Research Article

Psychological Science
2019, Vol. 30(1) 92–104

DOI: 10.1177/0956797618808474
www.psychologicalscience.org/PS
SAGE

Temporal Proximity Links Unrelated News Events in Memory

Mitchell G. Uitvlugt and M. Karl Healey
Department of Psychology, Michigan State University

Abstract
Some memories are linked such that recalling one can trigger the retrieval of another. What determines which memories are linked? Some models predict that simply occurring close together in time is sufficient for links to form between memories. A competing theory suggests that temporal proximity is generally not sufficient, and existing evidence for such links is an artifact of using chainlike lists of items in artificial laboratory tasks. To test these competing accounts, we asked subjects to recall news stories that they had encountered over the past 2 years (Experiment 1) or 4 months (Experiment 2). In both experiments, subjects showed a strong bias to successively recall stories that appeared in the news within days of each other—even after accounting for the fact that stories that occur close in time tend to be semantically related. By moving beyond laboratory tasks, this research solidifies the foundation of contemporary memory theory.

Keywords
episodic memory, free recall, temporal contiguity

Received 1/16/18; Revision accepted 8/25/18

Memory links our experiences in a complex network of associations: Recalling one event can cue us to retrieve other linked events. What determines which memories are linked and which are not? According to an influential class of human memory models (Howard, Shankar, Aue, & Criss, 2015; Lohnas, Polyn, & Kahana, 2015) with deep roots in the history of memory theory (Estes, 1955; McGeoch, 1932), simply occurring close in time should be sufficient to link events in memory. Under these models, experiencing an event activates a cognitive representation that persists after the event has ended. These echoes of past events then become associated with new events, even if they are not deeply connected by meaning. Thus, links between events that occur around the same time should be ubiquitous in memory.

The clearest evidence of these temporal links comes from experiments that simulate memory for events by having a subject study a list of sequentially presented but unrelated items (e.g., 15 nouns). When the subject successfully recalls one item from the list, that item tends to call to mind other items that were studied nearby in time (Cortis Mack, Cinel, Davies, Harding, & Ward, 2017; Kahana, 1996; Kintsch, 1970). This temporal-contiguity effect (TCE) has been very useful in evaluating theories of memory search (Cortis Mack, Dent, Kennett, & Ward, 2015; Davelaar, Goshen-Gottstein, Ashkenazi, Haarmann, & Usher, 2005; Farrell, 2012; Lehman & Malmberg, 2013; Unsworth, 2008) and is so robust in laboratory list-learning tasks that some researchers have claimed that the TCE is a universal principle of memory (Healey, Long, & Kahana, 2014, in press).

A competing perspective (Hintzman, 2011, 2016) points out that relying primarily on list learning to simulate real-world memory places claims of universality, and the validity of many modern memory models, on shaky ground. A list of items presented one after another has an obvious chainlike structure that could encourage subjects to adopt a deliberate strategy of memorizing, and then recalling, the list as a chain. Such ad hoc strategies would produce a TCE in list-learning tasks but would tell us nothing about how real-world experiences become linked in memory.

Corresponding Author:
M. Karl Healey, Michigan State University, Department of Psychology, 316 Physics Rd., East Lansing, MI 48824
E-mail: khealey@msu.edu

Some work has tested this strategy account of the TCE by removing the impetus to deliberately encode items. When subjects view a list of words under the guise of some judgment task (e.g., "Does the word refer to a living thing?") and later receive a surprise memory test, the TCE is reduced (Nairne, Cogdill, & Lehman, 2017) but not eliminated (Healey, 2018). But even incidental encoding studies use lists of simple stimuli presented in a chainlike structure. A more incisive test would be to look for a TCE when subjects search memory for actual events and information they experienced outside of the laboratory.

There is an extensive literature on the recall of autobiographical and semantic memories that provides some evidence that memories are temporally organized (Altmann, 2003; Friedman, 2004; Moreton & Ward, 2010; Roediger & Crowder, 1976; Rubin, 1982; Shum, 1998; Whitten & Leonard, 1981). The most direct test for a TCE in autobiographical recall comes from Moreton and Ward (2010), who asked subjects to freely recall life events from the past 5 weeks, months, or years. Regardless of the targeted time period, subjects showed a TCE. A limitation of this study is that it potentially confounds semantic similarity with temporal distance (Hintzman, 2016). For example, events that occur during temporally proximate periods of one's life are more likely to involve similar people and places than events that occur during more temporally distant periods. Thus, it remains unclear whether temporal associations influence memory search outside list-learning tasks.

To more directly address these issues, we tested for a TCE in the recall of memories that were formed outside the laboratory: newsworthy events. In Experiment 1, subjects were asked to recall events related to the 2016 U.S. presidential campaign, and we controlled for semantic similarity by using the judgments of independent raters. In Experiment 2, subjects were asked to recall any news events, political or otherwise, from the previous 4 months, and we controlled for semantic similarity using latent semantic analysis (LSA).

Experiment 1

Method

Subjects. Measures of the TCE in recall of news stories have never been reported in the literature; therefore, effect-size estimates are not available to aid in determining appropriate sample sizes. Instead, we sought to ensure adequate power to detect an effect much smaller than that observed in the laboratory. Sederberg, Miller, Howard, and Kahana (2010) reported a meta-analysis of the TCE in laboratory list-learning studies. Power calculations revealed that a sample size of 1,000 would provide a $1 - \beta$ power of .99 to detect (via a one-tailed, one-sample t test) an effect one tenth the size of the average effect they reported.

A total of 1,051 subjects participated online on Amazon's Mechanical Turk (MTurk) and were paid $1.00 each. After exclusion criteria (described in detail below) were applied to remove subjects who did not provide enough data to measure the TCE, the final sample size was 855. The average age of subjects was 35.77 years (SD = 11.52), 50.47% were female, and 96.61% reported living in the United States. When subjects were asked how closely they followed the current election, the average response on a scale from 1 (*not at all*) to 5 (*very closely*) was 4.14 (SD = 0.93).

Data were collected in two phases. The first phase occurred 2 days after Election Day (i.e., November 10, 2016; n = 459), and the second phase began a month later (i.e., December 20, 2016; n = 592). Identical procedures were followed for both phases. No differences were found between the two collection phases; therefore, the data were collapsed across phase.

Procedure. The experiment was composed of a recall task, followed by an ordering task, and ending with a demographic questionnaire.

Recall task. The recall task provided the data for our main measure of the TCE. Subjects were instructed to "try to recall as many election-related news stories as you can." Subjects were told to ignore stories that had appeared in the news after the election results were announced (although we included all recalled election-related stories, even if they appeared after Election Day). After subjects recalled a story, they were asked to "come up with a short, descriptive headline of the sort you would read in a newspaper. It doesn't have to be long, just something that makes it clear to us which news story you are thinking of." Headlines were typed individually into a text box and submitted by pressing the enter key. Subjects were given 7 min to submit as many headlines as they could recall, after which they progressed to an ordering task.

