Multi-trial Free Recall for Evaluating Memory

Adrogue, R.T., Herz, N., Halpern, D., Tracy, J., and Kahana, and M.J.

University of Pennsylvania

Author Note

The authors express their gratitude to the patients and participants who selflessly volunteered to participate in this experiment. This work was supported by the Department of Defense DARPA Restoring Active Memory (RAM) program (Cooperative Agreement N66001-14-2-4032) and by National Institutes of Health grant R01 NS 106611 awarded to MJK. 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. Correspondence concerning this should be addressed to Michael J. Kahana (kahana@psych.upenn.edu). M.J.K. holds a greater than 5% equity interest in Nia Therapeutics Inc., a company intended to develop and commercialize brain stimulation therapies for memory restoration.

Abstract

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 multi-trial free recall of semantically organized and unrelated word lists. **Methods:** We compared memory performance in multi-trial free recall tasks with the Rey Auditory Verbal Learning Task (RAVLT) and the California Verbal Learning Task (CVLT), two common neuropsychological tests aimed at evaluating memory function in clinical settings. We compared predictive validity between the tasks by (a) evaluating deficits in a patient sample, (b) examining age-related declines in memory. We also compared test-retest reliability, convergent validity and the emergence of common recall dynamics between the tasks. **Results:** We demonstrate that both laboratory free recall tasks have better predictive validity, as well as 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, but we demonstrate the benefits of repeated trials for evaluating the dynamics of memory search and their neuropsychological sequelae. **Conclusion:** The results provide evidence for the clinical validity of lab-based multi-trial 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.

Keywords: free recall, categorized free recall, RAVLT, CVLT

Multi-trial Free Recall for Evaluating Memory

Impact Statement

Question: How does the reliability and validity of laboratory-based multi-trial recall paradigms compare with more traditional neuropsychological measures? **Findings:** Multi-trial free recall paradigms show strong convergent validity with gold-standard neuropsychological scores, show strong internal validity, and have similar or higher reliability scores than neuropsychological tests. **Importance:** Evidence supports the clinical validity of multi-trial free recall paradigms. **Next Steps:** Identifying the neural correlates of behavioral deficits in multi-trial recall can offer novel insights into the biological mechanisms underlying patterns of preserved and impaired memory performance.

Introduction

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; Wright & Persad, 2007; Weissberger et al., 2017), neuroscientific studies of memory contrast good and poor memory states within a given individual, across trials or conditions (e.g., Paller & Wagner, 2002; Sederberg, Kahana, Howard, Donner, & Madsen, 2003; Hanslmayr et al., 2007). To identify neural correlates of variable memory, a subject must contribute data across many trials. Sorting these trials into remembered and non-remembered categories allows researchers to identify the neural correlates of mnemonic success. Although this repeated testing methodology represents the norm in cognitive neuroscience, neuropsychological measures of memory typically entail just one or two distinct word lists. In the Rey Auditory Verbal Learning Task (RAVLT), subjects attempt to freely recall a single list of 15 auditorally 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 30 minute delay. The California Verbal Learning Test (CVLT) follows a very similar procedure but

with lists of semantically categorized items (Delis, Kramer, Kaplan, & Ober, 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, Kulisevsky, Boltes, Otermín, & García-Sánchez, 2003). These neuropsychological measures therefore tend to emphasize a less granular approach relative to the those used in neuroscientific studies, but their rich normative basis makes their quantification of performance the gold standard for determining day-to-day functional health status or the presence of a clinically significant memory disorder.

The discord in methodologies used across these allied disciplines – cognitive neuroscience and neuropsychology – limits inferences about the clinical value of experimental studies. We therefore sought 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 considered delayed free recall of lists comprising either 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 (Weidemann et al., 2019; Kragel et al., 2017) 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., Submitted). Establishing the equivalence of memory measures obtained by the two approaches 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 repeated-list

delayed recall tasks with the RAVLT and the CVLT performed by the same subjects.

