Modeling Retest Effects in a Longitudinal Measurement Burst Study of Memory

Adam W. Broitman

Cornell University

Michael J. Kahana

University of Pennsylvania

M. Karl Healey

Michigan State University

Author Note

Correspondence should be addressed to Karl Healey at khealey@msu.edu.

Abstract

Background. 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 show that a measurement burst design combined with a model of retest effects (Sliwinski, Hoffman, & Hofer, 2010) can be used to study age-related change with modest sample sizes.

Methods. A combined model of age-related change and retest-related effects was applied to data from a measurement burst study in which eight subjects completed a burst of seven sessions of free recall every year for 5 years. Six additional subjects completed a burst only in years 1 and 5, and should, therefore, have a smaller retest effect but equal age effects.

In a simulation experiment, we show that with sample sizes as small as $n = 8$, the model can reliably detect age effects of the size reported in the longitudinal literature.

Results. The raw data suggested slight improvement in memory over 5 years. However, applying 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.

Conclusion. 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 Study of Memory

Introduction

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 and retest effects in typical longitudinal designs where each outcome variable is measured once per subject at each wave of the study (e.g., Nilsson, 2003; Salthouse, 2016). This typical longitudinal design is not appropriate, however, when the outcome variable of interest cannot be reliably assessed with a single measurement from each subject. For example, episodic memory performance is notoriously variable within a single individual due to endogenous fluctuations over time in the processes that support memory function (Kahana, Aggarwal, & Phan, in press), and therefore a single measurement does not provide an accurate assessment of a subject's ability. This within-subject variability can be overcome by collecting multiple measurements from each subject spread across several days of testing sessions.

In our cross-sectional work on age-related memory impairment (Healey & Kahana, 2016), we have taken exactly this multi-session approach by having subjects complete 112 lists of the free recall task spread over 7 sessions. Extending this multi-trial 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) with successive bursts being 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 et al. (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). Our goal is to establish whether a burst design combined with a model of retest effects can be used to study age-related change in a multi-session episodic memory study with modest sample sizes.

We begin by fitting a model to the initial results of a measurement burst longitudinal study in which eight subjects completed seven sessions of the free recall task each year for four–five years. Next, we report a series of simulations which show that the model provides over 90% power to detect realistically sized age effects with sample sizes as small as $n = 8$. Finally, we apply the model to a second group of subjects who received less task experience (only two bursts of free recall) but had aged by the same amount. The results show that the model is sensitive to differences in level of retest experience.

Method

The data are from the Penn Electrophysiology of Encoding and Retrieval Study (PEERS, Healey, Crutchley, & Kahana, 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.

Subjects

Original cross-sectional PEERS sample. The full PEERS older adult sample includes 39 individuals who completed an initial cross-sectional study (Healey & Kahana, 2016). All subjects were recruited from the Philadelphia area. Potential subjects were excluded if they suffered from any medical conditions, or regularly took medications, that might affect cognitive performance.

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

Practice-Control Sample. During the 5^{th} year of data collection we recruited six additional older adults from the original sample to return for a 5-year follow up, allowing us to measure performance in subjects who were less well practiced. Subjects were selected for enrollment based on their availability to return for additional testing. Although subjects were not randomly assigned to the yearly-testing and practice-control samples from the outset of the study, this control sample still provides a useful comparison. These Practice-control subjects ranged from 62 to 79 years ($M = 66.83$) at the start of the experiment, and they completed each burst in 1.1–6.3 weeks ($M = 3.7$).

PEERS Experiment

Each measurement burst was comprised of 7 sessions of the free recall task. At the beginning of each burst, the Recent Life Changes Questionnaire (M. A. Miller & Rahe, 1997) was administered to collect information about any potential changes in each subject's health or personal lives. No subjects 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 a 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 subject was then given 75 seconds to recall aloud any of the just-presented items. Trained experimenters scored recall accuracy from audio recordings of subject's recalls.

Results

Behavioral Results: Measurement Burst Study

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 subject 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 subject. Figure 1B shows that the average slope was 0.0058 (i.e., on a 0 to 1.0 scale, performance increased by 0.0058 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.

Model-Based Analysis of Age and Retest Effects

Several existing models have been applied to quantify the accumulation of retest effects in multi-session studies, such as those described in Anderson, Fincham, and Douglass (1999) and in Sliwinski et al. (2010). Both of these models provided good fits and similar results when applied to our data set during preliminary analyses. 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 subject'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 subject by minimizing the χ^2 difference value between the model predictions and observed data using the equation $\chi^2 = \Sigma_{i=1}^{n} \left(\frac{p_i - \hat{p}_i}{SE_{\bar{p}_i}}\right)^2$, where n is the total number of sessions completed by the subject, p_i the actual performance on day i, $\hat{p}_i$ is the model's prediction for day i, and $SE_{\bar{p}_i}$ is the standard error of p_i calculated across the lists of day i. To minimize χ^2, for each subject 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, -0.025– 0.025 change in percent recall per year for β_{age}, -0.5–0.5 for β_{retest}, and 0.1–1.0 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 subject's best fitting parameter values were used to derive model-predicted performance across sessions. These predictions (averaged across subjects) 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 retest 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 a subject, 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:

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

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 subject, 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.

