

The Penn Electrophysiology of Encoding and Retrieval Study

Michael J. Kahana, Lynn J. Lohnas, M. Karl Healey, Ada Aka, Adam W. Broitman, Patrick Crutchley, Elizabeth Crutchley, Kylie H. Alm, Brandon S. Katerman, Nicole E. Miller, Joel R. Kuhn, Yuxuan Li, Nicole M. Long, Jonathan Miller, Madison D. Paron, Jesse K. Pazdera, Isaac Pedisich, Joseph H. Rudoler, and Christoph T. Weidemann
Department of Psychology, University of Pennsylvania

The Penn Electrophysiology of Encoding and Retrieval Study (PEERS) aimed to characterize the behavioral and electrophysiological (EEG) correlates of memory encoding and retrieval in highly practiced individuals. Across five PEERS experiments, 300+ subjects contributed more than 7,000 memory testing sessions with recorded EEG data. Here we tell the story of PEERS: its genesis, evolution, major findings, and the lessons it taught us about taking a big scientific approach in studying memory and the human brain.

Keywords: memory, free recall, scalp EEG, cognitive modeling

Although Herman Ebbinghaus gained fame for his herculean 1,800 hr of self-experimentation on list learning (Ebbinghaus, 1885/1913), much of what we have subsequently learned about memory has derived from single-session studies performed by fewer than 100 naive individuals. This approach contrasts with research areas such as perception, where a small handful of observers often contribute data across a dozen or more experimental sessions. Whereas larger samples permit broader inference, intensive study of a small number of subjects can enable detailed analyses and model-fitting of individual behavior.¹

The Penn Electrophysiology of Encoding and Retrieval Study (PEERS) sought to obtain high-resolution, within-subject data from a large number of subjects performing a variety of episodic memory tasks. We pursued three primary aims across five experiments: (a) to obtain sufficient trial-level data so that we could apply models to individual subject performance measures, (b) to obtain sufficient data across subjects to permit the analysis of individual differences, and (c) to obtain high-quality continuous electrophysiological (EEG) data during memory encoding and retrieval, thus allowing us to relate brain measures to indices of performance with high reliability. To study individual differences, we also collected a variety of psychometric measures, including scales of intelligence, personality, mood, and anxiety.


We achieved these goals through a 10-year data collection effort in which more than 300 subjects contributed data from more than 7,000 sessions of memory testing. Although findings emerging from these studies have appeared in earlier publications, the present article presents the overarching motivation, methods, behavioral and electrophysiological results, and implications of this large memory study.

Background and Motivation

For science to advance humanity's most noble goals, society must trust the work of scientists. Failures to replicate high-profile scientific findings have captivated the attention of both scientists and the lay public, calling into question the enterprise of scientific inquiry. Although research on human memory has fared better than some subdisciplines, our field faces the same forces that impede replicability. One such force is the extreme variability of human cognition, behavior, and physiology (e.g., Kahana et al., 2018; Kahneman et al., 2021). Empirical patterns can differ reliably across individuals and even within an individual, and variability in these effects does not arise solely due to external variables, such as the memorability of items, or the conditions of encoding and retrieval. Rather, it appears that variability results from endogenous factors within each individual (Kahana et al., 2018; Weidemann & Kahana, 2021).

Cognitive neuroscience faces even greater challenges to replicability than does cognitive psychology (Poldrack et al., 2017; Szucs & Ioannidis, 2017). This is because variability in task performance must give rise to variability in brain activity, but measured brain activity includes additional sources in variability beyond that seen in overt behavior. In the case of EEG, these sources of variability include electromyographic (EMG) signals produced by eye and muscle movements, as well as other sources of electrical noise outside of the subject. In the case of functional magnetic resonance imaging, noise can come from head movements and non-cognitive predictors of intracerebral blood flow (Liu, 2016). In addition, these measurement modalities record only a small fraction of brain activity, with the precise neural activity recorded varying across individuals and recording sessions. Moreover, many features of brain activity that vary in a session have little to do with task performance but may be correlated with other brain signals for uninteresting reasons. Thus, studies seeking to establish brain-behavior relations may require more trials to achieve the same level of statistical power than do studies relying on purely behavioral measures. Yet, the cost of obtaining neural measures forces researchers to economize on data

Michael J. Kahana  <https://orcid.org/0000-0001-8122-9525>

Jesse K. Pazdera  <https://orcid.org/0000-0002-4913-9236>

Michael J. Kahana and Lynn J. Lohnas contributed equally to this work which was funded by National Institutes of Health Grant (MH55687).

Correspondence concerning this article should be addressed to Michael J. Kahana, or Lynn J. Lohnas, Department of Psychology, University of Pennsylvania, Suite 302C, 3401 Walnut Street, Philadelphia, Pennsylvania 19104, United States. Email: kahana@psych.upenn.edu or ljlohnas@syr.edu

¹ Smith and Little (2018) articulated the benefits of small *N* studies.

collection, thus fueling variability across experiments. The highly multivariate nature of neural data encourages researchers to look at their data in myriad ways (curse of dimensionality). This increases the chance of false positives unless every step of an analysis pipeline has been “preregistered” (Simmons et al., 2011). On the other hand, neural recordings typically produce many more observations per unit of time than behavioral measures, which in some cases may increase statistical power to detect reliable brain–behavior relations. Future research with large open data sets may help to unravel the complex relations between these variables. To probe the upper bound of what could be learned using scalp EEG, we assembled the PEERS data sets, which comprise millions of encoding and retrieval events. The large number of trials and sessions contributed by each subject allowed us to conduct analyses at the individual trial level and validate these analyses across sessions, people, and task manipulations.

Given that nearly all cognitive neuroscience data sets encompass fewer than 100 hr of experimental data collection, increasing our data set by a factor of 10 would have been sufficient to address replicability. In PEERS, we exceeded this benchmark by a factor of 100. Beyond replicability, PEERS sought to provide adequate data for the study of individual differences in both overt behavior and physiology, and also to provide data with sufficient resolution for individual subject modeling (e.g., Healey & Kahana, 2014, 2016). The ultimate goal of this program, which we have yet to achieve, is a detailed model-based analysis of subject-specific electrophysiology. Rather than waiting until all of the work is done, we have embraced an open-science approach and have disseminated the PEERS data for several years. We feel that it is now time to pull together the as-yet-incomplete PEERS story, both in the interests of fostering a discussion of big science approaches to the study of human cognition and to help further disseminate PEERS data to interested scientists.

Method Overview

Each of the five PEERS experiments involved multiple sessions of memory tasks with EEG recording. Although one might hope that this major experimental endeavor resulted from a careful planning exercise, with scholars representing diverse interests advising as to the optimal choice of tasks, the initial choice of tasks arose from a desire to obtain high-resolution data on a set of tasks that had fueled our lab’s theoretical work on human memory. Each experiment involved some variant of a free recall task. Experiments 1 to 3 included encoding task manipulations, variation in distractor conditions, and end-of-session recognition and final-recall tests. Subjects completed these first three PEERS experiments across 20 sessions. In addition, this cohort also completed two sessions of standardized cognitive and emotional assessments including the Wechsler Adult Intelligence Scale, the NEO Five Factor Personality Inventory (McCrae & Costa, 2010), the California Verbal Learning Test, and several subtests of the Wechsler Memory Scale. PEERS Experiment 4 sought to maximize the statistical power of data collected in a delayed recall-task without any encoding task manipulations (we estimated, based on earlier PEERS studies, that subjects could maximally complete 24 2-hr-long sessions in a single term). Because human speech is a natural medium for recalling information, we collected vocal responses which we annotated (offline) for accuracy and response times. But because vocalization causes EMG artifacts in EEG data, we conducted a fifth PEERS experiment to control for premotor correlates of retrieval. This last PEERS study concluded just before the COVID-19 pandemic.

Due to the very substantial investment of time and resources, each subject first participated in a screening session to ensure that they understood the demands of the experiment before signing on for the full experiment. Below we provide a concise description of the experimental methods. Table 1 gives the number of subjects who completed each experiment. Additional procedural details appear in online appendix at <https://memory.psych.upenn.edu/PEERS>.

PEERS Experiments 1 and 3

Because Experiments 1 and 3 were virtually identical, we describe their methods together. As illustrated in Figure 1A, each session comprised a series of 16 immediate free recall trials, each involving a unique list of 16 visually presented words. Each session ended with a recognition test. Half of the sessions were randomly chosen to include a final free recall test before recognition (in final-free recall [FFR], subjects attempt to recall as many words as they can remember from all 16 lists). Experiment 3 differed from Experiment 1 in that a subset of subjects received externalized free recall (EFR) instructions. In externalized recall (Kahana et al., 2005), subjects verbalized all words that came to mind at the time of test (even if they thought those words did not occur in the most recent list or had already been recalled during the current recall period) and pressed the spacebar to indicate awareness of any such error.

Subjects encountered three types of lists: (a) no-task lists, which they studied with the generic instruction of trying to learn the items for a subsequent test, (b) task lists, where each item appeared concurrently with a cue indicating one of two judgments (size or animacy) the subject should make for that word, and (c) task shift lists, where subjects alternated between size and animacy tasks every two to six items within each list. The size task asked subjects “Will this item fit into a shoebox?”; the animacy task asked subjects “Does this word refer to something living or not living?” The color, font, and case of the presented item indicated the current task. Each session included 12 task lists and four no-task lists. The first session of PEERS Experiment 1 included equal numbers of size, animacy, and task-shift lists; subsequent sessions included three size, three animacy, and six task-shift lists. We constructed a pool of 1,638 words for use in PEERS Experiments 1 to 3. Based on the results of a prior norming study, only words that were clear in meaning and that could be reliably judged in the size and animacy encoding tasks were included in the pool.

PEERS Experiment 2

Experiment 2 introduced a within-subject, within-session, distractor manipulation (Figure 1B). In addition to immediate free recall trials, as in Experiments 1 and 3, this experiment introduced delayed free recall and continual distractor-free recall, with distractor intervals of varying duration. In each distractor interval, subjects solved math problems of the form $A + B + C = ?$, where A , B , and C were positive, single-digit integers. When a math problem appeared, subjects typed the sum as quickly as possible consistent with high accuracy (they received a monetary bonus based on the speed and accuracy of their responses). For the distractor intervals in the first two lists, one list had a distractor period following the last word presentation for 8 s and the other had an 8 s distractor period prior to and following each word presentation. In the remaining 10 lists, subjects performed free recall with five possible durations for the between-item and end-of-list distractor tasks,

Table 1
Demographic Information for PEERS Studies

PEERS experiment	<i>N</i>	Sessions	Dates
Preliminary experiment	~730	1	2010–2019
Exp. 1: Immed. recall + task manip. Final-free recall. Recognition.	172	7	2010–2014
Exp. 2: Recall + distractors. Final-free recall. Recognition.	157	7–9	2010–2014
Exp. 3: Exp. 1 + externalized recall. Final-free recall. Recognition	60 (IFR), 92 (EFR)	4 (IFR), 6 (EFR)	2010–2014
Exp. 4: Delayed recall	98	24	2014–2018
Exp. 5: Long-delay recall + premotor control	57	10	2019–2020

Note. PEERS = Penn Electrophysiology of Encoding and Retrieval Study; Exp. = experiment; Immed. = immediate; IFR = immediate free recall; EFR = externalized free recall.

such that two lists had each of the five conditions. As listed here, the first number indicates the between-list distractor duration and the second number indicates the end-of-list distractor, both in seconds: 0–0, 0–8, 0–16, 8–8, and 16–16. A 0 s distractor refers to the typical, non-filled duration intervals as described for Experiments 1 and 3. Subjects encoded all items using either a size or an animacy judgment task. Session 1 included seven size-judgment lists and seven animacy judgment lists. Subsequent sessions included six task-shift lists, three size-task lists, and three animacy-task lists.

PEERS Experiment 4

This experiment sought to simplify the methodology used in previous experiments, focusing exclusively on delayed free recall. Here each of 98 subjects completed 24 sessions of delayed free recall. Each session consisted of 24 trials, with each trial containing a list of 24 individually presented words followed by a 24-s distractor period (see Figure 1C). A random half of the lists (excluding the first list) were preceded by a 24-s, distractor-filled delay. A free recall test followed the postlist distractor on each list.

The word pool for this experiment consisted of a 576-word subset of the 1,638-word pool used in a previous PEERS experiment, and subjects saw the same 576 words (24 lists × 24 items) on each of sessions 1 through 23 with the ordering of words randomized for each session. The 24th session introduced a set of novel words. Subjects were given a short break (~5 min) after every eight lists in a session.

PEERS Experiment 5

The fifth PEERS experiment sought to contrast neural correlates of retrieval following a very long delay, with neural correlates of retrieval of a just presented single item. During each of the first five sessions, subjects quietly read each of the 576 words used in Experiment 4. After reading each word, they waited 1 s (or longer) before saying it aloud. These 576 immediate recall trials occurred in 24 blocks of 24 items, each preceded by a countdown, thus mimicking the 24-list structure of Experiment 4.

At the start of session six, subjects received a surprise free recall task in which they were instructed to recall as many words as possible from the previous sessions in any order, while also vocalizing any additional words that come to mind in their attempt to recall these items (externalized recall instructions: Kahana et al., 2005; Lohas et al., 2015; Zaromb et al., 2006). We administered this long-delay recall task at the start of Sessions 6 through 10, giving subjects 10 min to recall as many of the 576 words as they could remember. After this free recall test, subjects continued with the same immediate recall task as in earlier sessions.

Compensation and Performance-Based Bonus

In each of the PEERS experiments, subjects received a base salary for their participation. In addition, they received a modest bonus for performance and a separate bonus for completing all of the sessions. The performance bonus varied slightly across experiments, but it incentivized subjects to achieve high levels of performance on both the memory tasks and the arithmetic distractor tasks. In addition, we provided a bonus to subjects for maintaining a low eyeblink rate during critical item presentation events.