Ordering task. In the ordering task, subjects were shown all of the headlines they had written in the recall task and were asked to place them in the order in which they thought the headlines had originally appeared in the news. Headlines were presented in a drop-down menu, and subjects were to select the headline that occurred the earliest and then press enter. They then selected the second earliest headline, and so on. Each time a headline was selected, it was removed from the drop-down menu and appended to an ordered list that was visible below the drop-down menu. Subjects continued ordering the headlines until there were no headlines remaining.

Demographics questionnaire. After the ordering task, subjects answered a series of demographic questions about themselves (gender, age, level of education, where they live), their politics (who they voted for, political party, interest in politics, interest in this election, when they voted), and the news sources they used regularly. These data were not analyzed for this article.

Date assignment. To test for a TCE, we needed to estimate the original order in which subjects encountered the news stories they recalled. We did this by identifying the actual news stories corresponding to each subject-authored headline and finding the date on which it had originally appeared in the news (for a complete description of how dates were assigned from subject-authored headlines, see the Supplemental Material available online). We then placed the headlines on a timeline starting with the earliest story recalled by any subject: "HRC Has Conference About Emails and Claims Convenience and One Device." This story appeared on March 10, 2015. Therefore, March 10, 2015, was defined as Day 1 of the timeline. The next earliest story, "Hillary Clinton Announces She's Running for President," occurred 33 days later on April 12, 2015, and was therefore defined as Day 34. Headlines that were too vague to be assigned a unique date (e.g., poll results: "Hillary Is Pulling Ahead"), that included an opinion (e.g., "Donald Trump Will Help All Americans"), or that were unrelated to the election (e.g., "Cubs Win World Series") were deemed invalid and were not included in the analyses.

The most recent headline recalled by any subject, "Trump Won't Pursue Case Against Clinton," referenced a statement that Trump made on November 22, 2016, in which he reversed his campaign pledge to seek a new criminal investigation into Hillary Clinton. It occurred 623 days after the first headline—Day 624 on the timeline.

Quantifying temporal proximity and semantic similarity. The key unit of analysis is the transition: recalling one story and then moving on (transitioning) to recall another. As reported below, there was a total of 5,707 valid transitions across subjects. To measure the temporal proximity between the stories recalled in each transition, we calculated a *lag* as the difference, in days, between when the two stories had originally appeared in the news. For example, if a subject recalled one story that occurred on Day 86 and next recalled another story from Day 86, the lag would be 0 (86 – 86). If the next story that this subject recalled was from Day 50, the lag would be −36 (50 – 86). When lags were calculated, both stories had to be valid (i.e., have dates). Transitions that either originated from or ended at an invalid story were not included in the analyses.

As we will discuss, it is also useful to know the semantic similarity between the stories involved in a transition. Therefore, raters recruited on MTurk judged the similarity of the two stories involved in each transition on a scale from 1 to 10.

To begin, each rater in the first set (n = 1,235) rated the similarity of 40 transitions that were selected pseudorandomly from the 5,707 total transitions such that each transition would be rated by at least 4 independent raters. Because of this crowdsourcing approach, in which a large number of raters each rated a small number of transitions, we could not compute traditional interrater reliability scores: Any 2 raters had on average only 0.14 (SD = 0.01) headlines in common. Instead, we ensured quality ratings by replacing raters who performed the task too quickly (< 1 min) and who provided ratings that deviated from the across-rater distribution by more than 2.5 standard deviations (for full details, see the Supplemental Material). After these exclusions were made, some headline pairs did not have at least four ratings. Therefore, a second set of raters (n = 6) was required. These raters each rated 28 pairs of headlines. After all ratings had been collected, each transition had been rated by at least 4 raters (M = 5.28, range = 4–8). The semantic-similarity score for each transition was then averaged across raters.

Raters also were able to mark headlines as not election related rather than assessing a similarity score for the pair; we detail in the next section how we used this judgment to measure whether subjects followed task instructions.

Exclusion criteria. To ensure that subjects followed task instructions, we excluded those who produced a large number of stories that were not election related. For each subject, we counted how many of the stories were marked as unrelated to the election by at least two raters. If the count was 2.5 standard deviations or more above the across-subject average, the subject was excluded from analysis. This criterion led to the exclusion of 37 subjects. To ensure that each subject contributed a reasonable amount of data to the analyses, we also excluded any subject who made fewer than two transitions between stories that could be assigned a date. This criterion led to the exclusion of 159 subjects. Thus, the final sample size after all exclusions was 855 (1,051 – 37 – 159). As shown in the Supplemental Material, very similar results were found if all subjects were included.

Results

A total of 9,535 headlines was recalled—an average of 11.15 per subject (SD = 5.39). After eliminating invalid headlines (those that were not related to the election

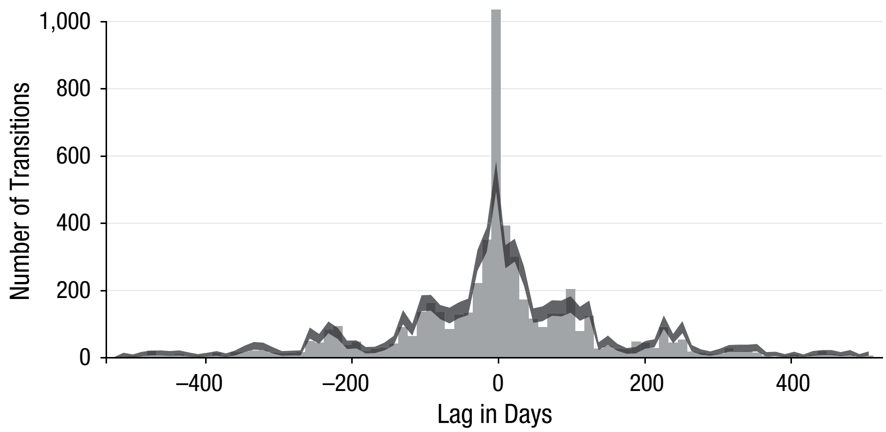

Fig. 1. Distribution of transition lags in Experiment 1. Subjects recalled as many news stories related to the 2016 U.S. presidential election campaign as possible, in whatever order they came to mind. Each time a subject recalled one story and transitioned to recalling another story, we defined the lag of the transition as the difference, in days, between when the two stories had originally appeared in the news. The light gray bars are a histogram showing the distribution of these lags across subjects, and the darker gray line shows the middle 95% of the distribution of lags from a null model in which transitions are random.

or could not be dated), we found that 7,579 headlines were recalled (M = 8.86 per subject, SD = 4.31). Across subjects, 5,707 transitions were made (M = 6.67 per subject, SD = 3.98).

