Testing Effects of Personalized Settings on Memory Recall Through Flashcard Retention Testing

JAKE RITZ, Student at Colorado State University, USA
ZACHARY FULLER, Student at Colorado State University, USA
HUNTER HODEN, Student at Colorado State University, USA
JAY SNELLINGS, Student at Colorado State University, USA

This paper explores the effects of personalized and optimized universal settings on memory recall. To compare a user's personal preference with standardized "best" settings gathered from previous testing, a prototype to present and test visual (text) recall with customization features was created and presented to a variety of subjects. Features whose impact was tested include: background color, text color, text size, text boldness, text italics, and background music. Testing efficacy was computed based on recall accuracy and speed. Additionally, a correlative value was drawn between the overlap of the user's preferential and previously tested "best" settings.

Additional Key Words and Phrases: Memory recall, Memory improvement, Memory testing, Generation Effect, Forward Testing Effect

ACM Reference Format:
Jake Ritz, Zachary Fuller, Hunter Hoden, and Jay Snellings. 2018. Testing Effects of Personalized Settings on Memory Recall Through Flashcard Retention Testing. *J. ACM* 37, 4, Article 111 (August 2018), 11 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Memory is the basis of understanding, knowledge, and learning. Thus, preserving and improving it is of the utmost importance. This article will discuss how memorization can be optimized alongside the comparative difference between personalized and optimized settings as previously documented, focusing on characterizing human memory using small-scale controlled experiments [34], for memory recall test retention, accuracy, and speed.

Memory tests have proven an effective technique for increasing recall and retention both immediately following as well as extended periods after the test has been taken [16]. Many tests have been conducted attempting to isolate and optimize certain settings within textual recall such as the font color red being recalled more quickly than black or blue [19] This article seeks to determine whether individual preference in customizable settings outweighs (and outperforms) these results.

Authors' addresses: Jake Ritz, Student at Colorado State University, Fort Collins, Colorado, USA, 80523-1801, jritz11@colostate.edu; Zachary Fuller, Student at Colorado State University, Fort Collins, Colorado, USA, 80523-1801, zrfuller@colostate.edu; Hunter Hoden, Student at Colorado State University, Fort Collins, Colorado, USA, 80523-1801, hodenh@colostate.edu; Jay Snellings, Student at Colorado State University, Fort Collins, Colorado, USA, 80523-1801, jaysnell@colostate.edu.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2018 Association for Computing Machinery.
Manuscript submitted to ACM

2 RELATED WORKS

Many different studies have been conducted on the improving memory recall [3, 18, 22];however, the subject is far too broad to discuss in a single article. Instead, related articles tend to focus on enhancing specific features within memory testing to produce the best results for a general audience in terms of recollection of terms presented. Two main methods are used to alter the memory recall test: direct alteration of the exam and application of external forces. Within these constraints, some optimization for recall settings emerge.

Adding Virtual Reality to this mix reveals a subsection of memory recall testing that has just scratched the surface, allowing us to adapt various methods to this new form of testing.

2.1 Direct Alteration

One method commonly used to test memory recall is to directly alter the presentation or viewing experience for the subjects. This is achieved by adjusting items such as font type, font size, font color, and background color. Various combinations of all these attributes have been studied to find the most effective combination within each category.

As previously mentioned in the study by Khan, the color red may improve recall speed over less vibrant colors such as black or blue [19]. The researchers also found that using colored paper, such as yellow, over plain white paper in addition to the red colored font added to the recall speed of their subjects. Another way of improving the recall of subjects was to alter the font size. One study found that there was statistical significance when the font size had been reduced [13], whereas another found that font size made no difference over its three different experiments [26].