Methods

Participants

The present study included a cohort recruited through Amazon's Mechanical (MTurk), an online crowd-sourcing platform, and a sample of patients undergoing neurosurgical evaluation for treatment of their pharmacoresistant epilepsy. The MTurk cohort comprised 1,076 individuals (aged $19 - 69$ years; mean age = 39.8 years; 48.9% males; 70.0% with bachelors degrees) who contributed one or more sessions of data across two different experiments: 1. A four-session experiment comprising two sessions of RAVLT and two sessions of FR, 2. A single session experiment involving one of three tasks: FR, CatFR, or CVLT. We programmed these tasks using the JsPsych library (de Leeuw, 2015) and administered them using psiTurk (Gureckis et al., 2016; Eargle, Gureckis, Alexander, McDonnell, & Martin, 2021), an open-source library created to interface with Amazon's MTurk (Buhrmester, Kwang, & Gosling, 2011). All subjects in the MTurk cohort reported English as their native language and reported residing in the United States.

The clinical cohort consisted of 456 epilepsy patients who enrolled in the DARPA-sponsored restoring active memory (DARPA-RAM) project at collaborating clinical centers (see Weidemann et al., 2019; Adamovich-Zeitlin et al., 2021, 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 patients. 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).

Tasks

FR and CatFR

The FR and CatFR delayed free-recall tasks followed the same procedure as described in Ezzyat et al. (2018) and Weidemann et al. (2019). Each trial of the task consisted of a ten-second countdown, followed by a visual presentation of 12 words. Words were displayed on the screen sequentially for 1600 ms, followed by a 750 ms inter-stimulus interval, with a random jitter of 0 to 250 ms. Following list encoding, subjects performed a 20 second math distractor consisting of the summations of random three integers ranging from 1-9 (A+B+C=?). Participants then had 30 seconds to freely-recall all items presented in the preceding list. On the online version, participants were required to type their responses, while in the patient sample their vocal responses were recorded and annotated offline. Words in the FR task were semantically unrelated to one another, whereas in the CatFR task words were drawn from three different semantic categories. Each session of the task included 25 word lists. The tasks included the same word pool as the one used in the DARPA-RAM 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 participants completing more than one session of FR, the original FR word pool was randomly divided into two word pools, to avoid repetition of items across sessions. This procedure was meant to match the FR task to the RAVLT and CVLT which included a standard and an alternative forms, each including a different word set. To create the two FR forms, an algorithm randomly divided the word pool into 46 lists, consisting of 12 words each, using the pseudo-random number generator. Standard and alternative forms were then created by randomly dividing the 46 lists into two sets (FR^s and FR^a), each including 23 lists per session.

RAVLT and CVLT

Our PsiTurk implementation of RAVLT attempted to replicate the procedures described in the RAVLT Handbook (Schmidt, 1996). Immediately after hearing a list of 15 words (List A) subjects attempted to recall as many words as they could remember by typing each word into a text box. This immediate free recall procedure repeated five times with List A, which appears each time in the same order. Following these five learning trials involving List A, subjects complete a single encoding and recall trial of a 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 minutes before recalling List A one final time ('long delay recall'). The words presented in the RAVLT were taken from either the standard (RAVLT^s) or the alternate (RAVLT^a) form of the RAVLT manual (Schmidt, 1996). 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

From the 1,076 participants that took part in the MTurk study, 949 participants participated in the single session experiment on Amazon's MTurk (either FR, CatFR, or CVLT).

The remaining 127 participants completed four sessions of the experiment, separated a week and a half apart. Each participant 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^a, RAVLT^a, FR^s). After completing the first session and passing exclusion criteria[1], subjects were invited to complete three additional sessions of the experiment.

[1] We excluded subjects who met any of the following criteria: 1. recalling $< 10\%$ or $> 95\%$ of the words, 2.

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 words repetition.

Data and Code Availability

Anonymized data and analysis code may be freely downloaded from the public website: `http://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 multi-trial free recall of semantically organized (CatFR) and unrelated word lists (FR) by comparing these laboratory tasks with standard neuropsychological measures: the Rey Auditory Verbal Learning Test (RAVLT) and the California Verbal Learning Test (CVLT). For each task we report measures of predictive (internal) validity, convergent validity, and test-retest reliability.