If temporal contiguity influenced subjects' recall, then after recalling one story, the subject should transition to another that appeared in the news around the same time. To test this prediction, we calculated the number of days (i.e., the lag) separating the two stories involved in each of the transitions made by subjects. The longest possible lag was between Day 1 and Day 624; therefore, lags could range from −623 to +623 (positive lags indicate moving forward in time, and negative lags indicate moving backward). The histogram of lags in Figure 1 shows that transitions were not equally distributed across this range of possible lags. Instead, there was a clear peak at a lag of 0 days, and the frequency of transitions decreased steeply for larger absolute values of lag. That is, subjects showed a TCE.

But this does not necessarily mean that time, per se, forges the links between events. If a jar of marbles contains 80% red marbles and you grab a random handful, you will probably get mostly red marbles. Recall transitions could be like a jar of red marbles: If many stories cluster around particular days (e.g., the presidential debates) with fewer stories in between, there would be more ways to make short-lag transitions than to make long-lag transitions.

Figure 2 shows that such story clustering does indeed occur. To quantify the artificial contiguity effect that story clustering would produce on its own, we ran a simulation in which temporal order could not influence recall order. Each simulated subject recalled the same number of stories, k, as did an actual subject by randomly drawing k items from a probability distribution that matched the distribution in Figure 2. Critically, because each draw from the distribution is independent, all links between successive recalls are broken, and transition lags depend only on story clustering. Running this random-transition model 1,000 times for each simulated subject provided a distribution of lags expected by chance. Looking at this null distribution (the darker gray line in Fig. 1), we see that its peak is much lower than the peak of the actual data.

The difference between the actual and null distributions in Figure 1 is largest at short lags. To better visualize the difference, we zoomed in on these short lags by grouping lags into bins, using wider bins for longer lags (for details on how the bins were constructed, see the Supplemental Material). For each bin, we used the actual and null distributions to calculate a *temporal-bias score* on the basis of the number of times each subject actually made a transition falling into that bin and the number of times the subject would be expected to make such a transition under the null model:

$$\text{temporal-bias score} = \frac{\text{actual count} - \text{expected count}}{\text{expected count}}.$$

The solid line in Figure 3 shows the result: Zero-lag transitions occurred significantly more often than

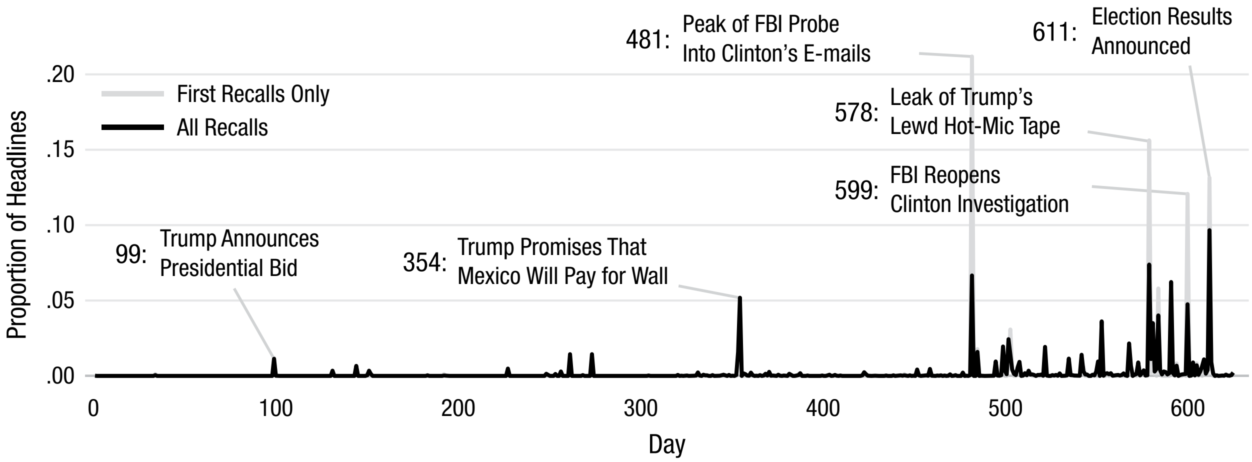

Fig. 2. Distribution of recalled stories across days in Experiment 1. For each recalled headline, we found the date on which it had originally appeared in the news. Day 1 was defined as the date of the earliest recalled stories. The darker bars represent all recalled stories. The lighter bars represent only the first story recalled by each subject. The height of the bars for a particular day shows what proportion of all included stories occurred on that day. Note that more than one unique event can be associated with each day; the examples included are the most frequently recalled event from each peak.

expected by chance. Near-lag transitions of 1 to 10 days were also more frequent than chance, and far-lag transitions were less frequent than chance. This pattern shows a bias for near lags—a TCE—which cannot be accounted for by clustering of stories around particular dates. These temporal-bias scores also reveal an asymmetry: Subjects were more likely to make transitions forward in time than backward in time: Temporal-bias scores were higher for the positive 1- to 10-day bin than the negative 1- to 10-day bin (mean difference = 0.32; 95% confidence interval, or CI = [0.11, 0.54]). This finding parallels the asymmetry effect found in laboratory list-learning tasks (Healey, Long, & Kahana, in press; Kahana, 1996).

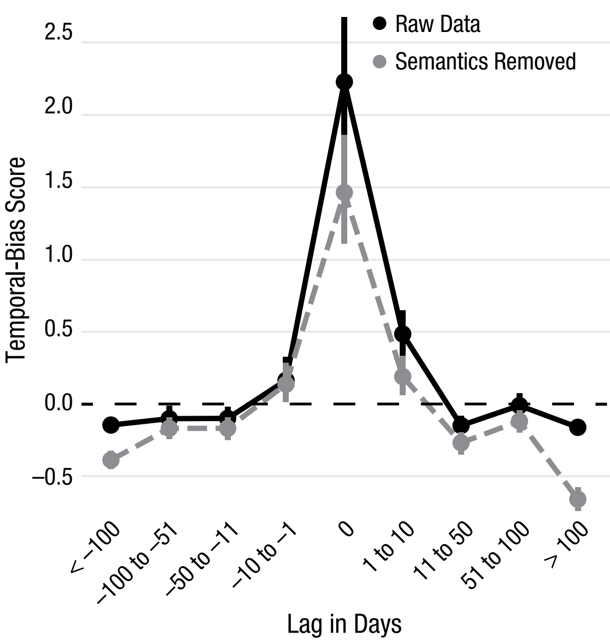

Fig. 3. Mean temporal-bias score as a function of lag in Experiment 1. For each lag bin, we counted the number of times that the subject actually made a transition corresponding to that bin as well as the number of times that such a transition would be expected by chance. We defined temporal bias as the difference between the actual counts and the expected counts expressed as a proportion of the expected counts. The solid line shows the raw data; the dotted line shows the residuals after removing the influence of semantic similarity between successively recalled stories. Error bars indicate 95% bootstrapped confidence intervals computed across subjects.

