a supporting role in formal analysis, investigation, resources, and writing–review and editing and an equal role in validation. M. J. Kahana played a lead role in conceptualization, funding acquisition, methodology, resources, supervision, validation, writing–original draft, and writing–review and editing; a supporting role in data curation, formal analysis, project administration, software, and visualization; and an equal role in investigation.

Correspondence concerning this article should be addressed to M. J. Kahana, Department of Psychology, University of Pennsylvania, Levin Building, Room 263, 425 South University Avenue, Philadelphia, PA 19104, United States. Email: kahana@psych.upenn.edu

Memory exhibits marked variability both within and across individuals. Whereas clinical studies focus on differences in memory across individuals and their relation to brain injury or neurological disease (e.g., Vakil, 2005; Weissberger et al., 2017; Wright & Persad, 2007), neuroscientific studies of memory contrast good and poor memory states within a given individual, across trials or conditions (e.g., Hanslmayr et al., 2007; Paller & Wagner, 2002; Sederberg et al., 2003). To identify neural correlates of variable memory, a subject must contribute data across many trials. Sorting these trials into remembered and unremembered categories allows researchers to identify the neural correlates of mnemonic success. Although this repeated testing methodology represents the norm in cognitive neuroscience, neuropsychological memory measures typically entail just one or two distinct word lists. In the Rey Auditory Verbal Learning Test (RAVLT), subjects attempt to freely recall a list of auditorily presented items across five learning trials. After studying and recalling an "interference" list, they attempt to recall the original list both immediately and following a delay. The California Verbal Learning Test (CVLT) follows a very similar procedure but with lists of semantically categorized items (Delis et al., 2000). Neuropsychologists use the scores derived from these tests to evaluate the severity of a patient's disability and/or their response to treatment (Estévez-González et al., 2003).

The discord in methodologies used across these allied disciplines—cognitive neuroscience and neuropsychology—limits inferences about the clinical value of experimental studies. We therefore seek to evaluate the psychometric properties and determine the clinical validity of two variants of delayed free recall commonly used in neuroscientific investigations of memory. Specifically, we consider delayed free recall of lists comprising unrelated common words (similar to the RAVLT) or words drawn from a small number of taxonomic categories (similar to the CVLT). Prior studies have established clear neural correlates of successful memory encoding and retrieval in both of these tasks (Fernández et al., 1999; Kim, 2011; Kragel et al., 2017; Strange et al., 2002; Weidemann et al., 2019; Xue et al., 2010) and have indicated that closed-loop electrical stimulation during the encoding phase of these tasks can boost subsequent recall (Ezzyat et al., 2018; Kucewicz et al., 2018; Kahana et al., 2023). Establishing the clinical validity of memory measures obtained by these methods could elucidate the utility of such brain stimulation in the clinical realm.

To bridge the gap between clinical and laboratory approaches to the measurement of episodic memory, we used an online platform to collect data on both sets of tasks from a large community-based cohort of individuals varying in age and educational status. We also analyzed data from epilepsy patients who had thorough neuropsychological assessments and who also took part in repeated experiments involving delayed free recall of unrelated (FR) and categorized word lists (CatFR). This allowed us to directly compare the same subjects' performance on the delayed free recall tasks with their performance on the RAVLT and CVLT.

# Method

## Subjects

The present study includes two samples: (a) a community cohort recruited through Amazon's Mechanical Turk (MTurk), an online crowd-sourcing platform, and (b) a sample of patients undergoing neurosurgical evaluation for treatment of their pharmacoresistant epilepsy. Before participating, both cohorts completed an informed consent form that was approved by the University of Pennsylvania institutional review board. The MTurk cohort comprises 1,076 individuals who contributed one or more data sessions across two different experiments: (a) a four-session experiment with two sessions of RAVLT and two sessions of FR and (b) a single-session experiment involving one of three tasks: FR, CatFR, or CVLT. We programed these tasks using the JsPsych library (de Leeuw, 2015) and administered them using psiTurk (Eargle et al., 2021; Gureckis et al., 2016), an open-source library created to interface with Amazon's MTurk (Buhrmester et al., 2011). All subjects in the MTurk cohort reported English as their native language and reported residing in the United States. Subjects were not asked about their medical or mental status and thus represent a nonselected community sample.

The clinical cohort consists of 479 epilepsy patients who enrolled in the Defense Advanced Research Projects Agency-sponsored Restoring Active Memory project at collaborating clinical centers (see Adamovich-Zeitlin et al., 2021; Weidemann et al., 2019, for details on the patient sample) 281 subjects completed the FR task and 242 subjects completed the CatFR task. In addition, neuropsychologists at collaborating sites reported CVLT scores for 158 subjects and RAVLT scores for 92 subjects. A subset of patients performed both the laboratory and neuropsychology tests (RAVLT and FR: $n = 27$, RAVLT and CatFR: $n = 30$, CVLT and FR: $n = 94$, CVLT and CatFR: $n = 79$). We determined the administration of FR or CatFR based on research priorities at the time that patients were admitted to the hospital. The neuropsychologist evaluating the patient determined whether to administer the RAVLT or the CVLT. The institutional review board at the University of Pennsylvania approved the research studies reported here.

Table 1 includes demographic information for both the clinical and the community cohorts. Table 1 illustrates that both age and gender were similarly distributed across the two samples: age; $\chi^2(5, N = 1{,}509) = 7.19$, $p = .21$ and gender; $\chi^2(1, N = 1{,}514) = 0.91$, $p = .34$. However, MTurk subjects achieved significantly higher educational levels compared with hospital patients, $\chi^2(7, N = 1{,}398) = 675.64$, $p < .001$. The table also shows differences between the two samples in the representation of different races and ethnicities, $\chi^2(6, N = 1{,}473) = 113.17$, $p < .001$, with a larger percentage of Hispanic subjects in the MTurk study and a larger percentage of Black subjects in the hospital study. Table 2 includes a description of the clinical sample's epileptic foci.

## Justification of Sample Demographics

We used samples of convenience in this study. Any differences in demographics or epilepsy statuses are attributable to the subjects available to our study at the time.

## Tasks

### FR and CatFR

The FR and CatFR delayed free recall tasks follow the same procedures as described in Ezzyat et al.'s (2018) study and Weidemann et al.'s (2019) study. Each trial of the task consists of a 10-s countdown, followed by a visual presentation of 12 words.

**Table 1**
*Demographic Information*

| Characteristic | MTurk | Hospital |
|---|---|---|
| *n* | 1,076 | 479 |
| Gender | | |
| Male | 45.6% | 51.8% |
| Female | 47.4% | 48.2% |
| Not reported | 7.0% | 0.0% |
| Race and ethnicity | | |
| White Hispanic | 12.2% | 4.6% |
| White non-Hispanic | 70.0% | 72.0% |
| Black Hispanic | 1.5% | 0.2% |
| Black non-Hispanic | 2.4% | 14.8% |
| Asian | 3.3% | 1.9% |
| American Indian/Alaskan Native | 1.8% | 0.6% |
| Native Hawaiian/Pacific Islander | 0.4% | 0.2% |
| Not reported | 8.3% | 5.6% |
| Degrees (years of education) | | |
| Less than diploma (<12) | 0.2% | 7.1% |
| High school diploma (12) | 1.0% | 24.2% |
| Some colleges (13) | 2.9% | 7.1% |
| Associate degree (14+) | 1.9% | 14.6% |
| Bachelor's degree (16) | 65.1% | 15.2% |
| Some graduate education (17) | 1.2% | 1.3% |
| Master's degree (18) | 20.4% | 4.8% |
| Doctorate degree (>20) | 0.4% | 1.3% |
| Not reported | 6.9% | 24.4% |
| Age group | | |
| 18–20 | 2.7% | 2.7% |
| 20–30 | 33.2% | 29.4% |
| 30–40 | 26.9% | 29.6% |
| 40–50 | 19.9% | 22.3% |
| 50–60 | 8.4% | 11.7% |
| 60–70 | 2.1% | 2.9% |
| Not reported | 6.9% | 1.3% |

*Note*. Gender, race, ethnicity, educational level, and age distributions of the community (MTurk) and clinical (hospital) samples. MTurk = Mechanical Turk.

