OXFORD
UNIVERSITY PRESS
Journal of Gerontology: Psychological Sciences

Modeling Retest Effects in a Longitudinal Measurement Burst Design Study of Episodic Memory

Journal:	*Journal of Gerontology: Psychological Sciences*
Manuscript ID	JGPS-2017-221
Manuscript Type:	Original Research Report
Date Submitted by the Author:	24-Aug-2017
Complete List of Authors:	Broitman, Adam; University of Pennsylvania, Psychology Kahana, Michael; University of Pennsylvania, Psychology Healey, M.; Michigan State University, Psychology
Keywords:	Memory, Longitudinal Change, Quantitative Methods
Alternate Keyword:	Free Recall

Modeling Retest Effects in a Longitudinal Measurement Burst Design Study of Episodic Memory

Adam W. Broitman
University of Pennsylvania; Philadelphia, Pennsylvania

Michael J. Kahana
University of Pennsylvania; Philadelphia, Pennsylvania

M. Karl Healey
Michigan State University; East Lansing, Michigan

Word Count: 2,733

Corresponding Author:
M. Karl Healey
Department of Psychology
Michigan State University
316 Physics Road, Room 240D
East Lansing, MI 48824
khealey@msu.edu
Phone: (517) 432 – 3107

Abstract

Objectives. Longitudinal designs must deal with the confound between increasing age and increasing task experience (i.e., retest effects). Most existing methods for disentangling these factors rely on large sample sizes and are impractical for smaller scale projects. Here, we provide a method for separating aging and retest effects with a modest sample size.

Method. We conducted a measurement burst study in which eight participants completed a burst of seven sessions of free recall every year for 5 years. Six control participants completed a burst only in years 1 and 5, and should, therefore, have a smaller retest effect but equal age effects. We modeled memory performance as a combination of age-related change and accumulating test experience.

Results. The raw data suggested slight improvement in memory over 5 years. But fitting the model to the yearly-testing group revealed that a substantial positive retest effect was obscuring stability in memory performance. Supporting this finding, the control group showed a smaller retest effect but an equal age effect.

Discussion. Measurement burst designs combined with models of retest effects allow researchers to employ longitudinal designs in areas where previously only cross-sectional designs were feasible.

Keywords: free recall; memory models; stability

Modeling Retest Effects in a Longitudinal Measurement Burst Design Study of Episodic Memory

Inferring age-related cognitive change from cross-sectional designs is fraught with well-known inferential problems (Baltes, 1968). Longitudinal designs, in principle, provide a more direct measure of within-individual cognitive change and are therefore an important complement to cross-sectional research (Hoffman, Hofer, & Sliwinski, 2011). But longitudinal studies generally introduce retest effects (e.g., practice effects), which can obscure age-related effects (Hoffman et al., 2011; Salthouse, 2016).

Techniques have been developed to disentangle age-related and retest-related effects in typical longitudinal designs in which a very large sample of participants is tested once on each measure (at each time point; e.g., Salthouse, 2016). This typical design is not appropriate, however, when the constructs of interest cannot be reliably measured with a single test. For example, in cross-sectional designs we have had participants complete seven sessions of free recall to provide sufficiently reliable measures to study individual (Healey, Crutchley, & Kahana, 2014) and age (Healey & Kahana, 2016) differences in the dynamics of episodic memory search.

Extending this multi-session design to a longitudinal study would constitute what has been termed a “measurement burst” design (Nesselroade, 1991; Sliwinski, 2008): A burst is composed of multiple tests separated by a short time (e.g., days) and successive bursts are separated by a longer time (e.g., a year). This intensive testing makes it impractical to undertake a longitudinal study with a sample large enough to apply most existing methods of estimating retest effects.

Sliwinski, Hoffman, and Hofer (2010) introduced a method to separate age and retest effects in measurement burst designs. This method involves modeling changes in performance across retests as the combined output of a linear function of age and a non-linear function of number of retests (e.g., Munoz, Sliwinski, Scott, & Hofer, 2015). Here, we report the initial results of a measurement burst longitudinal study in which six participants completed seven sessions of the free recall task each year for five years. To establish this as a methodologically feasible approach to longitudinal research with modest sample sizes, we attempt to separately model retest-related and age-related effects.