A null age effect combined with a small sample size naturally raises concerns about statistical power. In the next section we report a series of analyses that measure the power and type I error rate of our model-based analysis.

Establishing Power and Type I Error Rate

The aim of this simulation experiment was to determine whether the model-based analysis we developed above can detect realistic levels of age related-change with small sample sizes. To do so, we created simulated datasets with known levels of age-related change, retest effects, and noise and then tested the model's ability to detect the age effects given different sample sizes.

To set a realistic level of age-related change in our simulations, we used the data from the Betula project (Nilsson et al., 1997), which reported that the mean age-related change for adults over 60 in was $-.0375$ SD units per year. We translated this value into a change in free recall performance by multiplying $-.0375$ by the standard deviation of recall performance for all 39 older adult subjects who completed the original cross-sectional sample ($SD = .0872$). This produced a β_{age} coefficient of -0.00327, meaning that a normally aging subject who recalls 40% of the study items in a session of free recall can be expected to recall 38.37% of the items in a similar free recall test after 5 years, assuming there are no practice effects. We created two other levels of simulated age effect: a "high" condition where β_{age} 130% of the Betula project mean, and a "no effect" condition where β_{age} was set to zero (i.e., to test the false positive rate of the model).

For each level of age-related change we created three different data sets with varying numbers of simulated subjects: one with $n = 4$, one with $n = 8$, and one with $n = 12$. Each data set consisted of simulated subjects who completed 45 sessions of free recall over 5 time waves (9 sessions per year). Each simulated subject was assigned a testing date vector in which the distance between time waves (400 days), as well as the distance between sessions within each time wave (5 days), were set to the mean values observed in the PEERS data reported above. Baseline memory performance and practice accumulation effects were generated using the mean β_0, β_{retest}, and d parameter values reported for the Yearly Group in Table 1. To add realistic levels of noise to the simulated data, a random perturbation was added to each simulated data point. This noise was set to have a mean and standard deviation equal to the difference between each observation in our data set and each data point created by the optimized parameters.

We fit the model to each simulated subject by minimizing the Root-Mean Squared Deviation (RMSD) between the model predictions and the observed data (we could not use χ^2 because whereas for actual subjects we can calculate $SE_{\bar{p_i}}$ across lists in a session, the model provides a single p_i for each session, preventing us from estimating $SE_{\bar{p_i}}$). To make the total run-time of the simulations tractable, we used MATLAB's GlobalSearch minimization algorithm rather than the more costly grid search technique employed for the actual data. To determine if the model detected the presence of an age effect, we tested whether the across-subject distribution of recovered β_{age} parameter values was significantly different from zero using a one-tailed t-test with $\alpha = 0.05$.

We repeated this entire procedure 100 times for each combination of simulated age effect (Hi, Medium, Zero) and sample size (4, 8, 12). Thus we can estimate power as the number of times, out of 100 simulations, an age effect was detected, via the t-test, in the Hi and Medium conditions, and we can estimate the type I error rate as the number of false positives, out of 100, in the zero effect conditions.

Figure 2 shows that power was over 90% for sample sizes of 8 and 12, but less than 50% for sample sizes of 4. Across all sample sizes in the zero effect condition the number of significant results was 4.7%, indicating that the Type I error rate was successfully set to approximately $\alpha = 0.05$. These simulations show that a sample size as small as 8 provides sufficient power to detect age effects of the size reported in the literature.

Behavioral Results: Replicating Age-Related Stability

As a final 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 subjects who received *less* test experience but had aged by the same amount. Whereas the original sample completed 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 3 shows the results from the practice-control group. The raw slope across years was slightly negative, but this disguises a 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 ($t(12) = -.01$, *n.s.*).

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 (Nilsson, 2003; 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 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 some 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. A series of simulations revealed that rates of age-related memory change comparable to those reported in the literature can be detected with 90% power in samples as small as $n = 8$.

This result demonstrates that longitudinal research need not be limited to projects that follow hundreds of subjects for decades. It is possible to conduct longitudinal studies with smaller samples for shorter periods of time, 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. doi: 10.1037/rev0000015

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.

