The Penn Electrophysiology of Encoding and Retrieval Study

Kahana, M.J., Lohnas, L.J., Healey, M.K., Aka, A., Broitman, A.W., Crutchley, P., Crutchley, E., 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.
University of Pennsylvania

Abstract

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 90 minute 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 science approach to the study of memory and the human brain.

Introduction

Although Herman Ebbinghaus is known to students of memory for his herculean investigations of list learning comprising nearly 2,000 hours of self experimentation (Ebbinghaus, 1885/1913), nearly all of what we have come to know about human memory in the last century has been gleaned from single-session experiments typically performed by samples of fewer than 100 students, often in fulfillment of a psychology course requirement. This approach contrasts with research areas such as perception where a small handful of observers contribute data across 5-20 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: (1) to obtain sufficient trial-level data so that we could apply models to individual subject's performance measures, (2) to obtain sufficient data across subjects to permit the analysis of individual differences, and (3) to obtain high-quality continuous EEG data during memory

[1] Smith and Little (2018) articulate the benefits of small N studies.

The authors gratefully acknowledge support from National Institutes of Health grant MH55687. Correspondence concerning this article should be addressed to Michael J.Kahana, kahana@psych.upenn.edu.

encoding and retrieval, thus allowing us to relate brain measures to indices of performance with high reliability. For the purpose of studying 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 recall and recognition tasks. Although findings emerging from these studies have appeared in more than 20 scientific publications, the present paper 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 sub-disciplines, our field faces the same forces that impede replicability. One such force is the extreme variability of human cognition, behavior, and physiology (e.g., Kahneman, Sibony, & Sunstein, 2021; Kahana, Aggarwal, & Phan, 2018). Empirical patterns can differ reliably across individuals and even within an individual and variability in these effects do not arise simply 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. This is because variability in task performance must give rise to variability in brain activity, but measured brain activity includes additional sources of 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, it should not surprise anyone that finding robust results using brain recording methods should require substantially *greater* numbers of observations than those relying only on measures of task performance to achieve the same level of statistical power. Yet, the cost of obtaining neural measures forces researchers to economize on data collection, thus fueling variability across experiments. Finally, the highly multivariate nature of neural data encourages researchers to look at their data in myriad ways, thereby increasing the chance of false positives unless every step of an analysis pipeline has been "preregistered" (Simmons, Nelson, & Simonsohn, 2011).

To advance our understanding of these methods in the face of the above challenges, we need some way of estimating the power of our neural measurements. To probe the upper-bound of what could be learned using scalp EEG, we assembled the PEERS datasets, which comprise millions of encoding events and recall responses. 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 datasets encompass fewer than 100 hours of experimental data collection, increasing our dataset 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, 2016, 2014). 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 been disseminating 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 facilitate dissemination of the PEERS data to interested scientists.

Methods Overview

Below we provide a concise summary of the methods for each of the five PEERS experiments. Each experiment involved multiple sessions of memory tasks with EEG recording. 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 prior to signing on for the full experiment. Subjects completed the first three PEERS experiments across a series of 20 sessions involving word list recall and recognition tasks. In addition, this cohort also completed two sessions of neuropsychological tests. Table 1 gives the number of subjects who completed each experiment. Whereas Experiments 1-3 included encoding task manipulations, variation in distractor conditions, and end-of-session recognition and final-recall tests, 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 the earlier PEERS studies, that subjects would not be able to complete more than 24 two-hour long sessions in a single term. With the goal of maximizing our statistical power we thus recruited subjects for a 24 session experiment, striving to enroll 10 subjects at the start of each term (we typically completed around 7). Because human speech is the most natural medium for recalling information, we collected vocal responses which we annotated (offline) for accuracy and response times. Because vocalization causes significant EMG artifact in EEG data, we conducted a fifth PEERS experiment designed specifically to control for pre-motor correlates of retrieval. This was the last PEERS study completed prior to the start of the COVID-19 pandemic. Below we provide a concise description of the experimental methods, with additional details provided in an online appendix at `memory.psych.upenn.edu`.

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

PEERS experiment	N	Sessions	Dates
Preliminary Experiment	∼700	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	2015-2018
Exp. 5: Long-delay recall + pre-motor control	57	10	2019-2020

with a recognition test (yellow box). Half of the sessions were randomly chosen to include a final free recall test before recognition (in final free recall, 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* instructions. In externalized recall (Kahana, Dolan, Sauder, & Wingfield, 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 to press the spacebar following any such error.

Subjects encountered three types of lists: (1) No-task lists, which they studied with the generic instruction of trying to learn the items for a subsequent test, (2) 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 (3) task shift lists, where subjects alternated between size and animacy tasks every 2-6 items within each list. The size task asked subjects "Will this item fit into a shoebox?"; the animacy task asked "Does this word refer to something living or not living?" The current task was indicated by the color, font and case of the presented item. 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 PEERS1-3. 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 5 possible durations for the between-item and end-of-list distractor tasks, such that 2 lists had each of the 5 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, 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 one 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-second distractor period (see Figure 1C). A random half of the lists (excluding the first list) were preceded by a 24-second, distractor-filled delay. A free recall test followed the post-list distractor on each list.

The word pool for this experiment consisted of a 576-word subset of the 1638-word pool used in a previous PEERS experiment, and subjects saw the same 576 words (24 lists $\times$ 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, as described in the Appendix. Subjects were given a short break (approximately 5 minutes) after every 8 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 sec (or longer) before saying the word 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 were given 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; Lohnas, Polyn, & Kahana, 2015; Zaromb et al., 2006). We administered this long-delay recall task as the start of each of the sessions 6 through 10, giving subjects 10 minutes 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 for achieving high levels of recall while maintaining a high-level