Although story clustering cannot account for the TCE, we must consider the possibility that semantic similarity can. Semantic similarity could produce a peak at near lags if news stories that occur near in time to one another tend to be semantically related such that one story reminds a subject of other stories that happened around the same time. This reminding process could operate both when the stories are originally experienced (i.e., semantically mediated study-phase retrieval; Hintzman, 2016) and during memory search (i.e., semantic-recall clustering; Howard & Kahana, 2002b). In either case, semantic similarity would give the illusion of a true TCE. For example, if several stories about e-mail servers appeared over the course of a few days, they may be recalled successively even if subjects ignored temporal proximity and relied only on semantic similarity. The only other study to test for the TCE in a real-world, non-list-learning task (Moreton & Ward, 2010) did not account for this confound.

Figure 4 shows that semantic-similarity scores (for details on these scores, see the Method section) were indeed higher for near-lag transitions. To determine whether semantic similarity can account for the TCE, we

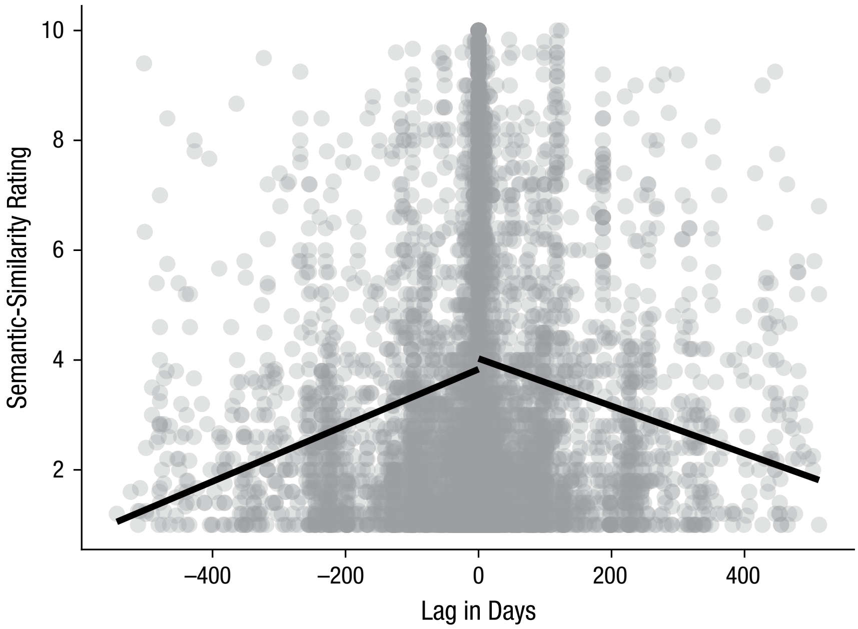

Fig. 4. Scatterplot showing the relationship between temporal lag and semantic similarity in Experiment 1. To help visualize the relationship, we added separate linear regression lines for negative lags and positive lags (lag 0 is included in both regression lines; excluding lag 0 from both yields shallower slopes).

statistically removed the effect of similarity from the binned temporal-bias scores in Figure 3. We did this by running a hierarchical linear regression in which temporal bias for each lag bin was predicted from the average semantic similarity for that lag bin (allowing the model to select a different intercept for each subject). Specifically, the regression equation was as follows:

$$y_{is} = \beta_{00} + \beta_1 x_{is} + b_{0s},$$

where y_{is} is the temporal-bias score for lag bin i of subject s, β_{00} is a common intercept applied to all subjects, β_1 is the slope, x_{is} is the average semantic similarity for lag bin i of subject s, and b_{0s} is a subject-specific intercept fitted for subject s.

The residuals from this regression measured the portion of the temporal-bias scores that could not be predicted by semantic similarity. As the dashed line in Figure 3 shows, the TCE remained strong even after we removed the effect of semantic similarity. Controlling for similarity did, however, eliminate the asymmetry between the positive 1- to 10-day bin and the negative 1- to 10-day bin (mean difference = 0.05; 95% CI = [–0.12, 0.23]). Some models attribute asymmetry to activity related to the meaning of one experience persisting over time and becoming associated with later experiences (e.g., Howard & Kahana, 2002a). Future work should test whether such models can account for the disappearance of asymmetry when semantic associations are statistically removed.

Discussion

Experiment 1 showed a TCE in the recall of real-world news stories encountered outside the laboratory. The unique contribution of this experiment was to test whether this effect could be due to the fact that stories occurring nearby in time tend to be semantically related (cf. Moreton & Ward, 2010). We did this by statistically controlling for ratings of similarity. One could argue that this analysis did not fully account for semantic similarity for at least two reasons. First, because subjects were asked to focus on a set of events (election stories) with strong semantic associations and recurring themes (e.g., Clinton's e-mails, Trump's wall), it is difficult to fully remove the effect of semantics. Second, having raters judge similarity on a 10-point scale may not capture all of the relevant aspects of semantic similarity. Therefore, in a second experiment, we asked subjects to recall any news story (political or otherwise) that had occurred over the past 4 months, and we quantified semantic similarity using the same LSA (Landauer & Dumais, 1997) techniques that are commonly used in the literature to study the interaction of temporal and semantic associations (Healey & Kahana, 2014; Howard & Kahana, 2002b).

Experiment 2

Method

Subjects. A total of 621 subjects participated online on MTurk and were paid \$1.00 each. Data were collected in one phase, beginning on May 2, 2018. After exclusion criteria (described in detail below) were applied, the final sample size was 561. The average age of included subjects was 35.12 years (*SD* = 10.35), 50.36% were female, and 100% reported living in the United States. When subjects were asked how closely they followed the news, the average response on a scale from 1 (*not at all*) to 5 (*very closely*) was 3.59 (*SD* = 1.02).

Procedure. The experiment was composed of a recall task, followed by a reference-lookup task and a demographic questionnaire.

Recall task. In the recall task, subjects were instructed to recall as many news stories as possible from 2018 (i.e., a 4-month period). Subjects were told to ignore stories that had appeared in the news prior to 2018. Once subjects recalled a story, they were to write a concise headline summarizing it. Headlines were typed individually into a text box and submitted by pressing the enter key. Subjects were given 3 min to submit as many headlines as they could recall, after which they progressed to the lookup task.

Lookup task. In the lookup task, we showed the subjects each headline they had written in the same order in which they had originally recalled them. For each, we asked them to search online and find a hyperlink to a story reporting this event from a credible news source (e.g., NPR, *The Wall Street Journal*). We asked them to find a source that was published as close as possible to when they learned about the event. Specifically, we stated, "We are interested in when *you* first encountered the story, so try to find a source **dated as close as possible to when you first learned about the story**" (italics and boldface were included in the instructions). Such subjective dating of memories has been shown to be quite accurate (Rubin & Baddeley, 1989). For each headline, we provided a link to a Google News search with the headline as the search text, to make the task easier and more uniform across subjects. Subjects were also asked to enter the date on which the story they selected had been published. These dates were used to calculate transition lags.