Kahana, M. J., Aggarwal, E., & Phan, T. D. (in press). The variability puzzle in human memory. *Journal of Experimental Psychology: Learning, Memory and Cognition*.

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.

Nilsson, L.-G. (2003). Memory function in normal aging. *Acta Neurologica Scandinavica*, *107*, 7–13.

Nilsson, L.-G., BÄCkman, L., Erngrund, K., Nyberg, L., Adolfsson, R., Bucht, G., ... Winblad, B. (1997). The betula prospective cohort study: Memory, health, and aging. *Aging, Neuropsychology, and Cognition*, *4*(1), 1–32.

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. Retrieved from `+http://dx.doi.org/10.1093/geronb/gbu063` 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

β_0	.51 (.39)	.38 (.36)
β_{age}	-0.0014 (0.0055)	-0.0014 (0.0058)
β_{retest}	.14 (.05)	.09 (.10)
d	.35 (.22)	.46 (.22)

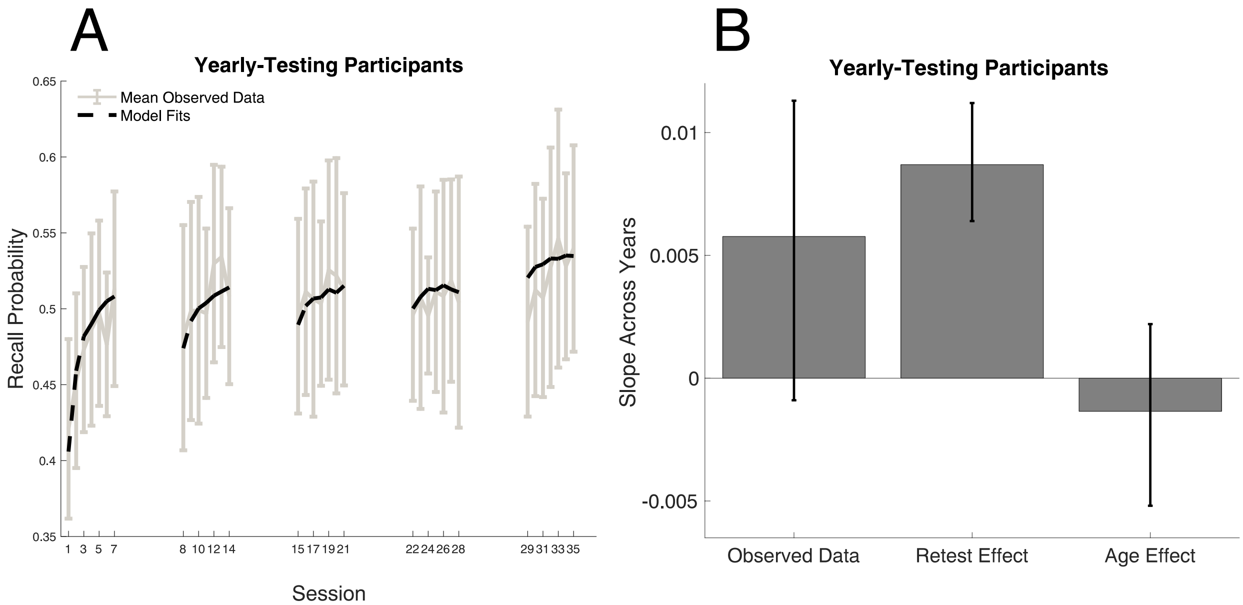

Figure 1. Yearly-testing Sample. A) Mean observed performance by session (gray) along with mean model fits (black) across the five years of the study. $N = 8$ for years $1 - 4$. $N = 6$ for year 5. B) Slopes reflecting change per year in observed free recall performance, model-estimated practice effects,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