Words are displayed on the screen sequentially for 1,600 ms, followed by a 750 ms interstimulus interval, with a random jitter of 0–250 ms. Following list encoding, subjects have a 1-s rest period in which they are required to look at a fixation cross until a math distractor task appears on the screen. Subjects then perform a 20-s math distractor consisting of the summations of three random integers ranging from 1 to 9 (A + B + C=?). Subjects then have 30 s to freely recall all items presented in the preceding list. On the online version, subjects are required to type their responses, while the patient sample speak their responses into a microphone. Laboratory staff annotate these audio files offline. We ask subjects to recall by following these instructions: "When you have completed these math problems, you will see a fixation (+) flash on the screen. At this time, type as many words as you can remember from the list, IN ANY ORDER. You will have a fixed amount of time in which to recall the list. Please try hard throughout the recall period, as you may recall some words even when you feel you have exhausted your memory." For tasks administered in the hospital, we replace the word "type" with the word "say." Words in the FR task are semantically unrelated to one another, whereas in the CatFR task words are drawn from 25 different semantic categories. To form each CatFR list, an algorithm randomly selects three semantic categories to provide four words each. Each session of the task includes 25 word lists. These tasks include the same word pools as the ones used in the Defense Advanced Research Projects Agency Restoring Active Memory project (see Ezzyat et al., 2018, for details on word pool formation). Figure 1A illustrates the flow of events in the FR and CatFR tasks.

For subjects completing more than one session of FR on MTurk, a larger word pool was randomly divided into two, to avoid repetition of items across sessions. This division imitates the standard and alternative forms of the RAVLT and CVLT. To create the two FR forms, an algorithm randomly divided the word pool into 46 lists, consisting of 12 words each, using a pseudorandom number generator. Standard and alternative forms were then created by randomly and evenly dividing the 46 lists into two sets ($FR^s$ and $FR^a$).

**Table 2**
*Epilepsy Seizure Onset Zone*

| Seizure onset zone | *n* = 479 |
|---|---|
| Left temporal lobe | 23.6% |
| Right temporal lobe | 18.4% |
| Left prefrontal cortex | 6.7% |
| Right prefrontal cortex | 3.5% |
| Left frontal gyrus | 3.8% |
| Right frontal gyrus | 2.9% |
| Left parietal lobe | 5.6% |
| Right parietal lobe | 2.9% |
| Left occipital lobe | 2.1% |
| Right occipital lobe | 1.3% |
| Other | 16.1% |
| Not reported | 13.4% |

*Note*. This table reports the breakdown of our patient sample's epileptic foci.

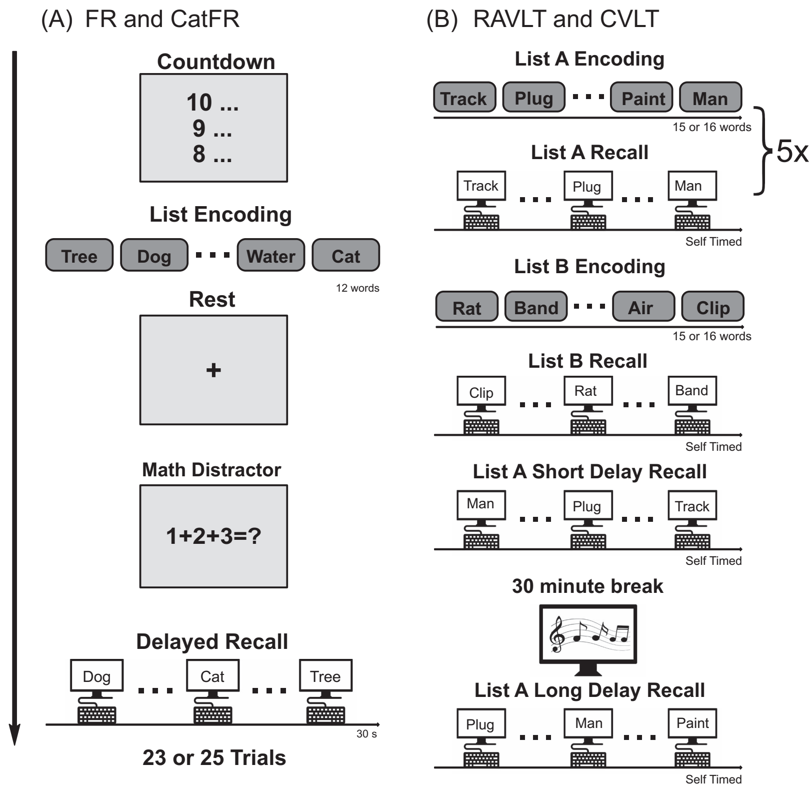
### *RAVLT and CVLT*

Our psiTurk implementation of the RAVLT attempts to replicate the procedures described in the RAVLT Handbook (Schmidt, 1996). Immediately after hearing a list of 15 words (List A) subjects attempt to recall as many words as they can remember by typing each word into a text box. This immediate free recall procedure repeats five times with List A, which appears in the same order each time. Following these five learning trials involving List A, subjects complete a single encoding and recall trial of the second list (List B), after which the task asks them to attempt to recall List A again. Following this "short delay recall" of List A, subjects view music videos for 30 min before recalling List A one final time (long delay recall). The words presented in our implementation of the RAVLT are taken from either the standard ($RAVLT^s$) form or the alternate ($RAVLT^a$) form of the RAVLT (Schmidt, 1996). Our psiTurk implementation also attempts to replicate the procedures described in the CVLT-II manual (Delis et al., 2000). Our collaborating neuropsychologists routinely perform neuropsychological evaluations and they administer either the CVLT-II, the RAVLT, or the WMS as part of this clinical workup. The CVLT follows the same procedure as the RAVLT but with a different set of word lists. Lists A and B in the CVLT contain 16 words, drawn from four different semantic categories. Figure 1B illustrates the flow of events in the RAVLT and CVLT tasks.

## Procedure

Of the 1,076 subjects that took part in the MTurk study, 949 subjects participated in the single-session experiment (either FR,

**Figure 1**
*Schematic Illustration of the Study Tasks*




*Note*. (A) Repeated-trial free recall of unrelated (FR) and categorized lists (CatFR), which differ only in the categorical organization of the items. (B) Rey Auditory Verbal Learning Test (RAVLT) and California Verbal Learning Test (CVLT). The two tests differ in list length (15 and 16 items, respectively) and in the categorical organization of items (see Method section).

CatFR, or CVLT). When a HIT is launched, MTurk posts a link to our experiment to the entire MTurk population in the United States at the same time. For single-session experiments, we launched several HITs over a week until we achieved a target sample size. For example, we launched the CVLT over 1 week and then separately launched CatFR over the following weeks.

