Examining the Independence of Scales in Episodic Memory using Experience Sampling Data

Hyungwook Yim (hyungwook.yim@unimelb.edu.au)
School of Psychological Sciences, The University of Melbourne, Australia
Division of Psychology, School of Medicine, University of Tasmania, Australia

Paul M. Garrett (paul.garrett@uon.edu.au) & Megan Baker (megan.s.baker@uon.edu.au)
School of Psychology, The University of Newcastle, Australia

Simon J. Dennis (simon.dennis@gmail.com)
School of Psychological Sciences, The University of Melbourne, Australia

Abstract

We investigated whether memories of different time scales (i.e., week, day, hour) are used independently (i.e., independence of scales). To overcome the limitations of previous studies that have low ecological validity in selecting the test stimuli, we used experience sampling technology. Participants wore a smartphone around their neck for two weeks, which was equipped with an app.that automatically collected time, images, GPS, audio and accelerometry. After a one-week retention interval, participants were presented with an image that was captured during their data collection phase, and tested on their memory of when the event happened (i.e., week, day of week, and hour). We find that, in contrast to previous studies, memories of different time scales were not retrieved independently in everyday life. Additionally, we replicated previous laboratory findings such as correlations between confidence rating and memory performance, and patterns found between valence rating and memory accuracy.

Keywords: independence of scales; experience sampling; episodic memory

Introduction

The ability to remember *when* a past event happened is one of the crucial cognitive mechanisms we use in everyday life. We are also capable of remembering a past event in different time scales such as remembering the year, day of month, day of week, and time of day an event happened. Previous studies examining this ability have reported that remembering one time scale of an event (e.g., month of the event) is independent from remembering another time scale (e.g., day or hour) of the same event. For example, Friedman & Wilkins (1985) tested participants using news events (e.g., John F. Kennedy's assassination), where they were asked about the exact year, month, day of month, day of week, and hour of the events. A reasonable prediction would be that if a finer time scale such as the hour of the event is remembered, it would be likely that a larger time scale such as the year of the event would be remembered as well (e.g., if one remembers that J.F. Kennedy was assassinated around 12:30pm, it would be likely that he/she remembers that the year was 1963). However, Friedman found that in some cases remembering a finer time scale was more accurate than remembering a larger time scale (i.e., scale effects). The scale effects support the idea that people could use different cues to retrieve different time scales of the event rather than only relying on the overall memory strength of the event (i.e., contextual association theory and location-based process; Friedman, 1993).

Evidence for scale effects, which imply that memories of different time scales are used independently, have been reported in many other studies using different methods. For example, Friedman (1987) asked participants about when a local earthquake happened, Huttenlocher, Hedges, & Prohaska (1992) asked participants, who previously responded to an phone survey, the day of week and time of the phone survey, and Larsen & Thompson (1995) asked when events in participants' diary happened. These studies have all shown evidence for scale effects, but at the same time suffer from a lack of ecological validity. Historical and media events (e.g., John F. Kennedy's assassination) might have less self-relevance than our day to day events, or may be more salient than the typical events that occur on a daily basis (e.g., local earthquake). Diary studies have the issue of selection bias, where more salient events would be recorded by the participants than regular events (Sreekumar, 2015). Therefore, with these methodological drawbacks it is unclear whether the scale effects would be found in our day to day life events.

An alternative way to examine scale effects with high ecological validity is using experience sampling techniques. Experience sampling has the advantage of collecting each participant's day to day events automatically without selection-bias, and by utilizing modern smartphones, various modalities can be easily recorded such as time, images, sounds, GPS, and accelerometry. Previous memory studies using experience sampling techniques have been successful in showing interesting finding about human memory usage in real life. These findings range from the kinds of cues people use to remember when an event happened, to how time and space are represented in the brain (e.g., Dennis, Yim, Evans, & Garrett, 2017; Nielson, Smith, Sreekumar, Dennis, & Sederberg, 2015; Sreekumar, Dennis, Doxas, Zhuang, & Belkin, 2014; Chow & Rissman, 2017).