Demographics questionnaire. After the lookup task, subjects answered a series of demographic questions about themselves (gender, age, level of education, where they live) and how closely they follow the news in general.

Quantifying semantic similarity. When studying the influence of semantic similarity in free recall, researchers have typically used LSA or highly related techniques (e.g., word-association spaces; Steyvers, Shiffrin, & Nelson, 2005) to measure semantic similarity (Howard & Kahana, 2002b; Morton & Polyn, 2016; Polyn, Norman, & Kahana, 2009). LSA can represent any text as a vector in a high-dimensional space. The space is constructed by taking a corpus of written material, dividing it into documents (e.g., each paragraph can be considered a document), counting how many times each unique word occurs in each document, and subjecting this word-by-document matrix of count values to singular value decomposition (SVD). The SVD solution can then be used to convert any text into a vector representation. The similarity of two texts can be measured as the cosine of the angle separating their respective vectors in this space.

In free recall, the corpus used to create the representational space is typically common college-level reading material, and the texts that are converted into vector representations are individual words from the study lists. To measure the similarity of news stories, we used the text of all 3,656 web pages provided by subjects as the corpus (i.e., each web page is a document). After the LSA algorithm (described in detail in the Supplemental Material) is run, each news story is represented as a vector, and the similarity between different stories can be measured as the cosine of the angle between their vectors, $\cos(\theta)$.

Following the literature on free recall, we divided the distribution of similarity values into bins. This was necessary because, unlike temporal lag, which is a discrete variable and can take on only a small number of unique values, semantic similarity, measured as $\cos(\theta)$, is continuous. To ensure that we had a reasonable number of transitions per bin, we divided the distribution of semantic-similarity values across all transitions into five bins based on quintiles. The upper edges of these bins were as follows: $\cos(\theta) = .013$ (low similarity), $\cos(\theta) = .028$, $\cos(\theta) = .044$ (medium similarity), $\cos(\theta) = .071$, and $\cos(\theta) = 1.000$ (high similarity).

Exclusion criteria. To ensure that each subject contributed a reasonable amount of data to the analyses, we excluded any subject who made fewer than two transitions between valid stories. This criterion led to the exclusion of 60 subjects. Thus, the final sample size after all exclusions was 561.

Results

Across subjects, 3,656 valid stories were recalled (*M* = 6.52 per subject, *SD* = 2.68), and an additional 61 stories were excluded because subjects failed to provide a

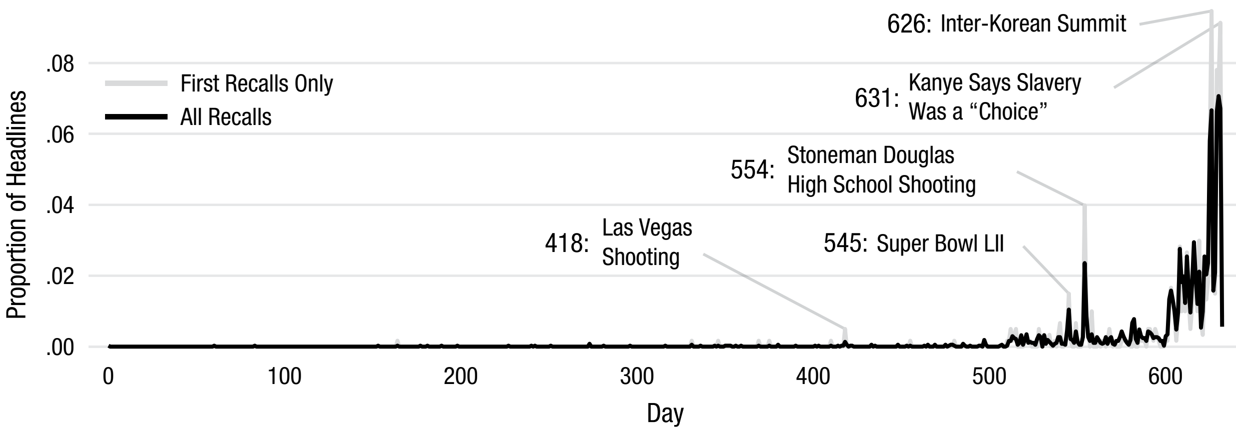

Fig. 5. Distribution of recalled stories across days in Experiment 2. For each recalled story, the subjects provided the date of a news article published as close as possible to when they learned of the event. Day 1 was defined as the date of the earliest recalled story. The darker bars represent all recalled stories. The lighter bars represent only the first story recalled by each subject. The height of the bars for a particular day shows what proportion of all included stories occurred on that day. Note that more than one unique event can be associated with each day; the examples included are the most frequently recalled event from each peak.

valid URL to the story. Figure 5 shows the distribution of these headlines across days—again, there was a clear recency effect. Overall, 3,054 transitions were made (M = 5.44 per subject, SD = 2.67). As we did in Experiment 1, we calculated the lag in days separating the stories involved in each of these transitions and plotted the distribution of lags in a histogram (see Fig. 6). Replicating the key finding of Experiment 1, this distribution had a peak at a lag of 0, which was significantly higher than the peak of the null distribution expected by chance (with chance determined by running 10,000 simulations per subject of the same random transition model described for Experiment 1).

Figure 7 shows temporal-bias scores as a function of lag bin (for how these scores and bins were created, see Experiment 1). Results again replicated those of the first experiment: Subjects made zero-lag and near-lag transitions of 1 to 10 days more frequently than expected by chance but made far-lag transitions less frequently than expected by chance. That is, there was a TCE.

Although the main finding of a TCE was replicated, comparing Figure 7 with Figure 3 does reveal a few important differences between the two experiments. First, whereas the temporal-bias score peaked at more than 2.0 in Experiment 1, it peaked at slightly more

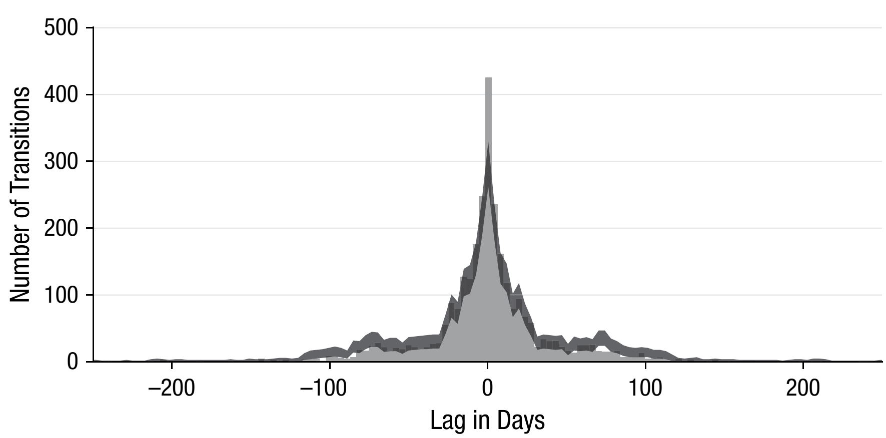

Fig. 6. Distribution of transition lags in Experiment 2. Subjects recalled as many news stories from the preceding 4 months as possible, in whatever order they came to mind. Each time a subject recalled one story and transitioned to recalling another story, we defined the lag of the transition as the difference, in days, between the dates that the subjects provided for the two stories. The lighter gray bars are a histogram, and the darker gray line shows the middle 95% of the distribution of lags from a null model in which transitions are random.