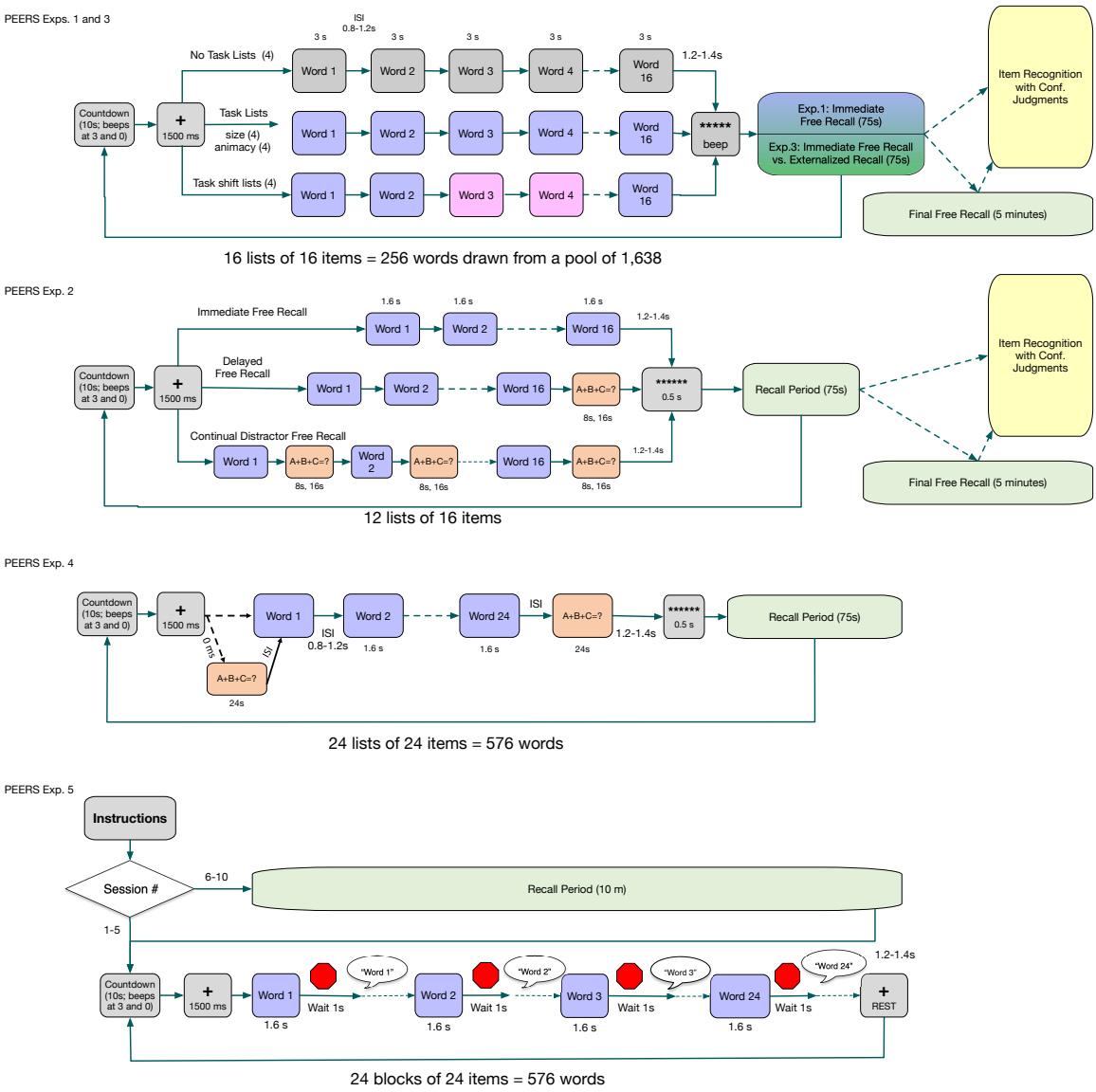

Figure 1. **Schematic of PEERS methods.** See text for details.

of performance on the arithmetic distractor tasks. In addition, we provided a bonus to subjects for maintaining a low blink rate during critical item presentation events.

PEERS Raw Data Repository and online Methods description

All PEERS data may be freely obtained from the Computational Memory Lab webpage, hosted by the University of Pennsylvania: `http://memory.psych.upenn.edu` The same website also provides a detailed methods description of each of the PEERS studies briefly described above.

Results

Here we present our results organized into five major sections. Section 1 provides an overview of the major behavioral findings. Section 2 discusses both experimental and endogenous sources of variability in recall of items and lists. Section 3 focuses on EEG correlates of successful memory encoding. Section 4 focuses on individual differences and model-based analyses of performance. Section 5 discusses EEG correlates of memory retrieval.

1. Overview of Major Behavioral Findings

The PEERS free-recall experiments replicated many classic findings, including serial position effects, temporal and semantic organization of memories, the exponential growth of inter-response times with output position, and subjects' tendency to commit extra-list and prior-list intrusions as a function of their temporal and semantic relation to the just-recalled items. Having subjects make size or animacy judgments during word encoding led to worse overall performance than free encoding (for similiar findings, see Polyn, Norman, & Kahana, 2009; Long et al., 2017; Mundorf, Lazarus, Uitvlugt, & Healey, 2021). Subjects exhibited strong temporal and semantic organization regardless of encoding task condition, but both size and animacy encoding tasks led to more semantic organization and less temporal organization as compared with no-task lists (for temporal organization see Figure 3A). On lists where subjects had to switch from size to animacy at random points, recall transitions were more likely between items encoded with the same task instruction (accounting for the lag between these items in the study list).

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 (inter-item) distractor (Kahana (2017); Lohnas and Kahana (2014); see, 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 first 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. Finally, we find striking effects of semantic similarity on free recall (e.g. Manning, Sperling, Sharan, Rosenberg, & Kahana, 2012), across all distractor conditions, as illustrated by Figure 2E (Kahana, 2017).

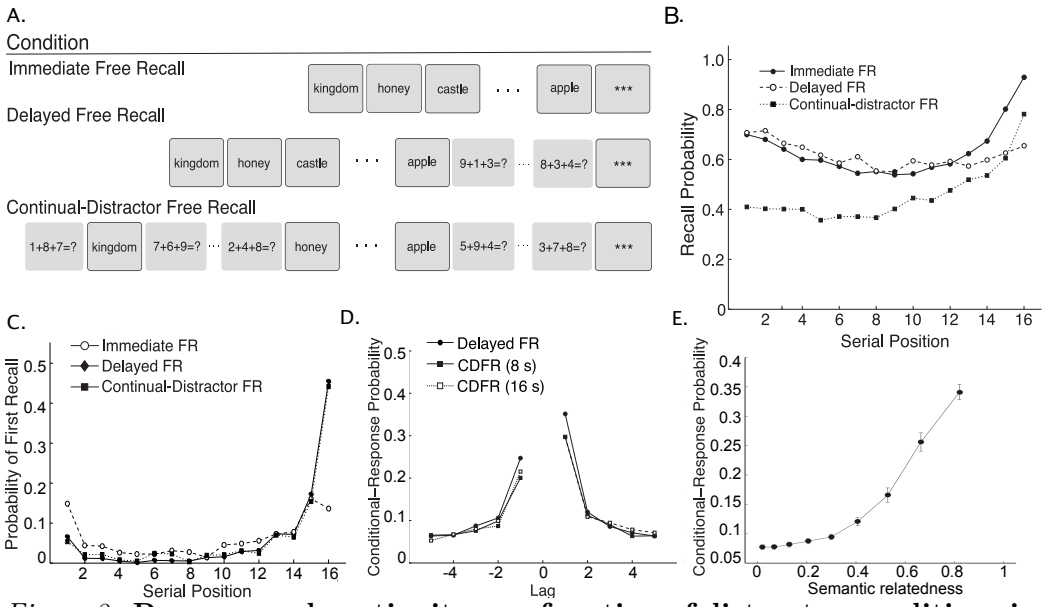

Figure 2. **Recency and contiguity as a function of distractor conditions in PEERS Experiment 2. A.** Illustration of immediate, delayed, and continual distractor free recall tasks (IFR, DFR and CDFR). **B.** Serial position analysis showing recency in IFR, attenuated recency in DFR, and long-term recency in CDFR. **C.** Recall initiation, as measured by the probability of first recall, shows that initiating with recenct 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 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 prior-list and extra-list intrusions, but have little or no effect on correct recalls (Lohnas et al., 2015). Inclusion of externalized recall instructions provided valuable data on intrusions which occur only rarely in standard free recall.

Because subjects participated in PEERS Experiments 1 – 3 as a series of experiments, data from PEERS Experiment 3 provides valuable information on free recall under conditions of high practice (i.e., in each session of PEERS Experiment 3, subjects will have been doing word-list free recall and recognition for a dozen or more prior sessions). Here we found a positive effect of practice on recall performance, but a large effect on temporal organization, with subjects increasingly exhibiting a tendency to make successive transitions among items studied in neighboring serial positions (see Figure 3B). This finding also appeared in PEERS Experiment 4, as described below.