In other studies, researchers investigated the connection between font style and memory retention to facilitate learning and retrieval of new information [40]. One study [36] concluded that although font size does not influence memory recall, words printed in a bigger 48-point font are perceived as being more remembered than words printed in a smaller 18-point font. Other studies [37, 38] examined the effect of serif and sans-serif typefaces on memory performance through a series of tests. Serif typefaces, such as Times New Roman, have extra strokes at the edges of letters. Sans-serif typefaces then are fonts without these extra strokes, like Calibri. The results demonstrated that serif typefaces performed better in memory recall and recognition tests than sans-serif fonts. This observation relates to other studies that aimed to tackle the easiest font for dyslexic users, resulting in the same conclusion that serif fonts are better [1]. This is due to the fact that each letter is unique in its presentation and easy to distinguish; however, other studies [4] suggest that other appearance-related elements, such as importance [8] or context of letters, could influence recollection more strongly than direct alteration.

All of this work on optimizing various parts of the viewing experience of the memory recall test has led us to create what is theoretically the most optimized testing experience for memory recall. Incorporating all of these elements together allowed us to create the optimized viewing settings; red serif font with a yellow background. Using classic Times New Roman, and a medium font size since the evidence there was inconclusive.

2.2 External Forces

So far, the works presented have aimed to manipulate the way the memory recall test is presented to the subject, but there are many other external forces that can impact the subject's ability to recall information. One study used music, or lack thereof, to give the subjects a cue that they should recall a set of words [2]. They found that playing the same music greatly improved short-term memory recall, but its effects got weaker as time went on. The type of music we listen to can also have a large impact on our ability to retain.

In one experiment, participants were asked to retain information while listening to both positive and negative music [17]. They found that positive music had a substantial increase in short-term retention, while negative music was more reliable for retaining information over time. However, the terms 'positive' and 'negative' for music were too abstract for our experiment. In another, participants' brainwaves were altered by music, with alpha and beta waves increasing and decreasing respectively [24]. This study aimed to prove that a person is more likely to have an alpha state of consciousness and improve his/her capacity for memory retention if they listen to relaxing music. Adding music to our experiment was important as it showed to have improved short-term recall. We offered subjects a choice of three genres of music: classical, lofi, or no music at all [7]. This allows the subject to pick a genre they prefer while also getting the benefit of increased retention rates.

People often assume that people recall information better by 'learning by doing' vs 'learning by viewing', but a study by Steffens found that there is no significant difference in memory recall of individuals who are writing words down to remember them rather than just looking at them [33]. Some research has gone to the extent of trying to blur the subject's vision to see if it forces them to focus more on their work and retain information better [31]. Despite this attempt, blurring did not have any significant improvement in the subject's memory recall. In a similar way, multiple studies have found that increasing the dis-fluency by alteration of font or addition of complexity via the generation effect may yield better recollection.[31]. The generation effect is an effect that causes subjects to generate information within their own mind rather than just reading from a paper. Using disfluency to alter and change the font size and color requires subjects to create a cohesive understanding in their mind, as opposed to the messy and disordered one on the paper. These two methods are both aimed at utilizing something known as the desirable difficulty effect [27], which attempts to make subjects put in more effort at the benefit of a higher likelihood of success.

The study presented here differs in many ways from all works presented. Our aim is not to see if any one piece presented here improves recall, as we have already seen that they do.

2.3 Virtual Reality

Virtual Reality (VR) provides the opportunity to directly alter the external forces and environment the user is presented and suggests that VR provides a viable medium for measuring learning and memory [25] [29]. We developed a virtual reality experience that we believe can help (or hinder) memory recall based on the customizable parameters, using virtual cards to test the user and generating a virtual environment, presenting users with virtual stimulus around them [15] [21]. A study conducted in virtual reality sought out to test the idea of external stimuli being an engaging way for learning [30]. Their goal was to teach individuals how to drive a car in a simulated city. The experiment described here differs in many ways to our own. Unfortunately, the study failed to test all individuals on whether or not they have driven a car or had any experience in virtual reality, both of which could alter the outcome of the experiment drastically.

Using technology to our advantage for memory recall has become the norm within our society, using applications such as Microsoft Word and Excel to easily store and send data to one another. When we look at note taking, how would virtual reality differ to these 2d desktop applications? One study found that there is significant benefit to using virtual reality for note keeping and information management [11]. The main idea they believe is causing this is experience and grandiose nature of being in virtual reality. This study helped us form our own virtual world, where we wanted a simple, beautiful scene that participants would find both enjoyable and soothing.