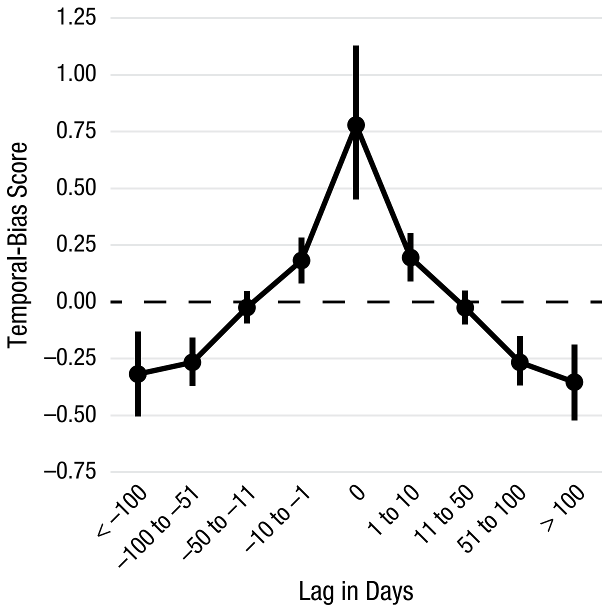

Fig. 7. Mean temporal-bias score as a function of lag in Experiment 2. For each lag bin, we counted the number of times that the subject actually made a transition corresponding to that bin as well as the number of times that such a transition would be expected by chance. We defined temporal bias as the difference between the actual counts and the expected counts expressed as a proportion of the expected counts. Error bars indicate 95% bootstrapped confidence intervals computed across subjects.

than 0.75 in Experiment 2. That is, whereas the TCE was qualitatively similar in the two experiments, it was quantitatively smaller in Experiment 2. This difference is difficult to interpret given that the events that subjects were asked to recall in the two experiments were quite different along a variety of dimensions, but it may have implications for the timescale invariance of the TCE. We return to this issue in the Discussion section. A second difference is that whereas a clear forward asymmetry was observed in Experiment 1, there was no evidence of asymmetry in Experiment 2. This apparent instability of the asymmetry effect may pose a challenge for models in which asymmetry arises naturally.

The critical question for this experiment is whether the TCE remains even when subjects are transitioning between stories that are not strongly semantically related. If the TCE is an artifact of a tendency for things that happen close in time to be semantically related, then the TCE should disappear when transitioning between stories with very low semantic relatedness. To jointly measure the influence of temporal lag and semantic similarity, we recorded both the lag in days and the semantic-similarity bin of each transition a subject made. This resulted in a Temporal Lag × Semantic Bin matrix of counts of how many times transitions were actually made.

For example, 1 subject transitioned between the following two stories: "U.S. Bombing in Syria" and "Van Attack in Toronto." The subject dated these stories as April 14, 2018, and April 23, 2018, respectively; thus, the transition had a lag of +9. These stories had an LSA similarity of $\cos(\theta) = .030$, which puts them in the medium-similarity bin. Therefore, for this transition, we would increment the actual count of the lag = +9/medium-similarity cell of the count matrix. When simulating the random-transition null model, we kept track of a corresponding matrix of expected counts.

The first step in understanding the role of semantic similarity was to examine the influence of semantics when ignoring temporal lag. We did this by summing the actual and expected count matrices across temporal lags and computing a semantic-bias score that was directly analogous to the temporal-bias score. The top row of Figure 8 shows these semantic-bias scores as a function of similarity bin. It reveals a semantic contiguity effect that is quite similar to that which has been widely reported in free recall of word lists (Healey & Kahana, 2014; Howard & Kahana, 2002b; Morton & Polyn, 2016): Subjects were more likely to transition between stories that had high versus low semantic similarity.

To test whether the TCE can be explained by this semantic-contiguity effect, we computed a temporal-bias-score curve separately for each semantic-similarity bin. If semantic similarity is driving the TCE, we would expect the temporal-bias curve to be strongly peaked when considering only transitions that fall into the high-semantic-similarity bin and to be flat when considering only transitions that fall into the low-semantic-similarity bin. The results are shown in the bottom row of Figure 8. At each level of semantic similarity, the temporal-bias scores peaked at 0.

The across-lag average level of these curves tends to increase across semantic-similarity bins, reflecting the fact that transitions between semantically related stories are more frequent overall. But even for the lowest similarity bin, 0-lag transitions are more frequent than expected by chance and far-lag transitions are less frequent than expected by chance. That is, the TCE occurs regardless of the level of semantic similarity.

General Discussion

Our results suggest that mere temporal proximity can cause memories to be linked. This is true even when events are not deliberately studied and do not occur one after another in a chainlike sequence. Moreover, we showed that the TCE cannot be explained by the clustering of stories appearing in the news around the same day. Nor is the effect an artifact of temporal