PEERS Experiments 1 – 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 final free recall (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 recall rates for the last few items in each list (Craik, 1970). Kuhn, Lohnas, and Kahana (2018) found that the negative recency effect critically depended on when subjects recalled those 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 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: 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 DFR and CDFR conditions of PEERS Experiment 2.

After FFR (or if absent, after the immediate free recall period of the last list), all 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). Given that retrieval is highly cue-dependent, we wanted to include additional assays of memory for the purpose of obtaining more information on the successful encoding of studied items and also to provide additional means of examining the neural correlates of retrieval (see, e.g., Weidemann & Kahana, 2019). Performance in these tasks replicated classic findings, including the relation between confidence and response times in recognition (Murdock & Dufty, 1972), and the shape of the ROC curves (Lockhart & Murdock, 1970; Van Zandt, 2000).

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 1-3. 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 dataset for the study of variability in memory, across items, lists and sessions (see Section 2). The large number of trials obtained in this study also allowed us to conduct detailed analyses of inter-response times during recall, as described in Goldman and Kahana (2022). A discussion of the EEG correlates of memory encoding and retrieval in each of the PEERS studies appears in later sections.

2. 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. By measuring EEG activity during memory encoding, for example, we can observe variability across items in the mnemonic processes that predict subsequent retrieval. Successful encoding of words, however, may also reflect psycholinguistic properties of items. Here we first examine how properties of words and lists relate to their subsequent memorability; we then consider the possibility that endogenous processes, unrelated to experimentally controlled factors, may also underlie variability in memory performance.

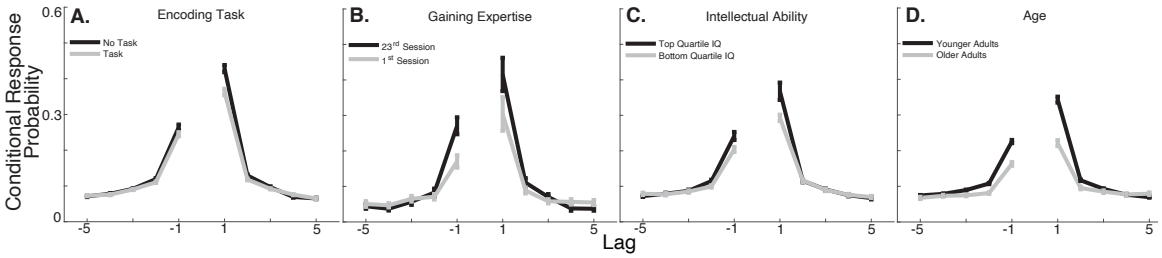

Figure 3. **Conditions influencing the contiguity effect. A.** The contiguity effect is smaller when assigned a task on how to encode the items (a size or animacy judgment) than when not given instructions on how to encode the list. **B.** Task experience amplifies the contiguity effect: a large contiguity effect is present in the 1st 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.

Consider how memory for a word varies with the word's frequency of occurrence in the English language. Here, classic studies report a large word frequency effect in recognition memory, with subjects exhibiting superior memory for rare words than for common words (Schulman, 1967; Shepard, 1967). In free recall, however, studies have reported inconsistent effects, with some researchers finding superior recall for rare words, and other researchers finding superior recall for common words. 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. In recognition memory, they found the expected pattern: 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 4).

A unique aspect of PEERS Experiment 4 is that subjects studied the same set of 576 words in each of the 23 experimental sessions. Results from this dataset showed us that some words, lists, and sessions are easier to recall than others. What is the source of this variability? Aka et al. took a psycholinguistic approach to answer this question and studied how word features relate to both word and list-level memorability. 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, Phan, and Kahana (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.

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 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 5 for data on list number,

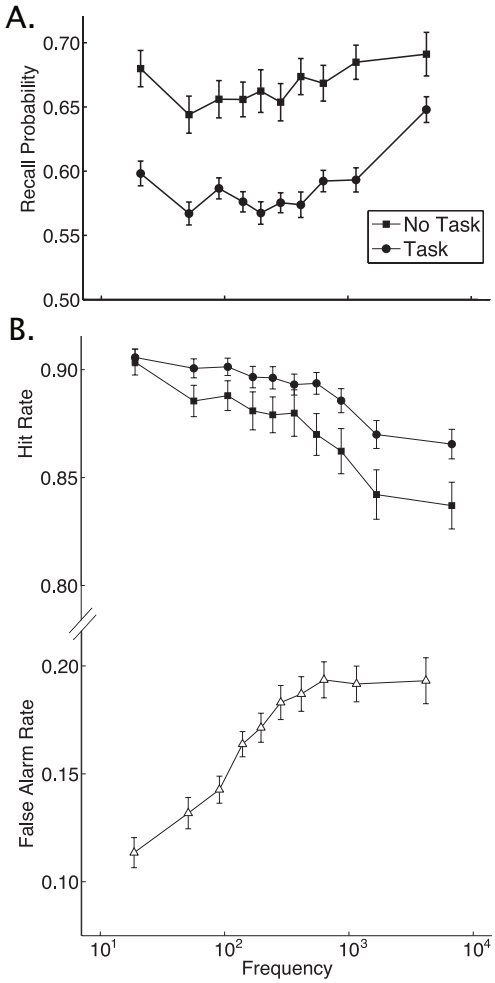

Figure 4. **Word frequency effects in recall and recognition. A.** Subjects recalled higher proportions of both low frequency and high frequency words as compared with intermediate frequency words, regardless of whether the item was presented without an encoding task (filled squares) or with an encoding 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) included in these analysis were partitioned into deciles on the basis of their word frequency counts in the CELEX2 database. Error bars represent 95% confidence intervals.

Table 1

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

	M β	*SE β*
Predictors of Word-Level Recall Model		
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 Model		
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

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$ Word Length, Valence, Arousal, and Animacy variables are residualized variables.

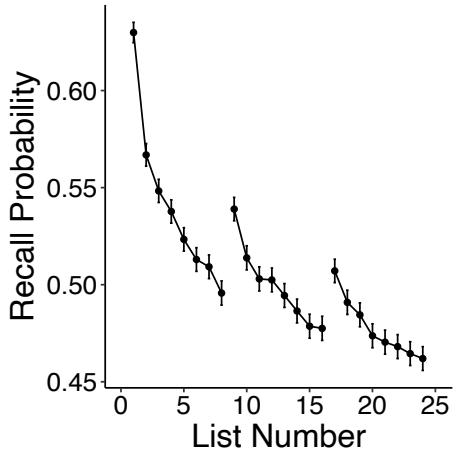

Figure 5. **Predictors of interlist variability.** Within each session, recall decreased across successive lists, but increased following the two breaks, consistent with a proactive interference account.

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 excessive variability. To test this hypothesis, they added performance on the prior list as an additional explanatory variable in their model and found that this was in fact the strongest predictor of performance on a given list. Converging evidence for the endogenous factors came from investigations of item and list-level subsequent memory effect described in the following section (Weidemann & Kahana, 2021).