The remaining 127 subjects completed four sessions of the experiment, with a week and a half between sessions. Each subject completed two sessions of FR and two sessions of RAVLT. The first session was randomly assigned as either FR or RAVLT (counterbalanced for order) and subsequent sessions alternated between the tasks (e.g., RAVLT$^{s}$, FR$^{s}$, RAVLT$^{a}$, FR$^{a}$). In order to complete any session on MTurk, subjects had to correctly complete an attention check. The attention check consisted of a long quote with a subtle instruction to reply “never” to the following question. If subjects correctly responded “never” to the presented question, they could begin the session. After completing the first session and passing exclusion criteria,[1] we invited subjects to complete three additional sessions of the experiment. Bonus payments, given upon completion of each session, incentivized subjects to complete the four sessions of the experiment. Subjects that took part in the four sessions completed one standard and one alternative form from each task to compute test–retest correlation without the confound of repeated items.

## Data and Code Availability

Anonymized data and analysis code may be freely downloaded from the public website: https://memory.psych.upenn.edu.

## Results

We first present analyses of the dynamics of memory in each of the tasks, evaluating the degree to which each task reveals established principles of memory search, including the effects of primacy, recency, contiguity, and similarity. We then establish the clinical validity of multitrial free recall of semantically organized (CatFR) and unrelated (FR) word lists by comparing these laboratory tasks with standard neuropsychological measures: the RAVLT and the CVLT. For each task, we report measures of predictive (internal) validity, convergent validity, and test–retest reliability.

[1] We excluded subjects who met any of the following criteria: (a) recall rate below 10% or above 95%, (b) providing more than one list without any correct recalls, and (c) reporting taking notes during the task.

## Recall Dynamics for Unrelated and Categorized Word Lists

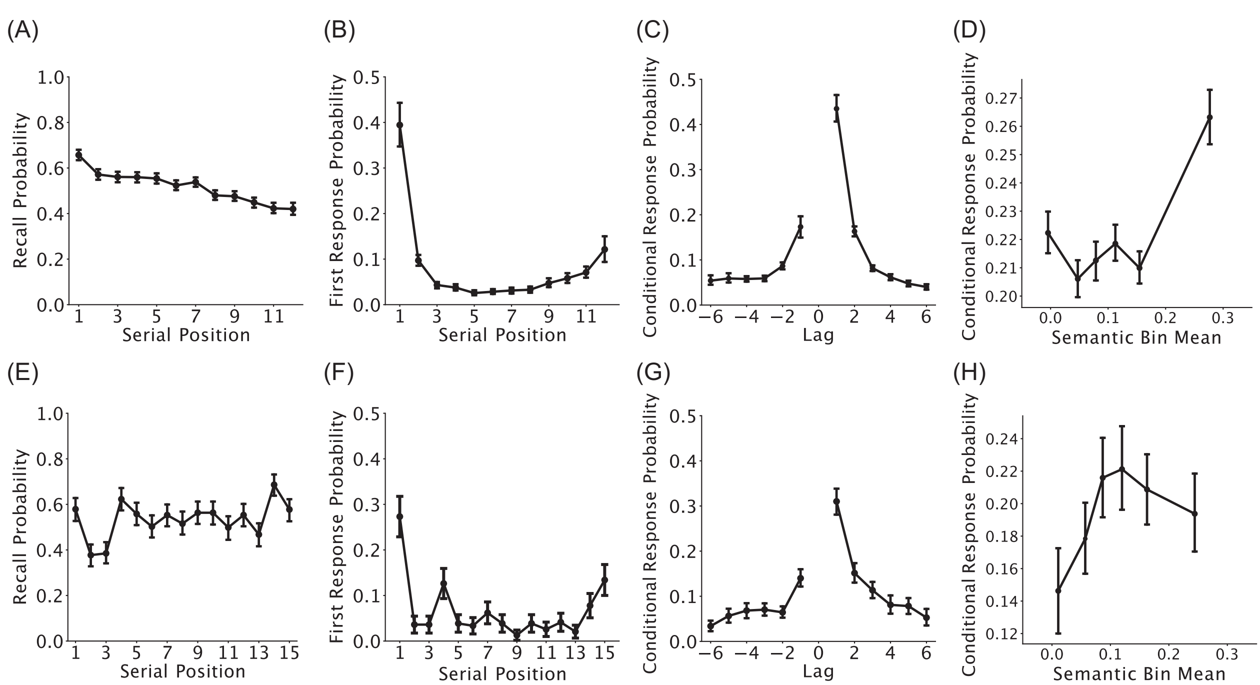
Analysis of the relationship between the order of learning and the order of recall reveals four major principles of memory: primacy, recency, contiguity, and similarity. Here we report data collected from our MTurk sample performing the FR, CatFR, RAVLT, and CVLT tasks. We first consider FR and RAVLT, as these tasks both involve lists of unrelated items. For the FR task, we observed a strong monotonic primacy effect, demonstrating that subjects exhibited superior recall of early list items (Figure 2A). The interpolation of a demanding arithmetic distractor task between the final item presentation and the beginning of the recall period eliminated the recency effect typically seen in immediate free recall (Glanzer & Cunitz, 1966). We can better understand the serial position effect by examining the probability of first recall as a function of serial position (Howard & Kahana, 1999). As shown in Figure 2B, we see that subjects exhibit a strong tendency to initiate recall with the first list item (nearly 50%), and a small elevated probability of initiated recall with one of the final list items, indicating a residual recency effect that does not appear in the serial position curve.

Analyses of subsequent recall transitions illustrate the effects of contiguity and similarity on recall. Figure 2C shows the probability that successively recalled study items came from positions $i$ and $i$ + lag $i$ as a function of lag ranging from −6 to +6, and aggregating over all list positions indexed by $i$. Here we see the classic asymmetric contiguity effect indicating that subjects tend to make transitions among neighboring items with a forward bias (Healey et al., 2019; Kahana, 1996). Figure 2D shows the probability that successively recalled study items have similar meanings as measured using Google's word2vec algorithm (Mikolov et al., 2013). This curve illustrates the classic effect of semantic clustering, wherein subjects will be more likely to transition to an available item that is semantically related to the just recalled item (Howard & Kahana, 2002; Manning et al., 2012).

The first trial data from the RAVLT shows similar tendencies to initiate recall with early and final list items and a strong contiguity effect, as seen in Figure 2F and 2G. Because the RAVLT involves a fixed list of items presented to each subject in the same predetermined order, variation in the memorability of individual words appears to obscure the effect of list position on recall probability, making it hard to see whether subjects show the expected primacy and recency effects usually seen in free recall tasks (see Figure 2E). The RAVLT appears to show a modest effect of semantic organization but here, too, the lack of variability in items and the stagnant presentation order across subjects and trials undermines the ability to infer organizational principles from recall order (see Figure 2H).