Therefore, in the current study we used experience sampling techniques to examine whether memories of different time scales are independently used (i.e., independence of scales), and whether scale effects are present in everyday real life. In the experiment, participants collected their day to day

events for two weeks using a smartphone which collected various modalities including images and time of the event. Then participants were presented with images that they had collected and were asked what week, day of week, and time of day the event depicted by the image happened. Additionally, we asked how confident the participants were for making each judgment, and the valence of the event.

Experiment

Methods

Participants Eighteen adults participated in the study (nine females, M = 26.38 yrs, SD = 6.50 yrs) who were recruited from flyers posted at the University of Newcastle. Participants were paid AU$100 for their time and effort.

Materials The stimuli used for each participant's experiment were images that were captured by each participant during the data collection period. Images were individually prepared after the data collection phase and before the test phase. First, blurry images were filtered out that had an entropy values below 17.0 and a variation of the Laplacian (Pech-Pacheco, Cristobal, Chamorro-Martinez, & Fernandez-Valdivia, 2000) below 7.0. Then, one image was selected for each one-hour time slot based on its distance to images in other time slots (i.e., highest MIN-distance), where the distance was measured by Euclidean distance using *gist* representation of each image (Oliva & Torralba, 2001). Due to individual differences in collecting the images, the number of images at test differed across participants ($M = 67.5$, $SD = 27.72$, $range = 22 - 122$).

Procedure There was a two-week data collection phase followed by a one-hour test phase. The phases were separated by approximately seven days. The data collection phase started on a Monday and ended on the following Friday. During the data collection phase, participants were told to wear an Android smartphone around their neck during the weekdays when they were awake (see Figure 1A). The phone was equipped with the 'Unforgettable' app. (Unforgettable Technologies, 2017), which collected image, time, audio (obfuscated), GPS, accelerometer and orientation information at approximately five minute time intervals (see Figure 1B). Participants could turn off the app anytime they needed privacy. When the phone detects WiFi and is charged, it sends the stored data automatically to a remote server, which usually happened once per day at the end of the day when users charge the phone overnight.

Seven days after the data collection phase (i.e., on the third Friday), participants were asked to login to an online webpage for the test phase. Participants were presented with a selection of their images collected during the data collection phase. The images were presented one at a time on the left side of the screen with related questions on the right side (see Figure 1C). Participants were asked in which week, day, and time the event captured in the image happen, along with a five scale confidence rating for each response. The valence of the event was also elicited (i.e., "rate how you felt about the event that was occurring when the photo was taken.") using a five point scale (i.e., very negative, negative, neutral, positive, very positive). The number of test trials differed based on the number of images that were collected by each participant during the data collection phase (see Materials).

Additionally, a study-test memory task, which was irrelevant to the current investigation, was administered using the same pool of images on the third Monday (i.e., approximately four days before the current test phase). The results of this task will be reported elsewhere.

Results

All analyses involving statistical inference were conducted using bootstrapping methods (Efron & Tibshirani, 1997), unless stated otherwise. The group data was taken where each subject's raw data was re-sampled 500,000 times with replacement.

We first examined the accuracy for each time scale (see Table 1). Although chance level for $P(hour)$ would theoretically be 1/24, we assumed the chance level as 1/12 since the average time range of the collected images across the two-week interval was 12.88 ($SD = 2.11$, $range = 8 - 16$). Results show that performance in all time scales were above chance, which indicate that participants in the current study were capable of recalling when an event happened with reasonable precision.

Table 1: Accuracy for each time scale with mean accuracy, chance-level for each time scale, 95% confidence interval (95% CI), and Bonferroni corrected empirical p-value against each chance-level derived from bootstrapping.