Method

The data are from the Penn Electrophysiology of Encoding and Retrieval Study (PEERS, Healey et al., 2014; Healey & Kahana, 2014, 2016; Lohnas & Kahana, 2013, 2014; J. F. Miller, Kahana, & Weidemann, 2012), an ongoing project aiming to assemble a large database on memory ability in older and younger adults. The full methods of the PEERS study, which include some manipulations that we do not consider in this paper, are described in the supplemental materials; here, we focus on the details relevant to our analyses.

Participants – Original cross-sectional PEERS sample

The full PEERS older adult sample includes 39 individuals who completed an initial cross-sectional study (Healey & Kahana, 2016). As described below, 18 of these participants were recruited to return for longitudinal testing (12 were retested yearly, 6 were retested after 5 years). All participants were recruited from the Philadelphia area. Potential participants were excluded if they suffered from any medical conditions or regularly took medications that might affect their cognitive performance.

Yearly-testing Sample. Twelve older adults were recruited for annual testing. The age of participants ranged from 62 to 73 years (*M* = 66.87) at the start of the experiment, and the participants completed each yearly burst ranging from 1.6 to 19.0 weeks (*M* = 3.9). Four of these participants have been excluded from the current analyses due to insufficient data (3 participants decided to leave the study, and 1 has passed away). Of the 8 participants included in the present analyses, 2 have completed four annual waves of testing and 6 have completed five waves.

Practice-Control Sample. Six additional older adults from the original sample were recruited to return 5 years after their first burst. Their ages ranged from 62 to 79 years (*M* = 66.83) at the start of the experiment, and they completed each yearly burst ranging from 1.1 to 6.3 weeks (*M* = 3.7).

PEERS Experiment

Once recruited, participants completed 7 sessions of the free recall task each year. At the beginning of each wave, the Recent Life Changes Questionnaire (M. A Miller & Rahe, 1997) was administered to collect information about any potential changes in each participant's health or personal lives. No participants included in the current analyses developed a medical condition that would have excluded them from initial participation.

Each session included 16 free recall lists. For each list, 16 words were presented one at a time on a computer screen followed by an immediate free recall test. Each stimulus was drawn from a pool of 1638 words. Lists were constructed such that varying degrees of semantic relatedness occurred at both adjacent and distant serial positions.

For each list, there was a 1500 ms delay before the first word appeared on the screen. Each item was on the screen for 3000 ms, followed by jittered (i.e., variable) inter-stimulus interval of 800 – 1200 ms (uniform distribution). After the last item in the list, a tone sounded, and a row of asterisks appeared. The participant was then given 75 seconds to recall aloud any of the just-presented items.

Results

The solid gray lines in Figure 1A show changes in free recall performance (proportion of words recalled) across sessions and years for the yearly-testing sample. The data show little sign of declining memory performance across years. In fact, there is a modest increase from year 1 to year 5. To quantify this trend, we began by conducting a linear regression for each participant using the number of days that had elapsed since their first session (defining session 1 as day 1) to predict their memory performance in individual sessions. This provided us with a slope (which we report as change in memory performance per year) for each participant. Figure 1B shows that the average slope was .0058 (i.e., on a 0 to 1.0 scale, performance increased by .0058 per year), with 95% confidence intervals that include zero. Thus, there is a small, non-significant, increase across years.

Although performance increased slightly across years, examining performance within each measurement burst (i.e., the seven sessions for a given year in Figure 1A) shows large increases from the first to the last session, suggesting strong retest effects. To quantify these retest effects, we simultaneously modeled age related change and the accumulation of task experience. Several existing models have been applied to the accumulation of retest effects in multi-session studies (e.g., Anderson, Fincham, & Douglass, 1999; Sliwinski et al., 2010). We selected the Anderson et al. (1999) model because it includes a single term that allows retest effects to accumulate when sessions are close together in time (i.e., within a measurement burst) and then dissipate when there are long gaps between sessions (i.e., in the months between measurement bursts).