PEERS Raw Data Repository and Online Method Description

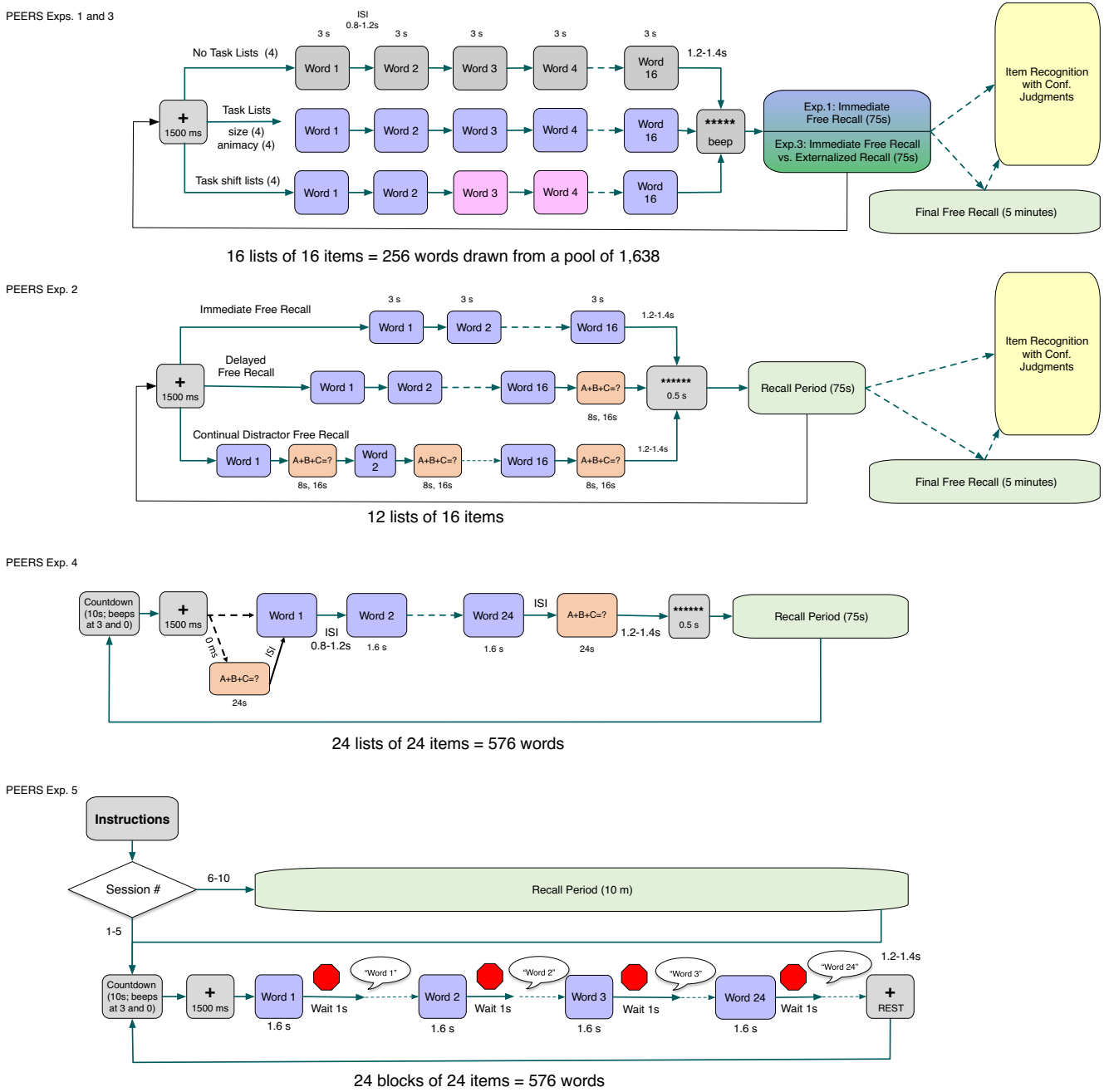
Data sharing is critical to the scientific impact of any large-scale data collection effort like PEERS. Sharing data increases the transparency of our research, and we hope that it can enable other research groups to replicate and extend our work in new directions. For maximal impact and integrity, best practice is to follow the FAIR data principles (Wilkinson et al., 2016)—data must be findable, accessible, interoperable, and reusable. We chose to format our data according to the increasingly popular and exceptionally well-documented Brain Imaging Data Structure (BIDS; Gorgolewski et al., 2016), which was initially developed for the functional magnetic resonance imaging (fMRI) community and has been extended to support EEG in recent years (Pernet et al., 2019). As a community standard BIDS has gained tremendous momentum and financial support, leading to the development of online data repositories like OpenNeuro (Markiewicz et al., 2021) and tools for formatting and parsing data with popular scientific programming languages like Python (we used MNE-BIDS; see Appelhoff et al., 2019) and MATLAB.

The PEERS data, both behavior and electrophysiology, are freely available as OpenNeuro Data Set ds004395 (Kahana et al., 2023). Data can be downloaded directly through the OpenNeuro web interface (<https://openneuro.org/datasets/ds004395/>) or by using their command line utility tool. The data set has its own digital object identifier (DOI) and citation tools are available on the data set webpage. The Computational Memory Lab website also provides a detailed methods description of each of the PEERS studies described above: <https://memory.psych.upenn.edu/PEERS>.

Results

Here we present our results organized into five major sections. The Overview of Classic Behavioral Findings section provides an overview of the basic behavioral findings. The Individual Differences and Cognitive Modeling section discusses both experimental and

Figure 1
Schematic Diagram of PEERS Methods



Note. The same group of subjects took part in Experiments 1 to 3, across 20 experimental sessions. Experiments 4 and 5 involved separate subject groups, recruited in later years of the project. Each experiment involved some form of a free recall task, and Experiments 1 to 3 also included recognition and final-free recall tasks. Experiment 5 only included final-free recall. For Experiments 1 to 3, subjects either studied items without a specific encoding task, or judged items' size or animacy. The color of the word bubbles in the first row of the schematic indicates the encoding task. Experiment 2 also included an encoding-task manipulation not shown in the schematic diagram. The Method section provides many details omitted here. PEERS = Penn Electrophysiology of Encoding and Retrieval Study; Exp. = experiment; Conf. = confidence; ISI = interstimulus interval. See the online article for the color version of this figure.

endogenous sources of variability in recall of items and lists. The Variability in Recall Across Items and Lists section focuses on EEG correlates of successful memory encoding. The EEG Correlates of Successful Memory Encoding section focuses on individual

differences and model-based analyses of performance. The Spectral Markers of Memory Retrieval section discusses EEG correlates of memory retrieval. Whereas the Overview of Classic Behavioral Findings section summarizes basic behavioral findings, including

This document is copyrighted by the American Psychological Association or one of its allied publishers. This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

those replicating prior work, subsequent sections highlight novel findings that would not have been easily discerned without the substantially greater statistical power afforded by the PEERS data sets.

Overview of Classic Behavioral Findings

The PEERS free-recall experiments replicated many classic findings, including serial position effects (Deese & Kaufman, 1957; Murdock, 1962), temporal clustering (Healey et al., 2019; Kahana, 1996; Ward et al., 2010), semantic clustering (Bousfield, 1953; Howard & Kahana, 2002; Jenkins & Russell, 1952; Polyn et al., 2009), the exponential growth of interresponse times with output position (Murdock & Okada, 1970; Patterson et al., 1971; Pollio et al., 1968; Rohrer & Wixted, 1994; Unsworth, 2007), and subjects' tendency to commit extra-list intrusion (ELI) and prior-list intrusion (PLI) as a function of their temporal and semantic relation to the just-recalled items (Zaromb et al., 2006). Long et al. (2017) discussed results of the encoding task manipulation in PEERS Experiment 1. Briefly, subjects exhibited better recall and stronger temporal clustering under free encoding conditions than when asked to make size or animacy judgments during word encoding.

PEERS Experiment 2 replicated all of the classic findings concerning distractor effects, including the reduction in recency with increased length of an end-of-list distractor, but recovery of recency with increased length of a within-list (interitem) distractor (Bjork & Whitten, 1974; Kahana, 2017; Lohnas & Kahana, 2014) (also, see Figure 2). Here we can also see the striking similarity in recall initiation across immediate and continual-distractor free recall, and the substantial attenuation in recency in delayed free recall (Figure 2C). As demonstrated by Howard and Kahana (1999), the contiguity effect does not differ across the distractor conditions, indicating that whatever enables subjects to make transitions between neighboring items depends on the relative and not the absolute distances between the items (Figure 2D). Finally, we find striking effects of semantic similarity on free recall (e.g., Manning et al., 2012), across all distractor conditions (Figure 2E).

PEERS Experiment 3 compared free recall under standard and externalized recall instructions. In externalized recall, the experimenter instructs subjects to recall any item that comes to mind as they are trying to remember the lists, even if they realize that it was not a studied item, or if it is an item that they have already recalled. In these cases, we instruct subjects to press the space bar to "reject" the item they just recalled. As expected from prior work (Kahana et al., 2005; Zaromb et al., 2006) externalized instructions elicit many more PLI and ELI, but have little or no effect on correct recalls (Lohnas et al., 2015). Inclusion of externalized recall instructions provided valuable data on intrusions that rarely occur in standard free recall.

Because subjects participated in PEERS Experiments 1 to 3 as a series of experiments, data from PEERS Experiment 3 provides valuable information on free recall under conditions of high practice (i.e., after performing 12 or more sessions of PEERS Experiments 1 and 2). Practice exerted a large effect on temporal organization, with subjects exhibiting a stronger tendency to make transitions among neighboring items in later experimental sessions (see Figure 3B). This finding also appeared in PEERS Experiment 4.

PEERS Experiments 1 to 3 included two additional measures of memory following all of the lists in a given session: on a random half of sessions, subjects performed a FFR test on all prior lists. This FFR test came immediately after the recall period for the

final list (see Figure 1). In FFR, subjects exhibited a long-term recency effect, seen in the much higher recall rates for items on the last few lists. Subjects also exhibited a negative within-list recency effect, as seen in worse FFR rates for the last few items in each list (Craik, 1970; Kuhn et al., 2018). After FFR (or if absent, after the recall period of the 16th study list), subjects performed a recognition memory task, with confidence judgments, on a percentage of items studied across all of the lists (see, Lohnas & Kahana, 2013; Weidemann & Kahana, 2016, for details). Performance in these tasks replicated classic findings concerning the relations between confidence, accuracy and response times, as well as yielding novel insights into the relation between response time and confidence (see Weidemann & Kahana, 2016).

PEERS Experiment 4 created a much simpler experimental scenario in which to examine the electrophysiology of memory encoding and retrieval. Free encoding instructions simplified item presentation and minimized eye movements evoked by the task cue in PEERS Experiments 1 to 3. Delayed free recall facilitated aggregation across list items by reducing the size of the recency effect. Owing to its simplicity and repetitive structure, PEERS Experiment 4 provides a particularly rich data set for the study of variability in memory, across items, lists, and sessions (see Individual Differences and Cognitive Modeling section). PEERS Experiment 5 aimed to test novel hypotheses regarding the electrophysiological correlates of memory retrieval at short and long delays. A discussion of the EEG correlates of memory encoding and retrieval in each of the PEERS studies appears in later sections.

Individual Differences and Cognitive Modeling

PEERS data provided a unique window into individual differences in both behavior and physiology. Healey and Kahana (2014) examined the effects of primacy, recency, temporal contiguity, and semantic clustering at the level of individual subjects. They found that 90% of Experiment 1 subjects showed recency, 93% showed primacy, at least 96% showed a forward-asymmetric contiguity effect, and 100% showed semantic clustering. Despite this remarkable level of consistency, the magnitude of these effects varied widely across individuals. Analyzing PEERS Experiments 1 and 2, Healey et al. (2014) found that these four effects represent statistically distinct sources of variability among individuals. Of these, only temporal contiguity and semantic clustering correlated with overall recall performance, suggesting that associative organization processes contribute to successful memory search (see also Sederberg et al., 2010; Spillers & Unsworth, 2011). Moreover, variation in the temporal contiguity effect (but not the other effects) correlated positively with full-scale Wechsler Adult Intelligence Scale Intelligence Quotient (WAIS-IV IQ) (see Figure 4). These findings suggest that the ability to control the drift of mental context representations may be critical not just to memory, but to general intellectual ability (Healey & Uitvlugt, 2019).

We designed the PEERS experiments with the goal of modeling individual-subject data and of using the estimated model parameters to help understand individual differences. Healey et al. (2014) showed one clear reason for the importance of subject-level analysis and modeling: When averaged across subjects, it would appear that in immediate recall, subjects mostly initiate with the final (recency) items, but occasionally initiate with early (primary) items. In this case, aggregation disguised the true nature of the data, wherein most subjects almost always initiate with the final list item but

Figure 2*Recency and Contiguity as a Function of Distractor Conditions in PEERS Experiment 2*

A.

Condition

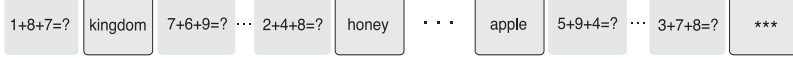
Immediate Free Recall



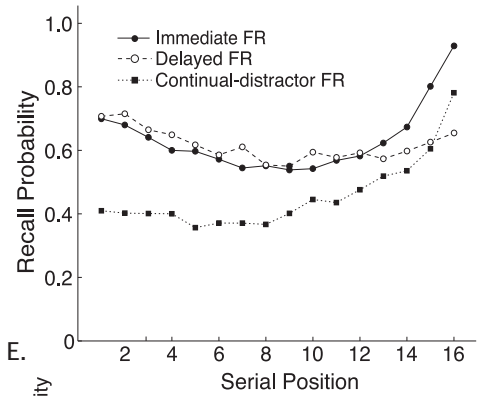
Delayed Free Recall



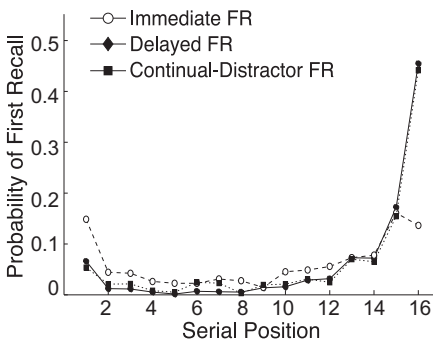
Continual-Distractor Free Recall



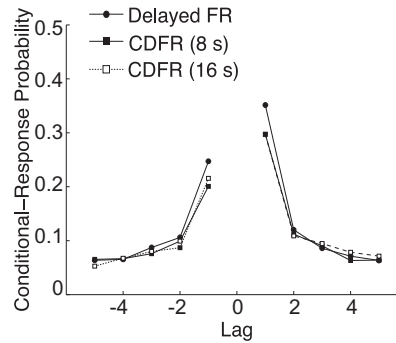
B.



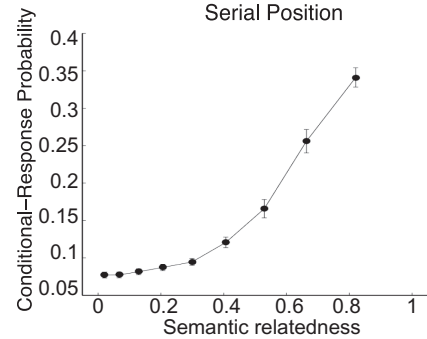
C.



D.



E.



Note. (A) Illustration of IFR, DFR, and CDFR tasks. (B) Serial position analysis showing recency in IFR, attenuated recency in DFR, and long-term recency in CDFR. (C) Recall initiation, as measured by the probability of first recall, shows that initiating with recent items does not differ between DFR and CDFR. (D) Contiguity is generally preserved in all three conditions. (E) Subjects are more likely to recall items that are semantically related to the just-recalled item. PEERS = Penn Electrophysiology of Encoding and Retrieval Study; IFR = immediate free recall; DFR = delayed free recall; CDFR = continual-distractor free recall; FR = free recall.

some subjects almost always initiate with the first list item. Here the average data did not provide an accurate representation of each individual. It seems unlikely that one could discern this pattern of results with a more typically powered memory experiment.

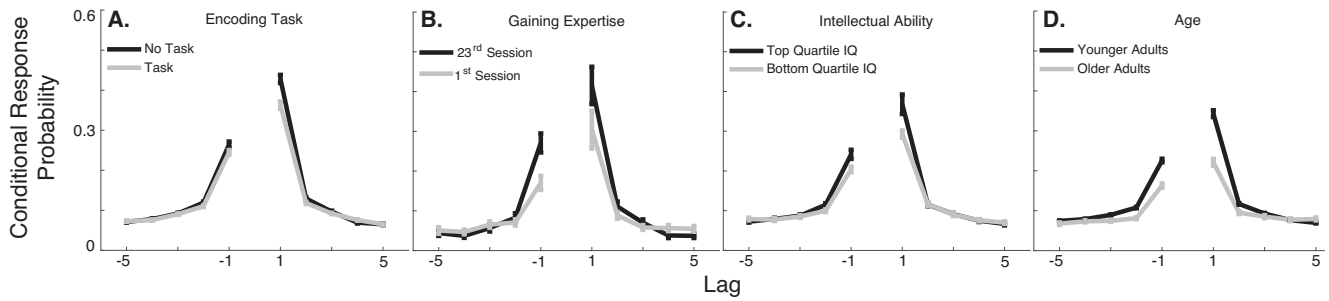
Before conducting individual-level modeling, however, Lohnas et al. (2015) used the PEERS data to extend retrieved-context theory beyond single-list behavioral data. They specifically sought to address a fundamental and frequently neglected problem: How can a model simultaneously account for the gradual accumulation of memories over a lifetime and the specificity with which we are able to retrieve memories learned in a given context? Unlike earlier implementations of retrieved-context theory (RCT) that reset the memory system at the start of each list (e.g., Polyn et al., 2009; Sederberg et al., 2008), they extended the theory to continuously accumulate associations in memory. Their model, termed the context maintenance and retrieval (CMR2) model, inherited basic assumptions of earlier RCT implementations, including the core idea of a slowly drifting representation of temporal context (Manning, in press). The evolution of context follows the standard formalism of RCT in which features of the currently experienced item (recursively) retrieve their associated past contexts, which in turn update the state of context.