An unpublished study by Kreiger, Aggarwal, and Kahana (2019) offered further support for 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 prior to the start of the list. The end-of-list distractor serves the role of disrupting active rehearsal and thereby diminishing the recency effect (see Figure 6). 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, Hockley, and Murdock (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 below.

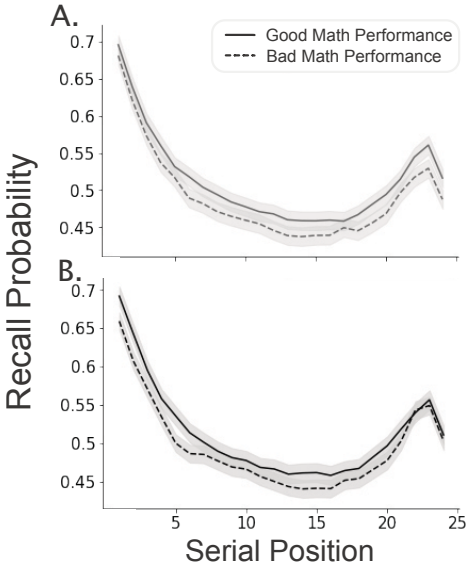

Figure 6. **Recall and Distractor Task Performance. A.** When a math distractor task follows a study list, there is a greater difference in recall probability between good and bad math performance for later serial positions. **B.** When a math distractor task precedes a study list, this difference is greater for earlier serial positions.

3. EEG Correlates of Successful Memory Encoding

Our first set of electrophysiological investigations sought to examine the EEG features during memory encoding that predict subsequent recall (the so-called subsequent memory effect, or SME). Long et al. (2014), analyzing a subset of the PEERS Experiment 1 data, discovered 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 sought to determine whether these scalp EEG biomarkers of successful encoding overlap with spectral biomarkers determined using direct brain recordings in neurosurgical patients with drug-resistant epilepsy (e.g., Sederberg, Kahana, Howard, Donner, & Madsen, 2003; Sederberg et al., 2007). To answer this question, they conducted a careful comparison between the scalp topography and the frequency specificity of the PEERS Experiment 1 EEG data and a large intracranial dataset reported by Burke et al. (2014). This comparison revealed striking commonalities in both the regions and time courses of the LFA and HFA effects, supporting the conclusion that scalp EEG can resolve similar signals, albeit with far poorer spatial resolution.

Long and Kahana (2017) tested the hypothesis that these HFA/LFA biomarkers track not only 'whether' a stimulus will be subsequently remembered, but 'how' a stimulus is later recalled. Specifically, the authors assessed 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 (Figure 7). 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 suggest that temporal vs. semantic based organization may trade off, given their dependence on the same biomarkers.

The preceding results illustrate some ways in which EEG activity during item encoding relate to whether and how it is subsequently recalled. One may ask, however, whether these EEG correlates of subsequent recall reflect properties of the item or perhaps slowly changing brain states that 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 regression models to predict recall performance from a range of neural features. To account for the effects of external factors (such as properties of individual words or their positions within a list or experimental session) they first regressed out the effects of these factors and then calculated a "corrected" subsequent memory effect, by using neural features to predict the residual recall performance. To assess the extent to which neural features that predict subsequent recall performance

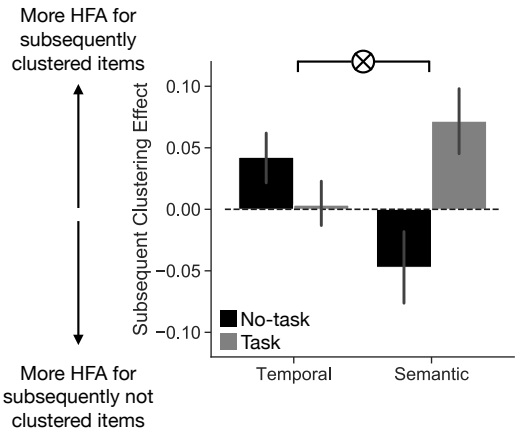

Figure 7. **Neural Subsequent Clustering Effect.** Difference in study-phase high frequency activity (HFA; 44-100Hz) over left prefrontal cortex between words that are later temporally or semantically clustered, separately for no-task (black) and task (grey) lists. HFA is greater for subsequently temporally clustered words studied with no task and greater for subsequently semantically clustered words studied with a task. There is a significant interaction between the type of clustering (temporal, semantic) and whether subjects performed a semantic orienting task ($p < 0.001$). Error bars represent standard error of the mean.

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. Figure 8 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).

PEERS included a cohort of 39 older adults who each took part in 10 experimental sessions (the preliminary screening session, seven sessions of PEERS Experiment 1, and the two sessions of psychometric testing described previously) The EEG data collected from the 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 above 14 Hz, there were virtually no age differences in these neural gradients—both age groups showed a reduction in power across the list. Moreover, older adults who showed the smallest age-related behavioral memory deficits showed the largest departures from the younger adult pattern of neural activity. These results suggest that age differences in the dynamics of neural activity across

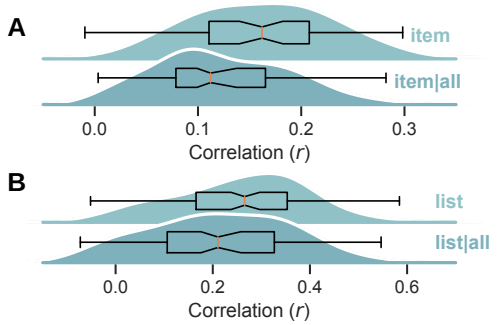

Figure 8. **Item-level and list-level subsequent memory effects (SMEs).** Distributions of correlations between multivariate model predictions and item (A) and list-level (B) free-recall performance. Each panel shows the full SME (labeled "item" and "list" respectively) as well as a corrected SMEs after effects from a range of external factors have been removed ("item|all" and "list|all" respectively). This figure is adapted from Weidemann and Kahana (2021).

an encoding period reflect changes in cognitive processing that compensate for age-related decline.

4. 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, Crutchley, and Kahana (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 contributes to successful memory search (see also Sederberg, Miller, Howard, & Kahana, 2010; Spillers & Unsworth, 2011). Moreover, variation in the temporal contiguity effect (but not the other effects) correlated positively with full-scale WAIS-IV IQ (see Figure 9). 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 some subjects almost always initiate with the first list item. Here the average data did not provide an accurate representation of each individual.

Before conducting individual-level modeling, however, we used the PEERS data to extend retrieved context theory to multi-list experiments. We briefly summarize the resulting

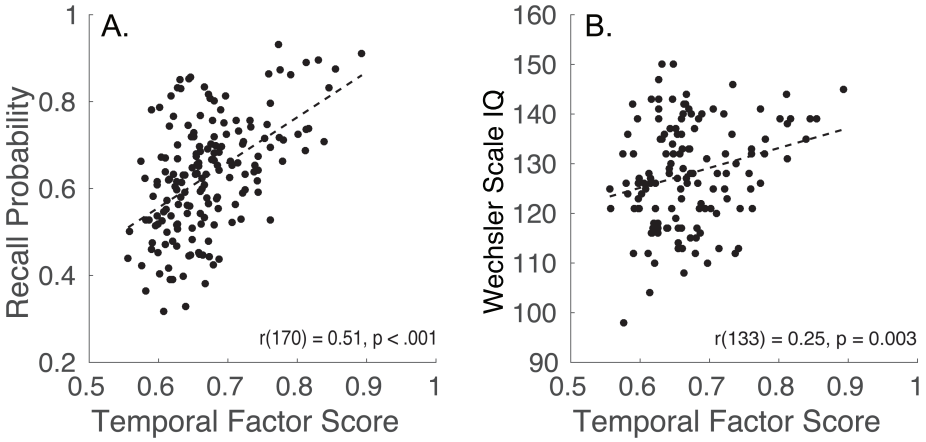

Figure 9. **Individual differences in contiguity predict memory performance and IQ A.** The correlation between temporal factor scores and overall recall probability. Temporal factor scores give the average percentile ranking the temporal lag of each actual transition with respect to the lags of all transitions that were possible at that time. **B.** The correlation between temporal factor scores and intelligence as measured by the Wechsler Adult Intelligence Scale IV.