Recall dynamics for unrelated and categorized word lists

Analysis of the relation 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,

---

having more than one list with 0 correct recalls, and 3. reporting taking notes during the task.

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 $i + lag$ 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 (Kahana, 1996; Healey, Long, & Kahana, 2019). Figure 2D shows the probability that successively recalled study items have similar meanings as measured using Google's Word2vec algorithm (Mikolov, Chen, Corrado, & Dean, 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, Sperling, Sharan, Rosenberg, & Kahana, 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 Figures 2F,G. 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 immediate 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 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 that 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 words pairs in each list (compare Figures 2D and 3D). For the 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 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 that of 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 pronounced impairment of verbal episodic memory. By comparing RAVLT and CVLT scores obtained from neuropsychological reports to normative scores (Schmidt, 1996; Delis et al., 2000), 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.

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. The greater degree of impairment evident in FR and CatFR may reflect the buildup of proactive interference across multiple list presentations.

To further establish the internal validity of the FR and CatFR tasks, we investigated the well-known age-related decline in verbal free recall (Wingfield & Kahana, 2002; Kahana, Dolan, Sauder, & Wingfield, 2005). 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 on the RAVLT and CVLT, but these were not a large as the declines seen on 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 had significantly stronger effects than the CVLT ($b = -0.163, p < 0.05$ and $b = -0.283, p < 0.001$), and stronger effects than the RAVLT ($b = -0.143$, $p = 0.153$ and $b = -0.264, p < 0.05$).

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 that performed two (or more) of the tasks of interest. For patients in our clinical sample we almost always 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 vs RAVLT: $r(125) = 0.28$, $p < 0.01$, Hospital FR1 vs RAVLT: $r(25) = 0.62$, $p < 0.001$, Hospital CatFR vs RAVLT: $r(28) = 0.50$, $p < 0.01$, Hospital FR1 vs CVLT: $r(92) = 0.33$, $p < 0.01$, Hospital CatFR1 vs CVLT: $r(77) = 0.30$, $p < 0.01$). For comparison, we obtained published data on the RAVLT and CVLT (Schmidt, 1996), which exhibited a similar inter-task correlation ($r(58) = 0.47$, $p < 0.001$) 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 for 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 multi-session community cohort provided data for the "MTurk" correlations, and Schmidt (1996) and Delis et al. (2000) provided values for the test-retest "Metanorm" data.

Figure 7 shows the test-retest Pearson r correlation value for each test. In all cases

we observed reliability values of around ~ 0.8, indicating a high degree of consistency in performance across repeated test administration. Overall, we found that the FR and CatFR tasks exhibited somewhat higher reliability coefficients than the standard neuropsychological tests (MTurk FR vs. MTurk RAVLT: $r_1(114) - r_2(114) = 0.125$, $p < 0.05$, Hospital FR vs. Metanorm RAVLT: $r_1(25) - r_2(85) = 0.0562$, $p = 0.465$, Hospital CatFR vs. Metanorm CVLT $r_1(48) - r_2(286) = 0.0691$, $p = 0.109$). This increase in reliability likely reflects the greater number of lists in FR and CatFR as compared with RAVLT and CVLT, and also the randomization of items across repeated lists.

Discussion

Modern research on human memory generally entails repeatedly evaluating a subject's memory for multiple lists of trial-unique items. This approach entails multiple advantages over the single-list assessments typically used in neuropsychological assessments. Here we provide evidence for the clinical validity of repeatedly evaluating memory for distinct lists of unrelated item (FR) or categorically-organized items (CatFR). We show that each of these tasks compares favorably with standard neuropsychological tests: the Rey Auditory Verbal Learning Task (RAVLT) and the California Verbal Learning Task (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 in a sample of patients with drug-resistant epilepsy and clear age-related declines in memory performance in a large online sample. 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 inter-task 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

than those of the established neuropsychological tests. For the measures of convergent validity and test-retest reliability we report data from both our clinical sample and a large community sample recruited through an online platform, Amazon's Mechanical Turk.