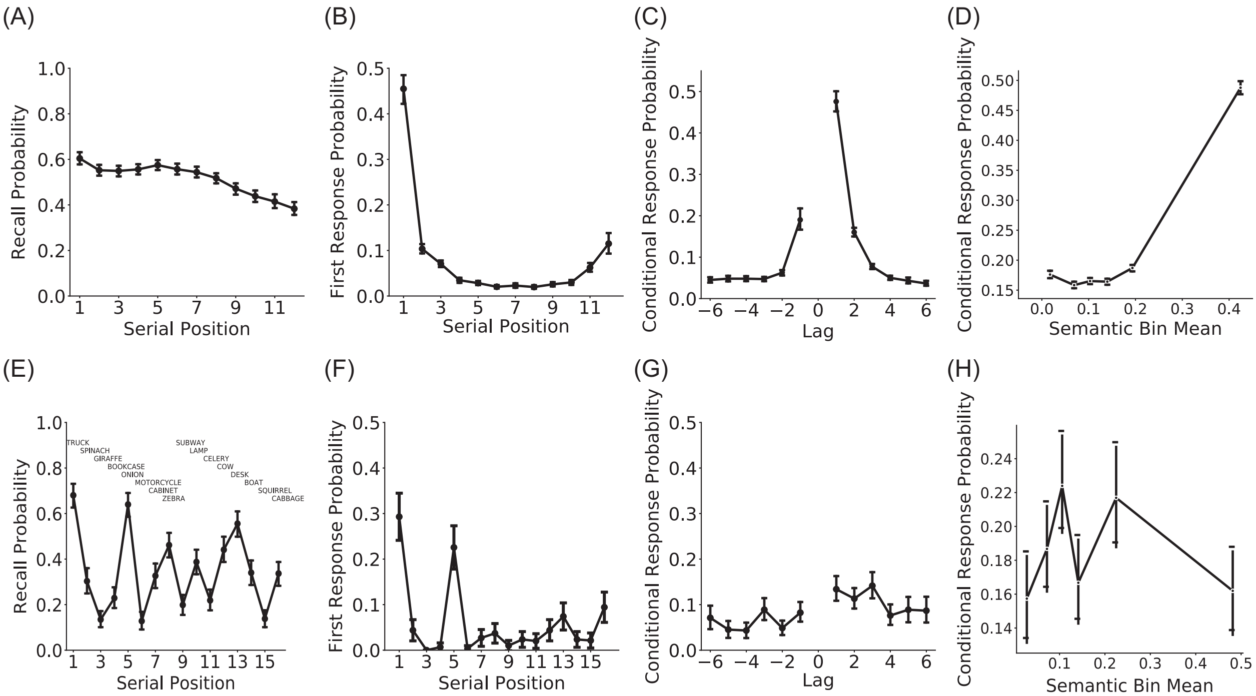
We next examined recall dynamics in CatFR and CVLT, which both involved lists of semantically organized items. For the CatFR task, we found nearly identical effects of primacy, recency, contiguity, and similarity to those seen in the FR task (see Figure 3A–D). The semantic similarity effect in CatFR grows to even larger values than in FR owing to the presence of more highly similar word pairs in each list (compare Figures 2D and 3D). For the

**Figure 2**
*Recall Dynamics for Unrelated Word Lists*




*Note.* Panels A–D show data for delayed free recall of unrelated word lists (FR task, see text). Panels E–H show data for the Rey Auditory Verbal Learning Test (RAVLT). All graphs illustrate data obtained from our online sample recruited through Amazon's Mechanical Turk. A and E: serial position curves. B and F: probability of first recall curves. C and G: Lag-CRP curves illustrating the likelihood of a recall transition from study item $i$ to study item $i$ + lag conditional on transition availability. D and H: Semantic-CRP curves illustrating the likelihood of a recall transition from study item $i$ to study item $j$ as a function of the similarity between $i$ and $j$, conditioned on the availability of that similarity bin. CRP = Conditional–Response Probability.

**Figure 3**
*Recall Dynamics for Categorized Recall*




*Note*. Panels A–D show data for delayed free recall of categorized word lists (CatFR task, see text). Panels E–H show data for the California Verbal Learning Test (CVLT). All graphs illustrate data obtained from our online sample recruited through Amazon's Mechanical Turk. A and E: serial position curves. B and F: probability of first recall curves. C and G: Lag-CRP curves illustrating the likelihood of a recall transition from study item $i$ to study item $i$ + lag conditional on transition availability. D and H: Semantic-CRP curves illustrating the likelihood of a recall transition from study item $i$ to study item $j$ as a function of the similarity between $i$ and $j$, conditioned on the availability of that similarity bin. CRP = Conditional–Response Probability.



first trial of CVLT, we find little evidence for either temporal or semantic organization and highly irregular serial position effects. As in the RAVLT, this task involves a single presentation order of a fixed set of items. As such, item-specific effects will make it difficult to observe the effects of the temporal or semantic organization of the items on recall dynamics (see Figure 3E–H).

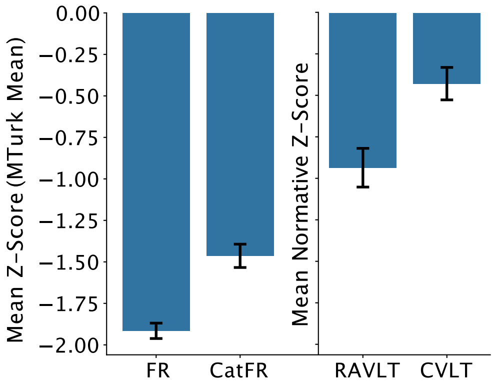
Subsequent analyses, which focus on the psychometric properties of the tasks, use subjects' overall recall performance. In the case of FR and CatFR, we analyzed the average number of correctly recalled items per list; in the case of RAVLT and CVLT, we analyzed overall recall across the initial five recall trials of List A.

## Internal Validity

We next sought to establish the internal validity of our laboratory tasks and to compare them with the RAVLT and CVLT. To gauge internal validity, we compared memory performance across two groups: a patient sample presumed to have memory loss and a sample of age-matched controls. Our patient cohort comprised individuals with drug-resistant epilepsy undergoing neurosurgical evaluation for potential resection of epileptogenic tissue. Prior studies have established that these patients frequently suffer from impaired verbal episodic memory (Helmstaedter, 2002; Langfitt et al., 2007). By comparing RAVLT and CVLT scores obtained from neuropsychological reports to age-matched normative scores found in Schmidt's (1996) study and Delis et al.'s (2000) study, we assessed the degree of memory impairment in our patient cohort. For the FR and CatFR measures, we used the mean and standard deviation of performance in our community sample to obtain standardized scores for our patients. We acknowledge that this comparison is less than ideal. To the extent that online subjects pay less careful attention, this comparison would overestimate patients' performance. Conversely, if the online subjects took notes or otherwise compromised the integrity of the test, this would underestimate patients' mean performance. For these reasons, we followed best practices to insure the highest degree of compliance from online subjects through the use of an attention check and exclusion criteria aimed at removing any subjects whose data indicated noncompliance with task instructions (see Method section).

Patients with drug-resistant epilepsy demonstrated significant memory loss on all four tasks as compared with their respective control cohorts (see Figure 4). This degree of impairment mirrors that seen in other neurological conditions, such as moderate-to-severe traumatic brain injury (Jacobs & Donders, 2007). Overall, we show that both FR and CatFR measure memory loss among patients with neurological disease relative to a community sample. In addition, the greater degree of impairment seen in the RAVLT relative to the CVLT closely relates to the findings of Loring et al. (2008), which shows that the RAVLT is more sensitive to verbal memory deficits, presumably due to the lack of semantic relatedness in this task, which may require subjects to rely on more effortful strategies.

