Using Emails to Quantify the Impact of Prior Exposure on Word Recognition Memory

Anonymous CogSci submission

Abstract

Recognition memory studies have reliably demonstrated the word frequency effect (WFE), where low-frequency words are more accurately recognized than high-frequency words. The context noise account of WFE argues that pre-experimental exposure to stimuli generates interference that compromises high-frequency words more than low-frequency words. Because the representations of the contexts associated with more recent exposures are assumed to overlap more with the representation of the study context, stimuli that have been seen more recently are thought to generate the most interference. We asked participants to log their daily email for two months. Based on the participant's email corpus, we constructed an individualized study-test recognition memory task to investigate the effect of recency. Results show that recency has a graded effect on recognition memory that extends for at least two months providing support for the context noise account.

Keywords: word frequency effect; recency; context noise; item noise; recognition memory

Introduction

The word frequency effect (WFE) in recognition memory refers to the phenomenon that low-frequency (LF) words are better discriminated than high-frequency (HF) words with LF words having higher hit rates and lower false alarm rates (Glanzer et al., 1993; Glanzer & Adams, 1990). The WFE has been of theoretical interest since (1) even though LF words are less well represented in memory they are recognized better, and (2) that the pattern is the opposite in a recall task, where high-frequency words are better recalled than low-frequency words.

There are two views of the mechanism of the WFE in recognition memory. The retrieving effectively from memory (REM) model claims that the effect stems from the different properties of LF and HF words (Shiffrin & Steyvers, 1997). Considering that items have different features with some being more common and some being less common, REM assumes that the uncommon features are more diagnostic. Additionally, REM, and most of the recognition memory models, assume a global matching process for the recognition process. In a global matching process, a to-be-recognized item is compared to all stored items in one's memory storage and when a similarity signal is above a certain criterion, a decision that the item was presented during the study list (YES) will be made, whereas if the signal strength is below the criterion a NO response will be made. Therefore, when making a decision whether a word was presented during the study list, it will be easier to distinguish LF words from other memories than HF, since LF words will have more uncommon features, which are more diagnostic. The model was supported by studies that control the word frequency while manipulating the frequency of the orthographic features, where words that contained uncommon orthographical features were better distinguished than words that contained common orthographic features (Malmberg et al., 2002). REM relies on the interference between the words itself and has been termed as an ***item noise account*** of the WFE.

While the REM model assumes that a decision in a recognition task involves discriminating between the items that were only presented in the test list (i.e., lures) from the items that were presented in the study list (i.e., targets), there is evidence to suggest that the decision also depends on distinguishing between pre-experimental contexts and the experimental context, where context refers to information from the environment peripheral to the item that is stored in memory with the item (Dye et al., 2017). This information can come from a number of places: the physical environment (Godden & Baddeley, 1975), the semantic context (Steyvers & Malmberg, 2003), or the temporal context (Hintzman & Summers, 1973).

The Bind-Cue-Decide Model of Episodic Memory (BCDMEM) model argues that an item presented on a test list triggers memory traces of the prior contexts where the word was experienced (Dennis & Humphreys, 2001). Therefore, the WFE stems from the fact that the LF words have been seen in fewer contexts (i.e., less context variability) than the HF words have been and are easier to distinguish since there are fewer contexts to be confused about. BCDMEM relies on the interference between different contexts and has been called a ***context noise account*** of the WFE**.**

The two models that explain the WFE introduces a fundamentally different explanation. The item noise account argues that the distinguishability of the item itself causes interference, while for the context noise account the

contexts that the item is experienced in cause interference. There have been studies that attempt to untangle the sources of interference by controlling for item distinctiveness and manipulating context variability in different ways (Chalmers et al., 1997; Chalmers & Humphreys, 2003; Malmberg et al., 2002; Reder et al., 2002). However, it is not trivial to completely test the idea in a laboratory experiment since the distinctiveness of the items tend to be correlated with the frequency (or diversity) of the contexts (e.g., the more times the item is seen the more familiar it becomes and thus becoming less distinctive).