	M	chance-level	95% CI	p-value
$P(week)$	.61	.50 (= 1/2)	[.56, .65]	$< .001$
$P(day)$	.34	.20 (= 1/5)	[.28, 40]	$< .001$
$P(hour)$	.22	.08 (= 1/12)	[.19, .25]	$< .001$

In Figure 2A, we present memory accuracy for each week. There was no performance difference between the first week ($M = .63$, $SD_{bootstrapped} = .03$) and second week ($M = .58$, $SD_{bootstrapped} = .02$; $p = .143$). Figure 2B presents memory accuracy for each day of week combined over the two weeks (M: $M = .28$, $SD_{bootstrapped} = .04$; Tu: $M = .18$, $SD_{bootstrapped} = .02$; W: $M = .19$, $SD_{bootstrapped} = .04$; Th: $M = .23$, $SD_{bootstrapped} = .03$; F: $M = .31$, $SD_{bootstrapped} = .07$;). Results shows a trend for recency and primacy effects as shown in traditional serial position effects (Ebbinghaus, 1913). The serial position effect was examined by fitting a quadratic polynomial to the data. Results showed an excellent fit as shown in Figure 2B in a red curved line (i.e., $ACCURACY = .029 \cdot DAY^2 - .163 \cdot DAY + .410$; $root - mean - squareerror = .001$, $R^2 = .963$), where there was a statistically significant coefficient for the second order term (i.e., DAY^2; $p = .002$) and the intercept ($p < .001$). The higher performance for Monday and Friday is likely to

Figure 1: Apparatus and test-trial example of the study. (A) participant wearing the Android phone for the data collection phase, (B) layout of the Unforgettable app which was used for data collection, and (C) an example of a trial for the test phase.

be contributed by the salient anchor points created by the weekends since memories for weekends are better than weekdays (e.g., Huttenlocher, Hedges, & Prohaska, 1992). Using salient cues from Saturday and Sunday may have aided participants' ability to retrieve the events.

To evaluate the independence between different time scales, we used pointwise mutual information (PMI) as in Equation 1:

$$PMI(A;B) = \ln\left(\frac{P(A,B)}{P(A)\cdot P(B)}\right) \quad (1)$$

where, $P(A,B)$ is the probability of correctly recalling both time scale A and B (e.g., week and day) of an event whereas $P(A)$ and $P(B)$ are correctly retrieving time scale A (e.g., week) and B (e.g., day) respectively. PMI ranges from $-\infty$ to $min(-logP(A), -logP(B))$, and when $P(A)$ and $P(B)$ are independent PMI is zero. Using bootstrapping methods, we examined whether PMI(week; day), PMI(week; hour), and PMI(day; hour) were statistically different from zero. Results are shown in Table 2, where all three PMI values were statistically different from zero. In contrast to previous studies Friedman & Wilkins (e.g., 1985), the results provide evidence for dependence between all scales (i.e., week and day, week and hour, and day and hour). The difference in the results may have stemmed from the difference in the stimuli that where presented to the participants. We will discuss this more in the General Discussion section.

Confidence ratings for each time scale were analyzed. The average confidence rating was 1.89 ($SD_{bootstrapped} = .24$) for the week response, 1.61 ($SD_{bootstrapped} = .21$) for the day response, and 1.89 ($SD_{bootstrapped} = .18$) for hour response. The confidence ratings for the three time scales were not statistically different ($F = 0.75$, $p = .94$), which indicates that the participants' subjective ratings for remembering each time scale was not different.

Table 2: Pointwise mutual information (PMI) between different time scales with mean PMI (M), 95% confidence interval, and empirical p-value against $PMI = 0$ derived from bootstrapping.

	M	95% CI	$p-value$
$PMI(week;day)$	.12	[.050, .193]	$< .001$
$PMI(week;hour)$	.09	[.016, .161]	.018
$PMI(day;hour)$	.26	[.113, .398]	$< .001$

The relationships between accuracy and response confidence at each time scale were also examined by calculating point biserial correlation coefficient (r_{pb}; see Figure 3). r_{pb} for the week (.18), day (.38), and hour scale (.27) all showed significant correlations ($p < .001$) replicating previous studies that show positive correlations between confidence and accuracy performance (Roediger & DeSoto, 2014).