Because a theory of multilist memory must account for how subjects target retrieval of items on the most recent list, data on recall errors (intrusions) from prior lists place tight constraints on theory.

However, the sparsity of such PLIs has impeded theory development. PEERS Experiments 1 and 3 provided Lohnas et al. with a sufficiently large dataset to precisely quantify PLI trends and use those to constrain their CMR2 model. Lohnas et al. (2015) proposed that subjects internally generate more recalls than they report, and omit recall of a generated item if it is not recognized as having been studied in the current list. Each generated item retrieves its associated context state from study, and CMR2 only recalls a generated item if it's retrieved temporal context resembles the current context. Although CMR2 can query which items are generated but not recalled, subjects require additional instruction. In the EFR paradigm, subjects attempt to recall all items that come to mind (e.g., Kahana et al., 2005; Roediger & Payne, 1985; Unsworth & Brewer, 2010; Unsworth et al., 2010; Wahlheim et al., 2019). If the subject perceives that they have recalled an item in error, they may “reject” such an item by pressing the spacebar immediately afterwards.

Lohnas et al. (2015) tested the generate-recognize mechanism using data from PEERS Experiment 3 (as shown in Figure 1, some subjects performed EFR, while others studied lists with the same structure, but performed standard free recall). Although subjects engaging in EFR produced more errors, the PFRs and SPCs were nearly identical between the two groups, suggesting that EFR relies on similar cognitive mechanisms to IFR. Buttressing this

Figure 3
Contiguity Modifying Variables



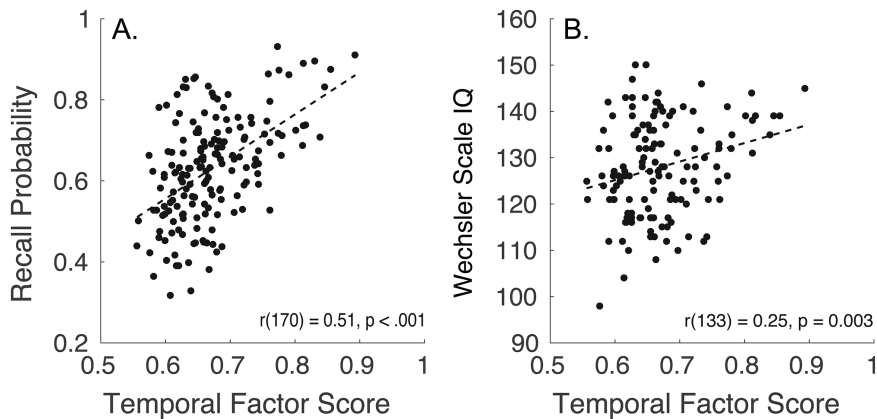
Note. (A) The contiguity effect is smaller when subjects perform an encoding task (making a size or animacy judgment) during study than when freely encoding items. (B) Task experience amplifies the contiguity effect: a large contiguity effect appears on the first session and grows larger by the 23rd session. (C) The contiguity effect also increases with intellectual ability, as measured by WAIS IQ. (D) Contiguity is preserved across the lifespan, but is larger for younger adults than for older adults. WAIS IQ = Wechsler Adult Intelligence Scale Intelligence Quotient.

account, Lohnas et al. (2015) found that CMR2 predicted the proportion and probability of rejection for PLIs in the EFR group, as well as reduced PLIs for the IFR group, using a single set of parameters for fitting data from both subject groups.

Having established that CMR2 accounts for group-level recall data, including rarely occurring PLIs, Healey and Kahana (2014) used the multisessions PEERS data to evaluate the model’s ability to account for variability in individual subject effects. Fitting each individual who participated in PEERS Experiment 1, they found that CMR2 provided a good fit to multiple behavioral effects in ~95% of individual subjects. As PEERS Experiment 1 included data from older adults, Healey and Kahana (2016) further asked whether the model could account for age differences in correct recalls and intrusions. They first fit CMR2 to data from individual younger and older adults using Kahana et al. (2002) as an independent data set for model development, allowing all model parameters to vary, and then identified the smallest subset of parameter changes

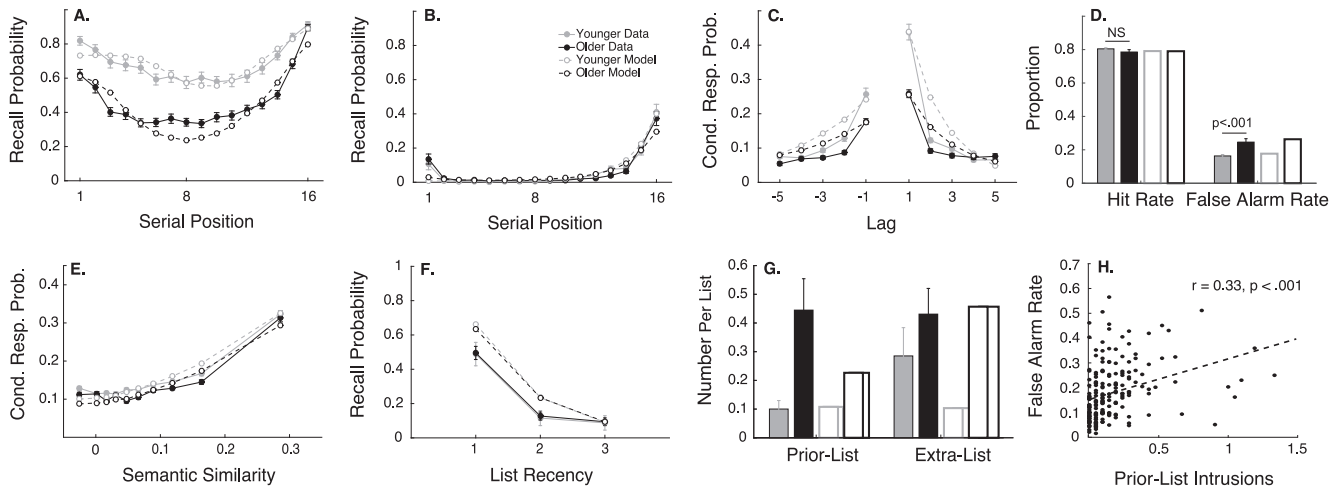
required to capture age-related differences. This method identified four components of putative age-related impairment: (a) contextual retrieval, (b) sustained attention (related to the primacy gradient), (c) error monitoring (related to rejecting intrusions), and (d) decision noise. Figure 5A–G shows that when this full model is applied to the PEERS Experiment 1 data, it provided a reasonable account of younger adults recall dynamics and that adjusting the four components mentioned above enabled the model to account for age-related changes in serial position effects, semantic and temporal organization, and intrusions. They then extended CMR2 to provide a context-similarity model of recognition judgments and age differences therein, based on the same mechanism used to filter intrusion errors. This joint model of free recall and recognition makes the novel prediction that the number of intrusions a subject makes in free recall should correlate positively with the number of false alarms they make in recognition. As shown in Figure 5H, the PEERS data confirmed this prediction.

Figure 4
Individual Differences in Contiguity Predict Memory Performance and General Intelligence



Note. (A) The correlation between temporal factor scores and overall recall probability. Temporal factor scores give the average percentile ranking of the temporal lag of each actual transition with respect to the lags of potential transitions. (B) Those subjects who exhibit greater temporal clustering during verbal free recall (high temporal factor score) also exhibit higher scores on the Wechsler Adult Intelligence Scale IV. IQ = intelligence quotient.

Figure 5
Age-Related Changes in Recall and Recognition



Note. Panels (A) to (C) illustrate serial position, probability of first recall and contiguity effects; Panel (D) illustrates recognition memory hits and false alarms; Panel (E) illustrates semantic organization; Panels (F) and (G) illustrate intrusion errors, and Panel (H) illustrates the correlation between intrusions and false alarms. Black lines/bars indicate data from older adults; gray lines indicate younger-adult data. Solid lines with filled symbols or filled bars show subject data and broken lines with open symbols or unfilled bars show context maintenance and retrieval model simulations. Cond. = Conditional; Resp. = response; Prob. = probability. Adapted from “A Four-Component Model of Age-Related Memory Change,” by M. K. Healey and M. J. Kahana, 2016, *Psychological Review*, 123(1), pp. 23–69 (<https://doi.org/10.1037/rev0000015>). Copyright 2019 by the American Psychological Association.

Cohen and Kahana (2022) further evaluated the individual-difference modeling approach by examining the role of emotional information in the organization of memory. Analyzing data from PEERS Experiment 1, Long et al. (2015) demonstrated that after recalling a word with positive affective valence, subjects were more likely to recall an item of the same valence (positive) as compared with a negative or affectively neutral item (controlling for available of these categories of items). Because similarities among same-valence words are likely greater than among words from different valence classes, Long et al. went beyond the basic emotional clustering result by showing that subjects exhibited reliable affective clustering even after controlling for item similarity. Cohen and Kahana (2022) took the same approach as Healey and Kahana (2016), modeling individual-level data on the organization of memory, including temporal, semantic, and emotional clustering. They found that the 24-experimental sessions contributed by each PEERS Experiment 4 subject provided sufficient statistical power to obtain reliable parameters at the subject level. They then used parameters fitted to individual subjects in PEERS Experiment 4 to generate and test novel predictions about how emotional disorders relate to memory performance for emotional materials.

Variability in Recall Across Items and Lists

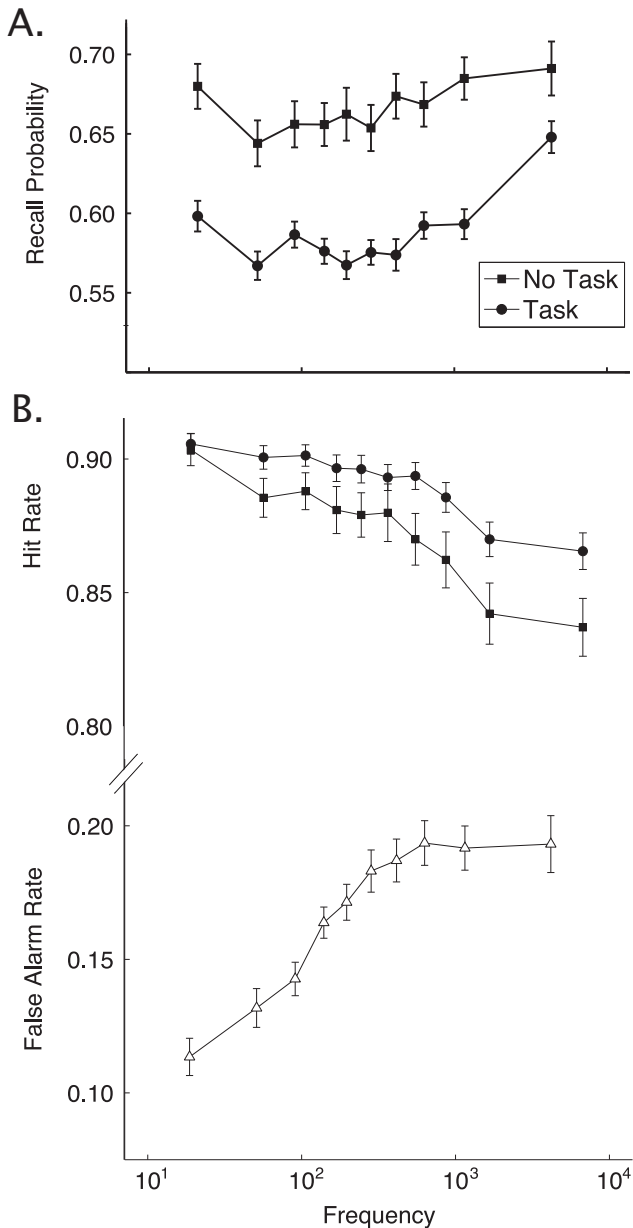
Cognitive processes that unfold during the encoding, retention, and retrieval of an item all contribute to performance in recall and recognition memory tasks. As such, neural measurements during these phases can help disentangle their respective contributions to subsequent memory. Successful encoding and retrieval of words, however, may also reflect their psycholinguistic properties. By amassing a very large number of trials involving recall and recognition of word lists, the PEERS data sets have permitted a detailed

view of the relation between word and list properties and their subsequent memorability.

As a first case study, consider how memory for a word varies with the word’s frequency of occurrence in the English language. Here classic studies report a mirror effect in recognition memory, in which subjects exhibit superior memory for low-frequency (rare) words compared to high-frequency (common) words (Glanzer, 1976; Hall, 1954; Schulman, 1967; Shepard, 1967). Furthermore, the mirror effect demonstrates how these low-frequency words produce a higher hit rate and a higher correct rejection rate than high-frequency words (Gorman, 1961). In free recall, however, studies using mixed lists have reported inconsistent effects, with some researchers finding superior recall for rare words (DeLosh & McDaniel, 1996; Merritt et al., 2006; Ozubko & Joordens, 2007), and other researchers finding superior recall for common words (Balota & Neely, 1980; Hicks et al., 2005). Lohnas and Kahana (2013) sought to clarify this issue by analyzing the effects of word frequency on both free recall and recognition in PEERS Experiment 1. Unlike prior studies analyzing groups of low-frequency versus high-frequency words, they analyzed memory performance as a continuous function of word frequency. Using multiple sessions per subject from PEERS Experiment 1 provided a sufficient number of words at each frequency for every subject thereby allowing this high-resolution view of the word-frequency effect. In recognition memory, they found a pattern harmonious with previous results: with increasing word frequency, hit rates declined, and false alarm rates increased. However, in free recall, they found a U-shaped pattern of results: subjects exhibited superior recall for both rare and common words (see Figure 6).

As a second case study, we consider the broader question of why some words and lists lead to better recall than others. Aka et al. (2021) addressed this question by analyzing data from PEERS

Figure 6
Word Frequency Effects in Recall and Recognition



Note. (A) Subjects recalled higher proportions of both low-frequency and high-frequency words as compared with intermediate-frequency words, regardless of whether the item was presented without an encoding task (filled squares) or with an encoding task (filled circles). (B) Subjects were more likely to incorrectly accept lures with increasing word frequency (open symbols) and less likely to correctly recognize targets with increasing word frequency (filled symbols), regardless of whether the items were presented with an associated encoding task (circles) or no task (squares). Data from Peers Experiment 1 (984 words) were partitioned into deciles on the basis of their word frequency counts in the CELEX2 database. Error bars represent 95% confidence intervals.

Experiment 4, as each subject in that experiment studied the same 576 words in each of the 23 experimental sessions. A multivariate model fit to word-level recall data revealed positive effects of animacy,

contextual diversity, valence, arousal, concreteness, and semantic structure (listed in descending order of importance) on recall of individual words. In their list-level recall model, Aka et al. (2021) examined how the average word features in each list influenced the average recall probability of that list. Here, average contextual diversity, valence, animacy, semantic similarity (weighted by temporal distance), and concreteness (listed in descending order of importance) emerged as significant predictors of list-level recall (see Table 2).

Although psycholinguistic variables, such as those examined by Aka et al. (2021), can account for significant variability in item recall, these factors account for a surprisingly small fraction of the overall variability in recall performance at the list level. Kahana et al. (2018) asked whether this variability in list-level recall could be due to experimentally determined factors, including both average item difficulty and list number. Although each of these factors explained significant variability in list-level recall (see Figure 7) for data on list number, Kahana et al. found the overall explanatory power of these factors to be quite limited. In view of the tremendous variation across lists and the limited explanatory power of their multivariate model, Kahana et al. speculated that endogenous, autocorrelated, neural activity may account for the unexplained variability. They hypothesized that if an autocorrelated latent factor underlies mnemonic variability then prior-list performance should serve as a significant predictor of subsequent list recall. Indeed, they found that this factor was the strongest predictor of recall in their multifactorial model. Evidence that this endogenous variability appears as variable neural activity came from investigations of item and list-level subsequent memory effects described in the following section (Weidemann & Kahana, 2021).