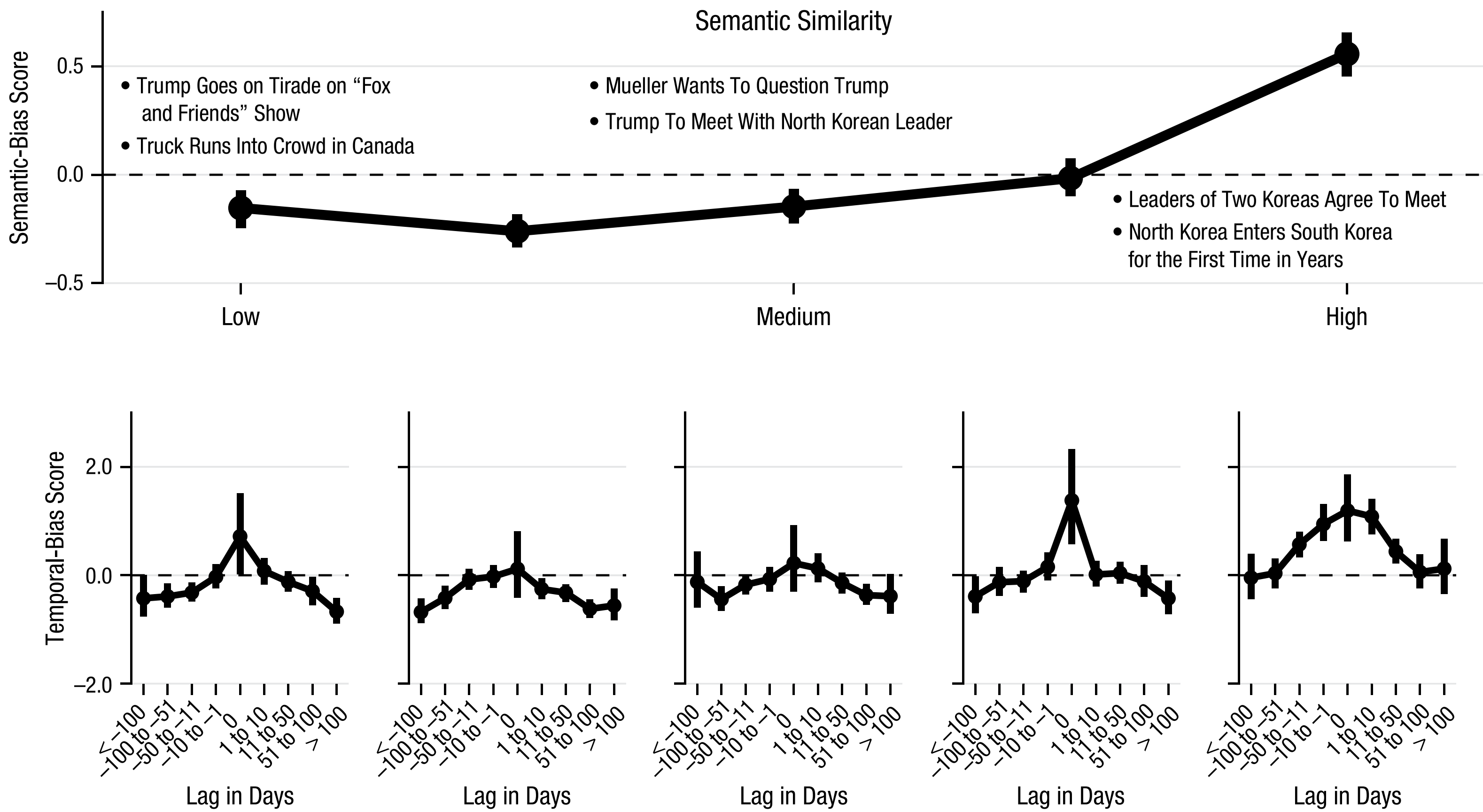

Fig. 8. Bias scores as a function of lag and semantic bin in Experiment 2. The top panel shows the influence of semantic similarity. For each similarity bin, we counted the number of times that subjects actually made a transition corresponding to that bin as well as the number of times that they would have been expected to by chance. We defined semantic bias as the difference between actual counts and the expected counts expressed as a proportion of the expected counts. The upper bounds on the similarity bins were as follows: cos(θ) = .013 (low), cos(θ) = .028, cos(θ) = .044 (medium), cos(θ) = .071, and cos(θ) = 1 (high). Example story pairs are shown for the low, medium, and high bins. The bottom panel shows the effect of temporal proximity at each similarity bin by computing temporal-bias scores only on transitions between stories that fell within that similarity bin. Error bars are 95% bootstrapped confidence intervals computed across subjects.

proximity being confounded with semantic similarity. These findings help us adjudicate among different models of human memory.

Because our subjects were not preparing for a memory test when they experienced the news stories, the data speak against theories that attribute the TCE to a task-specific mnemonic strategy (Hintzman, 2016), at least one that operates at encoding. To the contrary, the data suggest that the memory system naturally encodes information about temporal distance and uses that information during memory search.

Because the news stories that subjects recalled were separated by long time spans, the data speak against a model in which a TCE will arise only when events occur very close in time (Healey, Long, & Kahana, in press). For example, associative-chaining models (e.g., Lewandowsky & Murdock, 1989; Solway, Murdock, & Kahana, 2012) and dual-store models (e.g., Kimball, Smith, & Kahana, 2007; Raaijmakers & Shiffrin, 1981) can form associations between events that co-occupy short-term memory. Such models could produce a TCE at longer time scales if a semantically based reminding mechanism (e.g., study-phase retrieval; Hintzman, 2016; Hintzman, Summers, & Block, 1975) causes a newly experienced event to bring an earlier event to mind via semantic associations, allowing the two events to form a link in short-term memory. However, such an account predicts that the TCE should disappear when the influence of semantic similarity is removed, as was the case for the stories we considered here.

Two caveats to these conclusions should be noted. First, one cost of moving beyond laboratory list learning is that we cannot be certain when or how many times a subject was exposed to coverage of a story. Events that occur around the same time are likely to be covered in the news together, providing the basis for a true TCE, but older events can receive renewed coverage (e.g., Clinton's e-mail scandal), which could allow temporally distant events to be linked. It is clear that this "fuzzy dating" will add noise to the data, but it is vastly more likely to attenuate a genuine TCE than to create an artificial one. Second, because semantic similarity cannot be measured directly, it is always possible that its influence has not been fully removed and could still account, at least partly, for the TCE. Here, we attempted to control for semantic similarity in two quite different ways: human ratings in Experiment 1 and LSA in Experiment 2. Laboratory work has controlled for semantic similarity by using LSA (Howard & Kahana, 2002b) and experimental manipulation of the semantic structure of lists (Polyn, Erlikhman, & Kahana, 2011). In all cases, a TCE has been observed. This convergence of evidence makes it unlikely that the TCE can be fully explained by semantic similarity.

With these caveats in mind, the data point to a model that can produce a TCE at a variety of timescales. Several classes of models may be able to do this. These include retrieved context models (Lohnas et al., 2015), which achieve approximate timescale invariance by combining a drifting mental context representation with a competitive decision rule; the scale-indepenent memory, perception, and learning (SIMPLE) model (Brown, Neath, & Chater, 2007), which achieves timescale invariance by using a logarithmic temporal representation; and clustering models, such as the one described by Farrell (2012), which can produce contiguity at multiple timescales by associating items with a hierarchy of representations with increasingly coarse timescales.

Further work using the methods we have developed here to measure contiguity in the real world could pit these models against each other. For example, some mechanisms would tend to produce a TCE that is truly timescale invariant (Brown et al., 2007; Howard et al., 2015), whereas other mechanisms would tend to produce an effect that is only approximately timescale invariant and would decrease in magnitude as the timescale increases (Howard, 2004).

Conclusion

We tested the assumption that temporal links guide memory search by having subjects recall news stories from the 2016 U.S. presidential election cycle (Experiment 1) or the most recent 4 months (Experiment 2). The data revealed that when subjects recalled one story, they went on to next recall a story encountered close in time more often than would be expected by chance—that is, their memory search was guided by temporal links. These findings provide solid empirical ground for the claim that temporal links are a fundamental organizing principle of human memory.