**Figure 4**
*Epilepsy Versus Healthy Population*

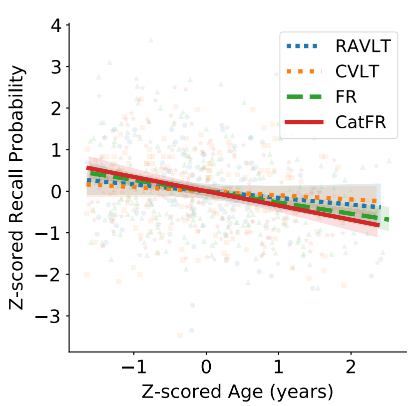



*Note.* Performance of the epileptic population as compared to healthy controls (MTurk workers for free recall and categorized free recall, age-corrected standard scores for the California Verbal Learning Test (CVLT) and the Rey Auditory Verbal Learning Test (RAVLT) on each of the four tasks. Means and standard deviations from healthy populations were used to Z-score each epileptic patient's performance. Error bars represent ±1 standard error of the mean. See the online article for the color version of this figure.

To further establish the internal validity of the FR and CatFR tasks, we investigated the well-known age-related decline in verbal free recall (Kahana et al., 2005; Wingfield & Kahana, 2002). We first evaluated the correlation between age and recall performance in our patient sample. Because ages did not range very widely in this cohort (our oldest patients were in their mid-60s), we only expected to find modest age-related declines in memory performance.

Figure 5 elucidates the negative relationship between age and recall performance in each task. In both FR and CatFR, we observed strong negative correlations between age and recall performance. We also observed reliable declines in the RAVLT and CVLT, but these were not as large as the declines seen in the laboratory tasks. Using a mixed-effects linear regression model, predicting recall as a function of age, test type, and their interaction and allowing the intercept to vary by subject, we found that both FR and CatFR measured significantly stronger age effects than the CVLT ($b = -0.163$, $p < .05$ and $b = -0.283$, $p < .001$ respectively) and the RAVLT ($b = -0.143$, $p = .153$ and $b = -0.264$, $p < .05$ respectively).

**Figure 5**
*Age-Related Decline in Recall Performance*




*Note.* The expected negative correlation between age and overall recall performance appeared in multitrial free recall of unrelated and categorized word lists (FR and CatFR) as well as in the California Verbal Learning Test (CVLT) and the Rey Auditory Verbal Learning Test (RAVLT). In each task, we computed the correlation using data from a large study of memory performance in patients with neurological disease (drug-resistant epilepsy). Error bands represent bootstrapped 95% confidence intervals. See the online article for the color version of this figure.

## Convergent Validity

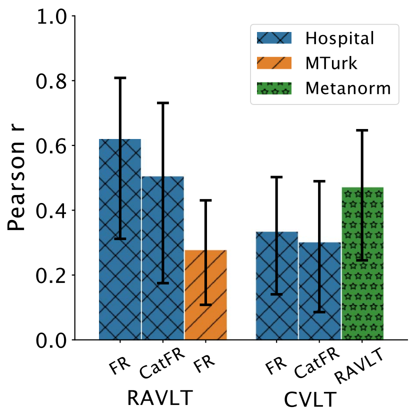
While our results thus far suggest strong internal validity for FR and CatFR, we next sought to establish convergent validity with the neuropsychological exams by comparing across-task correlations. We therefore limited our sample to subjects who performed two (or more) of the tasks of interest. Patients in our clinical sample had data on either the RAVLT or the CVLT (but not both) as well as data on either FR or CatFR (and occasionally both). The blue bars in Figure 6 show the correlations between each of the clinical tasks and each of the laboratory experiments. In all cases, the correlations were moderately positive and significantly greater than zero, MTurk FR versus RAVLT: $r(125) = .28$, $p < .01$, hospital FR versus RAVLT: $r(25) = .62$, $p < .001$, hospital CatFR versus RAVLT: $r(28) = .50$, $p < .01$, hospital FR versus CVLT: $r(92) = .33$, $p < .01$, hospital CatFR versus CVLT: $r(77) = .30$, $p < .01$. For comparison, we obtained published data on the RAVLT and CVLT (Schmidt, 1996), which exhibited a similar intertask correlation, $r(58) = .47$, $p < .001$, as that observed for comparisons between the clinical measures and our laboratory measures. There were no significant differences between any of these correlation coefficients. Thus, FR and CatFR show strong convergent validity with the widely used RAVLT and CVLT.

## Test–Retest Reliability

For our final analysis, we examined the test–retest reliability of all four tasks. As in the previous analysis, we used both the patient sample and the MTurk sample, depending on the available data for each task comparison. Hospital patients often ran "half-sessions" consisting of two sets of 12 lists on separate days. Half-sessions of FR and CatFR provided the method for calculating the test–retest values in the hospital. The multisession community cohort provided data for the "MTurk" correlations, and Schmidt (1996) and Delis et al. (2000) provided values for the test–retest "meta-norm" data.

Figure 7 shows the test–retest Pearson $r$ correlation value for each test. In all cases, we observe reliability values of ~0.8, indicating a high degree of consistency in performance across repeated test administration. Overall, we find that the FR and CatFR tasks exhibit somewhat higher reliability coefficients than the standard neuropsychological tests, MTurk FR versus MTurk RAVLT: $r_1(114) - r_2(114) = .125$, $p < .05$, hospital FR versus meta-norm RAVLT: $r_1(25) - r_2(85) = .0562$, $p = .465$, hospital CatFR versus meta-norm CVLT: $r_1(48) - r_2(286) = .0691$, $p = .109$. This increase in reliability likely reflects the greater number of lists in FR and CatFR as compared with RAVLT and CVLT, as well as the randomization of items across repeated lists.

**Figure 6**
*Task Performance Correlations*

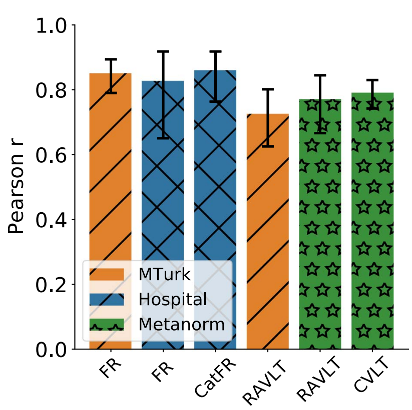



*Note.* This figure shows positive correlations for between-task comparisons. Laboratory tasks correlate strongly to the two neuropsychological tasks. This analysis calculates Pearson *r* values for the within-subject correlations between two sessions of different tasks (i.e., one session of RAVLT and one session of FR). Error bars represent 95% confidence intervals, calculated using standard Fisher transformations. FR = free recall; RAVLT = Rey Auditory Verbal Learning Test; CatFR = free recall of categorized word lists; CVLT = California Verbal Learning Test. See the online article for the color version of this figure.

## Discussion

Modern research on human memory generally entails repeatedly evaluating subjects' memory for multiple lists of trial-unique items. This approach has multiple advantages over the single-list assessments typically used in neuropsychological batteries. Here we provide evidence for the clinical validity of repeatedly evaluating memory for distinct lists of unrelated items (FR) or categorically organized items (CatFR). We show that each of these tasks compares favorably with standard neuropsychological tests: the RAVLT and the CVLT. We first demonstrate that key principles of recall dynamics, including temporal and semantic clustering, appear in all four tasks. We then provide evidence for the predictive validity of both variants, finding clear memory deficits and age-related declines in memory performance in a sample of patients with drug-resistant epilepsy. We find that both FR and CatFR tasks exhibit equivalent or larger effect sizes than the CVLT and RAVLT. We next examine convergent validity by comparing the correlation between the laboratory measures and the clinical tasks. Here we find significant intertask correlations for all tasks and similar levels of correlations for the laboratory tasks and the clinical neuropsychological tests. Finally, we demonstrate that both laboratory recall tasks exhibit high test–retest reliability, with higher or similar values compared to those of the established neuropsychological tests.