Potential issues with memory recall in virtual words is explored in a study relating to an effect known as context-dependency. The study allowed individuals to learn information in a virtual world, and tested them both in and out of VR to see if there was any significant differences in their memory [20, 32]. In changing the test from virtual to real world they saw a significant decline in participant's results, believing that their knowledge was tied to the context of Virtual Reality. Our experiment only tests user in virtual reality, and doesn't have the option to come back and be tested again 24 hours after so we are able to avoid context-dependency entirely.

3 METHODOLOGY

First a prototype memory testing software was built to facilitate customization of distinctive features using Unity. Within this prototype upon startup an interface allowing users to input their preferred settings [28] will be used to collect data on the user's preferred settings. After starting, the user will be given a brief description of the study before the test commences. The test is formed from a memory recall evaluation consisting of 3 sets of 10-15 words and their definitions. Each set of words will be presented one at a time and the order within each set will be randomized between tests; a visualisation of this is shown in figure number 1. The word font, position, size, and color alongside the time each word and its definition is present on the screen will be based on user input for Set 1. For Set 2. An amalgamation of user preferences and optimized settings will be used. For Set 3, optimized settings based on previous research will be used. Set 1, 2 and 3 will in essence serve to answer the questions 3.1 and 3.2 from Problem Formulation. These changes in features will be our independent variables designed to evoke change in the dependent variables mentioned later.

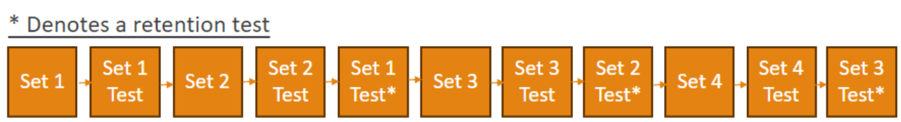

Fig. 1. The sequence of the words within each set will be randomized across tests, and each set will be shown one at a time.

The order in which sets 1,2,3 is shown will be randomized to ensure a valid correlation outside of increased performance over time. Once the words from a single set have been shown and looked over by the user, they will be tested on 2 separate features: Correct identification of definition based on a word, and time necessary to recall for each feature. These are our dependent (measured) variables. Additionally, after the Nth set is completed and the user has been tested on it, they will also be subjected to a recall test for the N-1th set. Since there is no N+1th set to retest the Nth set in order to stimulate working memory and visual imagery [14], a dummy set with different settings will be displayed but not tested for the final set. Once completed, results will be calculated to determine whether preferential, mixed, or optimized settings outperform each other. Additionally, as the time allotted to each card may make a substantial difference a penalty coefficient may be implemented to reduce the score for tests in which a high allotted time per card is used. Finally a correlative value can be calculated between the user chosen features, the ones they performed best on, and the optimized features.

3.1 Participants

In this study, we recruited 16 participants consisting of 13 males and 3 females who were friends and family members of the researchers. The participants' age distribution in this research was as follows: 12 individuals were aged between 18 and 24, 2 participants were over the age of 50, and 2 participants were in the 25–35 age range. This age distribution offers important details regarding the study's data collection. A large majority of the participants were between the ages of 18 and 24, suggesting that the outcomes of the study are likely more characteristic of this age group. Studies have shown that in terms of both the quantity of correct answers and the amount of time, younger adults often do better when recalling words [5]. But the study also benefits from having individuals between the ages of 25 and 35 and those greater than 50. It allows for the inclusion of a larger range of perspectives and experiences in the data and may be able to demonstrate memory deterioration as people age [9].

Regarding education, the participants came from a variety of academic fields, including engineering, computer science, computer information systems, communication, construction management, and finance. This intentionally diverse group of participants was chosen since the study's within-subject design was aimed to assess performance across various contexts. Adults' vocabulary sizes and structures are likely to vary greatly due to variances in education, interests, and skills [23] [6]. The outcomes of the study could be generalized to a wider variety of individuals thanks to this approach, which made sure they weren't restricted to a particular area of expertise [10].