In our adaptation of this model, memory performance on day i ($i = 1$ for the first session), denoted by p_i, is a function of both the linear effects of age-related episodic memory change and the power-law effects of test experience:

$$p_i = \beta_0 + \beta_{age}(Age) + \left(\beta_{retest} - \frac{\beta_{retest}}{\sum_{j=1}^{i} t_j^{-d}}\right) + \varepsilon_i. \tag{1}$$

In the model, β_0 is an intercept which represents the participant's performance in the absence of any age-related change or test experience. β_{age} is the amount by which performance changes daily as a result of aging. Performance on day i improves as a result of previous test experience up to a maximum retest benefit of β_{retest}. However, benefit from a session on any previous day, j, dissipates as the amount of time separating days j and i increases, with the exact benefit given by t_j^{-d}, where $t = 1 + i - j$ (i.e., how far back in time day j is), and d modulates the rate at which retest effects dissipate with the passage of time. t_j^{-d} is calculated for the session on day i and all previous sessions and then summed — the larger the sum, the closer the actual retest effect is to the maximum of β_{retest}. To summarize the determinants of the total retest effect, it increases as the number of previous sessions increases, it decreases as the amount of time separating previous sessions from day i increases, and it decreases as the value of the d parameter increases. Finally, an error term, ε_i, captures the deviation of the model from the data.

We fit the model separately to the free recall performance of each individual participant by minimizing the χ^2 difference value between the model predictions and observed data using the equation $\chi^2 = \sum_{i=1}^{n}(\frac{p_i - \hat{p}_i}{SE_{\overline{p}}})$, where n is the total number of sessions completed by the participant, p_i the actual performance on day i, and $\hat{p}_i$ is the model's prediction for day i. To minimize χ^2, for each participant we first ran a grid search by selecting 120 values for each of the four model parameters (evenly spaced between 0 – 1 for β_0, −.025 – .025 change in percent recall per year for β_{age}, $-.5 - .5$ for β_{retest}, and .1 – 1.0 for d). We then evaluated the parameter sets defined by the intersections of the grid, for a total of 120^4 parameter sets. Then for each of the 1000 best fitting sets from the grid search, we used the Interior Point method to find the local minimum and took the best of these local minima as the overall best fitting parameter set.

Each participant's best fitting parameter values were used to derive model-predicted performance across sessions. These predictions (averaged across participants) are shown by the black lines in Figure 1A. The means of the best fitting parameter values are shown in Table 1.

To determine the extent to which age and retest effects influence performance, we directly compared the model predictions to the across-session slope observed in the raw data (Figure 1B). To do so, we used the model fits to statistically isolate retesting effects on the one hand and aging effects on the other hand by using one component of the model at a time (the age component or the practice component) to predict performance. To isolate retest effects for each participant, we used their fitted values of the intercept, β_0, and the retest-related parameters β_{retest} and d to compute the component of performance, $\hat{p}_i^{retest}$, that can be predicted by test experience alone:

$$\hat{p}_i^{retest} = \beta_0 + \left(\beta_{retest} - \frac{\beta_{retest}}{\sum_{j=1}^{i} t_j^{-d}}\right). \quad (2)$$

To provide a comparison with the raw slope across sessions (which reflects retest effects and age effects), we computed a slope across sessions for the $\hat{p}_i^{retest}$ values predicted from retest effects alone. This slope, shown in Figure 1B is positive with 95% confidence intervals far above zero, suggesting that practice effects contribute to the positive slope in the raw data.

Similarly, to isolate the age effect for each participant, we used their fitted values of the intercept β_0 and the age parameter β_{age} to compute the component of performance, $\hat{p}_i^{age}$, that can be predicted by age alone:

$$\hat{p}_i^{age} = \beta_0 + \beta_{age}(Age). \tag{3}$$

We then computed a slope across sessions for the $\hat{p}_i^{age}$ values predicted from age alone, which is shown in Figure 1B. This age effect slope is not different than zero (the 95% confidence interval extends well below zero) and is significantly lower than the $\hat{p}_i^{retest}$ slope, $t(7) = -6.48$, $p < .01$. These results confirm that positive retest effects were obscuring age-related stability.