CMR2 model, and then discuss application of this model to individual differences, including age-related changes in memory performance. CMR2 sought to address a fundamental problem long neglected by memory modelers: 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., Sederberg, Howard, & Kahana, 2008; Polyn et al., 2009), CMR2 allowed the associative structures that store memories to continuously accumulate. The 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 retrieve their associated past contexts, which in turn update the state of context. This recursive notion of contextual retrieval endows the model with dynamics that match many details of list recall tasks.

In CMR2, Lohnas et al. extended these earlier ideas to the situation where information from prior lists impacts memory for information on the current, target, list. While many circumstances entail interactions between new and old memories, most list memory experiments create an artificial situation in which the rememberer seeks to focus their retrieval exclusively on the target list. Nonetheless, information from prior lists can impact current list recalls, as evidenced by subjects tendency to make prior-list intrusions from recent lists. Yet, such intrusion errors occur infrequently, indicating that subjects can control their search of memory to the current list, possibly by filtering out recalls that come from inappropriate contexts (e.g. Bahrick, 1970; Jacoby & Hollingshead, 1990; Raaijmakers & Shiffrin, 1980).

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 the current list. Each generated item retrieves its associated context state from study, and CMR2 only recalls a generated item if its retrieved temporal context is similar to the current context. Although CMR2 can query which items are generated but not recalled, subjects require additional instruction. In the externalized free recall (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, Brewer, & Spillers, 2010; Wahlheim, Alexander, & Kane, 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. With this set-up, subjects recall a large proportion of prior-list intrusions, extra-list intrusions, and repeats, and rarely reject correct items (Kahana et al., 2005). Although these results indicate that subjects *can* reject items successfully, it still leaves open the question of whether this is actually what subjects are doing during immediate free recall.

Lohnas et al. (2015) tested the generate-recognize mechanism using data from PEERS Experiment 3 (as shown in Figure 1, some subjects performed externalized recall, 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 account, Lohnas et al. (2015) found that CMR2 predicted the proportion and probability of rejection for prior-list intrusions 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. To further test CMR2's assumption of the role of temporal context in the generate-recognize mechanism, Lohnas et al. (2015) examined rejections of prior-list intrusions. In both CMR2 predictions and PEERS data, rejections of PLIs increased as a function of list recency.

Having established CMR2 as a successful model of free recall phenomena for mean data, we now return to the question of individual differences. Healey and Kahana (2014) fit CMR2 to individual subject data in PEERS Experiment 1 and found that the model provided a good fit to data from ∼95% of individual subjects. Healey and Kahana (2016) further tested the individual-subject modeling approach by asking whether CMR2 could account for key differences between younger and older adults in PEERS Experiment 1 data, including the elevated intrusion rates exhibited by older adults. They first fit CMR2 to data from individual younger and older adults using the Kahana, Howard, Zaromb, and Wingfield (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: 1) contextual retrieval, 2) the sustained attention (related to the primacy gradient), 3) error monitoring (related to rejecting intrusions), and 4) decision noise. Fig. 10 (A-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

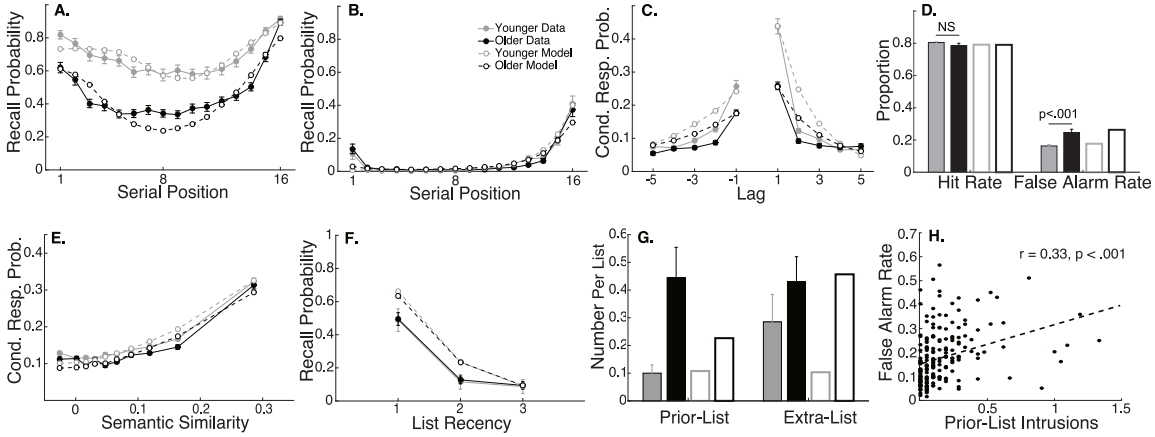

Figure 10. **Age-related changes in Recall and Recognition.** Panels A-C illustrate serial position, probability of first recall and contiguity effects; Panel D illustrates recognition memory hits and false alarms; Panel E illustrates semantic organization; Panels F-G illustrate intrusion errors, and 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 CMR2 simulations from Healey and Kahana (2016).

a subject makes in free recall should correlate positively with the number of false alarms they make in recognition. As shown in Fig. 10H, the PEERS data confirmed this prediction.

Analysis of data from the 39 older adults who took part in PEERS Experiment 1 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 Howard, Kahana, & Wingfield, 2006; Wahlheim & Huff, 2015). Older adults also exhibited a greater tendency to commit prior and extra-list intrusions (Zaromb et al., 2006; Wahlheim, Ball, & Richmond, 2017), but the were no age differences in the tendency for prior list intrusions to come from recent versus remote lists.

To investigate longitudinal age-related change in memory, we recruited a subgroup of the 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 five years of repeat testing sessions. This extensive within-subject data allowed us to evaluate age-related changes in performance while factoring out potential effects of repeated testing. Broitman, Kahana, and Healey (2019) 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, Fincham, & Douglass, 1999) and the effects of aging, which we assumed to be approximately linear across this five-year period. When applied to our annual-testing sample, the model uncovered both significant practice effects (increase of 0.72% annually) and a modest age-related decline in recall probability (0.14% annually). These model-based analyses illustrate how one can use data from multi-session experiments with small numbers

of subjects to address questions normally studied in large-scale individual difference studies.

Cohen and Kahana (in press) 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, Danoff, and Kahana (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 replicated Long et al.'s emotional clustering effect in the larger PEERS Experiment 4 dataset. They then 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 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.