**Figure 7**
*Hospital, Meta-Norm, and MTurk Test–Retest Correlations*




*Note.* Test–retest correlation for all tasks and administration types, calculated as the correlation coefficient between the first and second session of the same task. Laboratory tasks (FR and CatFR) administered on MTurk and in the hospital show nonsignificantly stronger correlations than neuropsychological tasks (RAVLT and CVLT) on MTurk or in meta-normative data. Error bars represent 95% confidence intervals, calculated using standard Fisher transformations. MTurk = Mechanical Turk; FR = free recall; CatFR = free recall of categorized word lists; RAVLT = Rey Auditory Verbal Learning Test; CVLT = California Verbal Learning Test. See the online article for the color version of this figure.

Impaired episodic memory constitutes one of the most disturbing aspects of healthy aging, neurological disease, such as Alzheimer's, and brain injury (Adamovich-Zeitlin et al., 2021; Craik, 2000; Huang & Mucke, 2012). As the human life span increases, so will the need to find new interventions to prevent, slow, or reverse memory decline. Achieving success in developing such interventions will require the development of better tools for evaluating human memory and its neural correlates. Although memory takes many forms, we have focused on evaluating measures of episodic memory; a form of memory that requires the rememberer to associate information representing an event within a spatiotemporal context. Episodic memory allows us to remember where we parked our car today and what we did last Saturday; it places us in our memories, marking each memory's position on our autobiographical timeline. Tasks that require subjects to freely recall lists of studied items in the absence of specific retrieval cues place strong demands on the episodic memory system (Kahana, 2020). Episodic memory exhibits marked declines in normal aging (Buchler et al., 2011; Naveh-Benjamin, 2000) and in a variety of neurological conditions (Dickerson & Eichenbaum, 2010; Vakil, in press). The pattern of deficits evident in free recall tasks predicts conversion from mild cognitive impairment to Alzheimer's disease (Trenkle et al., 2007). For the reasons described below, discerning these patterns of deficits benefits from having subjects engage in repeated trials involving unique lists of memoranda.

The CVLT and RAVLT derive their format from early 20th-century research within the verbal learning tradition. Until the 1960s, memory researchers had subjects learn a list of items via the method of repeated study–test trials. Then, they measured the rate of forgetting under conditions of interpolated learning (i.e., learn a new list, and then test the original list). Both the CVLT and the RAVLT have this basic structure. Subjects learn a single list across a series of repeated study–test trials. Then, after a delay (and following the learning of an "interfering" list), they try to recall the originally studied list. Early studies of memory applied this method to the acquisition of very long lists (e.g., 32 items) across 10 or more learning trials. Researchers would then trace the forgetting curve, evaluating memory for the mastered list after minutes, hours, and even days. The CVLT and the RAVLT attempt to miniaturize this procedure, fitting it into a 30–45-min testing session.

The cognitive revolution of the 1960s saw the emergence of new methods in the study of memory. Researchers embraced the study of memory on shorter time scales, with subjects repeatedly studying and subsequently recalling unique word lists. Rather than examining learning and forgetting across hours, researchers studied these processes by analyzing item-level memory within individual lists. The buildup of proactive interference across lists sped up the analysis of forgetting as subjects had to target memory for a single list among the many competing lists stored in memory (see Kahana, 2012, for a review of this work). Examination of individual-level data also revealed that naïve subjects adapt their strategies rather quickly across the first few trials of any such experiment (Murdock, 1974). It takes subjects several tries to become familiar with a memory task, and as a result their behavior will change, sometimes rapidly, across the first few trials of an experiment. Typically, after about five lists, behavior stabilizes. As such, many researchers discard the first few lists of a session as practice trials, only using the subsequent trials to evaluate hypotheses about memory. We therefore believe that neuropsychologists should be particularly wary of making determinations of an individual's memory based on how that person learns a single list of memoranda.

Beyond changes in behavior over the first few trials, subjects' ability to remember recently learned information can vary considerably across subsequent trials. Kahana et al. (2018) attempted to model this variability using numerous established variables in the memory literature, including item and list difficulty, proactive interference, and other variables (see, also, Aka et al., 2021). They found that even after controlling for these factors there was marked excess volatility in subjects' mnemonic ability. Indeed, the best predictor of performance on a given list is the subject's performance on the preceding list, suggesting that endogenous factors underlie mnemonic variability across trials. By recording neural activity across many trials of a memory task, researchers have found neural signals that predict variability in performance across both items and lists (Kragel et al., 2017; Weidemann & Kahana, 2021).

Here we examined the psychometric properties of delayed free recall of unrelated and categorized lists administered across repeated study–test trials involving list-unique items. During a ∼45-min session, each subject attempted recall on ∼24 trials. Aggregating performance across these repeated trials provided reliable measures of recall performance as well as temporal and semantic organization of the studied items (see Figures 2 and 3). Taking the number of correctly recalled words as a measure of mnemonic ability yielded favorable psychometric properties. This measure possessed numerically higher test–retest reliability and stronger correlations with age and neurological disease than two established neuropsychological measures, the RAVLT and the CVLT.

Although our results favor the multilist approach for measures based on overall recall, the RAVLT and CVLT provide other indices of performance that could be particularly meaningful in comparisons involving specific memory-impaired individuals or populations. The main advantage we see of the present approach is that it allows us to bridge the clinical literature on memory disorders with the modern literature on the cognitive neuroscience of human memory where researchers often rely on repeated observations to establish reliable relationships between behavioral and brain measures.

## Constraints on Generality

Whereas collecting data online allowed us to efficiently perform a large scale, longitudinal study, our inability to directly observe subjects as they performed the memory tasks somewhat limits the generality of our findings. However, a few factors help to mitigate these concerns. To diminish the potential effects of inattention or subjects not following the requested procedures, we used standard attention checks and exclusion criteria (Abbey & Meloy, 2017; see the Procedure subsection, for a full description of exclusion criteria and attention checks). In addition, the pattern of behavioral recall dynamics evident in the data collected on MTurk recapitulated the classic effects of temporal and semantic clustering seen in laboratory studies (e.g., Figures 2 and 3); these patterns would not have emerged if subjects deviated markedly from our instructions. Prior studies have also shown that memory performance as measured in the lab replicate in online studies (see, e.g., Arnold & McDermott, 2013; Leding, 2019; Mundorf et al., 2021; Uitvlugt & Healey, 2019).

In addition, the neuropsychology reports given to us by our collaborative hospitals did not regularly record formal performance validity measures, potentially hindering validity of the observed memory performance. However, overall scaled performance levels of patients in our study are consistent with scaled performance levels seen on similar word list memory tasks in patient populations, suggesting that patients performed the task with adequate effort (see Supplemental Materials, for a comparison between age-normed recall in our sample and those obtained from epilepsy patients in a prior study).

Finally, a few sample characteristics may also limit the generality of the observed results. First, our collaborating neuropsychology departments assigned patients to either RAVLT or CVLT based on clinical needs, potentially leading to a biased representation of subjects between tasks. Second, our community cohort included a nonselected sample of subjects and thus included subjects with varied cognitive and mental abilities. Finally, our clinical sample included epilepsy patients with diverse etiologies and epileptogenic foci, which may have deferentially influenced performance in the tasks. However, reanalyses of these data using a more homogeneous group of patients with temporal lobe epilepsy led to similar results, supporting the validity of the present analyses (see Supplemental Materials).