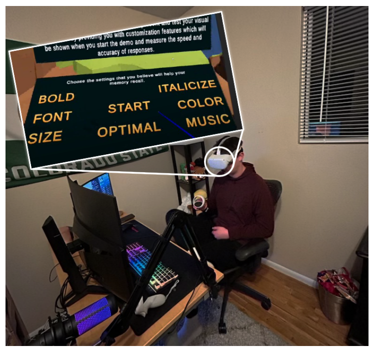

Fig. 2. A participant wearing a virtual reality headset while engaging in the experimental task.

Various methods were used to gather our participants. The most common way individuals were gathered was via word-of-mouth methods such as calls or text; asking friends and family members to participate in our experiment seen in figure number 2. Since our experiment is within-subject, we aimed for a high variety in our subject pool, as opposed to just having everyone participating be a computer science student at the university. This prevents any consistent external or preexisting knowledge from affecting the results of the experiment.

3.2 Experiment Design

This experiment was conducted with a sole focus on within-subject design. There are four sets of 12 terms consisting of two parts: questions and definitions. Since participants are meant to be tested on each of the three different types of settings a within subject design allows for a performance evaluation on an individual basis. The total parameters of the experiment are as follows: subject number, selected memorization settings, total elapsed time per test, and total number of correct responses. Within the memorization settings the adjustable values are as follows: font, color, background color, font size, choice of music and font style (i.e. italicized, bold, default) and music preference. The measured effects of this experiment were selected as total elapsed time per test and total number of correct responses per participant. The experimental variables on the other hand include the subset of adjustable values as well as the broader memorization settings.

Each experiment is split into four sections each representing a word set of 12 questions and definitions. Each section is composed of two modes: training and testing. The training mode of each section presents first a question and then a definition from the respective set to the participant as a flashcard at intervals of 3s per term 7s per definition. Once all questions and answers from the set have been exhausted the testing mode will begin and participants will be presented a question from the previously examined word set and 4 definitions to pick from. Upon clicking a given answer their response will be recorded and another randomly selected question will be tested. Once all questions and definitions from the first set have been tested the settings will be changed to one of the two other types of settings being evaluated. To ensure no learning effects order of the tested settings will be counterbalanced using a Latin Square design. After the settings have been adjusted to the next choice of memorization settings, the following word set will be presented in training mode. Upon completion, the testing mode will begin for the second tested word set. At this point in the test a retention test will occur testing the previously trained and tested set. This cycle will continue until the fourth word set. The fourth word set will also be presented in training and testing mode however no results will be collected from it. This is to allow for a collection of the retention set of the third set while avoiding an infinite cycle trying to hold a retention test for the last tested set of a given group. This concludes each participant's testing experience and results will be collected.

4 RESULTS

Each participant's speed, accuracy, and settings for each test were recorded and compiled once they had all finished the trials. (Table 1) displays the average time for each participant across all tests. The speed metric was recorded in milliseconds (Ms) and reflects just the time taken each testing phase of the trial. Similar to this, each participant's accuracy was evaluated independently for each test, with (Table 4) displaying the corresponding averaged data. Before examining the results of our ANOVA test, several observations may be made based on these tables. It should be noted that the mixed settings appeared to perform best in terms of accuracy, while the user-selected settings generally maintained to offer the best speed metric for users.