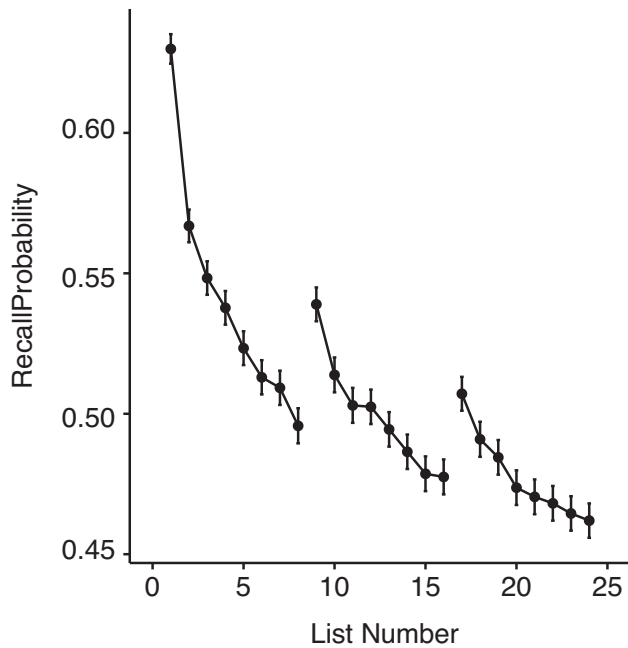
Kreiger et al. further supported the endogenous variability hypothesis. In PEERS Experiment 4, each subject performed a math distractor task between the end of the study list and the recall period, and on half of lists, subjects also performed a math task before the start of the list.

Table 2
Fixed Effects of Variables Predicting Probability of Word-Level and List-Level Recall in Multivariate Analyses

Predictors	<i>M</i> β	<i>SE</i> β
Predictors of word-level recall		
Concreteness	0.03***	0.004
Contextual diversity	0.06***	0.005
Word length	-0.003	0.003
Valence	0.05***	0.004
Arousal	0.04***	0.004
Animacy	0.09***	0.006
Meaningfulness	0.005*	0.005
Session number	-0.009***	0.0003
Predictors of list-level recall		
Concreteness	0.002*	0.0008
Contextual diversity	0.008***	0.001
Word length	-0.0004	0.0008
Valence	0.005***	0.0008
Arousal	0.001	0.0009
Animacy	0.004***	0.0008
Meaningfulness	0.002**	0.0008
Session number	-0.002***	0.0001
Trial number	-0.005***	0.0001

Note. Word length, valence, arousal, and animacy are residualized variables. * *p* < .05. ** *p* < .01. *** *p* < .001.

Figure 7
Predictors of Interlist Variability

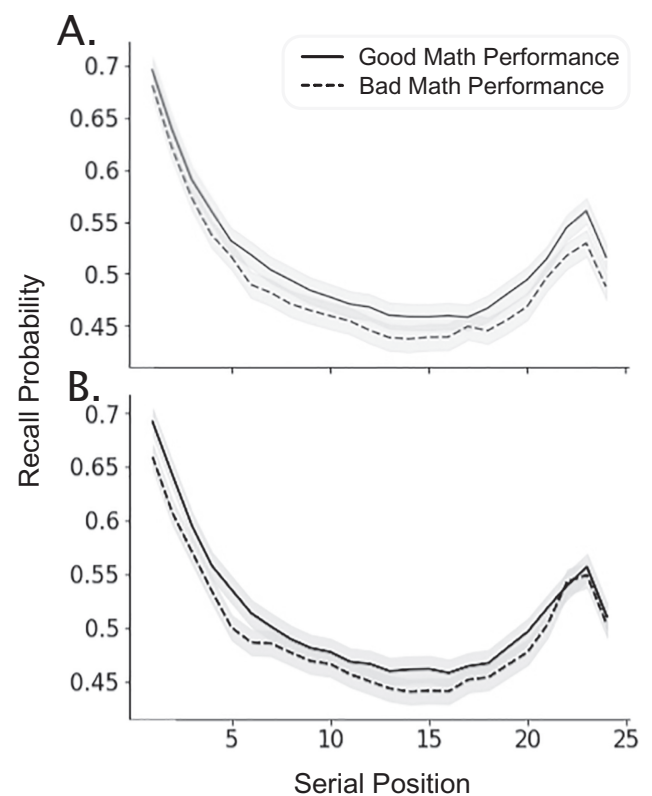


Note. Within each session, recall decreased across successive lists, but increased following the two five-min breaks, consistent with a proactive interference account.

The end-of-list distractor serves the role of disrupting active rehearsal and thereby diminishing the recency effect (see Figure 8). Kreiger et al. asked whether subjects might be sneaking rehearsals into the distractor period and thereby boosting recall performance (akin to the rehearsal borrowing analysis of Yonelinas et al. (1992)). Contrary to their prediction, they found that trials with above-average math performance (for a given subject) had stronger rather than weaker recency. Applying the same analysis to the math task given before the start of each list, they found that trials with above-average subject-specific math performance predicted strong primacy effects on those trials. Both findings align with the hypothesis that cognitive functions supporting both memory and math fluctuate over time and that periods of good cognitive ability lead to better math performance and better recall. We returned to this question in our analysis of the neural correlates of memory encoding at the item and list level, described in the EEG Correlates of Successful Memory Encoding section.

The PEERS studies have also revealed how encoding mechanisms during initial recall can influence performance on subsequent FFR and final recognition tests. Whereas end-of-session memory tests typically yield just a single trial per subject, the multiple-session nature of PEERS allowed for more detailed analyses of final recall and recognition data. As initially reported by Craik (1970) and replicated in our studies, subjects exhibit a negative recency effect in FFR, with recall performance declining over the last few list positions. Analyzing data from PEERS Experiment 1, Kuhn et al. (2018) found that negative recency critically depended on when subjects recalled terminal list items during their initial free recall. Specifically, negative recency arose primarily due to subjects recalling terminal list items at the start of the recall period. When the lag

Figure 8
Recall and Distractor Task Performance



Note. (A) When a math distractor task follows a study list, there is a greater difference in recall probability between good and bad math performance for later serial positions. (B) When a math distractor task precedes a study list, this difference is greater for earlier serial positions.

between studying and recalling an item was short, subjects were significantly less likely to recall the item in final recall than when the lag was long. Kuhn et al. (2018) interpreted this finding in relation to the well-known spacing effect (Madigan, 1969; Melton, 1970): the greater the spacing between two encoding events (in this case, the second being the retrieval of an item) the better the memory for those events. As further support for their interpretation, Kuhn et al. (2018) found greater evidence of negative recency in earlier than later output positions of the delayed free recall (DFR) and continual-distractor free recall (CDFR) conditions of PEERS Experiment 2. Reanalyzing PEERS data, Sheaffer and Levy (2022) found analogous results of spacing on negative recency in the final recognition data.

EEG Correlates of Successful Memory Encoding

Prior to the PEERS studies, a large body of research had already elucidated EEG correlates of successful memory encoding and retrieval, both as measured in the time domain (e.g., event-related potentials) and the frequency domain (e.g., EEG oscillations at various frequencies). Typically, however, these studies would entail having each of several dozen research subjects contribute one session of data, often across multiple experimental conditions. The

sparsity of data at the individual subject level, however, generally precluded analyses of individual subject-level EEG data. With PEERS, we assembled enough data to identify subject-level EEG correlates of memory. However, before describing these subject-level findings, we briefly mention some of the basic results that emerged from analyses of aggregated data.

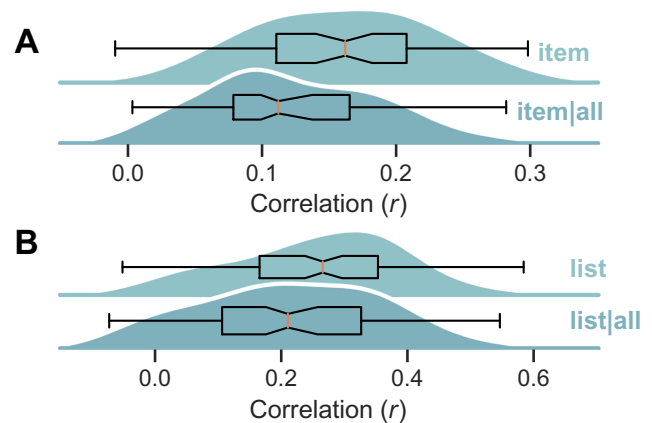
Analyses of PEERS Experiment 1 revealed that increases in broadband high-frequency activity (HFA, defined here as 44–100 Hz) and decreases in low-frequency activity (LFA, centered around the 8–12 Hz alpha band) marked periods of successful memory encoding, as defined based on the subsequent recall of those items (Long et al., 2014). Long and Kahana (2017) further asked if these HFA/LFA biomarkers track not only “whether” a stimulus will be subsequently remembered but “how” a stimulus is later recalled. The high-resolution of the PEERS data set allowed Long et al. to compare spectral signals during the study of words that were subsequently temporally clustered (recalled immediately before or after an item studied in a neighboring list position) or subsequently semantically clustered (recalled immediately before or after an item with a high degree of semantic similarity). They found that both forms of clustering can be predicted by HFA increases during study, but in a task-dependent manner. HFA over left prefrontal cortex predicted subsequent temporal clustering, specifically during no-task lists, when subjects freely encoded the presented words. HFA over left prefrontal cortex also predicted subsequent semantic clustering, but only during task lists, when subjects made a semantic judgment (size or animacy) on each word. These findings reveal a common mechanism that underlies different forms of memory organization and further suggests that temporal versus semantic organization may trade off, given their dependence on the same biomarkers.

Each of the 98 subjects who completed the 24 sessions of Experiment 4 contributed EEG data during each of 13,824 item encoding events. This uniquely large dataset allowed us to evaluate questions about the neural correlates of memory processes at the level of individual subjects while still providing an adequate sample size for across-subject inference. Building on the modeling of item-level memorability described in Variability in Recall Across Items and Lists section, we set out to determine whether the EEG activity during encoding that predicts subsequent recall better reflects item properties or slowly changing brain states hypothesized to support successful memory formation. This latter possibility aligns with our findings that prior list performance and performance in a math distractor task predicted recall of items whose study was separated from these tasks by many seconds.

To test this endogenous variability hypothesis, Weidemann and Kahana (2021) computed multivariate subsequent memory effects by training subject-specific regression models to predict recall performance from a range of neural features. To account for the effects of external factors, they regressed out the “recallability” of each item (determined from recall performance in an independent data set), the serial position of each item within the study list, the position of the corresponding study list within the study session (this factor also captures effects of interference or fatigue from prior lists) and the position of the current session within the series of sessions (this factor captures anything specific to each testing session, such as the time of testing, and any training effects from prior sessions). Together these broad factors captured a wide range of

properties of individual words and the context in which they were studied. By using neural features to predict the residual recall performance, Weidemann and Kahana (2021) calculated a “corrected” subsequent memory effect (SME) that statistically removed the effects of these external factors in order to understand the extent to which remaining endogenous factors can predict recall performance. To assess the extent to which neural features that predict subsequent recall performance persist beyond the individual item presentations they also introduced a list-level SME that uses average neural activity across the entire study list to predict list-level performance. This list-level SME can also be corrected by regressing out remaining external factors that apply to entire lists (i.e., the position of the list within the experimental session and the position of the experimental session within the series of sessions). Figure 9 shows the full and corrected item-level (A) and list-level (B) SMEs as correlations between model predictions and recall performance. Whereas correcting for external factors reduced the SMEs somewhat, substantial SMEs remained even when accounting for external factors, suggesting that a large proportion of SMEs are due to endogenous factors. Additionally, it was possible to predict list-level performance from list-averaged neural activity, supporting the conclusion that endogenous factors related to cognitive function vary slowly (at least on the order of many seconds). Separate analyses on intracranial recordings in neurosurgery patients participating in a free recall task have confirmed these conclusions (Rubinstein et al., 2023). What these analyses do not reveal, however, is the nature of the endogenous processes driving these SMEs. General fluctuations in arousal or attention likely play

Figure 9
Item-Level and List-Level SMEs



Note. Distributions of correlations between multivariate model predictions and free-recall performance at the item-level (A) and at the list-level (B). Each panel shows the full SME (labeled “item” and “list,” respectively) as well as a corrected SME which accounts for a range of external factors that predict recall performance (“item | all” and “list | all,” respectively). SME = subsequent memory effects. Adapted from “Neural Measures of Subsequent Memory Reflect Endogenous Variability in Cognitive Function,” by C. T. Weidemann and M. J. Kahana, 2021, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 47(4), pp. 641–651 (<https://doi.org/10.1037/xlm0000966>). Copyright 2019 by the American Psychological Association. See the online article for the color version of this figure.

a role (a hypothesis supported by the fact that lower levels of alpha power, which is generally regarded as inversely related to attention, predict higher levels of recall), but the extent to which the identified SMEs also reflect activity that is specific to encoding processes remains an open question.

Spectral Markers of Memory Retrieval

Li et al. (2024) examined the spectral correlates of successful retrieval in Experiment 4. The large number of trials contributed by each subject enabled comparison of EEG activity immediately preceding correct recalls and intrusion errors, at the level of individual subjects. This analysis revealed marked increases in HFA in the 500 ms period leading up to successful recall. Accompanying these HFAs increases, they also found decreases in 8–12 Hz alpha activity, with the degree of these two effects exhibiting considerable variability across subjects in both magnitude and frequency ranges. The majority of subjects also exhibited modest increases in theta activity preceding successful recall, but this effect did not prove reliable in aggregate statistical comparisons. Figure 10 illustrates these results separately for each of the 98 subjects.

Katerman et al. (2022) further investigated the spectral correlates of memory retrieval after very long delays, using a prevocalization period in immediate-recall a control for premotor activity (in PEERS Experiment 5). In addition to demonstrating increased HFA and decreased alpha activity, as shown by Li et al. (2024), Katerman and colleagues also found a striking increase in frontal theta activity in the moments leading up to successful retrieval, mimicking the encoding results described above (see Figure 11, where black outlines indicate frequency-region pairs that met an FDR-corrected $p < .05$ threshold for the comparison between delayed vs. immediate recall). Given the far greater demands on episodic memory retrieval when recalling items after one or more days, Katerman et al. (2022) interpreted the increased theta (T^+), decreased alpha (A^-), and increased gamma/HFA (G^+) as a $T^+A^-G^+$ of context-dependent memory retrieval.