One way to resolve the issue is considering recency. Following the context noise account, if the context of an item is the source of interference, not only will the number of times the item has been seen in different contexts matter, but also the amount of time that has passed since the item has been seen (i.e., recency) will matter as well. This is because the representations of the contexts that are temporally close together will be harder to distinguish, thus, create more interference. For example, when trying to remember whether you parked your car by the tree today, it will be harder to answer the question if you parked your car by the tree yesterday compared to if you parked your car by the tree a couple of weeks ago.

A couple of recent studies have provided evidence that recency matters by pre-exposing the to-be-recognized stimuli prior to the main recognition memory task (Dye et al., 2017). However, pre-exposing the materials in these studies involve an unnatural way of manipulating recency and possibly introduces confounds. Moreover, the time scale that these studies examine (the time between pre-exposure to testing) is relatively short (e.g., 20mins).

In the current study, therefore, we propose a more ecologically valid way in examining the effect of recency in recognition memory using experience sampling methods. By utilizing an experience sampling platform (i.e., Unforgettable.me), we collected participant's daily email for two months. Then, participants went through a study-test recognition memory test using the words that occur in their emails. Therefore, recency can be observed by the researcher instead of being manipulated as in previous studies. Additionally, by having access to the words that an individual experience (i.e., personal corpus), it is possible to generate individualized frequency rather than relying on normative frequency measures. Since individualized frequency will provide a more customized window - looking into one participant's experience - it is possible that the individualized frequency will provide a better account for understanding the WFE compared to the normative frequencies, which rely on the assumption that the normative frequency is a good approximation to the frequency experienced by each participant.

Experiment

Methods

Participants Sixty-five participants (42 females) were recruited via online and flyers around the **[UNIVERSITY_NAME].** Participants were paid $30 for their email data and $15 for completion of the memory test. Approval for this research was obtained from the **[UNIVERSITY_IRB_ETHICS_COMMITEE_NAME]**.

Design and Materials The Unforgettable.me platform (Dennis et al., 2019) was used to collect two months of email from the participant's primary email account. The platform automatically saved the participant's received emails onto the server. The system preserved privacy ensuring the researcher was not able to see the actual content of the email. There had to be at least ten incoming emails a week to participate in the study. In the current study, participants had an average of 773 (*SD* = 1,024) emails, which on average had a total of 52,505 (*SD* = 73,080) words with an average of 2,843 (*SD* = 1,203) unique words after the preprocessing steps (see below).

Words were drawn from the participant's received emails. The recency and frequency of the words were taken and the normative frequency (per million counts) was calculated for the words from the SUBTLEX frequency database (Brysbaert & New, 2009). The individualized frequency was a raw count of how many times a word appeared in the participants' email, and recency was measured in hours since the word was last seen before the experiment. For words to be eligible to be selected for inclusion in the study and test list, they had to meet a number of criteria. First, they had to be in the SUBTLEX frequency database with a normative frequency between 1 and 300 counts per million, thus excluding both extremely common words and rarely used words. Second, words had to be between 3 and 10 letters long. Also, offensive words and names were removed using the Better Profanity Python package and using the database from the top 1000 US Baby Names from 1800 to 2009 (http://github.com/hadley/data-baby-names). Finally, a Porter stemmer algorithm implemented by the Natural Language Toolkit (NLTK; 2019) was used to normalize morphological and inflexional endings (e.g., plans, planning -> plan).

Then, the words were binned based on the SUBTLEX (normative) frequency (high/low), individualized frequency (high/low), and individualized recency (high/low) using a median split for each category. This resulted in 8 possible bins. Based on the number of emails that the participant collected, different numbers of experiment sessions were created with a maximum number of sessions being constrained to 18. On average, participants went through 13.88 sessions (*SD* = 3.76). For each session, six words were randomly selected from each bin to

construct the study list (i.e., 48 words), and another six words were randomly selected to construct the list of lure words which was presented during the test phase (i.e., 48 words).