Impaired episodic memory constitutes one of the most disturbing aspect of both healthy aging, neurological disease, such as Alzheimer's, and brain injury (Craik, 2000; Huang & Mucke, 2012; Adamovich-Zeitlin et al., 2021). As 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 which 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 (Naveh-Benjamin, 2000; Buchler, Faunce, Light, Gottfredson, & Reder, 2011) and in a variety of neurological conditions (Dickerson & Eichenbaum, 2010; Vakil, in press). The pattern of deficits evident in free recall tasks predict conversion from mild cognitive impairment to Alzheimer's disease (Trenkle, Shankle, & Azen, 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 minute 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. 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, Aggarwal, and Phan (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, Phan, & Kahana, 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 subjects 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 minute 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 multi-list 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 established reliable relations between behavioral and brain measures.

References

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

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

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.

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.

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

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

Delis, D., Kramer, J., Kaplan, E., & Ober, B. (2000). *California verbal learning test*. Psychological Corporation New York.

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

Eargle, D., Gureckis, T., Alexander, S., McDonnell, J., & Martin, J. (2021, October). *psiturk: An open platform for science on amazon mechanical turk (version v3.2.1).*

Zenodo.

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*, 1021-1028. doi: 10.1002/gps.1010

Ezzyat, Y., Wanda, P., Levy, D. F., Kadel, A., Aka, A., Pedisich, I., ... Kahana, M. J. (2018). Closed-loop stimulation of temporal cortex rescues functional networks and improves memory. *Nature Communications*, *9*(1), 365. doi: 10.1038/s41467-017-02753-0

Glanzer, M., & Cunitz, A. R. (1966). Two storage mechanisms in free recall. *Journal of Verbal Learning and Verbal Behavior*, *5*, 351-360.

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

Hanslmayr, S., Klimesch, W., Sauseng, P., Gruber, W., Doppelmayr, M., Freunberger, R., ... Birbaumer, N. (2007). Alpha phase reset contributes to the generation of erps. *Cerebral Cortex*, *17*(1), 1–8.

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

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. doi: 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*, 85–98.

Huang, Y., & Mucke, L. (2012). Alzheimer mechanisms and therapeutic strategies. *Cell*,

148(6), 1204–1222.

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. doi: 10.1016/j.acn.2006.12.002

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

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

Kahana, M. J. (2020). Computational models of memory search. *Annual Review of Psychology*, *71*(1), 107–138. doi: 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. doi: 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. doi: 10.1093/geronb/60.2.P92

Kahana, M. J., Wanda, P. A., Ezzyat, Y., Adamovich-Zeitlin, R., Lega, B., Jobst, B. C., ... Diaz-Arrastia, R. R. (Submitted). Biomarker-guided neuromodulation aids memory in traumatic brain injury. *MedRxiv*.

Kragel, J. E., Ezzyat, Y., Sperling, M. R., Gorniak, R., Worrell, G. A., Berry, B. M., ... Kahana, M. J. (2017). Similar patterns of neural activity predict memory function during encoding and retrieval. *NeuroImage*, *155*, 60–71. doi: 10.1016/j.neuroimage.2017.03.042

Kucewicz, M. T., Berry, B. M., Kremen, V., Miller, L. R., Khadjevand, F., Ezzyat, Y., ... Worrell, G. A. (2018, January). Electrical stimulation modulates high gamma activity and human memory performance. *eNeuro*, *5*(1). doi: 10.1523/ENEURO.0369-17.2018

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. doi: 10.1523/JNEUROSCI.5321-11.2012

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781v3*.

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*, 1170-1187.

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

Schmidt, M. (1996). *Rey auditory verbal learning test.* Printed. Torrance, CA: 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. doi: 10.1523/JNEUROSCI.23-34-10809.2003

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.