Recognition memory tests confer certain advantages over recall tests in the study of retrieval processes. Specifically, the recognition procedure provides experimental control over the arrival of the retrieval cue allowing for precise analyses of cue-dependent memory retrieval. In addition, the recognition procedure provides valuable information about retrieval processes when subjects have limited memory for a given target. PEERS Experiments 1 to 3 included a recognition phase at the end of each session in which subjects made yes–no responses, followed by confidence ratings. In addition to reducing uncertainty around timing of retrieval processes, recognition tests also provide data on the strength of the underlying memory signal (“memory strength”) usually from introspective judgments in the form of confidence ratings. Weidemann and Kahana (2016, 2019) examined the extent to which implicit measures, such as response speed or brain activity preceding the recognition decision, might reveal memory strength without the need to rely on introspection. We can assess different measures with respect to their ability to reveal memory strength by constructing receiver operating characteristic functions (ROC) that relate false alarm rates to hit rates across the range of the measures (Wickens, 2002). Confidence ratings can be arranged from “sure old” to “sure new” and the ROC function traces out the cumulative false alarm and hit rates corresponding

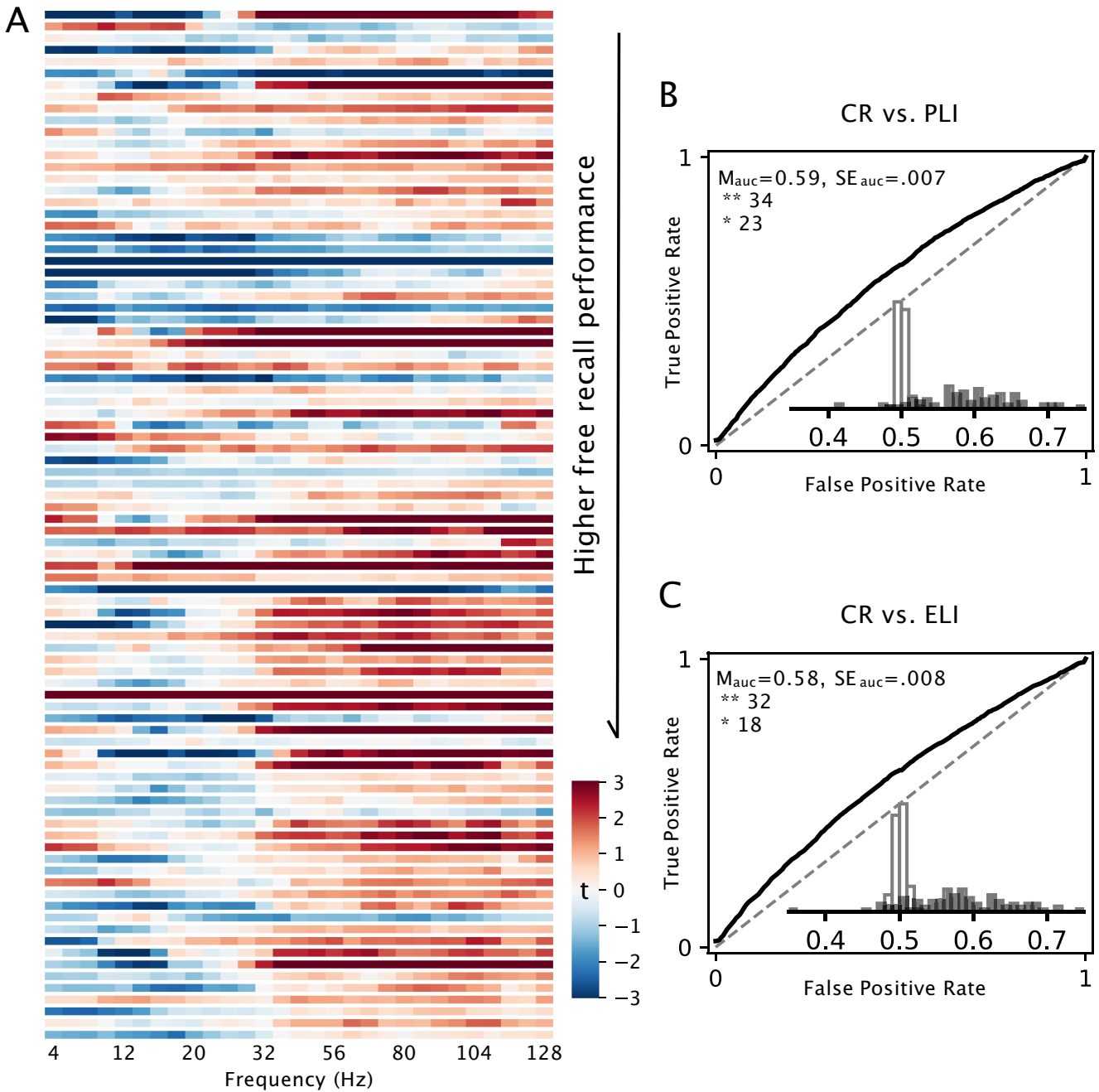
to these ratings. With the assumption that binary old/new responses are faster when response confidence is higher, individual responses can be similarly arranged according to response speed from fast “old” responses to slow “old” responses to slow “new” responses to fast “new” responses. The cumulative false alarm and hit rates trace out a latency-ROC function that does not depend on introspective judgments beyond the binary old/new response. When using neural activity as features for a classifier distinguishing between targets and lures, we can use the output of that classifier (a continuous measure related to classification “confidence”) to generate ROC functions that only depend on neural activity and do not reflect any overt response. The area under the corresponding ROC curve (AUC) indexes how much the corresponding measure is able to distinguish old from new items with an AUC of 0.5 indicating chance performance and an AUC of 1.0 indicating perfect discrimination between old and new items. Figure 12 shows the AUC for confidence ratings, response latencies, and EEG activity with qualitatively similar patterns across these measures and substantial correlations between the different AUCs. These results suggest that these measures all offer different views on the same memory strength signal underlying recognition decisions. Analyses on classifiers predicting an item’s old–new status using brain activity during different time windows in the lead-up to a recognition response also showed that evidence is integrated into a unitary memory signal giving rise to recognition decisions. This result contrasts with theories proposing that different kinds of evidence dominate individual recognition decisions (Weidemann & Kahana, 2019).

Neural Context Reinstatement

Free recall confers specific advantages in the study of episodic memory retrieval. In particular, the lack of external retrieval cues in free recall allows one to use neural similarity between encoding and retrieval to study reinstatement of the encoding activity in the subject’s mind. Analyzing data from Experiment 1, Lohnas et al. (2023) asked whether spectral features of scalp EEG would reveal evidence of neural context reinstatement as previously uncovered by intracranial EEG studies (Howard et al., 2012; Manning et al., 2011). Furthermore, they examined how task manipulations influenced the pattern of neural similarity between encoding and retrieval. Lohnas et al. defined a neural measure of temporal context using principles of RCT: studying an item should cause context to drift slowly, and recall of an item should reinstate its temporal context from study. They found that spectral features of scalp EEG activity demonstrate the reinstatement of temporal context preceding word recall (Figure 13A).

Having demonstrated a neural signature of context reinstatement in lists involving size and animacy encoding tasks, and in no-task lists, they then examined the dynamics of context in task shift lists (see Figure 1 for an illustration of the task manipulation). Lohnas et al. hypothesized that a change in task disrupts temporal context (Polyn et al., 2009) and, therefore, context should exhibit a greater change across successive words if they are studied with the different tasks than if they are studied with the same task. Consistent with this prediction, neighboring items had reduced neural similarity in temporal context when presented with different tasks. Lohnas et al. also found strong evidence for the novel RCT prediction that, during recall, the disruption to temporal context promotes increased temporal

Figure 10
Subject-Specific Spectral Markers of Successful Episodic Retrieval



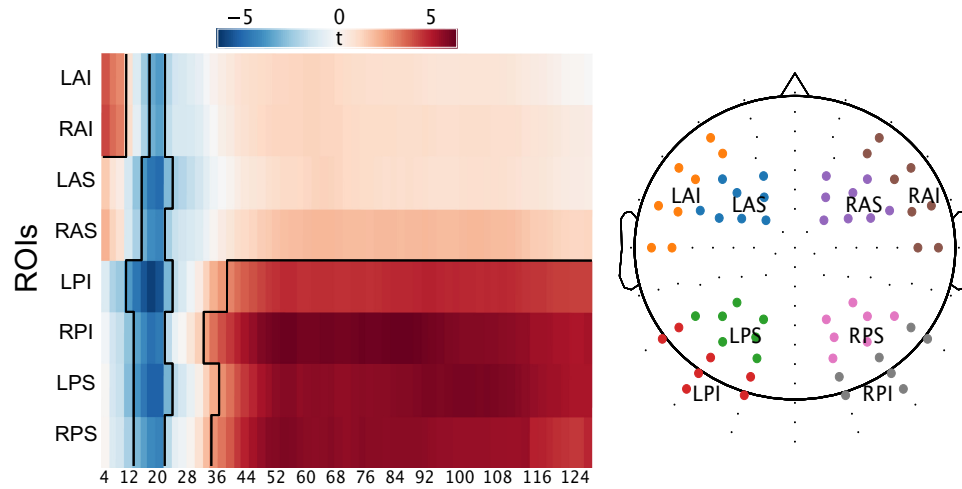
Note. (A) Subject-specific independent t statistics for comparing the 500-ms period immediately preceding correct recalls with the same period preceding intrusion errors (PLI and ELI) aggregated across all scalp electrodes. Each row shows the results from one subject, with rows sorted in order of ascending recall performance. Power increases and decreases are shown in red and blue, respectively. (B) ROC curves created by varying the threshold value of ΔEEG used to classify a retrieval as a CR or PLI. (C) ROC curve for classifying retrievals as a CR or ELI. PLI = prior-list intrusion; ELI = extra-list intrusion; ROC = receiver operating characteristic; EEG = electrophysiological; CR = correct recall. See the online article for the color version of this figure.

contiguity for same-task neighboring items, and decreased temporal contiguity for neighboring items studied with different tasks.

Lohnas et al. also examined individual differences in neural temporal disruption reinstated during recall. They defined each subject's neural similarity difference as the similarity of neighboring item

pairs with the same task minus neighboring item pairs with different tasks. Across subjects, the neural similarity difference during encoding was correlated with the neural similarity difference during recall. This provides further evidence that temporal context, including disruptions to context representations, reinstates during free recall.

Figure 11
Statistical Maps Illustrating Relative Increases (Red) and Decreases (Blue) in Spectral Power Across Key Memory Contrasts for Eight Regions of Interest



Note. Spectral power contrast for delayed recall versus immediate recall in PEERS Experiment 5. PEERS = Penn Electrophysiology of Encoding and Retrieval Study; L = left; R = right; A = anterior; P = posterior; I = inferior; S = superior; ROIs = regions of interest. See the online article for the color version of this figure.

Furthermore, across subjects, the neural similarity difference at recall correlated with the behavioral modulation of temporal contiguity, suggesting that the neural measure of temporal context contributed to subject behavior. Taken together, these results highlight the impact of task changes on temporal representations, having implications for neural activity and memory organization.

Data Requirements for Detecting Neural Context Reinstatement

One might ask whether the effects we have reported in the PEERS studies might have been detected using more typically powered data sets. Here, we address this question by reanalyzing the aforementioned EEG reinstatement analysis, resampling different subsets of the data reproduced in Figure 13A and repeating the key comparison reported by Lohnas et al. (2023). Specifically, we re-examined neural context reinstatement for subsets of data representing a factorial of subsampled subjects and trials. We constructed these subsets by sampling 25%, 50%, or 100% of trials crossed with the same three fractions of subjects (Resulting in Nine Subjects \times Trial Sampling Schemes). For each subset type, we randomly sampled the full data set to create 25 unique samples. For each of these samples, we estimated the neural context reinstatement effect by comparing neural similarity for lag = -1 to neural similarity of lags = -3 to -5 (Lohnas et al., 2023). Each subset comparison produced a single t -statistic, which we then averaged across all 25 samples to estimate the reliability of the reinstatement effect. Figure 13B illustrates how the reinstatement effect (the z transform of the p value) varies with the number of trials included in the analysis. Only the subsets that included at least half of the data provided adequate power to detect a reliable reinstatement effect. Furthermore, due to the variability of EEG recordings across sessions, each session needed to have some lists with task changes and some without task changes. Yet a single session could not provide an adequate number of observations to conduct the recall analyses. Thus, the large number of sessions

per subject, with varied task manipulations, supported the critical conclusions connecting neural measures of temporal context, event segmentation, and memory. Thus, we can conclude that this theoretically motivated EEG analysis relied on the large scale of the PEERS data set.

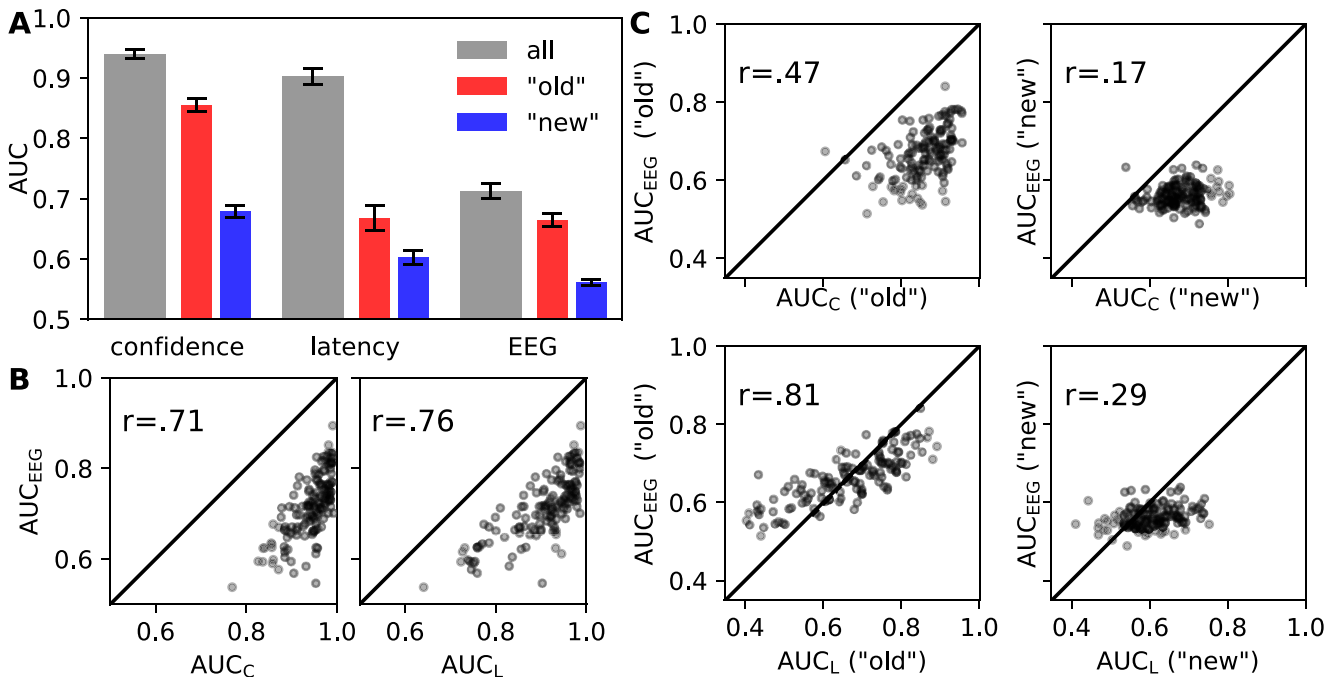
Lessons Learned

The PEERS project taught us many lessons, some of which we briefly review here:

Subject Recruitment, Retention, and Performance Monitoring

Each term we sought to recruit between eight and 12 subjects to participate in the full 22 sessions of PEERS Experiments 1 to 3, or the 24 sessions of PEERS Experiment 4. Because many potential subjects would either be unwilling or unable to make such a large time commitment, we first recruited subjects for a preliminary session, to ensure that they knew what the series of studies would entail. During this preliminary screening session, subjects performed a series of trials involving immediate free recall of 15 item lists. At the end of the screening session, we evaluated subjects' blink rate, recall performance, and any evidence of their inability to follow instructions. We invited subjects to enroll in the full study assuming that they met a very liberal criterion on these variables. The main value of this type of screening trial is to ensure that subjects who enroll know what they are "getting into" before committing to 20+ sessions of data collection.

During the main experiment, we provided a performance and a completion bonus (in addition to a base payment for each session). Nonetheless, we still experienced attrition rates of around 30%. We optimized the performance bonus for each study, generally rewarding subjects based upon a combination of low-blink rates during item presentations, high recall, accurate recognition, and distractor task performance.

Figure 12*Inferring Memory Strength From Confidence Ratings, Response Latencies, and EEG Activity*

Note. (A) The area under the ROC curve (AUC) for functions constructed from confidence ratings, response latencies, and EEG. These AUCs indicate the extent to which the corresponding measure reflects a memory signal. As detailed by Weidemann and Kahana (2019), we can calculate AUCs across all responses or calculate AUCs separately within “old” and “new” responses. We see a qualitatively similar pattern across modalities with a stronger memory signal for “old” than for “new” responses. (B) Scatterplots relating AUCs from confidence (*C*) and latency (*L*) ROC functions to those from EEG activity. We see strong relationships between these AUCs. As detailed by Weidemann and Kahana (2019), this strong correspondence is difficult to interpret because every ROC function based on all responses is constrained to pass through the point corresponding to the overall hit and false alarm rate and thus the corresponding areas are not independent. (C) As in (B), but for ROC functions only based on “old” or “new” responses, as indicated. These ROC functions are not constrained to pass through the same point, but the corresponding ROCs are nevertheless highly correlated. EEG = electrophysiological; ROC = receiver operating characteristic; AUC = area under the curve. Figure adapted from “Dynamics of Brain Activity Reveal a Unitary Recognition Signal,” by C. T. Weidemann and M. J. Kahana, 2019, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 45(3), pp. 440–451 (<https://doi.org/10.1037/xlm0000593>). Copyright 2019 by the American Psychological Association. See the online article for the color version of this figure.