Procedure Participants collected their received emails for two months through the Unforgettable.me platform (https://unforgettable.me/). Immediately after the two-month period, they were instructed to complete an online memory experiment in a quiet distraction-free environment. The experiment was created in the Unforgettable.me system using jsPsych (de Leeuw, 2015). The experiment did not last more than one hour.

In each experiment session, there was a study phase followed by a 45-second retention interval and a test phase. In the study phase, participants were presented with words that were extracted from their emails (see Design and Materials). The words were presented on the center of a white screen for 1sec with a font size of 2em and font color of black. There was no interstimulus interval. During the retention interval, participants played a card game (i.e., Egyptian Rat Screw) which was a combination of a 2-back working memory task (e.g., press button A if two hearts show in a row and B if two queens appear in a row) and a pattern-matching task (e.g., press J if you see a Joker). In the study phase, participants were randomly presented with 96 words one at a time, where half were from the study phase and the other half not, which served as lures. Test words were presented in the center of the screen until a response was made by the participant. Participants were instructed to indicate whether they had seen the word presented on the study list, by pressing the 'Y' key to indicate yes and the 'N' key to indicate no.

Results

Analyses were conducted on the pooled subject data. The words had an average recency (hour past from test) of 497.55 hours (*SD* = 402.26). The average normative frequency was 31.87 per million (*SD* = 51.42), and the average personalized frequency was 10.99 counts (SD = 57.36). Overall accuracy was .61 (*SD* = .002), which was statistically above chance level ($p < .001$).

Hit rates (HR) and false alarm rates (FAR) were analyzed using a logistic regression model. We first, binned the normative (SUBTLEX) frequency, individualised frequency[1] and recency into 10-quantiles. Based on the quantized bins the median value of the bin and the accuracy corresponding to the bin was fit to the logistic regression (binomial regression with a logit function).

Most importantly, for recency, HR increased (regression weight **b** = 7.5e-5, $p = .002$) and FAR (**b** = -1.15e-4, $p < .001$) decreased as the time since the word had been seen increased. (see Figure 1a). For normative frequency, HR decreased (**b** = -.002, $p < .001$) and FAR increased (**b** = .001, $p = .002$) as the frequency of the presented word at test increased (see Figure 1b). Similarly, for individualized frequency, HR decreased (**b** = -.09, $p = .001$) and FAR increased (**b** = .22, $p < .001$) as the frequency of the presented word at test increased (see Figure 1c).