5. EEG biomarkers of memory retrieval

Li, Pazdera, and Kahana (2022) examined the spectral correlates successful retrieval in PEERS Experiment 4. Comparing EEG activity immediately preceding correct recalls and intrusion errors they found marked increases in high frequency activity in the 500 ms period leading up to successful recall. Accompanying these HFA 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 (see Figure 11B). A majority of subjects also exhibited modest increases in theta activity preceding successful recall, but this effect did not prove reliable in aggregate statistical comparisons.

Conducting parallel analyses on successful encoding, Li et al. replicated the overall spectral patterns identified by Long et al. (2014) and Long et al. (2017); namely, increased HFA and diminished alpha activity accompanied successful memory encoding (see Figure 11A). Li et al. also found a striking positive theta effect overlying frontal regions of the brain, extending earlier findings implicating frontal theta in working memory and cognitive control (Cavanagh & Frank, 2014; Jacobs, Hwang, Curran, & Kahana, 2006).

Katerman, Li, Pazdera, Keane, and Kahana (2021) investigated the spectral correlates of memory retrieval after very long delays, using a pre-vocalization period in immediate-recall a control for premotor actvity (in PEERS Experiment 5). In addition to demonstrating increased HFA and decreased alpha activity, as seen in Li et al. (2022), 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 12, where black outlines indicate frequency-region pairs that met an FDR-corrected $p<0.05$ 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 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

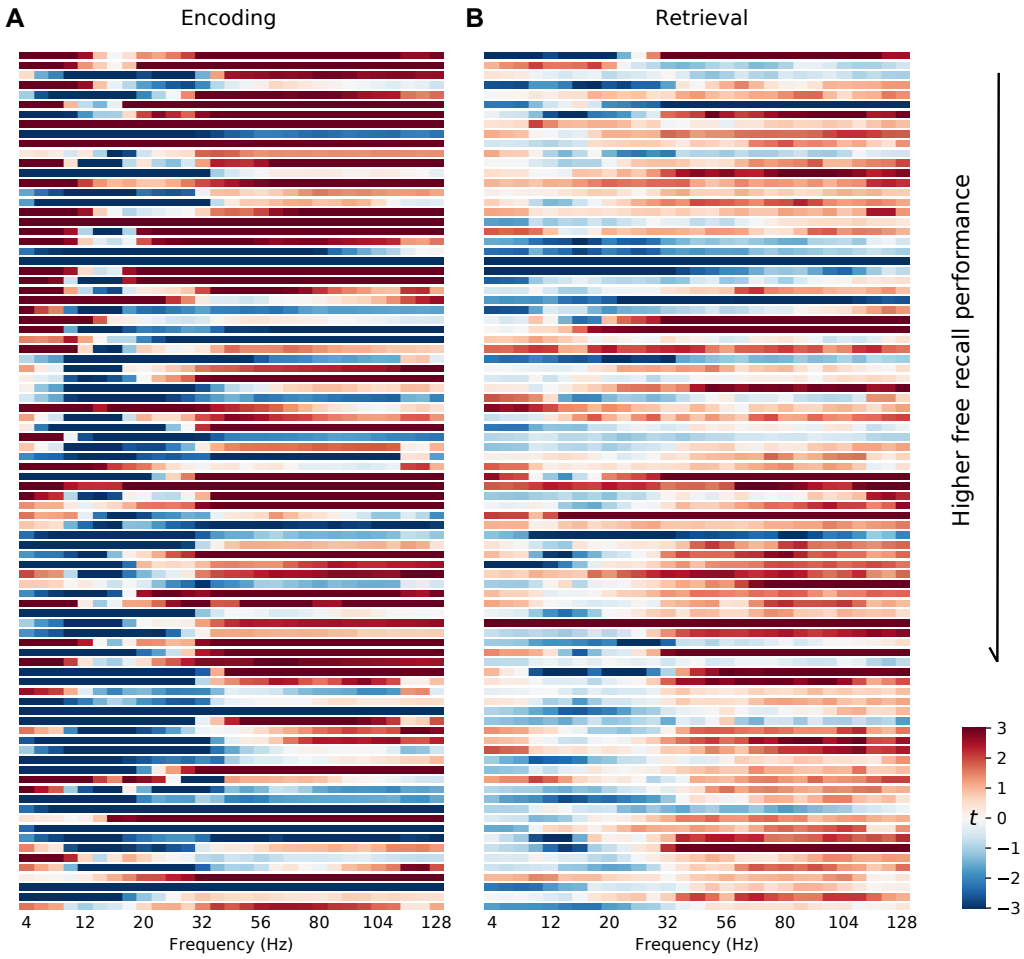

Figure 11. **Subject-specific spectral markers of successful episodic memory in encoding and retrieval.** Each row shows results from one subject, sorted in ascending recall performance. Subject-specific independent t-statistics for the successful and unsuccessful memory contrast are collapsed across eight ROIs. Power increases and decreases are shown in red and in blue, respectively.

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 1-3 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

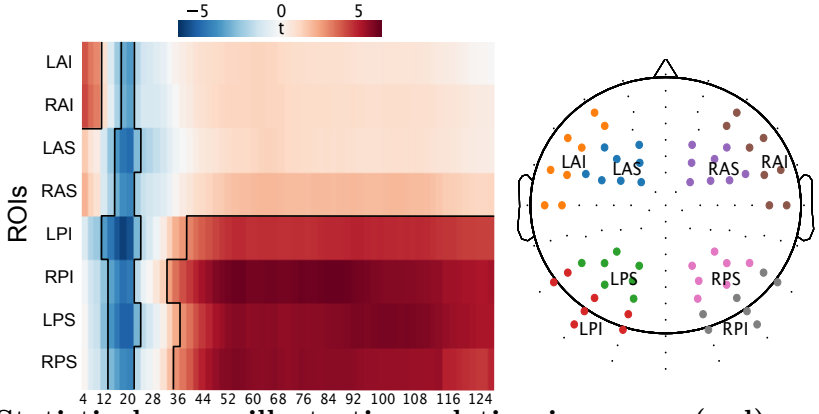

Figure 12. **Statistical maps illustrating relative increases (red) and decreases (blue) in spectral power across key memory contrasts for eight regions of interest.** Spectral power contrast for Delayed Recall vs. Immediate Recall in PEERS Experiment 5.

hit rates across the range of the measures (Wickens, 2002). 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 prefect discrimination between old and new items. Figure 13 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).

At the same time, free recall confers other advantages over recognition tests in the study of retrieval processes. In particular, when study items reappear as probe items on a recognition test, similarities in brain activity between encoding and retrieval may reflect item similarity rather than memory reinstatement. By contrast, 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 mind of the subject. Lohnas, Healey and Davachi (BioRxiv, 2021) examined the neural correlates of context reinstatement in scalp EEG, asking specifically how task manipulations influence 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 (Folkerts, Rutishauser, & Howard, 2018; Howard, Viskontas, Shankar, & Fried, 2012; Manning, Polyn, Baltuch, Litt, & Kahana, 2011). They then show that spectral features of scalp EEG activity demonstrate the reinstatement of temporal context prior to word recall (using data from PEERS Experiment 1). 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 methods). Lohnas et al. hypothesized that a change in task causes a disruption to 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 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 neural similarity of neighboring item pairs with the same task minus neighboring item pairs with different tasks. Across subjects, the neural similarity difference during encoding were correlated with the neural similarity difference during recall. This provides further evidence that temporal context states, including disruptions to context representations, are reinstated during free recall. Further, 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.