Our experience indicated that an experimenter should be present during a subject’s first session of each new experimental phase. In subsequent sessions, we allowed subjects to perform the tasks without overt monitoring. However, we observed the subjects’ performance remotely by monitoring their screen and in some cases with an experimenter video. We also provided subjects with a “call button” that they could use to ask for assistance from the experimenter.

Annotation of Vocal Responses

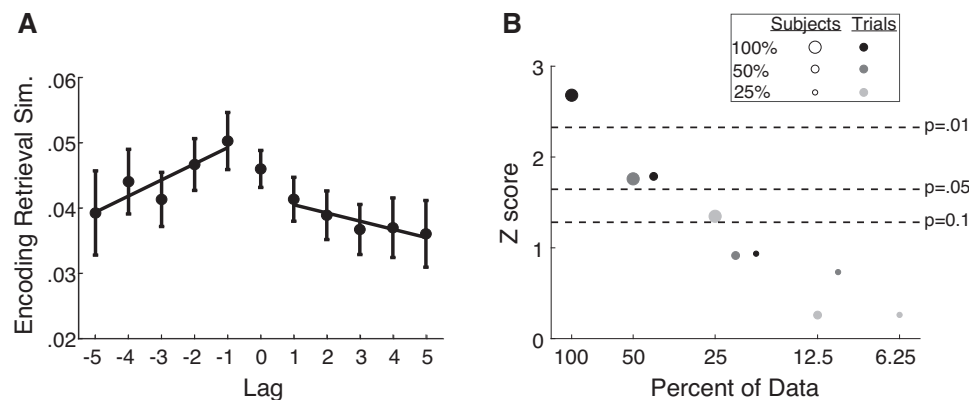
Although one can collect free recall responses using a keyboard, spoken recall remains the most natural mode of output for subjects. In addition, not all subjects are equally proficient at touch typing, and when allowed to type responses they may wish to backtrack and make changes before committing. This is an especially important consideration when comparing younger and older adults. Therefore, we allowed subjects to freely recall items by speaking them out loud to a microphone and a computer recorded their vocal responses. We developed custom software to help annotate subjects’ vocal responses. This software allowed a research assistant to listen to the recalled items and mark the identity and the onset time of each spoken response (<https://memory.psych.upenn>

.edu/TotalRecall). Over the course of this project, we refined the Penn Total Recall software, making it easier for researchers to process the recordings efficiently. Nonetheless, it requires considerable time and care to annotate a single session of vocal recall responses. In future work, it may be possible to fully automate voice detection and response identification using tools such as Google’s speech recognition engine. We experimented with these tools towards the end of the PEERS study, but never achieved the level of performance that would allow us to replace manual annotation.

Measuring Recognition

In line with our goal of making user responses as natural as possible, we also opted for vocal responses during our recognition test. Following a suggestion by our colleague, Professor Saul Sternberg, we asked our subjects to say “pess” or “po” instead of “yes” or “no.” Because the letter “P” is a stop consonant this would enable precise answer timing and remove any differences in measure of reaction time between yes and no responses. Weidemann and Kahana (2016, 2019) used these data in their analyses of ROC functions. We also collected confidence judgments and decided to take

Figure 13
Neural Context Reinstatement



Note. (A) Consistent with RCT, the similarity between a recalled item's neural context and its temporal neighbors from study decreases with absolute lag. Replotted from Lohnas et al. (2023). (B) The significance of the critical reinstatement effect decreases with subsets of subjects and/or trials. RCT = retrieved-context theory. Adapted from "Neural Temporal Context Reinstatement of Event Structure During Memory Recall" by L. J. Lohnas, M. K. Healey, and L. Davachi, 2023, *Journal of Experimental Psychology: General*, 152(7), pp. 1840–1872 (<https://doi.org/10.1101/2021.07.30.454370>). Copyright 2019 by the American Psychological Association.

confidence ratings after subjects made their recognition responses. To ensure the highest quality response time data, we incentivized subjects for their speed in responding "yes" or "no." We also incentivized them for their accuracy using their confidence judgments as an index of performance.

Data Quality Control

Early in the project, we discovered that data quality issues could emerge after a certain session (e.g., a problem with the testing equipment) or that subjects might become confused regarding the instructions for a particular phase of the task. To maintain data quality we began creating automated subject reports, using a cron job that ran overnight following annotation of the subjects' vocal responses (see below). These HTML or PDF reports, which could be accessed through a webpage, indicated various data quality metrics including word-presentation evoked potentials, blink rates, recall performance, and the testing room in which the session took place. The reports did not reveal any comparisons across conditions, or other results that could bias the research in any way. The research team reviewed these reports weekly and when they saw any anomalous data they presented these findings to the principal investigator. We found these reports to be so useful that we made them a standard part of all of our research both in our scalp EEG studies and in our intracranial EEG research.

When analyzing brain recordings, researchers rightly worry about EEG artifacts corrupting their data. Such artifacts could result from eye or muscle movements that create large electrical potentials or from changes in the recording system's ability to measure brain signals (e.g., resulting from electrodes losing their conductivity with the scalp). Throughout the PEERS Experiments, we developed and refined methods for identifying electrodes and trials with data quality problems. Specific methods appear in each of the papers reporting PEERS data.

Although we sought to minimize the impact of electrical artifacts in our EEG analyses, the main purpose of such procedures is to reduce the potential for large outliers to skew the distribution of observed values across subjects, sessions, and trials. If artifacts do not covary with physiological or behavioral effects, then having a very large dataset avoids the risk of having a small number of large-artifact trials dominate the aggregate results. The machine learning approaches made possible by large datasets automatically down-weight nondiagnostic electrical artifacts. Recent work has shown that with such methods, data cleaning reduces statistical power to observe established effects (Meisler et al., 2019).

Big Data Studies in Peer Review

We did not notice any striking difference on the part of reviewers or editors in the handling of papers involving novel analyses of an ongoing study, or retrospective analyses of established data sets, as compared with traditional studies reporting novel data.

Recruiting Diverse Populations

When a study is not specifically focused on a particular subgroup of individuals, one hopes the study's findings will generalize to society at large. Our experimental design, which required participants to return to the lab for 10 or more sessions, presents particular challenges to recruiting a representative subject population. Specifically, most of our subjects came from the undergraduate and graduate student community of the University of Pennsylvania. To increase diversity, we placed flyers in the surrounding neighborhood and at nearby institutions with more diverse student demographics. Nonetheless, the very fact that our study required subjects to return for so many repeated sessions introduced bias into our participant sample. We hope that future studies of this kind could be specifically designed to recruit participants from more diverse ages and educational backgrounds.

General Discussion

Beyond the specific lessons learned from PEERS, writing this article led us to reflect on three broad issues raised by our adventure into the world of big data, memory, and the human brain. First, we discuss the utility of scalp EEG recordings in the study of human memory. Many readers will assume that such data have great value for the science of memory. Yet, some of us had our doubts at the outset of this exploration. Second, we discuss the strengths and limitations of studying well-practiced subjects. Third, borrowing a term from the field of corporate finance, we introduce the idea of “Portfolio Choice” and discuss the question of how we ought to optimize the scientific portfolio in the study of human memory. We close with a discussion of the risks, rewards, and possibilities in the emerging era of big data.

On the Utility of Scalp EEG

Scalp EEG is among the oldest techniques available to cognitive neuroscientists. Beginning with the classic work of Berger (1929), EEG has become a staple of clinical neurology, with applications to detecting epileptic seizures, identifying sleep stages and abnormal sleep patterns, diagnosing perceptual disturbances, and many other indications. Although some early scalp EEG studies examined correlations between alpha activity and learning and memory, EEG became a commonly used method in the 1980s (e.g., Donald, 1980; Sanquist et al., 1980). With the advent of more recent modalities of neural imaging, one may wonder whether scalp EEG still has the potential to address important questions in the realm of human memory.

The PEERS study sought to answer this question in the domain of episodic memory. For example, intracranial EEG studies have uncovered striking correlates of behavior at relatively high frequencies (e.g., 80–150 Hz)—frequencies which are commonly filtered out in scalp EEG studies, particularly those averaging the EEG signal into event-related potentials, due to concerns about electromyographic signals. PEERS data demonstrated that spectral correlates of memory encoding and retrieval from noninvasive EEG recordings closely resemble those from intracranial recordings in patients with epilepsy. Whereas earlier EEG studies had documented effects in the alpha and theta frequency bands, PEERS data highlighted relevant signals in spectral activity at higher frequencies calling into question the standard practice of filtering out these signals.

The large number of trials contributed by each PEERS subject allowed us to evaluate scalp EEGs ability to forecast behavior, by training classifiers on either encoding or retrieval-related spectral activity. These classification studies required many more trials to achieve the same classification performance as intracranial EEG classifiers. Specifically, we found that scalp EEG training data from 500 24-item lists provided classification performance similar to that obtained with 50 12-item lists of intracranial EEG data. This 20-fold difference likely reflects the much higher spatial resolution of intracranial recordings as well as the ability to sample deeper brain structures. Although we do not have hard numbers to compare our PEERS results to other recording methods, such as MEG or fMRI, it is at least gratifying to know that given sufficient data, EEG can reliably perform the same classification tasks as intracranial recording studies.²

Because researchers can obtain scalp EEG data efficiently and at low cost from both healthy adults, and from diverse patient populations, it offers unique advantages over other recording modalities, at least at the time of this writing. The PEERS studies demonstrate how multisession

data collection allows for decoding at the individual subject level. Future work will illuminate the value of model-based electrophysiology for furthering our understanding of cognitive processes.

Strengths and Limitations of Studying Well-Practiced Subjects

Subjects in the PEERS experiments contributed between seven and 24 sessions of experimental data, thus ensuring that they were familiar with the tasks being performed. In later sessions, we would consider them highly practiced at performing our memory tasks. Indeed, subjects performed a preliminary session in which they became comfortable with having electrodes applied to the scalp and performing memory tasks while minimizing eye movements. Familiarizing subjects with our procedures reduced task-related anxiety and the variable rates at which naive subjects learn how to perform tasks, where such variability can introduce significant noise into measured behavior and physiology. On the other hand, our design choice runs the risk of confusing highly idiosyncratic task-specific strategies or control processes with more general memory phenomena. Indeed, analyses of the PEERS data by Romani et al. (2016) show that subjects alter their strategies across trials and sessions which, for some subjects, results in substantial performance improvements. This raises the question of whether the findings reviewed above generalize to naive participants. Examining practice effects in the PEERS data suggests that several core effects became apparent long before extensive practice. For example, contiguity effects appear in the very first intake session of PEERS (Healey et al., 2019).

Recent internet-based studies have complemented the PEERS approach by collecting a single trial’s worth of behavioral data from 1000’s of subjects. Such studies have found that key serial position and contiguity effects appear on the very first trial among naive subjects, and even with incidental encoding (Healey, 2018; Mundorf et al., 2021). Not surprisingly, then, the performance of well-practiced subjects reflects a complex interplay between general memory principles and task-specific control processes. In this regard, a great advantage of the high-resolution PEERS data is the ability to use modeling and detailed analyses to precisely disentangle task-general memory processes from task-specific strategies (Healey & Kahana, 2014).

PEERS data also facilitate the evaluation of machine learning approaches to predicting variability in memory behavior. In our experience, models relating EEG features to behavioral outcomes vary considerably across individuals such that data contributed by a given subject will typically yield a far better forecast of behavior in unseen sessions than data contributed by other subjects.

What Is the Optimal Scientific “Portfolio”?

When making decisions across multiple projects, investors and corporations face a fundamental “portfolio allocation” problem. This is the same problem that faces scientific investigators deciding to allocate resources across projects. Scientists usually have many more good projects than there is time or grant support to carry out the work; like the company, they face a budget constraint and must make wise decisions about their resource allocation. The problem of portfolio choice is relevant not only to an individual investigator but also to a scientific field

² We collected half of our PEERS Experiment 4 data with water-based and half with gel-based EEG systems (BioSemi and EGI) and we did not find any reliable difference in classification performance between the two systems.

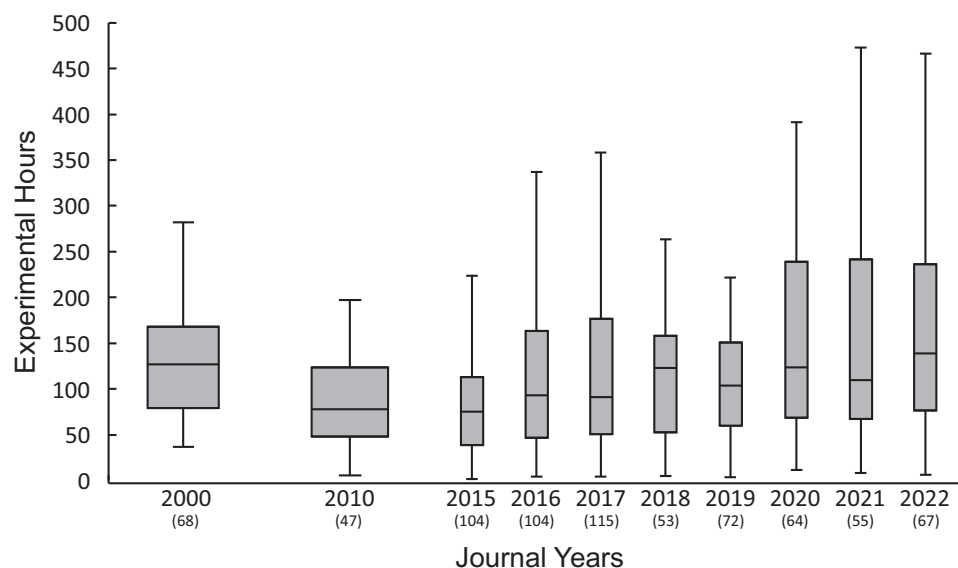
as a whole: granting agencies, for example, must decide how to allocate resources across projects.

Throughout its short history, experimental psychology has embraced a model of small science (Ebbinghaus, 1885/1913, is the famous first exception). Individual investigators, or more typically individual trainees, design and carry out small-scale experiments on humans or non-human animals. In the case of human research, each subject typically takes part in just 1 hr-long session. A survey of recent articles published in the *Journal of Experimental Psychology: Learning, Memory, and Cognition* reveals that a typical experiment entails fewer than 40 hr of data collection and a typical publication includes approximately three experiments. Figure 14 shows the median and interquartile range across articles published in each of the years 2000, 2010, and 2015–2022, with the number of articles included in the survey indicated below each year. As the figure indicates, although some publications reported several hundred hours of experimentation, the majority report fewer than 100 hr of human experimental data, with a slight upward trend across the years. Although these publications may afford the power required to support their main conclusions, many will lack adequate statistical power to support secondary analyses, where such analyses may offer insights into the higher-order structure of the data. Certainly, we are beginning to see an increase in the number of publications reporting crowdsourcing experiments often with large samples of convenience, but these studies generally provide very limited data on each individual subject, and we do not yet have the technology to crowdsource neural recording data (though this may change sooner than we expect). We are also beginning to see studies reporting secondary analyses of previously published data, which is a welcome trend in our view. Yet, the allocation of science to originally and singly published studies versus secondary analyses remains markedly lopsided.