Finally, we examined the valence of the events which the participants experienced during the data collection phase. As shown in Table 3, the proportion of extreme valences (i.e., very negative and very positive) was less than .04 (4%). The

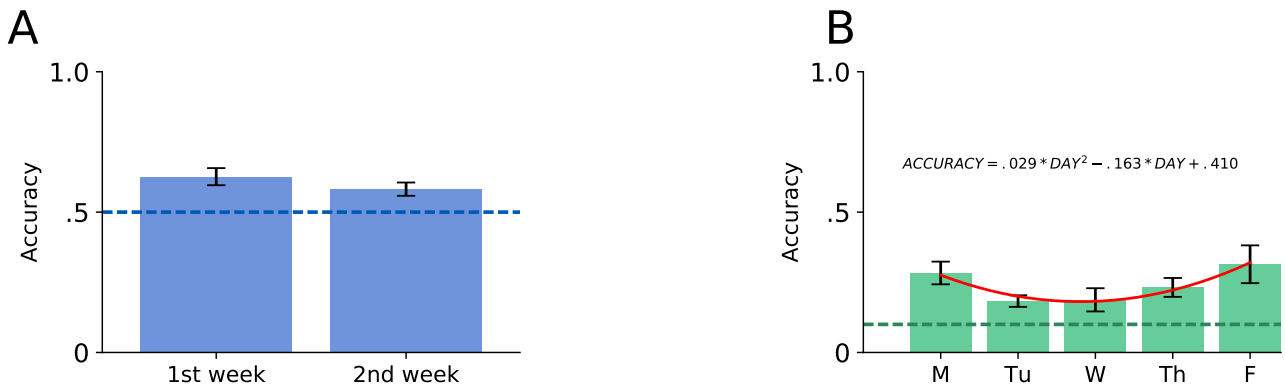

Figure 2: Accuracy for (A) each week and (B) each day. The red curved line on panel B represents the quadratic fit to the data with estimated coefficients above the curve. Dotted lines represent chance level for each time scale, error bars represent the standard deviation of the bootstrapped samples.

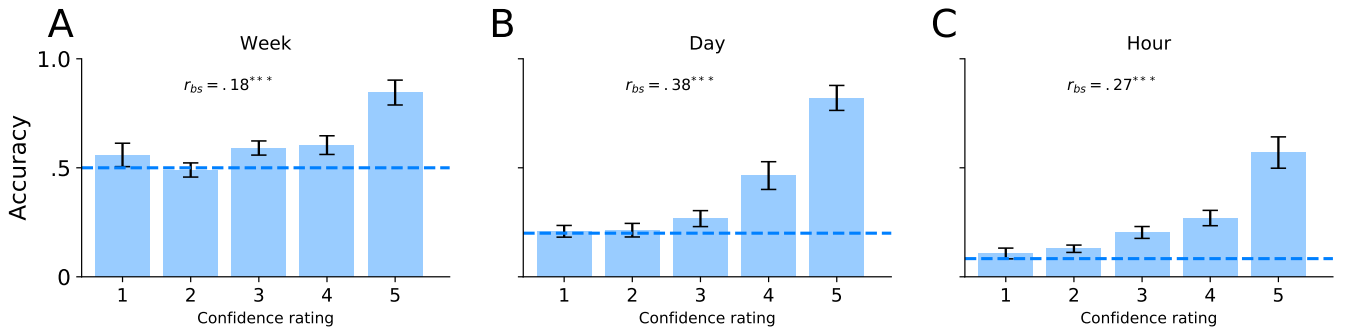

Figure 3: Accuracy by confidence rating for (A) week, (B) day, and (C) hour. Values on the x-axis represent confidence rating scores from 'Not at all confident' (1) to 'Very confident' (5). Dotted lines represent chance level for each time scale, error bars represent the standard deviation of the bootstrapped samples. Point biserial correlations (r_{bs}) are presented for each time scale, where *** represents $p < .001$.