## Conclusion

The present study compares memory performance in two common neuropsychological tests with memory performance on multitrial free recall paradigms commonly used in laboratory settings. Multitrial free recall demonstrates comparable internal validity to that of the neuropsychological tests, capturing expected memory deterioration with age and in a clinical population suffering from drug-resistant epilepsy. The laboratory tests also demonstrate high test–retest reliability and strong convergent validity with the neuropsychological tests. We further show the advantage of using multitrial free recall paradigms for measuring well-established recall dynamics. These results establish the clinical validity of multitrial free recall paradigms and provide a first step toward bridging the gap between the clinical and basic memory literature to provide a path toward better treatment options and measurement tools for patients suffering from memory disorders.

## References

Abbey, J. D., & Meloy, M. G. (2017). Attention by design: Using attention checks to detect inattentive respondents and improve data quality. *Journal* *of Operations Management*, *53–56*, 63–70. https://doi.org/10.1016/j.jom.2017.06.001

Adamovich-Zeitlin, R., Wanda, P. A., Solomon, E., Phan, T., Lega, B., Jobst, B. C., Gross, R. E., Ding, K., Diaz-Arrastia, R., & Kahana, M. J. (2021). Biomarkers of memory variability in traumatic brain injury. *Brain Communications*, *3*(1), Article fcaa202. https://doi.org/10.1093/braincomms/fcaa202

Aka, A., Phan, T. D., & Kahana, M. J. (2021). Predicting recall of words and lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(5), 765–784. https://doi.org/10.1037/xlm0000964

Arnold, K. M., & McDermott, K. B. (2013). Free recall enhances subsequent learning. *Psychonomic Bulletin & Review*, *20*(3), 507–513. https://doi.org/10.3758/s13423-012-0370-3

Buchler, N. G., Faunce, P., Light, L. L., Gottfredson, N., & Reder, L. M. (2011). Effects of repetition on associative recognition in young and older adults: Item and associative strengthening. *Psychology and Aging*, *26*(1), 111–126. https://doi.org/10.1037/a0020816

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5. https://doi.org/10.1177/1745691610393980

Craik, F. I. M. (2000). Age-related changes in human memory. In D. Park & N. Schwarz (Eds.), *Cognitive aging: A primer* (pp. 75–92). Psychology Press.

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in aWeb browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Delis, D., Kramer, J., Kaplan, E., & Ober, B. (2000). *California Verbal Learning Test-Second Edition (CVLT-II)* (2nd ed.). Psychological Corporation.

Dickerson, B. C., & Eichenbaum, H. (2010). The episodic memory system: Neurocircuitry and disorders. *Neuropsychopharmacology Reviews*, *35*(1), 86–104. https://doi.org/10.1038/npp.2009.126

Eargle, D., Gureckis, T., Alexander, S., McDonnell, J., & Martin, J. (2021). *psiTurk: An open platform for science on Amazon Mechanical Turk* (Version v3.2.1). Zenodo. https://doi.org/10.5281/zenodo.6311090

Estévez-González, A., Kulisevsky, J., Boltes, A., Otermín, P., & García-Sánchez, C. (2003). Rey verbal learning test is a useful tool for differential diagnosis in the preclinical phase of Alzheimer's Disease: Comparison with mild cognitive impairment and normal aging. *International Journal of Geriatric Psychiatry*, *18*(11), 1021–1028. https://doi.org/10.1002/gps.1010

Ezzyat, Y., Wanda, P., Levy, D. F., Kadel, A., Aka, A., Pedisich, I., Sperling, M. R., Sharan, A. D., Lega, B. C., Burks, A., Gross, R. E., Inman, C. S., Jobst, B. C., Gorenstein, M. A., Davis, K. A., Worrell, G. A., Kucewicz, M. T., Stein, J. M., Gorniak, R., … Kahana, M. J. (2018). Closed-loop stimulation of temporal cortex rescues functional networks and improves memory. *Nature Communications*, *9*(1), Article 365. https://doi.org/10.1038/s41467-017-02753-0

Fernández, G., Effern, A., Grunwald, T., Pezer, N., Lehnertz, K., Dümpelmann, M., Van Roost, D., & Elger, C. E. (1999). Real-time tracking of memory formation in the human rhinal cortex and hippocampus. *Science*, *285*(5433), 1582–1585. https://doi.org/10.1126/science.285.5433.1582

Glanzer, M., & Cunitz, A. R. (1966). Two storage mechanisms in free recall. *Journal of Verbal Learning & Verbal Behavior*, *5*(4), 351–360. https://doi.org/10.1016/S0022-5371(66)80044-0

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., Halpern, D., Hamrick, J. B., & Chan, P. (2016). psi-Turk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods*, *48*(3), 829–842. https://doi.org/10.3758/s13428-015-0642-8

Hanslmayr, S., Klimesch, W., Sauseng, P., Gruber, W., Doppelmayr, M., Freunberger, R., Pecherstorfer, T., & Birbaumer, N. (2007). Alpha phase reset contributes to the generation of ERPs. *Cerebral Cortex*, *17*(1), 1–8. https://doi.org/10.1093/cercor/bhj129

Healey, M. K., Long, N. M., & Kahana, M. J. (2019). Contiguity in episodic memory. *Psychonomic Bulletin & Review*, *26*(3), 699–720. https://doi.org/10.3758/s13423-018-1537-3

Helmstaedter, C. (2002). Effects of chronic epilepsy on declarative memory systems. *Progress in Brain Research*, *135*, 439–453. https://doi.org/10.1016/S0079-6123(02)35041-6

Howard, M. W., & Kahana, M. J. (1999). Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(4), 923–941. https://doi.org/10.1037/0278-7393.25.4.923

Howard, M. W., & Kahana, M. J. (2002). When does semantic similarity help episodic retrieval? *Journal of Memory and Language*, *46*(1), 85–98. https://doi.org/10.1006/jmla.2001.2798

Huang, Y., & Mucke, L. (2012). Alzheimer mechanisms and therapeutic strategies. *Cell*, *148*(6), 1204–1222. https://doi.org/10.1016/j.cell.2012.02.040

Jacobs, M. L., & Donders, J. (2007). Criterion validity of the California Verbal Learning Test-Second Edition (CVLT-II) after traumatic brain injury. *Archives of Clinical Neuropsychology*, *22*(2), 143–149. https://doi.org/10.1016/j.acn.2006.12.002

Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory & Cognition*, *24*(1), 103–109. https://doi.org/10.3758/BF03197276

Kahana, M. J. (2012). *Foundations of human memory*. Oxford University Press.