Action Editor

Caren Rotello served as action editor for this article.

Author Contributions

M. K. Healey developed the study concept. Both authors contributed to the study design. Data collection was conducted by M. G. Uitvlugt. Data analysis was conducted by M. G. Uitvlugt under the supervision of M. K. Healey. Both authors contributed to writing the manuscript and approved the final manuscript for submission.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797618808474

Open Practices

All data have been made available at the Computational Basis of Cognitive Control Lab website and can be accessed at https://cbcc.psy.msu.edu/data/newsevents_data.zip. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

References

Altmann, E. M. (2003). Reconstructing the serial order of events: A case study of September 11, 2001. *Applied Cognitive Psychology*, *17*, 1067–1080.

Brown, G. D. A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, *114*, 539–576.

Cortis Mack, C., Cinel, C., Davies, N., Harding, M., & Ward, G. (2017). Serial position, output order, and list length effects for words presented on smartphones over very long intervals. *Journal of Memory and Language*, *97*, 61–80.

Cortis Mack, C., Dent, K., Kennett, S., & Ward, G. (2015). First things first: Similar list length and output order effects for verbal and nonverbal stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*, 1179–1214.

Davelaar, E. J., Goshen-Gottstein, Y., Ashkenazi, A., Haarmann, H. J., & Usher, M. (2005). The demise of short-term memory revisited: Empirical and computational investigations of recency effects. *Psychological Review*, *112*, 3–42. doi:10.1037/0033-295X.112.1.3

Estes, W. K. (1955). Statistical theory of distributional phenomena in learning. *Psychological Review*, *62*, 369–377. doi:10.1037/h0046888

Farrell, S. (2012). Temporal clustering and sequencing in short-term memory and episodic memory. *Psychological Review*, *119*, 223–271. doi:10.1037/a0027371

Friedman, W. J. (2004). Time in autobiographical memory. *Social Cognition*, *22*, 591–605.

Healey, M. K. (2018). Temporal contiguity in incidentally encoded memories. *Journal of Memory and Language*, *102*, 28–40.

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., Long, N. M., & Kahana, M. J. (in press). Contiguity in episodic memory. *Psychonomic Bulletin & Review*.

Hintzman, D. L. (2011). Research strategy in the study of memory: Fads, fallacies, and the search for the "coordinates of truth." *Perspectives on Psychological Science*, *6*, 253–271.

Hintzman, D. L. (2016). Is memory organized by temporal contiguity? *Memory & Cognition*, *44*, 365–375.

Hintzman, D. L., Summers, J. J., & Block, R. A. (1975). Spacing judgments as an index of study-phase retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *104*, 31–40.

Howard, M. W. (2004). Scaling behavior in the temporal context model. *Journal of Mathematical Psychology*, *48*, 230–238. doi:10.1016/j.jmp.2004.03.004

Howard, M. W., & Kahana, M. J. (2002a). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*, 269–299.

Howard, M. W., & Kahana, M. J. (2002b). When does semantic similarity help episodic retrieval? *Journal of Memory and Language*, *46*, 85–98.

Howard, M. W., Shankar, K. H., Aue, W. R., & Criss, A. H. (2015). A distributed representation of internal time. *Psychological Review*, *122*, 24–53. doi:10.1037/a0037840

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

Kimball, D. R., Smith, T. A., & Kahana, M. J. (2007). The fSAM model of false recall. *Psychological Review*, *114*, 954–993. doi:10.1037/0033-295X.114.4.954

Kintsch, W. (1970). Models for free recall and recognition. In D. A. Norman (Ed.), *Models of human memory* (pp. 331–373). New York, NY: Academic Press.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

Lehman, M., & Malmberg, K. J. (2013). A buffer model of memory encoding and temporal correlations in retrieval. *Psychological Review*, *120*, 155–189. doi:10.1037/a0030851

Lewandowsky, S., & Murdock, B. B. (1989). Memory for serial order. *Psychological Review*, *96*, 25–57. doi:10.1037/0033-295X.96.1.25

Lohnas, L. J., Polyn, S. M., & Kahana, M. J. (2015). Expanding the scope of memory search: Intralist and interlist effects in free recall. *Psychological Review*, *122*, 337–363.

McGeoch, J. A. (1932). Forgetting and the law of disuse. *Psychological Review*, *39*, 352–370. doi:10.1037/h0069819

Moreton, B. J., & Ward, G. (2010). Time scale similarity and long-term memory for autobiographical events. *Psychonomic Bulletin & Review*, *17*, 510–515.

Morton, N. W., & Polyn, S. M. (2016). A predictive framework for evaluating models of semantic organization in free recall. *Journal of Memory and Language*, *86*, 119–140.

Nairne, J. S., Cogdill, M., & Lehman, M. (2017). Adaptive memory: Temporal, semantic, and rating-based clustering following survival processing. *Journal of Memory and Language*, *93*, 304–314.

Polyn, S. M., Erlikhman, G., & Kahana, M. J. (2011). Semantic cuing and the scale-insensitivity of recency and contiguity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 766–775.

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*, 129–156. doi:10.1037/a0014420

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, *88*, 93–134. doi:10.1037/0033-295X.88.2.93

Roediger, H. L., III, & Crowder, R. G. (1976). A serial position effect in recall of United States presidents. *Bulletin of the Psychonomic Society*, *8*, 275–278.

Rubin, D. C. (1982). On the retention function for autobiographical memory. *Journal of Verbal Learning and Verbal Behavior*, *21*, 21–38.

Rubin, D. C., & Baddeley, A. D. (1989). Telescoping is not time compression: A model of the dating of autobiographical events. *Memory & Cognition*, *17*, 653–661. doi:10.3758/BF03202626

Sederberg, P. B., Miller, J. F., Howard, W. H., & Kahana, M. J. (2010). The temporal contiguity effect predicts episodic memory performance. *Memory & Cognition*, *38*, 689–699. doi:10.3758/MC.38.6.689

Shum, M. S. (1998). The role of temporal landmarks in autobiographical memory processes. *Psychological Bulletin*, *124*, 423–442.

Solway, A., Murdock, B. B., & Kahana, M. J. (2012). Positional and temporal clustering in serial order memory. *Memory & Cognition*, *40*, 177–190.

Steyvers, M., Shiffrin, R. M., & Nelson, D. L. (2005). Word association spaces for predicting semantic similarity effects in episodic memory. In A. F. Healy (Ed.), *Experimental cognitive psychology and its applications* (pp. 237–249). Washington, DC: American Psychological Association.

Unsworth, N. (2008). Exploring the retrieval dynamics of delayed and final free recall: Further evidence for temporal-contextual search. *Journal of Memory and Language*, *59*, 223–236. doi:10.1016/j.jml.2008.04.002

Whitten, W. B., & Leonard, J. M. (1981). Directed search through autobiographical memory. *Memory & Cognition*, *9*, 566–579.