Table 3: Proportion of valence response for events participants experienced during the data collection phase.

Very negative	Negative	Neutral	Positive	Very positive
.004	.058	.649	.253	.035

majority of the events were rated as neutral or positive (.90). The relationship between memory accuracy and valence ratings was examined by (1) polarity (i.e., negative vs. neutral vs. positive) and (2) strength (i.e., neutral vs. weak vs. strong). Due to the lack of responses in the extreme categories (i.e., very negative and very positive) we merged the data in the following way. For the polarity analysis 'very negative' and 'negative' responses were merged into Negative response, 'very positive' and 'positive' responses were merged into Positive response. In a similar fashion, for the strength analysis, 'negative' and 'positive' responses were merged into Weak response, and 'Very negative' and 'Very positive' responses were merged into Strong responses. The neutral response in both cases remained as Neutral. Figure 4 shows the results from the valence strength analysis. The results show better memory performance for stronger valence across all time scales. Figure 5 shows the results from the valence polarity analysis. In general, the results showed better memory performance when the valence was positive. For the week and day scales a statistically significant difference was only shown between the positive and neutral valence (Bonferroni corrected ps $< .05$), whereas for the hour scale the positive valence was statistically different from both negative and neutral valence (Bonferroni corrected ps $< .001$).

General Discussion

The current study examined whether memories of different time scales are independently used in real life as suggested by previous studies which may have suffered from low ecological validity (e.g., Friedman & Wilkins, 1985). We used experience sampling techniques to overcome this issue. Most

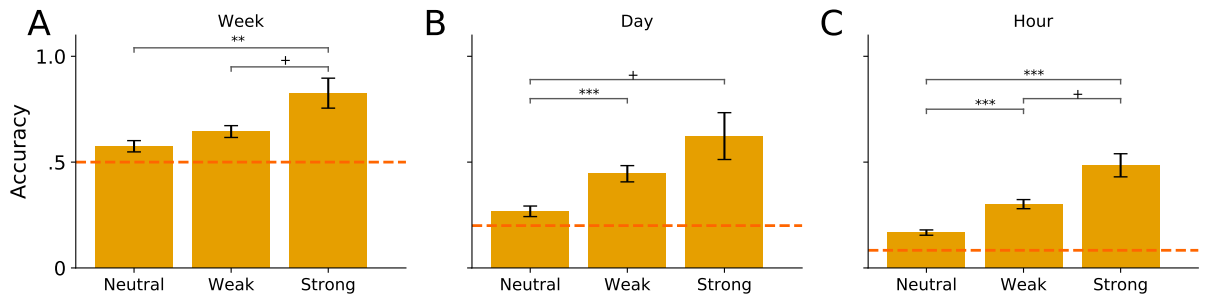

Figure 4: Accuracy by valence strength for (A) week, (B) day, and (C) hour. Dotted lines represent chance level for each time scale, error bars represent the standard deviation of the bootstrapped samples. $^{+}$, *, **, and *** represent Bonferroni corrected p-values that are $p < .05$, $p < .01$, $p < .005$, and $p < .001$ respectively.