One reason to avoid large allocations to single experiments, or single research teams, is to diversify the risk. This is a sensible approach, but if all of our knowledge depends on small experiments, this actually increases risk as these experiments cannot answer questions that require a large quantity of data. Over time, researchers will have answered most, if not all, of the major questions that can be answered with experiments involving fewer than, say, 10,000 trials (e.g., 200 trials \times 50 subjects). This may encourage researchers to form little cottage industries, promoting new phenomena that are often little more than rebranded variants of established paradigms and findings. New discoveries will then rely on new technologies, such as novel methods for recording brain data or manipulating brain activity (Ezzyat & Suthana, in press; Helfrich et al., in press). But for behavioral research, our knowledge will become stale, and without a way forward, much of what we know may be lost but for a few old textbooks rarely studied by the next generation of scientists. Yet big data can also be a new technology; by amassing large numbers of observations under varied conditions, researchers can exploit powerful new statistical techniques to find hidden structure in data that has long been in plain sight. Think of big data as a kind of microscope that allows us to zoom into a phenomenon and see structure that was previously obscured within the error bars of our small experiments.

When a field invests in big data, it can be a boon for early career researchers who have not yet established laboratories capable of generating large datasets. These researchers should be able to freely access data from many labs, answering questions that they could not easily answer by collecting new data of their own. However, if we are to invest in big data as a field, we must go beyond making the data publicly available; we have to also make the design of experiments and data collection a distributed process, where multiple researchers contribute to the planning of future studies.

Figure 14
Hours of Experimental Data per Publication (2000, 2010, and 2015–2022)



Note. We surveyed articles published in the *Journal of Experimental Psychology: Learning, Memory, and Cognition* that included sufficient information to estimate the number of experimental hours contributed by research subjects. For each year we report the median and interquartile range, based on the sum of hours across all experiments in each publication. The number of evaluated articles appears below each year.

Big Data: Risks, Rewards, and Future Challenges

We have heard colleagues voice several concerns about the big data approach exemplified in the PEERS project. One criticism that emerged early in our project concerned the law of diminishing returns. A distinguished colleague raised this objection, pointing out that as we collect more data the standard error will shrink as the ratio of the square root of the number of observations. Surely, it would be better to conduct a larger number of manipulations than to continually invest resources in the face of diminishing returns. This objection arose as one of us (Michael J. Kahana) presented some early PEERS findings. After being tongue-tied for a few moments (or longer), the presenter recalled many instances in his past research where additional data revealed some important result via a new “cut” of the data space. In essence, every time you think of an interesting new way to partition your data your sample size shrinks, and once again each additional observation provides valuable information. Just as fabricating a more powerful microscope or telescope allows you to see things that were invisible to previous generations of scientists, so too, the additional power provided by high-resolution data peers beneath the surface of our current knowledge, paving the way for new discoveries.

Another objection, highlighted by the current emphasis on replicability, is that perhaps some peculiar feature of a large study will generate results that do not generalize across diverse situations. Each PEERS experiment entailed myriad small decisions which could affect the data in unknown ways. Would it be smarter to diversify our research investment by having many smaller studies that vary these methodological choices? We appreciate the value of this objection and would not advocate for a cessation of small-science-style experiments. Rather, we see big data as an important addition to the scientific portfolio, complementing smaller studies. Indeed, the discoveries made possible with big data can inspire conceptual replications with smaller studies.

Although PEERS has already taught us a good deal about memory and its neural correlates, we see several exciting opportunities for future explorations. First, none of the analyses conducted thus far have delivered on the promise of using data on both memory and physiology to evaluate computationally explicit theories of memory. This challenging and rewarding endeavor stands in wait for the ambitious researcher. Second, ancillary data that we have collected on personality, mood, and intellectual abilities have yet to be studied in relation to neural measurements and their cognitive correlates. Such analyses could provide valuable new information on the neural basis of individual differences and the potential utility of neural recordings to inform our understanding of the relation between cognition and emotion. Finally, we have only begun to look at data from our aging subsample described in the Appendix. EEG data collected on older participants, some of whom contributed more than 20 experimental sessions, can help us understand the EEG correlates of variable memory performance in a participant group at risk for memory loss.

Looking beyond PEERS we see a variety of exciting uses of big data in the science of human memory. The internet represents a burgeoning modality for large N studies, both via traditional memory tasks delivered remotely (Mundorf et al., 2022) and through massive memory surveys administered via user interactions with products such as online trading platforms (Jiang et al., 2022). Large online gaming communities, such as the players of Sea Hero Quest or Chess, provide another valuable source of data on memory and cognitive processes (Coutrot et al., 2022; Russek et al., 2022; Spiers et al., 2023). Although the ability to capture human neural data remotely did not exist when we

conducted the PEERS experiments, the authors predict that future readers will see these technologies permeate their daily lives.

We see the PEERS project as a test case in applying big-data approaches to studying human memory. The strongest endorsement of our approach derives from other investigators using PEERS data to answer their own questions. We have begun to see this happen (Madan, 2021; Naim et al., 2019; Osth & Farrell, 2019; Popov & Reder, 2020; Romani et al., 2016; Sheaffer & Levy, 2022; Zhang et al., 2023) and hope that this article, in synthesizing key motivations, methods, and discoveries, will prompt additional investigators to consider the value of this approach.

References

- 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>
- Anderson, J. R., Fincham, J. M., & Douglass, S. (1999). Practice and retention: A unifying analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(5), 1120–1136. <https://doi.org/10.1037/0278-7393.25.5.1120>
- Appelhoff, S., Sanderson, M., Brooks, T. L., van Vliet, M., Quentin, R., Holdgraf, C., Chaumon, M., Mikulan, E., Tavabi, K., Höchenberger, R., Welke, D., Brunner, C., Rockhill, A. P., Larson, E., Gramfort, A., & Jas, M. (2019). MNE-BIDS: Organizing electrophysiological data into the BIDS format and facilitating their analysis. *Journal of Open Source Software*, 4(44), Article 1896. <https://doi.org/10.21105/joss.01896>
- Balota, D. A., & Neely, J. H. (1980). Test-expectancy and word-frequency effects in recall and recognition. *Journal of Experimental Psychology: Human Learning & Memory*, 6(5), 576–587. <https://doi.org/10.1037/0278-7393.6.5.576>
- Berger, H. (1929). Über das elektroencephalogramm des menschen [On the human electroencephalogram]. *Archiv für Psychiatrie und Nervenkrankheiten*, 87(1), 527–570. <https://doi.org/10.1007/BF01797193>
- Bjork, R. A., & Whitten, W. B. (1974). Recency-sensitive retrieval processes in long-term free recall. *Cognitive Psychology*, 6(2), 173–189. [https://doi.org/10.1016/0010-0285\(74\)90009-7](https://doi.org/10.1016/0010-0285(74)90009-7)
- Bousfield, W. A. (1953). The occurrence of clustering in the recall of randomly arranged associates. *Journal of General Psychology*, 49(2), 229–240. <https://doi.org/10.1080/00221309.1953.9710088>
- Broitman, A. W., Kahana, M. J., & Healey, M. K. (2020). Modeling retest effects in a longitudinal measurement burst study of memory. *Computational Brain & Behavior*, 3(2), 200–207. <https://doi.org/10.1007/s42113-019-00047-w>
- Cohen, R. T., & Kahana, M. J. (2022). A memory based theory of emotional disorders. *Psychological Review*, 129(4), 742–776. <https://doi.org/10.1037/rev0000334>
- Coutrot, A., Manley, E., Goodroe, S., Gahnstrom, C., Filomena, G., Yesiltepe, D., Dalton, R. C., Wiener, J. M., Hölscher, C., Hornberger, M., & Spiers, H. J. (2022). Entropy of city street networks linked to future spatial navigation ability. *Nature*, 604(7904), 104–110. <https://doi.org/10.1038/s41586-022-04486-7>
- Craik, F. I. M. (1970). The fate of primary memory items in free recall. *Journal of Verbal Learning and Verbal Behavior*, 9(2), 143–148. [https://doi.org/10.1016/S0022-5371\(70\)80042-1](https://doi.org/10.1016/S0022-5371(70)80042-1)
- Deese, J., & Kaufman, R. A. (1957). Serial effects in recall of unorganized and sequentially organized verbal material. *Journal of Experimental Psychology*, 54(3), 180–187. <https://doi.org/10.1037/h0040536>
- DeLosh, E. L., & McDaniel, M. A. (1996). The role of order information in free recall: Application to the word-frequency effect. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 22(5), 1136–1146. <https://doi.org/10.1037/0278-7393.22.5.1136>
- Donald, M. W. (1980). Memory, learning and event-related potentials. *Progress in Brain Research*, 54, 615–627. [https://doi.org/10.1016/S0079-6123\(08\)61681-7](https://doi.org/10.1016/S0079-6123(08)61681-7)

- Ebbinghaus, H. (1913). *Memory: A contribution to experimental psychology*. Teachers College, Columbia University (Original work published 1885).
- Ezzyat, Y., & Suthana, N. (in press). Brain stimulation. In M. J. Kahana & A. D. Wagner (Eds.), *Oxford handbook of human memory* (2nd ed.). Oxford University Press.
- Glanzer, M. (1976). Intonation grouping and related words in free recall. *Journal of Verbal Learning and Verbal Behavior*, *15*(1), 85–92. [https://doi.org/10.1016/S0022-5371\(76\)90009-8](https://doi.org/10.1016/S0022-5371(76)90009-8)
- Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., Flandin, G., Ghosh, S. S., Glatard, T., Halchenko, Y. O., Handwerker, D. A., Hanke, M., Keator, D., Li, X., Michael, Z., Maumet, C., Nichols, B. N., Nichols, T. E., Pellman, J., ... Poldrack, R. A. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data*, *3*(1), Article 160044. <https://doi.org/10.1038/sdata.2016.44>
- Gorman, A. M. (1961). Recognition memory for nouns as a function of abstractness and frequency. *Journal of Experimental Psychology*, *61*(1), 23–29. <https://doi.org/10.1037/h0040561>
- Hall, J. (1954). Learning as a function of word-frequency. *American Journal of Psychology*, *67*(1), 138–140. <https://doi.org/10.2307/1418080>
- Healey, M. K. (2018). Temporal contiguity in incidentally encoded memories. *Journal of Memory and Language*, *102*, 28–40. <https://doi.org/10.1016/j.jml.2018.04.003>
- Healey, M. K., Crutchley, P., & Kahana, M. J. (2014). Individual differences in memory search and their relation to intelligence. *Journal of Experimental Psychology: General*, *143*(4), 1553–1569. <https://doi.org/10.1037/a0036306>
- Healey, M. K., & Kahana, M. J. (2014). Is memory search governed by universal principles or idiosyncratic strategies? *Journal of Experimental Psychology: General*, *143*(2), 575–596. <https://doi.org/10.1037/a0033715>
- Healey, M. K., & Kahana, M. J. (2016). A four-component model of age-related memory change. *Psychological Review*, *123*(1), 23–69. <https://doi.org/10.1037/rev0000015>
- Healey, M. K., & Kahana, M. J. (2020). Age-related differences in the temporal dynamics of spectra power during memory encoding. *PLoS ONE*, *15*(1), Article e0227274. <https://doi.org/10.1371/journal.pone.0227274>
- 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>
- Healey, M. K., & Uitvlugt, M. G. (2019). The role of control processes in temporal and semantic contiguity. *Memory & Cognition*, *47*(4), 719–737. <https://doi.org/10.3758/s13421-019-00895-8>
- Helfrich, R. F., Knight, R. T., & D'Esposito, M. T. (in press). Methods to study human memory. In M. J. Kahana, & A. D. Wagner (Eds.), *Oxford handbook of human memory* (2nd ed.). Oxford University Press.
- Hicks, J. L., Marsh, R. L., & Cook, G. I. (2005). An observation on the role of context variability in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 1160–1164. <https://doi.org/10.1037/0278-7393.31.5.1160>
- 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>
- Howard, M. W., Kahana, M. J., & Wingfield, A. (2006). Aging and contextual binding: Modeling recency and lag-recency effects with the temporal context model. *Psychonomic Bulletin & Review*, *13*(3), 439–445. <https://doi.org/10.3758/BF03193867>
- Howard, M. W., Viskontas, I. V., Shankar, K. H., & Fried, I. (2012). Ensembles of human MTL neurons “jump back in time” in response to a repeated stimulus. *Hippocampus*, *22*(9), 1833–1847. <https://doi.org/10.1002/hipo.22018>
- Jenkins, J. J., & Russell, W. A. (1952). Associative clustering during recall. *Journal of Abnormal and Social Psychology*, *47*(4), 818–821. <https://doi.org/10.1037/h0063149>
- Jiang, Z., Liu, H., Peng, C., & Yan, H. (2022). Investor memory and biased beliefs: Evidence from the field. *SSRN*.
- 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. (2017). Memory search. In J. H. Byrne (Ed.), *Learning and memory: A comprehensive reference* (2nd ed., Vol. 2, pp. 181–200). Academic Press. <https://doi.org/10.1016/B978-0-12-809324-5.21038-9>
- 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), P92–P97. <https://doi.org/10.1093/geronb/60.2.P92>
- Kahana, M. J., Howard, M. W., Zaromb, F., & Wingfield, A. (2002). Age dissociates recency and lag recency effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 530–540. <https://doi.org/10.1037/0278-7393.28.3.530>
- Kahana, M. J., Rudoler, J. H., Lohnas, L. J., Healey, K., Aka, A., Broitman, A., Crutchley, E., Crutchley, P., Alm, K. H., Katerman, B. S., Miller, N. E., Kuhn, J. R., Li, Y., Long, N. M., Miller, J., Paron, M. D., Pazdera, J. K., Pedisich, I., & Weidemann, C. T. (2023). Penn Electrophysiology of Encoding and Retrieval Study (PEERS). *OpenNeuro*. <https://doi.org/10.18112/openneuro.ds004395.v2.0.0>
- Kahneman, D., Sibony, O., & Sunstein, C. (2021). *Noise: A flaw in human judgment*. Little, Brown Spark.
- Katerman, B. S., Li, Y., Pazdera, J. K., Keane, C., & Kahana, M. J. (2022). EEG biomarkers of free recall. *NeuroImage*, *246*, Article 118748. <https://doi.org/10.1016/j.neuroimage.2021.118748>
- Kuhn, J. R., Lohnas, L. J., & Kahana, M. J. (2018). A spacing account of negative recency in final free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(8), 1180–1185. <https://doi.org/10.1037/xlm0000491>
- Li, Y., Pazdera, J. K., & Kahana, M. J. (2024). EEG decoders track memory dynamics. *Nature Communications*, *15*, Article 2981. <https://doi.org/10.1038/s41467-024-46926-0>
- Liu, T. T. (2016). Noise contributions to the fMRI signal: An overview. *NeuroImage*, *143*, 141–151. <https://doi.org/10.1016/j.neuroimage.2016.09.008>
- Lohnas, L. J., Healey, M. K., & Davachi, L. (2023). Neural temporal context reinstatement of event structure during memory recall. *Journal of Experimental Psychology: General*, *152*(7), 1840–1872. <https://doi.org/10.1101/2021.07.30.454370>
- Lohnas, L. J., & Kahana, M. J. (2013). Parametric effects of word frequency in memory for mixed frequency lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(6), 1943–1946. <https://doi.org/10.1037/a0033669>
- Lohnas, L. J., & Kahana, M. J. (2014). Compound cuing in free recall. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *40*(1), 12–24. <https://doi.org/10.1037/a0033698>
- Lohnas, L. J., Polyn, S. M., & Kahana, M. J. (2015). Expanding the scope of memory search: Modeling intralist and interlist effects in free recall. *Psychological Review*, *122*(2), 337–363. <https://doi.org/10.1037/a0039036>
- Long, N. M., Burke, J. F., & Kahana, M. J. (2014). Subsequent memory effect in intracranial and scalp EEG. *NeuroImage*, *84*, 488–494. <https://doi.org/10.1016/j.neuroimage.2013.08.052>
- Long, N. M., Danoff, M. S., & Kahana, M. J. (2015). Recall dynamics reveal the retrieval of emotional context. *Psychonomic Bulletin and Review*, *22*(5), 1328–1333. <https://doi.org/10.3758/s13423-014-0791-2>
- Long, N. M., & Kahana, M. J. (2017). Modulation of task demands suggests that semantic processing interferes with the formation of episodic