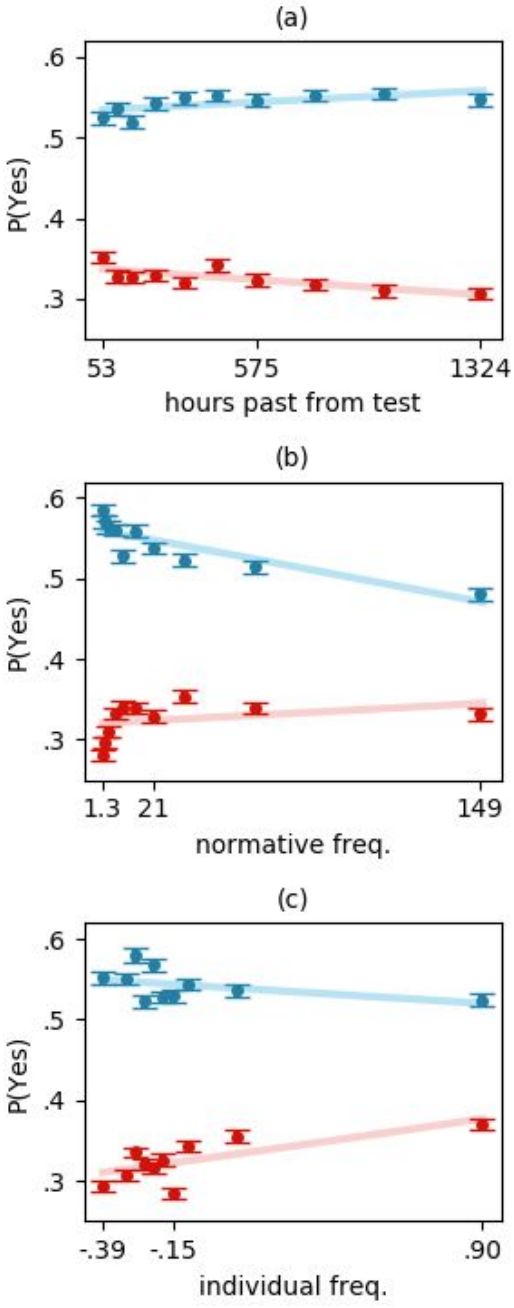

Figure 1. Hit rates (in blue) and False alarm rates (in red) against (a) recency, (b) normative frequency (SUBTLEX), and (c) individualized frequency. The dots represent data, error bars represent standard deviation, and the lines represent the fitted regression model.

[1] Individualized frequencies were standardized for each participant before being used in the analysis.

Next, we examined how discriminability (d-prime) changed in relation to the normative frequency, recency and individualized frequency. Since the words used in the HR model and FAR model were different, d-prime can only be simulated. Therefore, we first fit the HR and FAR with a logistic regression model, where all three variables were used as independent variables so that the model could fully fit the data as well as possible. The variables were standardized for the convenience of the simulation exercise. Using the estimated beta weights, HRs and FARs can be generated with a given set of independent variables where the corresponding equations are presented in equation 1.

$$HR = sigmoid(-.11 \cdot NF -.002 \cdot IF + .10 \cdot RC + .17) \quad (1)$$
$$FAR = sigmoid(.003 \cdot NF + .03 \cdot IF - .03 \cdot RC - .73)$$

where NF indicates normative frequency, IF indicates individualized frequency, and RC indicates recency. Then we submitted values for each independent variable which ranged from -5 to 5 to simulated HR and FAR (cf., note that the independent variables were standardized and the range of -5 to 5 represents a generous range of the independent variables). From the derived predicted values of HR and FAR, d-prime scores were calculated (i.e., Z(HR) - Z(FAR)). As shown in Figure 2, results are consistent with the analysis conducted with HR and FAR separately. d-prime increases as the test word was more distantly experienced in the past, and also increases with lower individual and normative frequencies. We also see a nice correlation between the individual and normative frequency (Figure 2c).

We further examined the correlation (Spearman's rho) between the three independent variables using a permutation test with 10,000 samples. Results are shown in Table 2 with all correlations being statistically significant. As predicted, there was a positive correlation between normative frequency and individualized frequency. The small size of this correlation suggests that they are two different measures and may include different information that does not overlap much. If the correlations had been large, this would have indicated that our measures of normative and individualised frequency were describing essentially the same thing. The use of experience sampling methods in this study therefore allowed us to truthfully capture participant's real experiences with words, independently of the words' normative frequency.