Lessons learned

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

Subject recruitment, retention and performance monitoring. Each term we sought to recruit between 8 and 12 subjects to participate in the full 22 sessions of PEERS Experiments 1 – 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 insure 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 for 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.

Our experience indicated that an experimenter should be present during a subject's first session of each new phase of the experiment. 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

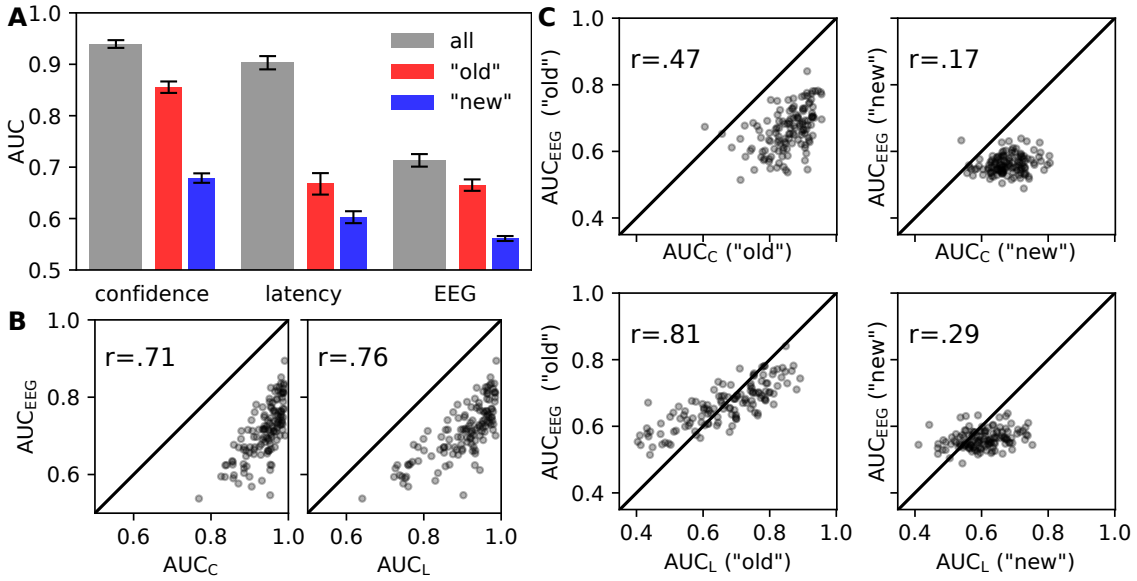

Figure 13. **Inferring memory strength from confidence ratings, response latencies, and EEG activity. A.** The area under the ROC curve (AUC) for ROC functions constructed from confidence ratings, response latencies, and EEG. These AUCs indicate the extent to which the corresponding measure reflects a memory signal. As detailed in 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 in Weidemann and Kahana (2019), this strong correspondence is difficult to interpret because every ROC functions 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. Figure adapted from Weidemann and Kahana (2019)

video. We also provided subjects with a "call button" that they could use to ask assistance from the experimenter.

Annotation of vocal responses. Although one can collect free recall responses using a keyboard, spoken recall still remains as 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 (`memory.psych.upenn.edu/TotalRecall`). Over the course of this project we refined the *Penn Total Recall* software, making it easier for researchers to efficiently process the recordings. 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 fully 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 made the decision to take 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.

Session-level report generation. Early in the project we discovered that data quality issues could emerge after a certain session (perhaps 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.

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., Sanquist, Rohrbaugh, Syndulko, & Lindsley, 1980; Donald, 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 non-invasive 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 don't 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 multi-session 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.

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 datasets, as compared with traditional studies reporting novel data. With the exception of a few reviewers and editors, most did not comment specifically on the large scale of our data collection effort, or our efforts to publicly disseminate our data.

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 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, is the famous first exception). Individual investigators, or more typically individual trainees, design and carry out small scale experiments either on humans or non-human animals. In the case of human research, each subject typically takes part in just one hour-long session. In a survey of articles published in the *Journal of Experimental Psychology* in 2015, the median number of hours of human experimentation was 40 (mean = 55±6.5 S.E.M.). By 2021 that number had increased to 66 (mean = 144 ±31 S.E.M.).

[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.

Although these numbers appear to be trending upward, and may even provide the power required to support the study's main conclusions, it is quite unlikely that they are powered to support interesting secondary analyses aimed at elucidating the higher-order structure within the data. Yet, until quite recently many of the field's leading journals expect papers reporting multiple original experiments (a median of three experiments per article in the years surveyed above). 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 crowd-source 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 vs. 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 (Helfrich, Knight, & D'Esposito, in press; Ezzyat & Suthana, 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.

Big Data: Risks and Rewards

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 (M.J.K.) 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, to complement smaller studies. Indeed, the discoveries made possible with big data can inspire conceptual replications with smaller studies.

In conclusion, we see the PEERS project as a test-case in applying big-data approaches to the study of 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 already (Romani, Katkov, & Tsodyks, 2016; Naim, Katkov, Recanatesi, & Tsodyks, 2019; Osth & Farrell, 2019; Popov & Reder, 2020; Madan, 2021; Zhang, Griffiths, & Norman, 2021; Sheaffer & Levy, 2022) and we hope that this paper, in synthesizing key motivations, methods and discoveries, will prompt additional investigators to consider the value of this approach (and the resulting data) for their own questions.

References

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

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.

Bahrick, H. P. (1970). Two-phase model for prompted recall. *Psychological Review*, *77*(3), 215–222. doi: 10.1037/h0029099

Berger, H. (1929). Über das Elektrenkephalogramm des Menschen (On the human electroencephalogram). *Archiv für Psychiatry und Nervenkrankheiten*, *87*, 527-570.

Broitman, A. W., Kahana, M. J., & Healey, M. K. (2019). Modeling retest effects in a longitudinal measurement burst study of memory. *Computational Brain & Behavior*, *3*(2), 200-207. doi: https://dx.doi.org/10.1007%2Fs42113-019-00047-w

Burke, J. F., Long, N. M., Zaghloul, K. A., Sharan, A. D., Sperling, M. R., & Kahana, M. J. (2014). Human intracranial high-frequency activity maps episodic memory formation in space and time. *NeuroImage*, *85*, 834–843. doi: 10.1016/j.neuroimage.2013.06.067

Cavanagh, J. F., & Frank, M. J. (2014). Frontal theta as a mechanism for cognitive control. *Trends in Cognitive Sciences*, *18*(8), 414-421. doi: 10.1016/j.tics.2014.04.012

Cohen, R. T., & Kahana, M. J. (in press). A memory based theory of emotional disorders. *Psychological Review*. doi: 10.1101/817486

Craik, F. I. M. (1970). The fate of primary memory items in free recall. *Journal of Verbal Learning and Verbal Behavior*, *9*(2), 658-664. doi: 10.1016/S0022-5371(70)80042-1

Donald, M. W. (1980). Memory, learning and event-related potentials. *Progress in Brain Research*, *54*, 615-627. doi: 10.1016/S0079-6123(08)61681-7

Ebbinghaus, H. (1885/1913). *Memory: A contribution to experimental psychology.* New York: Teachers College, Columbia University.