Kahana, M. J. (2020). Computational models of memory search. *Annual Review of Psychology*, *71*(1), 107–138. https://doi.org/10.1146/annurev-psych-010418-103358

Kahana, M. J., Aggarwal, E. V., & Phan, T. D. (2018). The variability puzzle in human memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(12), 1857–1863. https://doi.org/10.1037/xlm0000553

Kahana, M. J., Dolan, E. D., Sauder, C. L., & Wingfield, A. (2005). Intrusions in episodic recall: Age differences in editing of overt responses. *Journal of Gerontology: Psychological Sciences*, *60*(2), 92–97. https://doi.org/10.1093/geronb/60.2.P92

Kahana, M. J., Wanda, P. A., Ezzyat, Y., Adamovich-Zeitlin, R., Lega, B., Jobst, B. C., Gross, R. E., Ding, K., & Diaz-Arrastia, R. R. (2023). Biomarker-guided neuromodulation aids memory in traumatic brain injury. *Brain Stimulation*, *16*(4), 1086–1093. https://doi.org/10.1101/2021.05.18.21256980

Kim, H. (2011). Neural activity that predicts subsequent memory and forgetting: A meta-analysis of 74 fMRI studies. *NeuroImage*, *54*(3), 2446–2461. https://doi.org/10.1016/j.neuroimage.2010.09.045

Kragel, J. E., Ezzyat, Y., Sperling, M. R., Gorniak, R., Worrell, G. A., Berry, B. M., Inman, C., Lin, J. J., Davis, K. A., Das, S. R., Stein, J. M., Jobst, B. C., Zaghloul, K. A., Sheth, S. A., Rizzuto, D. S., & Kahana, M. J. (2017). Similar patterns of neural activity predict memory function during encoding and retrieval. *NeuroImage*, *155*, 60–71. https://doi.org/10.1016/j.neuroimage.2017.03.042

Kucewicz, M. T., Berry, B. M., Kremen, V., Miller, L. R., Khadjevand, F., Ezzyat, Y., Stein, J. M., Wanda, P., Sperling, M. R., Gorniak, R., Davis, K. A., Jobst, B. C., Gross, R. E., Lega, B., Stead, S. M., Rizzuto, D. S., Kahana, M. J., & Worrell, G. A. (2018). Electrical stimulation modulates high gamma activity and human memory performance. *eNeuro*, *5*(1). https://doi.org/10.1523/ENEURO.0369-17.2018

Langfitt, J.,Westerveld, M., Hamberger, M.,Walczak, T., Cicchetti, D., Berg, A., Vickrey, B. G., Barr, W. B., Sperling, M. R., Masur, D., & Spencer, S. S. (2007). Worsening of quality of life after epilepsy surgery effect of seizures and memory decline. *Neurology*, *68*(23), 1988–1994. https://doi.org/10.1212/01.wnl.0000264000.11511.30

Leding, J. K. (2019). Intentional memory and online data collection: A test of the effects of animacy and threat on episodic memory. *Journal of Cognitive Psychology*, *31*(1), 4–15. https://doi.org/10.1080/20445911.2018.1564756

Loring, D. W., Strauss, E., Hermann, B. P., Barr, W. B., Perrine, K., Trenerry, M. R., Chelune, G., Westerveld, M., Lee, G. P., Meador, K. J., & Meador, K. J. (2008). Differential neuropsychological test sensitivity to left temporal lobe epilepsy. *Journal of the International Neuropsychological Society*, *14*(3), 394–400. https://doi.org/10.1017/S1355617708080582

Manning, J. R., Sperling, M. R., Sharan, A., Rosenberg, E. A., & Kahana, M. J. (2012). Spontaneously reactivated patterns in frontal and temporal lobe predict semantic clustering during memory search. *Journal of Neuroscience*, *32*(26), 8871–8878. https://doi.org/10.1523/JNEUROSCI.5321-11.2012

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. PsyArXiv. https://doi.org/10.48550/arXiv.1301.3781

Mundorf, A. M. D., Lazarus, L. T., Uitvlugt, M. G., & Healey, M. K. (2021). A test of retrieved context theory: Dynamics of recall after incidental encoding. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(8), 1264–1287. https://doi.org/10.1037/xlm0001001

Murdock, B. B. (1974). *Human memory: Theory and data*. Lawrence Erlbaum and Associates.

Naveh-Benjamin, M. (2000). Adult-age differences in memory performance: Tests of an associative deficit hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(5), 1170–1187. https://doi.org/10.1037/0278-7393.26.5.1170

Paller, K. A., & Wagner, A. D. (2002). Observing the transformation of experience into memory. *Trends in Cognitive Sciences*, *6*(2), 93–102. https://doi.org/10.1016/S1364-6613(00)01845-3

Schmidt, M. (1996). *Rey Auditory Verbal Learning Test: A handbook*. Western Psychological Services.

Sederberg, P. B., Kahana, M. J., Howard, M. W., Donner, E. J., & Madsen, J. R. (2003). Theta and gamma oscillations during encoding predict subsequent recall. *Journal of Neuroscience*, *23*(34), 10809–10814. https://doi.org/10.1523/JNEUROSCI.23-34-10809.2003

Strange, B. A., Otten, L. J., Josephs, O., Rugg, M. D., & Dolan, R. J. (2002). Dissociable human perirhinal, hippocampal, and parahippocampal roles during verbal encoding. *Journal of Neuroscience*, *22*(2), 523–528. https://doi.org/10.1523/JNEUROSCI.22-02-00523.2002

Trenkle, D. L., Shankle, W. R., & Azen, S. P. (2007). Detecting cognitive impairment in primary care: Performance assessment of three screening instruments. *Journal of Alzheimer's Disease*, *11*(3), 323–335. https://doi.org/10.3233/jad-2007-11309

Uitvlugt, M. G., & Healey, M. K. (2019). Temporal proximity links unrelated news events in memory. *Psychological Science*, *30*(1), 92–104. https://doi.org/10.1177/0956797618808474

Vakil, E. (2005). The effect of moderate to severe traumatic brain injury (TBI) on different aspects of memory: A selective review. *Journal of Clinical and Experimental Neuropsychology*, *27*(8), 977–1021. https://doi.org/10.1080/13803390490919245

Vakil, E. (in press). The mnemonic consequences of moderate-to-severe traumatic brain injury. In M. J. Kahana & A. D. Wagner (Eds.), *Oxford handbook of human memory* (2nd ed.). Oxford University Press.

Weidemann, C. T., & Kahana, M. J. (2021). Neural measures of subsequent memory reflect endogenous variability in cognitive function. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(4), 641–651. https://doi.org/10.1037/xlm0000966

Weidemann, C. T., Kragel, J. E., Lega, B. C., Worrell, G. A., Sperling, M. R., Sharan, A. D., Jobst, B. C., Khadjevand, F., Davis, K. A., Wanda, P. A., Kadel, A., Rizzuto, D. S., & Kahana, M. J. (2019). Neural activity reveals interactions between episodic and semantic memory systems during retrieval. *Journal of Experimental Psychology: General*, *148*(1), 1–12. https://doi.org/10.1037/xge0000480

Weissberger, G. H., Strong, J. V., Stefanidis, K. B., Summers, M. J., Bondi, M. W., & Stricker, N. H. (2017). Diagnostic accuracy of memory measures in alzheimer's dementia and mild cognitive impairment: A systematic review and meta-analysis. *Neuropsychology Review*, *27*(4), 354–388. https://doi.org/10.1007/s11065-017-9360-6

Wingfield, A., & Kahana, M. J. (2002). The dynamics of memory retrieval in older adulthood. *Canadian Journal of Experimental Psychology*, *56*, 187–199. https://doi.org/10.1037/h0087396

Wright, S. L., & Persad, C. (2007). Distinguishing between depression and dementia in older persons: Neuropsychological and neuropathological correlates. *Journal of Geriatric Psychiatry and Neurology*, *20*(4), 189–198. https://doi.org/10.1177/0891988707308801

Xue, G., Dong, Q., Chen, C., Lu, Z., Mumford, J. A., & Poldrack, R. A. (2010). Greater neural pattern similarity across repetitions is associated with better memory. *Science*, *330*(6000), 97–101. https://doi.org/10.1126/science.1193125

Received August 17, 2022
Revision received March 11, 2023
Accepted April 5, 2023 ■