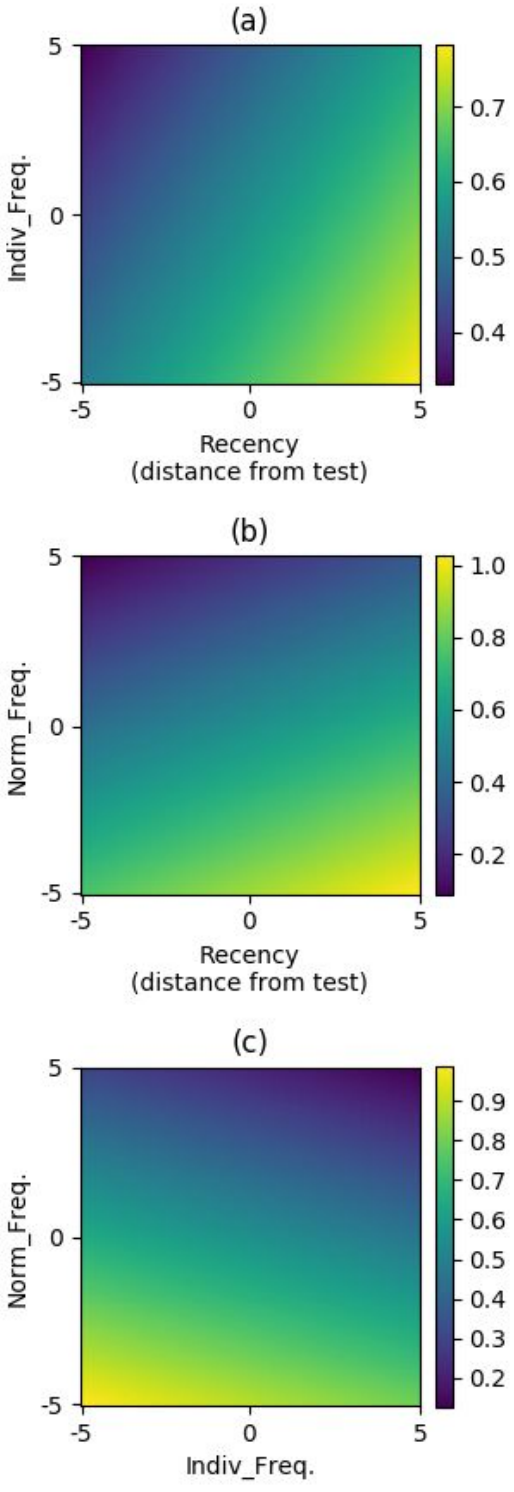

Figure 2. Simulated d-prime measures by two independent variables. (a) individualized frequency and recency, (b) normalized frequency and recency, and (c) normalized frequency and individualized frequency. The graded colors represent d-prime values, where the lighter yellow colors presents higher values than the darker blue colors.

Table 2: Correlations between independent variables. Values indicate Spearman's rho. The top table shows variables used in the hit rate model and the bottom table shows variables used in the false alarm rate model.

HR	*Norm. Freq.*	*Indivi. Freq.*	*Recency*
Norm. Freq.	·	.12	-.03
Indivi. Freq.	.12	·	-.34
Recency	-.03	-.34	·

FAR	*Norm. Freq.*	*Indivi. Freq.*	*Recency*
Norm. Freq.	·	.13	-.03
Indivi. Freq.	.13	·	-.35
Recency	-.03	-.35	·

Interestingly, the correlations of recency with both normative and individualized frequency were quite small relative to the logical prediction that since high-frequency words will have more chance of occurring they would have occurred more recently as well. The correlation between recency and normative frequency was particularly small. This observation is curious, as prior research and common sense would suggest that HF words are more likely to have been recently experienced, and vice versa for LF words (Scarborough et al., 1977), and as such we expected a moderate correlation between the two variables. However, there has been some work done in this area that may explain the observed relationship being smaller than expected. The small association may be attributable to what Albert-Laszlo Barabasi (2010) referred to as "bursts", which refers to the intermittent alternating between periods of low and high activity. This phenomenon was modelled by Burrell (1980) and elaborated upon by Anderson and Milson (1989), who posit that the probability of a memory being needed is dependent on its pattern of past use - that is, memories vary in desirability based on the environment. These levels of desirability can also be quite volatile, depending on the environment. Thus, according to this model, even though some words might have a high raw frequency count and therefore have been classed as HF, their presentations may have been clustered together towards the beginning of data collection. These words were probably momentarily very desirable but high volatility led to this desirability dropping off quickly, therefore creating clusters of high frequency words that were seen some time before the experiment. This explanation is particularly likely considering that words were drawn from emails, which are particularly "bursty" in nature (Barabasi, 2010).

Discussion

In the current study, we examined the word frequency effect in recognition memory using an experience sampling method in order to gain better ecological validity and overcome possible confounds. The experience sampling method, delivered a way to measure the recency of a given word in one's pre-experimental experience, where measuring the recency of a word provides a valuable way to distinguish different theories of recognition memory regarding the word frequency effect. In particular, the context noise account of the word frequency effect will predict that more recently experienced words will cause more interference than words that were experienced earlier. Notably, we find that recency has a graded effect on recognition memory that extends for at least two months (see Figure 1a). The results strongly support the ***context noise account*** of the word frequency effect since, and challenges the theories that support an item noise account.