As a test of the model's ability to discriminate practice and age effects (and to show the replicability of the main findings), we collected a second sample of data — from participants who received less test experience but had aged by the same amount. Whereas the original sample received seven sessions a year for 5 years, the practice-control sample completed seven sessions in year 1 but no further sessions until year 5. If the model is truly able to remove retest effects, providing a purer measure of age effects, then model estimates from the two samples should reveal different practice effects but equal age effects.

Figure 2 shows the results from the practice-control group. As seen in Figure 2A, little sign of decline between bursts is observed. Figure 2B shows that the slightly negative raw slope across sessions disguises a marginally significant positive retest effect (the 95% confidence interval is slightly above zero) and a non-significant age effect. Supporting the ability of the model to distinguish practice from aging, the retest effect in this practice-control sample was significantly smaller than the retest effect in the yearly-testing sample, $t(12) = -3.59$, $p < .01$, but the age effects in the two samples did not differ, $t(12) = -0.01$, $p = .99$.

Discussion

Precisely measuring within-individual age-related change requires a longitudinal design. But the repeated testing inherent in traditional longitudinal designs tends to increase performance such that the rate of age-related decline will be underestimated unless retest effects are taken into account (Salthouse, 2015, 2016). This retest problem is exacerbated if the construct of interest requires intensive testing to be reliably measured.

We attempted to overcome this problem in a study of episodic memory by using a measurement burst longitudinal design and applying a joint model of retest and age effects, as suggested by Sliwinski et al. (2010). The raw data showed a modest but non-significant increase in memory performance over the five-years of the study. But applying our model revealed significant and substantial retest effects. Indeed, once the retest effect was statistically removed, we found a slight (but non-significant) age-related decline in memory ability over five years, consistent with the results of traditional longitudinal studies (Salthouse, 2015, 2016). This finding of substantial practice effects and small age-related change was replicated in a second sample. Moreover, the model was also able to accurately detect that the second sample had received less test experience despite having aged by the same amount.

This result demonstrates that longitudinal research need not be limited to projects that follow hundreds of participants for decades. It is possible to conduct studies at a more practical scale, both in terms of sample size and number of years, provided one combines an intensive measurement burst design with a model of retest effects. The ability to conduct smaller longitudinal studies allows for designs that efficiently target specific research questions that have traditionally been the domain of cross-sectional work. Here, we applied the method to memory ability, and Munoz et al. (2015) applied a similar method to reaction time data. The method could easily be adapted to other research domains such as age-related change in social or personality factors and even neural measurements.

Funding

This work was supported by the National Institute on Aging at the National Institutes of Health (grant number AG048233) and the National Institute of Mental Health at the National Institutes of Health (grant number MH55687).

Acknowledgments

We thank Ada Aka, Elizabeth Crutchley, Patrick Crutchley, Kylie Hower, Joel Kuhn, Jonathan Miller, Logan O’Sullivan, and Isaac Pedisich for assistance conducting the study. All data collected as part of this study can be accessed by visiting memory.psych.upenn.edu.

References

Anderson, J., Fincham, J., & Douglass, S. (1999). Practice and retention: A unifying analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(5), 1120–1136.

Baltes, P. B. (1968). Longitudinal and cross-sectional sequences in the study of age and generation effects. *Human Development*, *11*(3), 145–171.

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*, 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.

Hoffman, L., Hofer, S. M., & Sliwinski, M. J. (2011). On the confounds among retest gains and age-cohort differences in the estimation of within-person change in longitudinal studies: A simulation study. *Psychology and Aging*, *26*(4), 778.

Lohnas, L. J., & Kahana, M. J. (2013). Parametric effects of word frequency effect 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

Miller, J. F., Kahana, M. J., & Weidemann, C. T. (2012). Recall termination in free recall. *Memory & Cognition*, *40*(4), 540–550. doi: 10.3758/s13421-011-0178-9

Miller, M. A., & Rahe, R. H. (1997). Life changes scaling for the 1990s. *Journal of Psychosomatic Research*, *43*(3), 279–292.