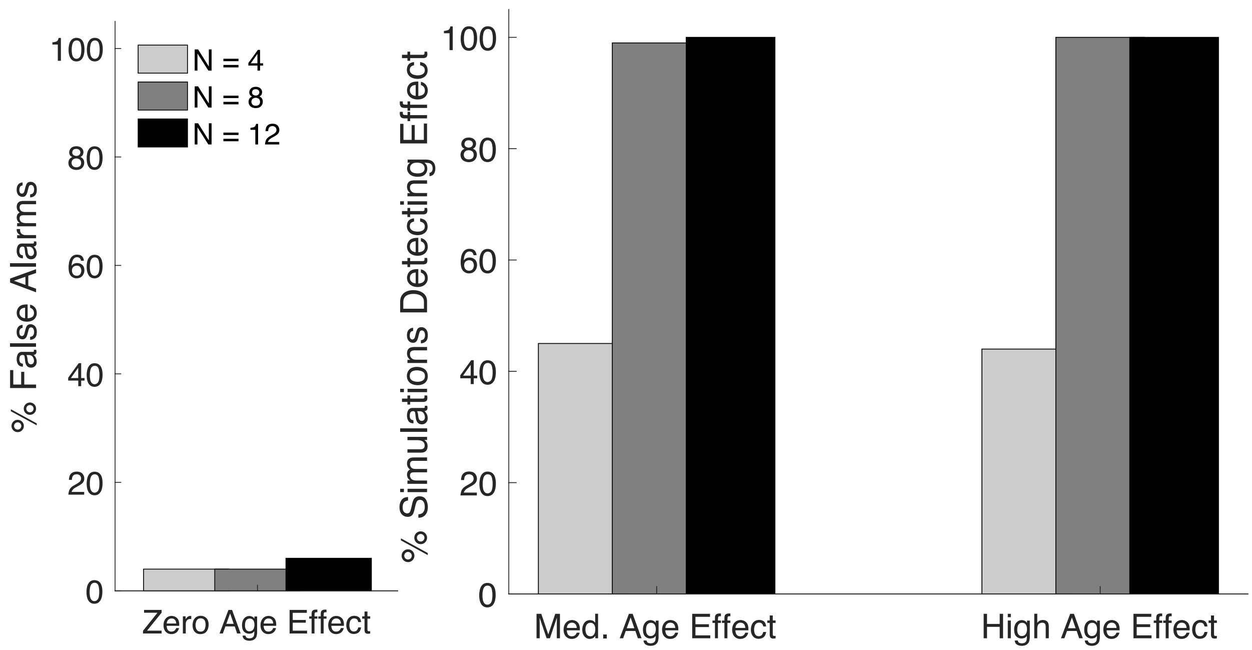

Figure 2. Percentage of simulated data sets showing significant aging effects as a function of sample size and the true degree of age-related memory decline in the simulated data.

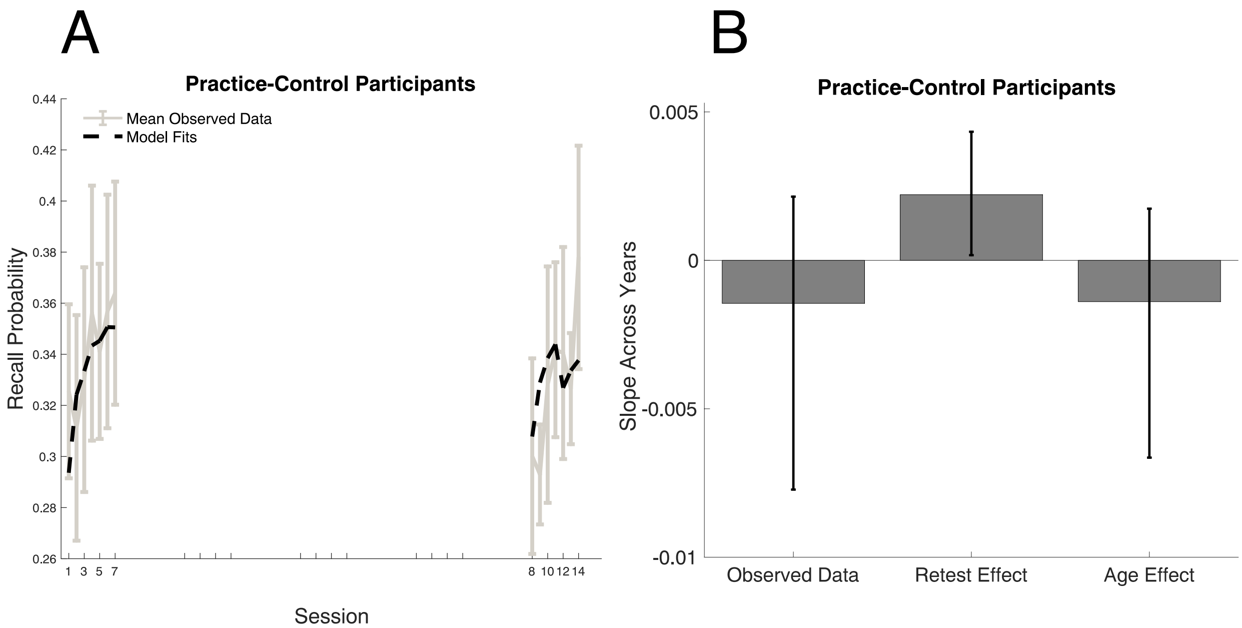

Figure 3. Practice-Control Sample. A) Mean observed performance by session (gray) along with mean model fits (black) across the five years of the study. $N = 6$ for years 1 and 5. B) Slopes reflecting change per year in observed free recall performance, model-estimated practice effects, and model-estimated aging effects. All error bars are 95% bootstrapped confidence intervals.