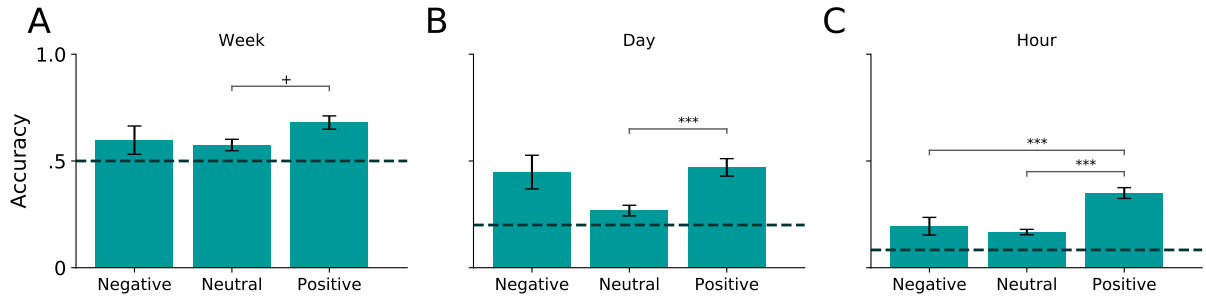

Figure 5: Accuracy by valence polarity for (A) week, (B) day, and (C) hour. Dotted lines represent chance level for each time scale, error bars represent the standard deviation of the bootstrapped samples. $^{+}$, *, **, and *** represent Bonferroni corrected p-values that are $p < .05$, $p < .01$, $p < .005$, and $p < .001$ respectively.

importantly, results showed that retrieving memories of all time scales were dependent on each other, supported by PMI values above zero.

Our result contradicts a series of previous studies that support independence of scales (e.g., Friedman, 1987; Friedman & Wilkins, 1985; Huttenlocher et al., 1992; Larsen & Thompson, 1995). It is very possible that this difference stems from the events that were used at test. Previous studies that tested the independence of scales used historical events (e.g., assassination of John F. Kennedy), infrequent events from the media (e.g., an earthquake), or events from participants' diaries. These events, in general, would be more salient and would happen more infrequently than events that happen in our everyday lives (c.f., events recorded in one's diary may have the selection-bias of being more salient/infrequent events). Additionally, the insignificant events that comprise our daily lives often demonstrate a repeated structure. As shown in the valence rating, the majority of the events (i.e., 90%) that the participants experienced were neutral or positive, whereas only a small portion (i.e., 4%) were rated as very negative or very positive. Moreover, considering that most of the participants were university students, many of the events they experience would repeat, and different time scales in these events would be correlated such as attending a cognitive psychology class every Monday and Wednesday at 9am, and going to work every Tuesday and Friday at 6pm. These repeated and correlated structures would make memories of different time scales more dependent, and as a result retrieving/using memories of different time scales from these events would be dependent. Therefore, results from the current study would be more representative of our real life. However, since the current study used a specific memory period (i.e., two weeks) and retention interval (i.e., one week), future study should examine whether using different length of memory period and retention interval would affect the results.

Another interesting finding is the serial position effect shown by the days of week data. It is unlikely that the effect originated from better encoding at the start of the week and lesser interference at the end of the week as in traditional explanation of the serial position effect. It is hard to imagine participants were more attentively encoding events on both Mondays and Fridays coupled with the fact that participants had a week long retention interval. Rather, it is more possible that the weekends acted as anchor points. As weekends are usually more memorable and contain salient events (c.f., Huttenlocher, Hedges, & Prohaska, 1992), it would have been easier for participants to retrieve the events which were near these anchor points. Higher accuracy near such salient anchor points have been shown in laboratory experiments (e.g., Hintzman, Block, & Summers, 1973; Nairne, 1991). For ex-

ample, Nairne (1991) presented lists of words and later asked participants to reconstruct the order of the presented lists and words in each list. Results showed a similar serial position effect where the position near the start and end of the each list and the list at the start and end of the experiment showed greater accuracy.

Additionally, using experience sampling techniques we have replicated memory phenomenon that have been examined previously. First, we find a statistically significant correlation between confidence rating and accuracy in all time scales. Although there are mixed results on the relationship between confidence rating and accuracy (Roediger, Wixted, & DeSoto, 2012), studies suggest that the non-significant correlations stem from the structure of non-studied items (i.e., similarity structure of lures), and when only testing studied items the correlations are preserved (Roediger & DeSoto, 2014). We also find that the valence of the event affects accuracy. In general, we find that events that were rated as having stronger valence and positive valence show greater accuracy. Although laboratory studies find evidence for enhanced accuracy for negative events (Holland & Kensinger, 2010), it is possible that the current data set does not contain enough negative ratings (i.e., .004 for very negative and .058 for negative ratings) and lack the power to detect the relationship (also see the larger error bars for the negative events in Figure 5 compared to other valence categories).