Munoz, E., Sliwinski, M. J., Scott, S. B., & Hofer, S. (2015). Global perceived stress predicts cognitive change among older adults. *Psychology and Aging*, *30*(3), 487.

Nesselroade, J. R. (1991). The warp and the woof of the developmental fabric. In R. Downs & L. Liben (Eds.), *Visions of aesthetics, the environment, and development: The legacy of Joachim F. Wohwill* (pp. 213–240). Hillsdale, N. J.: Erlbaum.

Salthouse, T. A. (2015). Test experience effects in longitudinal comparisons of adult cognitive functioning. *Developmental Psychology*, *51*(9), 1262.

Salthouse, T. A. (2016). Aging cognition unconfounded by prior test experience. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *71*(1), 49. doi: 10.1093/geronb/gbu063

Sliwinski, M. J. (2008). Measurement-burst designs for social health research. *Social and Personality Psychology Compass*, *2*(1), 245–261.

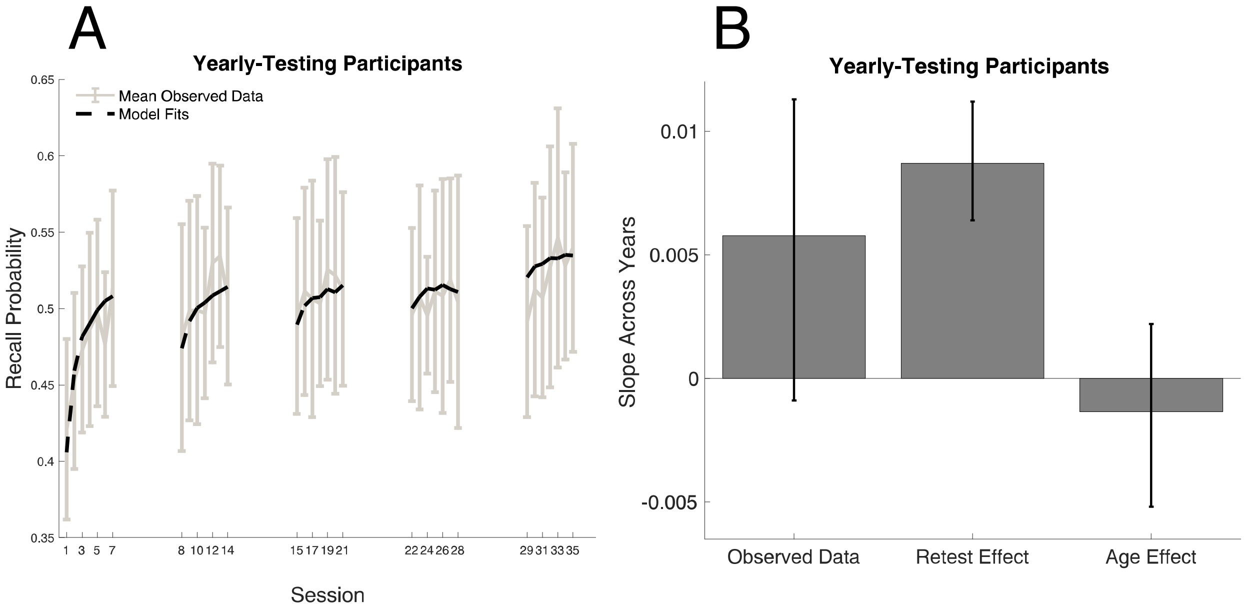
Sliwinski, M. J., Hoffman, L., & Hofer, S. (2010). Modeling retest and aging effects in a measurement burst design. In P. Molenaar & K. M. Newel (Eds.), *Individual pathways of change in learning and development* (p. 37-50). Washington, D.C.: American Psychological Association.