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. doi: 10.1080/13803390490919245

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

Weidemann, C. T., Kragel, J. E., Lega, B. C., Worrell, G. A., Sperling, M. R., Sharan, A. D., ... 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. doi: 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*, 354-388. doi: 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.

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. doi: 10.1177/0891988707308801

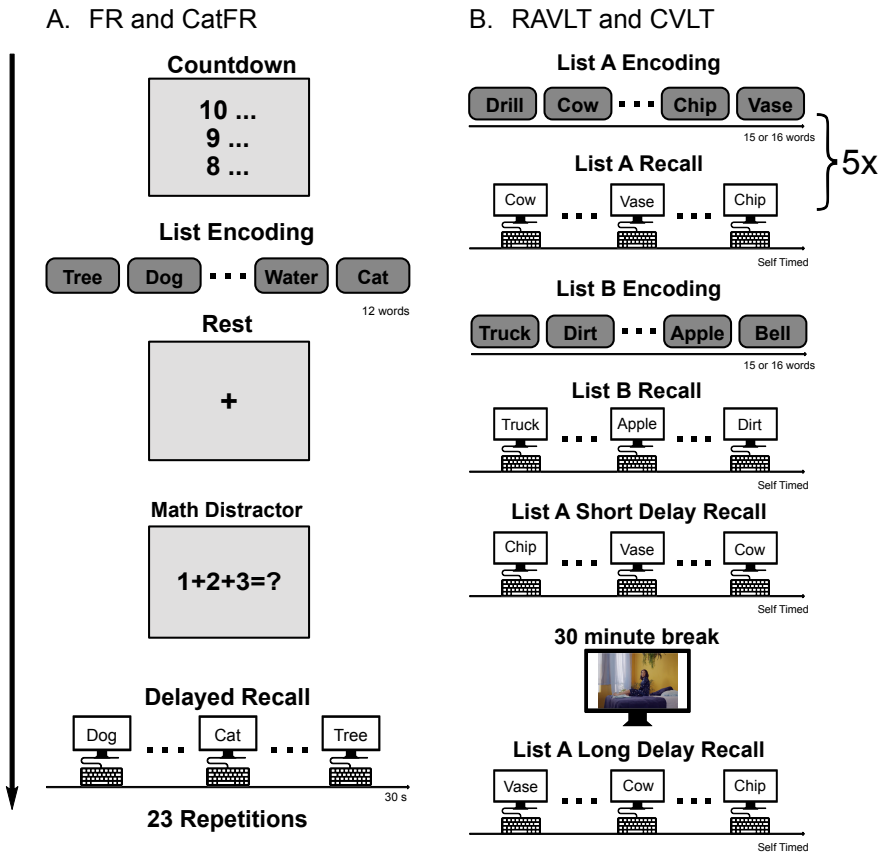

Figure 1

Schematic illustration of the study tasks: *A. Repeated-trial free recall of unrelated (FR) and categorized lists (CatFR). B. Rey Auditory Verbal Learning Task (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* Methods*).*