It is also interesting to see that the individualized frequency captures the WFE well. On one hand, it is an obvious result since the individualized frequency, which is calculated from an individualized corpus, is personalized and will handle the individual variability better. However, it is also worth noting that the individual corpus is based on (1) only two-month worth of emails, and (2) only from received emails. Considering how many words people experience in their daily life from diverse sources (e.g., TV, messenger, books, etc.) and also compared to the size of the normative corpus that has been used (50,000 words vs. 51 million words), the information that the individual frequency is providing is striking. Therefore, it is highly probable that having a longer length of data collecting period with a diverse source may provide more interesting information about one's pre-experimental word experience.

Finally, the study provides an interesting way to use experience sampling methods in memory research. Pure memory studies have been preferred controlled laboratory studies as they generally have less noise in the data and tighter manipulations for examining the effect of concern. However, the current study shows that some issues, some of which are greatly debated in pure memory literature, can not be achieved in the laboratory, and using experience sampling methods provides a promising way of conducting these studies.

References

Albert-Laszlo Barabasi (2010). *The Formula: The Universal Laws of Success*. Little, Brown and Company

Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review, 96*(4), 703.

Burrell, Q. (1980). A simple stochastic model for library

loans. *Journal of Documentation, 36*(2), 115-132.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Chalmers, K. A., & Humphreys, M. S. (2003). Experimental manipulation of prior experience: Effects on item and associative recognition. *Memory (Hove, England)*, *11*(3), 233–246. https://doi.org/10.1080/09658210244000009

Chalmers, K. A., Humphreys, M. S., & Dennis, S. (1997). A naturalistic study of the word frequency effect in episodic recognition. *Memory and Cognition*, *25*(6), 780–784. https://doi.org/10.3758/BF03211321

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Dennis, S. J., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review*, *108*(2), 452–478. https://doi.org/10.1037/0033-295X.108.2.452

Dennis, S. J., Yim, H., Garrett, P., Sreekumar, V., & Stone, B. (2019). A system for collecting and analyzing experience-sampling data. *Behavior Research Methods*. https://doi.org/10.3758/s13428-019-01260-y

Dye, M., Jones, M., & Shiffrin, R. (2017). *Vanishing the mirror effect: The influence of prior history & list composition on recognition memory*.

Glanzer, M., & Adams, J. K. (1990). The Mirror Effect in Recognition Memory: Data and Theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(1), 5–16. https://doi.org/10.1037/0278-7393.16.1.5

Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review*, *100*(3), 546–567. https://doi.org/10.1037/0033-295X.100.3.546

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of Psychology*, *66*(3), 325–331. https://doi.org/10.1111/j.2044-8295.1975.tb01468.x

Hintzman, D. L., & Summers, J. J. (1973). Long-term visual traces of visually presented words. *Bulletin of the Psychonomic Society*, *1*(5), 325–327. https://doi.org/10.3758/BF03334359

Malmberg, K. J., Steyvers, M., Stephens, J. D., & Shiffrin, R. M. (2002). Feature frequency effects in recognition memory. *Memory and Cognition*, *30*(4), 607–613. https://doi.org/10.3758/BF03194962

Reder, L. M., Angstadt, P., Cary, M., Erickson, M. A., & Ayers, M. S. (2002). A reexamination of stimulus-frequency effects in recognition: Two mirrors for low- and high-frequency pseudowords. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *28*(1), 138–152. PubMed. https://doi.org/10.1037//0278-7393.28.1.138

Scarborough, D. L., Cortese, C., & Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance, 3*(1), 1-17.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM-retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*(2), 145–166. https://doi.org/10.3758/BF03209391

Steyvers, M., & Malmberg, K. J. (2003). The effect of normative context variability on recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(5), 760–766. https://doi.org/10.1037/0278-7393.29.5.760