Table 1

Mean (standard deviation) of the fitted parameter values for each group

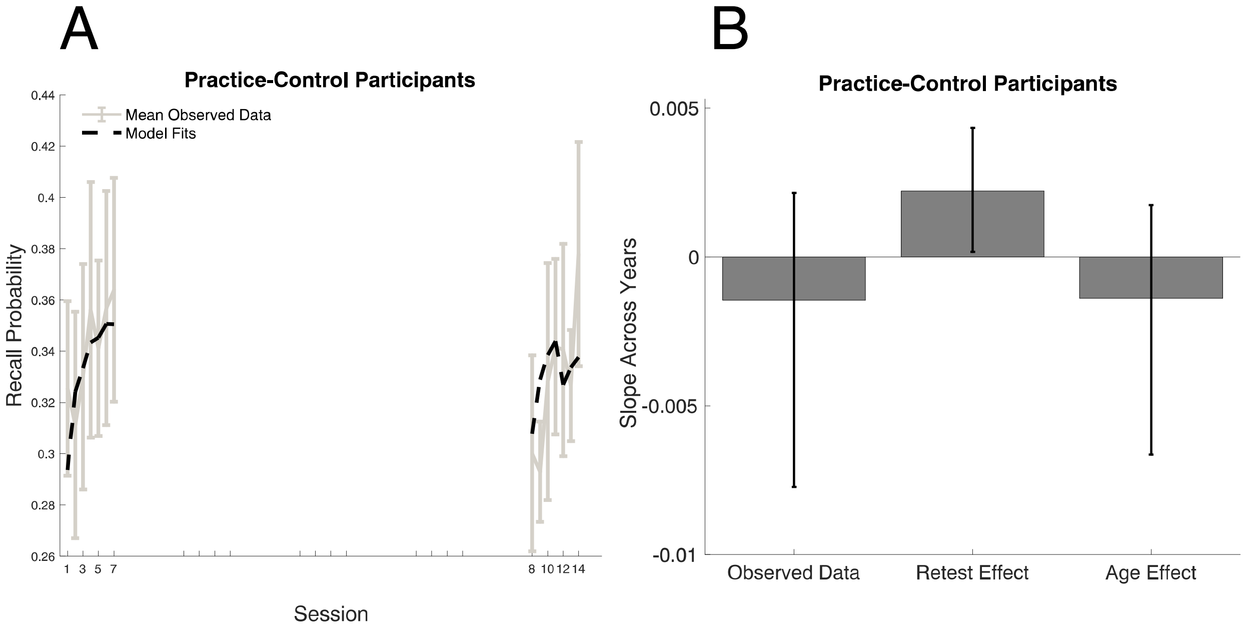
Parameter	Yearly Group	Control Group
β_0	.51 (.39)	.38 (.36)
β_{age}	−.0014 (.0055)	−.0014 (.0058)
β_{retest}	.14 (.05)	.09 (.10)
d	.35 (.22)	.46 (.22)

Figure 1



Yearly-testing Sample. A) Mean observed performance by session (gray) along with mean model fits (black) across the 5 years of the study. *N* = 8 for years 1 – 4. *N* = 6 for year 5. B) Slopes reflecting change per year in observed free recall performance, model-estimated retest effects, and model-estimated aging effects. All error bars are 95% bootstrapped confidence intervals.

Figure 2



Practice-Control Sample. A) Mean observed performance by session (gray) along with mean model fits (black) across the 5 years of the study. *N* = 6 for years 1 and 5. B) Slopes reflecting change per year in observed free recall performance, model-estimated retest effects, and model-estimated aging effects. All error bars are 95% bootstrapped confidence intervals.

Supplemental Materials for Modeling Retest Effects in a Longitudinal Measurement Burst Design Study of Episodic Memory

Adam W. Broitman
University of Pennsylvania; Philadelphia, Pennsylvania

Michael J. Kahana
University of Pennsylvania; Philadelphia, Pennsylvania

M. Karl Healey
Michigan State University; East Lansing, Michigan

Supplemental Materials for Modeling Retest Effects in a Longitudinal Measurement Burst Design Study of Episodic Memory

Additional Details on Methods of The Penn Electrophysiology of Encoding and Retrieval Study

The current analyses focus on the behavioral data from older adults in the Penn Electrophysiology of Encoding and Retrieval Study (PEERS, Healey, Crutchley, & Kahana, 2014; Healey & Kahana, 2014, 2016; Lohnas & Kahana, 2013, 2014; Miller, Kahana, & Weidemann, 2012) study.