Ezzyat, Y., & Suthana, N. (in press). Oxford handbook of human memory. In M. J. Kahana & A. D. Wagner (Eds.), (2nd ed., chap. Brain Stimulation). Oxford, U. K.: Oxford University Press.

Folkerts, S., Rutishauser, U., & Howard, M. (2018). Human episodic memory retrieval is accompanied by a neural contiguity effect. *Journal of Neuroscience*, *38*(17), 4200–4211. doi: 10.1523/JNEUROSCI.2312-17.2018

Goldman, S. T., & Kahana, M. J. (2022). In search of chunking. *Manuscript in preparation.*

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. doi: 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. doi: 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. doi: 10.1037/rev0000015

Healey, M. K., & Kahana, M. J. (2020). Age-related differences in the temporal dynamics of spectra power during memory encoding. *PLOSone*, *15*(1).

Healey, M. K., & Uitvlugt, M. G. (2019). The role of control processes in temporal and semantic contiguity. *Memory Cognition*, *47*(4), 719-737.

Helfrich, R. F., Knight, R. T., & D'Esposito, M. T. (in press). Oxford handbook of human memory. In M. J. Kahana & A. D. Wagner (Eds.), (2nd ed., chap. Methods to Study Human Memory). Oxford, U. K.: Oxford University Press.

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

Howard, M. W., Kahana, M. J., & 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. doi: 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*, 1833–1847.

Jacobs, J., Hwang, G., Curran, T., & Kahana, M. J. (2006). EEG oscillations and recognition memory: Theta correlates of memory retrieval and decision making. *NeuroImage*, *15*(2), 978–87.

Jacoby, L. L., & Hollingshead, A. (1990). Toward a generate/recognize model of performance

on direct and indirect tests of memory. *Journal of Memory and Language*, *29*, 433-454. doi: 10.1016/0749-596X(90)90065-8

Kahana, M. J. (2017). Memory search. In J. H. Byrne (Ed.), *Learning and memory: A comprehensive reference* (Second Edition ed., Vol. 2, p. 181-200). Academic Press. doi: 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. doi: 10.1037/xlm0000553

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

Kahana, M. J., 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. doi: 10.1037/0278-7393.28.3.530

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. (2021). EEG biomarkers of free recall. *NeuroImage*, *246*, 118748. doi: 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. doi: 10.1037/xlm0000491

Li, Y., Pazdera, J. K., & Kahana, M. J. (2022). EEG decoders track memory dynamics. *Submitted*.

Liu, T. T. (2016). Noise contributions to the fMRI signal: An overview. *Neuroimage*, *143*, 141–151. doi: https://doi.org/10.1016/j.neuroimage.2016.09.008

Lockhart, R. S., & Murdock, B. B. (1970). Memory and the theory of signal detection. *Psychological Bulletin*, *74*(2), 100–109.

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. doi: 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. doi: 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. doi: 10.1037/a0039036

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. doi: 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. doi: 10.1037/xlm0000300

Long, N. M., Sperling, M. R., Worrell, G. A., Davis, K. A., Gross, R. E., Lega, B. C., ...

Kahana, M. J. (2017). Contextually mediated spontaneous retrieval is specific to the hippocampus. *Current Biology*, *27*(7), 1074–1079. doi: 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*, 583-595. doi: https://doi.org/10.3758/s13423-020-01820-w

Manning, J. R. (in press). Oxford handbook of human memory. In M. J. Kahana & A. D. Wagner (Eds.), (2nd ed., chap. Context Reinstatement). Oxford, U. K.: 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. doi: 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. doi: 10.1523/JNEUROSCI.5321-11.2012

Mundorf, A., 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.

Murdock, B. B., & Dufty, P. O. (1972). Strength theory and recognition memory. *Journal of Experimental Psychology*, *94*, 284-290.

Naim, M., Katkov, M., Recanatesi, S., & Tsodyks, M. (2019). Emergence of hierarchical organization in memory for random material. *Scientific reports*, *9*(1), 10448.

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. doi: https://doi.org/10.1037/rev0000149

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. doi: 10.1037/a0014420

Popov, V., & Reder, L. M. (2020). Frequency effects on memory: A resource-limited theory. *Psychological Review*, *127*(1), 1-46. doi: https://doi.org/10.1037/rev0000161

Raaijmakers, J. G. W., & Shiffrin, R. M. (1980). SAM: A theory of probabilistic search of associative memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 14, p. 207-262). New York: Academic Press. doi: 10.1016/S0079-7421(08)60162-0

Roediger, H. L., & Payne, D. G. (1985). Recall criterion does not affect recall level or hypermnesia: A puzzle for generate/recognize theories. *Memory & Cognition*, *13*(1), 1-7.

Romani, S., Katkov, M., & Tsodyks, M. (2016). Practice makes perfect in memory recall. *Learning & Memory*, *23*, 169-173. doi: 10.1101/lm.041178.115

Sanquist, T. F., Rohrbaugh, J. W., Syndulko, K., & Lindsley, D. B. (1980). Electrocortical signs of levels of processing: Perceptual analysis and recognition memory. *Psychophysiology*, *17*(6), 568–576. doi: 10.1111/j.1469-8986.1980.tb02299.x

Schonfield, D., & Robertson, B. A. (1966). Memory storage and aging. *Canadian Journal of Psychology*, *20*, 228–236.

Schulman, A. I. (1967). Word length and rarity in recognition memory. *Psychonomic Science*, *9*(4), 211–212. doi: 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. doi: 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. doi: 10.1037/a0013396

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

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. doi: 10.3758/MC.38.6.689

Sederberg, P. B., Schulze-Bonhage, A., Madsen, J. R., Bromfield, E. B., McCarthy, D. C., Brandt, A., ... Kahana, M. J. (2007). Hippocampal and neocortical gamma oscillations predict memory formation in humans. *Cerebral Cortex*, *17*(5), 1190–1196. doi: 10.1093/cercor/bhl030

Sheaffer, R., & Levy, D. A. (2022). Negative recency effects in delayed recognition: Spacing, consolidation, and retrieval strategy processes. *Memory & Cognition*. doi: 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*, 156-163.

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.

Smith, P. L., & Little, D. R. (2018). Small is beautiful: In defense of the small-n design. *Psychonomic bulletin & review*, *25*(6), 2083 - 2101.

Spillers, G. J., & Unsworth, N. (2011). Variation in working memory capacity and temporal–contextual retrieval from episodic memory. *Journal Experimental Psychology: Learning, Memory and Cognition*, *37*(6), 1532–1539.

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. doi: 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. doi: 10.3758/MC.38.4.419

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(3), 582-600.

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*, 619-642. doi: 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. doi:

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. doi: https://doi.org/10.1037/pag0000049

Weidemann, C. T., & Kahana, M. J. (2016). Assessing recognition memory using confidence ratings and response times. *Royal Society Open Science*, *3*, 150670. doi: 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. doi: http://dx.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.

Wickens, T. D. (2002). *Elementary signal detection theory.* Oxford University Press.

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.

Zaromb, F. M., Howard, M. W., Dolan, E. D., Sirotin, 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. doi: 10.1037/0278-7393.32.4.792

Zhang, Q., Griffiths, T. L., & Norman, K. A. (2021). Optimal policies for free recall. *PsyArXiv*. doi: 10.31234/osf.io/sgepb