- associations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(2), 167–176. <https://doi.org/10.1037/xlm0000300>
- Long, N. M., Sperling, M. R., Worrell, G. A., Davis, K. A., Gross, R. E., Lega, B. C., Jobst, B. C., Sheth, S. A., Zaghoul, K., Stein, J. M., & Kahana, M. J. (2017). Contextually mediated spontaneous retrieval is specific to the hippocampus. *Current Biology*, 27(7), 1074–1079. <https://doi.org/10.1016/j.cub.2017.02.054>
- Madan, C. R. (2021). Exploring word memorability: How well do different word properties explain item free-recall probability? *Psychonomic Bulletin & Review*, 28(2), 583–595. <https://doi.org/10.3758/s13423-020-01820-w>
- Madigan, S. A. (1969). Intraserial repetition and coding processes in free recall. *Journal of Verbal Learning and Verbal Behavior*, 8(6), 828–835. [https://doi.org/10.1016/S0022-5371\(69\)80050-2](https://doi.org/10.1016/S0022-5371(69)80050-2)
- Manning, J. R. (in press). Context reinstatement. In M. J. Kahana, & A. D. Wagner (Eds.), *Oxford handbook of human memory* (2nd ed.). Oxford University Press.
- Manning, J. R., Polyn, S. M., Baltuch, G., Litt, B., & Kahana, M. J. (2011). Oscillatory patterns in temporal lobe reveal context reinstatement during memory search. *Proceedings of the National Academy of Sciences, USA*, 108(31), 12893–12897. <https://doi.org/10.1073/pnas.1015174108>
- 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>
- Markiewicz, C. J., Gorgolewski, K. J., Feingold, F., Blair, R., Halchenko, Y. O., Miller, E., Hardcastle, N., Wexler, J., Esteban, O., Goncalves, M., Jwa, A., & Poldrack, R. (2021). The OpenNeuro resource for sharing of neuroscience data. *eLife*, 10, Article e71774. <https://doi.org/10.7554/eLife.71774>
- McCrae, R. R., & Costa, P. T., Jr. (2010). The five-factor model, five-factor theory, and interpersonal psychology. In L. M. Horowitz & S. Strack (Eds.), *Handbook of interpersonal psychology: Theory, research, assessment, and therapeutic interventions* (pp. 91–104). Wiley.
- Meisler, S. L., Kahana, M. J., & Ezyat, Y. (2019). Does data cleaning improve brain state classification? *Journal of Neuroscience Methods*, 328, Article 108421. <https://doi.org/10.1016/j.jneumeth.2019.108421>
- Melton, A. W. (1970). The situation with respect to the spacing of repetitions and memory. *Journal of Verbal Learning and Memory*, 9(5), 596–606. [https://doi.org/10.1016/S0022-5371\(70\)80107-4](https://doi.org/10.1016/S0022-5371(70)80107-4)
- Merritt, P. S., DeLosh, E. L., & McDaniel, M. A. (2006). Effects of word frequency on individual-item and serial order retention: Tests of the order-encoding view. *Memory & Cognition*, 34(8), 1615–1627. <https://doi.org/10.3758/BF03195924>
- 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>
- Mundorf, A. M. D., Uitvlugt, M. G., & Healey, M. K. (2022). Does depth of processing affect temporal contiguity? *Psychonomic Bulletin & Review*, 29(6), 2229–2239. <https://doi.org/10.3758/s13423-022-02112-1>
- Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, 64(5), 482–488. <https://doi.org/10.1037/h0045106>
- Murdock, B. B., & Okada, R. (1970). Interresponse times in single-trial free recall. *Journal of Verbal Learning and Verbal Behavior*, 8(2), 263–267. <https://doi.org/10.1037/h0029993>
- Naim, M., Katkov, M., Recanatesi, S., & Tsodyks, M. (2019). Emergence of hierarchical organization in memory for random material. *Scientific Reports*, 9(1), Article 10448. <https://doi.org/10.1038/s41598-019-46908-z>
- Osth, A. F., & Farrell, S. (2019). Using response time distributions and race models to characterize primacy and recency effects in free recall initiation. *Psychological Review*, 126(4), 578–609. <https://doi.org/10.1037/rev0000149>
- Ozubko, J., & Joordens, S. (2007). The mixed truth about frequency effects on free recall: Effects of study list composition. *Psychonomic Bulletin & Review*, 14(5), 871–876. <https://doi.org/10.3758/BF03194114>
- Patterson, K. E., Meltzer, R. H., & Mandler, G. (1971). Inter-response times in categorized free recall. *Journal of Verbal Learning and Verbal Behavior*, 10(4), 417–426. [https://doi.org/10.1016/S0022-5371\(71\)80041-5](https://doi.org/10.1016/S0022-5371(71)80041-5)
- Pernet, C. R., Appelhoff, S., Gorgolewski, K. J., Flandin, G., Phillips, C., Delorme, A., & Oostenveld, R. (2019). EEG-BIDS, an extension to the brain imaging data structure for electroencephalography. *Scientific Data*, 6(1), Article 103. <https://doi.org/10.1038/s41597-019-0104-8>
- Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., Nichols, T. E., Poline, J.-B., Vul, E., & Yarkoni, T. (2017). Scanning the horizon: Towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience*, 18(2), 115–126. <https://doi.org/10.1038/nrn.2016.167>
- Pollio, H. R., Kasschau, R. A., & DeNise, H. E. (1968). Associative structure and the temporal characteristics of free recall. *Journal of Verbal Learning and Verbal Behavior*, 10(2, Pt. 1), 190–197. <https://doi.org/10.1037/h0025385>
- Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review*, 116(1), 129–156. <https://doi.org/10.1037/a0014420>
- Popov, V., & Reder, L. M. (2020). Frequency effects on memory: A resource-limited theory. *Psychological Review*, 127(1), 1–46. <https://doi.org/10.1037/rev0000161>
- Roediger, H. L., & Payne, D. G. (1985). Recall criterion does not affect recall level or hypermnnesia: A puzzle for generate/recognize theories. *Memory & Cognition*, 13(1), 1–7. <https://doi.org/10.3758/BF03198437>
- Rohrer, D., & Wixted, J. T. (1994). An analysis of latency and interresponse time in free recall. *Memory & Cognition*, 22(5), 511–524. <https://doi.org/10.3758/BF03198390>
- Romani, S., Katkov, M., & Tsodyks, M. (2016). Practice makes perfect in memory recall. *Learning & Memory*, 23(4), 169–173. <https://doi.org/10.1101/lm.041178.115>
- Rubinstein, D. Y., Weidemann, C. T., Sperling, M. R., & Kahana, M. J. (2023, June). Direct brain recordings suggest a causal subsequent-memory effect. *Cerebral Cortex*, 33(11), 6891–6901. <https://doi.org/10.1093/cercor/bhad008>
- Russek, E., Acosta-Kane, D., Van Opheusden, B., Mattar, M. G., & Griffiths, T. (2022). Time spent thinking in online chess reflects the value of computation.
- Sanquist, T. F., Rohrbaugh, J. W., Syndulko, K., & Lindsley, D. B. (1980). Electrocorical signs of levels of processing: Perceptual analysis and recognition memory. *Psychophysiology*, 17(6), 568–576. <https://doi.org/10.1111/j.1469-8986.1980.tb02299.x>
- Schonfield, D., & Robertson, B. A. (1966). Memory storage and aging. *Canadian Journal of Psychology*, 20(2), 228–236. <https://doi.org/10.1037/h0082941>
- Schulman, A. I. (1967). Word length and rarity in recognition memory. *Psychonomic Science*, 9(4), 211–212. <https://doi.org/10.3758/BF03330834>
- Sederberg, P. B., Gauthier, L. V., Terushkin, V., Miller, J. F., Barnathan, J. A., & Kahana, M. J. (2006). Oscillatory correlates of the primacy effect in episodic memory. *NeuroImage*, 32(3), 1422–1431. <https://doi.org/10.1016/j.neuroimage.2006.04.223>
- Sederberg, P. B., Howard, M. W., & Kahana, M. J. (2008). A context-based theory of recency and contiguity in free recall. *Psychological Review*, 115(4), 893–912. <https://doi.org/10.1037/a0013396>
- Sederberg, P. B., Miller, J. F., Howard, W. H., & Kahana, M. J. (2010). The temporal contiguity effect predicts episodic memory performance. *Memory & Cognition*, 38(6), 689–699. <https://doi.org/10.3758/MC.38.6.689>
- Sheaffer, R., & Levy, D. A. (2022). Negative recency effects in delayed recognition: Spacing, consolidation, and retrieval strategy processes. *Memory & Cognition*, 50(8), 1683–1693. <https://doi.org/10.3758/s13421-022-01293-3>
- Shepard, R. N. (1967). Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning and Verbal Behavior*, 6(1), 156–163. [https://doi.org/10.1016/S0022-5371\(67\)80067-7](https://doi.org/10.1016/S0022-5371(67)80067-7)
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366. <https://doi.org/10.1177/0956797611417632>

- Smith, P. L., & Little, D. R. (2018). Small is beautiful: In defense of the small-N design. *Psychonomic Bulletin & Review*, 25(6), 2083–2101. <https://doi.org/10.3758/s13423-018-1451-8>
- Spiers, H. J., Coutrot, A., & Hornberger, M. (2023). Explaining world-wide variation in navigation ability from millions of people: Citizen science project sea hero quest. *Topics in Cognitive Science*, 15(1), 120–138. <https://doi.org/10.1111/tops.12590>
- Spillers, G. J., & Unsworth, N. (2011). Variation in working memory capacity and temporal-contextual retrieval from episodic memory. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 37(6), 1532–1539. <https://doi.org/10.1037/a0024852>
- Szucs, D., & Ioannidis, J. P. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology*, 25(3), Article e2000797. <https://doi.org/10.1371/journal.pbio.2000797>
- Unsworth, N. (2007). Individual differences in working memory capacity and episodic retrieval: Examining the dynamics of delayed and continuous distractor free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(6), 1020–1034. <https://doi.org/10.1037/0278-7393.33.6.1020>
- Unsworth, N., & Brewer, G. (2010). Variation in working memory capacity and intrusions: Differences in generation or editing? *European Journal of Cognitive Psychology*, 22(6), 990–1000. <https://doi.org/10.1080/09541440903175086>
- Unsworth, N., Brewer, G., & Spillers, G. (2010). Understanding the dynamics of correct and error responses in free recall: Evidence from externalized free recall. *Memory & Cognition*, 38(4), 419–430. <https://doi.org/10.3758/MC.38.4.419>
- Wahlheim, C. N., Alexander, T. R., & Kane, M. J. (2019). Interpolated retrieval effects on list isolation: Individual differences in working memory capacity. *Memory & Cognition*, 47(4), 619–642. <https://doi.org/10.3758/s13421-019-00893-w>
- Wahlheim, C. N., Ball, B. H., & Richmond, L. L. (2017). Adult age differences in production and monitoring in dual-list free recall. *Psychology and Aging*, 32(4), 338–353. <https://doi.org/10.1037/pag0000165>
- Wahlheim, C. N., & Huff, M. J. (2015). Age differences in the focus of retrieval: Evidence from dual-list free recall. *Psychology and Aging*, 30(4), 768–780. <https://doi.org/10.1037/pag0000049>
- Ward, G., Tan, L., & Grenfell-Essam, R. (2010). Examining the relationship between free recall and immediate serial recall: The effects of list length and output order. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36(5), 1207–1241. <https://doi.org/10.1037/a0020122>
- Weidemann, C. T., & Kahana, M. J. (2016). Assessing recognition memory using confidence ratings and response times. *Royal Society Open Science*, 3(4), Article 150670. <https://doi.org/10.1098/rsos.150670>
- Weidemann, C. T., & Kahana, M. J. (2019). Dynamics of brain activity reveal a unitary recognition signal. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 45(3), 440–451. <https://doi.org/10.1037/xlm0000593>
- 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>
- Wickens, T. D. (2002). *Elementary signal detection theory*. Oxford University Press.
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1), Article 160018. <https://doi.org/10.1038/sdata.2016.18>
- Yonelinas, A. P., Hockley, W. E., & Murdock, B. B. (1992). Tests of the list-strength effect in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(2), 345–355. <https://doi.org/10.1037/0278-7393.18.2.345>
- Zaromb, F. M., Howard, M. W., Dolan, E. D., Sirotn, Y. B., Tully, M., Wingfield, A., & Kahana, M. J. (2006). Temporal associations and prior-list intrusions in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(4), 792–804. <https://doi.org/10.1037/0278-7393.32.4.792>
- Zhang, Q., Griffiths, T. L., & Norman, K. A. (2023). Optimal policies for free recall. *Psychological Review*, 130(4), 1104–1124. <https://doi.org/10.1037/rev0000375>

Appendix

PEERS Multisession Aging Study

PEERS included a cohort of 39 older adults who each participated in 10 experimental sessions (the preliminary screening session, seven sessions of PEERS Experiment 1, and the two sessions of psychometric testing described previously). Analysis of data from these older adults replicated several basic findings that suggest cognitive aging impacts some memory processes more than others (Healey & Kahana, 2016). For example, whereas there was substantial age-related impairment in free recall there was a more modest age-related impairment in item recognition (Schonfield & Robertson, 1966). Even within free recall, older adults showed a complex pattern of preserved and impaired functioning. Specifically, older adults showed no deficits in recall initiation (primacy and recency, Kahana et al., 2002) or semantic organization. They did, however, show a substantial reduction in temporal organization (a reduced contiguity effect, Figure 3, see Healey et al., 2019; Howard et al., 2006; Wahlheim & Huff, 2015). Older adults also exhibited a greater tendency to commit prior and ELIs (Wahlheim et al., 2017; Zaromb et al., 2006), but there were no age differences in the tendency for PLIs to come from recent versus remote lists.

The EEG data collected from older adults allowed us to investigate the biomarkers of this pattern of age-related behavioral change. Healey and Kahana (2020) found that age-related memory deficits are associated with differences in how neural activity changes across serial positions during study. Previous work had established that, among younger adults, oscillatory power changes in a highly consistent way from item-to-item across the study period (Sederberg et al., 2006). The PEERS aging data showed that at frequencies below 3 Hz and above 14 Hz there were virtually no age differences—at these frequencies power tended to decrease rapidly across serial positions, regardless of age. In contrast frequencies between 4 and 14 Hz showed very large age differences. Whereas for young adults, power at these frequencies tended to increase across serial positions, for older adults power decreased across serial positions in an almost complete crossover interaction. That is, at these frequencies, older adults showed higher power than younger adults early in a study list, but the age difference reversed at later serial positions. Moreover, older adults with the smallest behavioral memory deficits showed the largest departures from the younger adult pattern of neural activity. This result

(Appendix continues)

may suggest that age differences in the dynamics of neural activity across an encoding period reflect changes in cognitive processing that compensate for age-related decline.

To investigate longitudinal age-related change in memory, we recruited a subgroup of older adult subjects to return each year to repeat the seven PEERS Experiment 1 sessions. Among the original cohort of older adults, eight came back for 5 years of repeat testing sessions. This extensive within-subject data allowed us to evaluate age-related changes in performance while factoring out the potential effects of repeated testing. Broitman et al. (2020) fit a model to session-level changes in performance that included a term for the established power-law improvements in task performance resulting from practice (Anderson et al., 1999) and the

effects of aging, which we assumed to be approximately linear across this 5-year period. When applied to our annual-testing sample, the model uncovered both significant practice effects (an average increase of 0.72% annually) and a modest age-related decline in recall probability (an average of 0.14% annually). These model-based analyses illustrate how one can use data from multi-session experiments with small number of subjects to address questions normally studied in large-scale individual difference studies.

Received October 4, 2022

Revision received August 24, 2023

Accepted October 22, 2023 ■