Each year of the study consisted of 7 sessions, each of which included 16 free recall lists followed by 16 lists of recognition. For each recall list, 16 words were presented one at a time on a computer screen, followed by an immediate free recall test. The first session and half of the remaining sessions were randomly chosen to include a final free recall test before recognition, in which participants recalled words from any of the lists from the session. Each word was accompanied by a cue to perform one of two judgment tasks (“Will this item fit into a shoebox?” or “Does this word refer to something living or not living?”) or no encoding task. The current task was indicated by the color and typeface of the presented item. There were three conditions: no-task lists (participants did not have to perform judgments with the presented items), single-task lists (all items were presented with the same task), and task-shift lists (items were presented with either task). The first two lists were task-shift lists, and each list started with a different task. The next fourteen lists contained four no-task lists, six single-task lists (three of each of the task), and four task-shift lists. List and task order were counterbalanced across sessions and participants. The present analyses do not include performance data from the encoding task.

Each stimulus was drawn from a pool of 1,638 words. Lists were constructed such that varying degrees of semantic relatedness occurred at both adjacent and distant serial positions. Semantic relatedness was determined using the Word Association Space (WAS) model described by Steyvers, Shiffrin, and Nelson (2004). WAS similarity values were used to group words into four similarity bins (high similarity: $\cos\theta$ between words > .7; medium-high similarity, .4 < $\cos\theta$ < .7; medium-low similarity, .14 < $\cos\theta$ < .4; low similarity, $\cos\theta$ < .14). Two pairs of items from each of the four groups were arranged such that one pair occurred at adjacent serial positions and the other pair was separated by at least two other items. For each list, there was a 1500 ms delay before the first word appeared on the screen. Each item was on the screen for 3000 ms, followed by a jittered (i.e., variable) inter-stimulus interval of 800 – 1200 ms (uniform distribution). If the word was associated with a task, participants indicated their response via a keypress. After the last item in the list, there was a jittered delay of 1200 – 1400 ms, after which a tone sounded, a row of asterisks appeared, and the participant was given 75 seconds to attempt to recall aloud any of the just-presented items. If a session was selected for final free recall, then following the immediate free recall test from the last list, participants had 5 minutes to recall any item from the preceding lists. Final free recall data are not analyzed here.

A recognition test was administered following the free recall portion of the experiment. In this final recognition test, lures were selected from the remaining items not presented during the free recall phase, and target/lure ratio varied with session, where targets made up 80%, 75%, 62.5%, or 50% of the total items. In total, 320 words were presented one at a time on the computer screen. When a word was presented on the screen, participants were instructed to indicate whether the test word had been presented previously. Participants were told to respond verbally “pess” for old items and “po” for new items and to confirm their response by pressing the space bar. These responses ("pess" and "po") were chosen so that both response types would initiate with the same stop consonant (or plosive), thus assisting in automated detection of word onset times. Following the old-new judgment, participants made a confidence rating on a scale of 1 to 5, with 5 being the most confident. Although recognition was self-paced, participants were encouraged to respond as quickly as possible without sacrificing accuracy. Participants were given feedback on accuracy and reaction time.

References

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*, 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.

Lohnas, L. J., & Kahana, M. J. (2013). Parametric effects of word frequency effect 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

Miller, J. F., Kahana, M. J., & Weidemann, C. T. (2012). Recall termination in free recall. *Memory & Cognition*, *40*(4), 540–550. doi: 10.3758/s13421-011-0178-9

Steyvers, M., Shiffrin, R. M., & Nelson, D. L. (2004). *Word association spaces for predicting semantic similarity effects in episodic memory*. In A. F. Healy (Ed.), Cognitive psychology and its applications: Festschrift in honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer. Washington, D.C.: American Psychological Association.