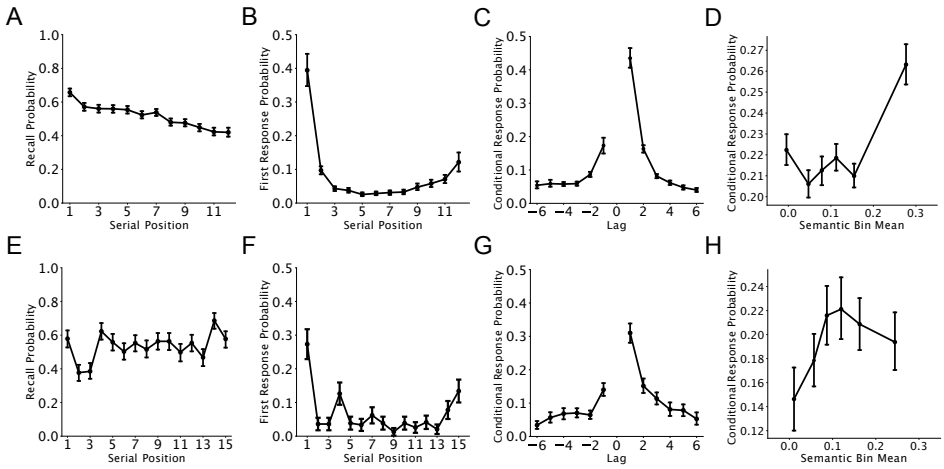

Figure 2

Recall dynamics for unrelated word lists. *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 Task (RAVLT). All graphs illustrate data obtained from our online sample recruited through Amazon's Mechanical Turk. A, E: Serial position curves. B, F: Probability of first recall curves. C, G: Lag-CRP curve illustrating the likelihood of a recall transition from study item i to study item $i + lag$ conditional on transition availability. D, H: Semantic CRP curve 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.*

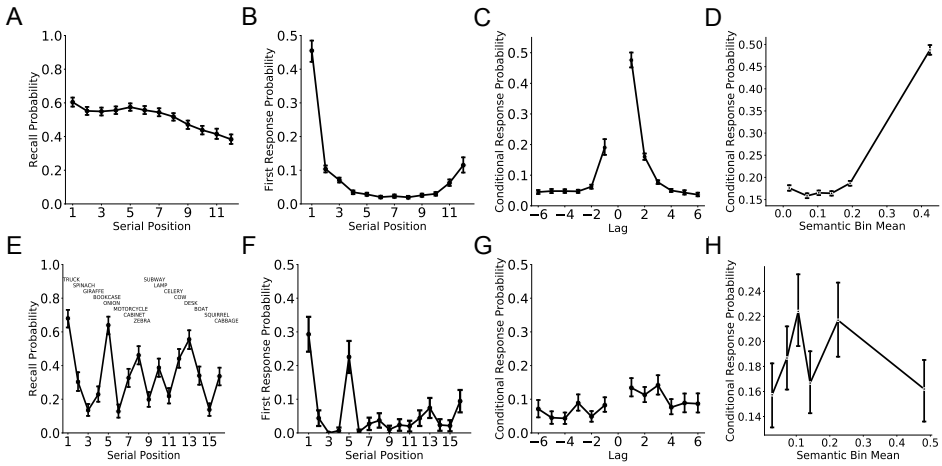

Figure 3

Recall dynamics for categorized recall. *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 Mechnical Turk. A, E: Serial position curves. B, F: Probability of first recall curves. C, G: Lag-CRP curve illustrating the likelihood of a recall transition from study item i to study item $i + lag$ conditional on transition availability. D, H: Semantic CRP curve 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.*

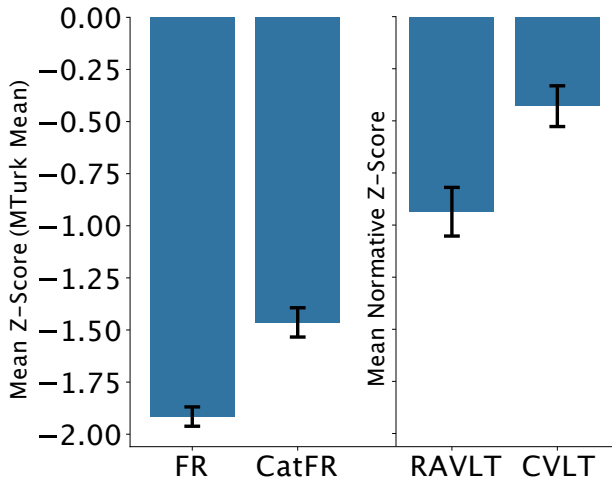

Figure 4

Epilepsy vs. Healthy Population. *Performance of the epileptic population as compared to healthy controls (MTurk workers for FR1 and CatFR1, metanorms for RAVLT and CVLT) 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 standard error of the mean.*