In sum, using experience sampling techniques, we provide evidence that memories of different time scales could be used and retrieved dependently in everyday life. Moreover, we replicated several memory phenomena using each participant's experience that was collected by experience sampling techniques, providing an example of how to extend memory experiments outside of the laboratory.

Acknowledgments

This research was supported under Australian Research Council's funding scheme to SJD (project number DP150100272).

References

Chow, T. E., & Rissman, J. (2017). Neurocognitive mechanisms of real-world autobiographical memory retrieval: insights from studies using wearable camera technology. *Annals of the New York Academy of Sciences*, *1396*(1), 202–221.

Dennis, S., Yim, H., Evans, N. J., & Garrett, P. (2017). A hierarchical Bayesian model of memory for when based on experience sampling data. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, (pp. 295–300).

Ebbinghaus, H. (1913). *On memory: A contribution to experimental psychology.*. New York: Teachers College.

Efron, B., & Tibshirani, R. (1997). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Friedman, W. J. (1987). A follow-up to "Scale Effects in Memory for the Time of Events": The earthquake study. *Memory & Cognition*, *15*(6), 518–520.

Friedman, W. J. (1993). *Distance and location processes in memory for the times of past events*, vol. 35. Elsevier Masson SAS.

Friedman, W. J., & Wilkins, A. J. (1985). Scale effects in memory for the time of events. *Memory & cognition*, *13*(2), 168–175.

Hintzman, D. L., Block, R. A., & Summers, J. J. (1973). Contextual associations and memory for serial position. *Journal of Experimental Psychology*, *97*(2), 220–229.

Holland, A. C., & Kensinger, E. A. (2010). Emotion and Autobiographical Memory. *Physics of Life Reviews*, *7*(1), 88–131.

Huttenlocher, J., Hedges, L. V., & Prohaska, V. (1992). Memory for day of the week: a 5 + 2 day cycle. *Journal of experimental psychology. General*, *121*(3), 313–325.

Larsen, S. F., & Thompson, C. P. (1995). Reconstructive memory in the dating of personal and public news events. *Memory & Cognition*, *23*(6), 780–790.

Nairne, J. S. (1991). Positional uncertainty in long-term memory. *Memory & Cognition*, *19*(4), 332–40.

Nielson, D. M., Smith, T. a., Sreekumar, V., Dennis, S. J., & Sederberg, P. B. (2015). Human hippocampus represents space and time during retrieval of real-world memories. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(35), 11078–11083.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*(3), 145–175.

Pech-Pacheco, J. L., Cristobal, G., Chamorro-Martinez, J., & Fernandez-Valdivia, J. (2000). Diatom autofocusing in brightfield microscopy: a comparative study. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, vol. 3, (pp. 314–317 vol.3).

Roediger, H. L., & DeSoto, K. A. (2014). Confidence and memory: Assessing positive and negative correlations. *Memory*, *22*(1), 76–91.

Roediger, H. L., Wixted, J. T., & DeSoto, K. A. (2012). The Curious Complexity between Confidence and Accuracy in Reports from Memory. In L. Nadel, & W. Sinnott-Armstrong (Eds.) *Memory and Law*, (pp. 84–118). New York: Oxford University Press.

Sreekumar, V. (2015). *Context in the wild: Environment, behavior, and the brain*. (Doctoral dissertation), The Ohio State University. Retrieved from https://etd.ohiolink.edu/.

Sreekumar, V., Dennis, S., Doxas, I., Zhuang, Y., & Belkin, M. (2014). The geometry and dynamics of lifelogs: Discovering the organizational principles of human experience. *PLoS ONE*, *9*(5), 1–8.

Unforgettable Technologies (2017). Unforgettable [Software].