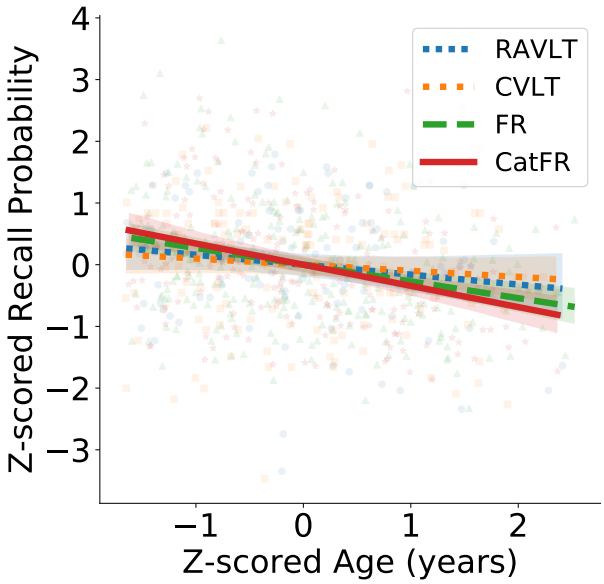

Figure 5

Age-related decline in recall performance. *The expected negative correlation between age and overall recall performance appeared in multi-trial 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*

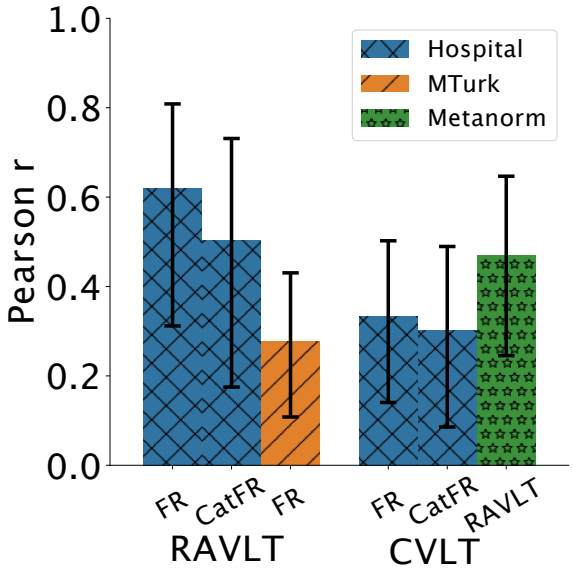

Figure 6

Task Performance Correlations. *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 FR1). Error bars represent 95% confidence intervals, calculated using standard Fisher transformations.*

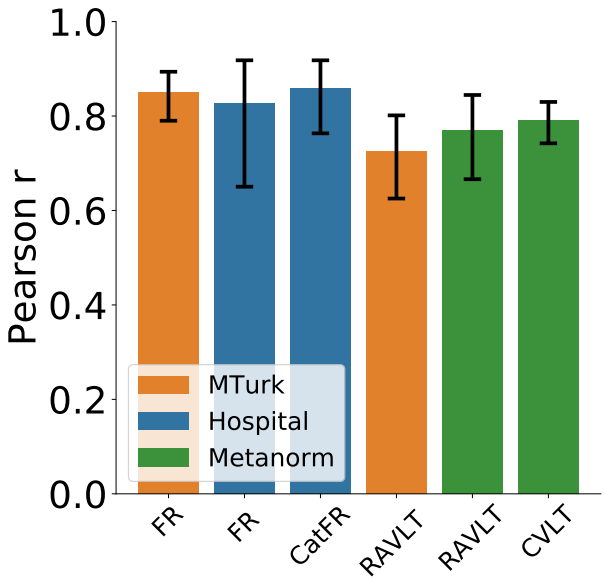

Figure 7

Hospital, Metanorm, and MTurk Test Retest Correlations. *Test-retest correlation for all tasks and administration types, calculated as the correlation coefficient between first and second session of the same task. Laboratory tasks (FR and CatFR) administered on MTurk and in the hospital show stronger correlations than neuropsychological tasks (RAVLT and CVLT) on MTurk or in metanormative data. Error bars represent 95% confidence intervals, calculated using standard Fisher transformations.*