No.	Type	Avg Number Correct	Avg Time Per Quiz (Ms)	Avg Elapsed Time (Ms)
1	Optimal	8	3418652	4716590
2	Optimal	4	1093318	1440909
3	Optimal	4	4052579	5202812
4	Optimal	5	1032103	1505021
5	Optimal	8	997462	1298748
6	Optimal	5	1516703	2103047
7	Optimal	7	3856658	4991506
8	Optimal	8	1914535	2683841
9	Optimal	7	5220508	6919855
10	Optimal	7	1637807	2155078
11	Optimal	6	2140444	2896144
12	Optimal	7	2797884	3731574
13	Optimal	3	3257085	4468489
14	Optimal	7	2123322	2881641
15	Optimal	10	2393364	3184385
16	Optimal	4	1758988	2317891

Table 1. Average Optimal Settings for each participant

However, two ANOVA tests were carried out to verify these findings. The results of the ANOVA test, which was only focused on the speed metric overall, indicated that the data had statistical significance, with a resulting p value of. 00005813. The effect size of f was recognized as substantial (0.48), and the F statistic was equal to 10.85. We can see that x1-x2 and x1-x3 do not have statistical significance based on the p-value when the Tukey HSD / Tukey Kramer technique of showing the results of an ANOVA is used, but x2-x3 does (Table 2). This demonstrates that the variations between the optimal and mixed settings utilized to derive the data's statistical significance come from those differences.

Pair	Difference	SE	Q	Lower CI	Upper CI	Critical Mean	p-value
x1-x2	2451391.876	1192093.336	2.0564	-1564046.328	6466830.079	4015438.204	0.3178
x1-x3	9397456.043	1123916.376	8.3613	5611664.598	13183247.49	3785791.444	1.66E-07
x2-x3	6946064.167	888533.9117	7.8174	3953133.239	9938995.095	2992930.928	8.81E-07

Table 2. ANOVA: Speed Metric (Ms): x1 represents user preference, x2 is optimized settings and x3 is mixed.

4.1 Mixed Settings

No.	Type	Avg Number Correct	Avg Time Per Quiz (Ms)	Avg Elapsed Time (Ms)
1	Mixed	12	7609578	17814493
2	Mixed	11	3236362	6901473
3	Mixed	9	7132929	19058432
4	Mixed	8	3643362	7457213
5	Mixed	11	4987976	8193402
6	Mixed	11	3271804	7761769
7	Mixed	8	7775329	18983883
8	Mixed	9	3983116	9740874
9	Mixed	11	13095038	30458117
10	Mixed	8	3698394	8531945
11	Mixed	11	3989913	10138333
12	Mixed	11	5941140	14280176
13	Mixed	9	7418514	16531541
14	Mixed	9	5656725	12148679
15	Mixed	12	4259967	10826618
16	Mixed	12	4302201	9478297

Table 3. Average Mixed Settings

Based solely on the accuracy metric, mixed settings performed the best in terms of total recalling accuracy, increasing accuracy by 39% compared to optimal settings and 24% compared to user settings. ANOVA was also performed on the time metric to see whether there was any significant relationship between the setting type and completion rate. The findings of the ANOVA were statistically significant in x1-x3, x2-x3, but not x1-x2 when just time was taken into consideration. The F statistic is 24.912 and the p value is 2.168 e-9.

Pair	Difference	SE	Q	Lower CI	Upper CI	Critical Mean	p-value
x1-x2	0.875	0.5292	1.6534	-0.9076	2.6576	1.7826	0.4745
x1-x3	1.6667	0.499	3.3403	-0.014	3.3473	1.6807	0.05244
x2-x3	2.5417	0.3945	6.4435	1.213	3.8704	1.3287	0.00004669

Table 4. ANOVA: Correct Response Metric: x1 represents user preference, x2 is optimized settings and x3 is mixed.

4.2 User-Preferred Settings

The user-preferred settings significantly outperformed the mixed settings, and the ideal settings also had the same consequence, but to a lesser extent, when the data was evaluated. The observed effect size was substantial (0.73). In this case, the user-selected parameters performed the best with an improvement of 400% over the ideal settings and 1200% over the mixed settings when time is the only factor considered. But when both time and accuracy are taken into consideration, user settings indicate a statistically significant improvement in both fields, with an average accuracy of 7.375 (as opposed to 9.04) and an average time of 830382.4375 (as opposed to 10227838.48) milliseconds.

No.	Type	Number Correct	Time Per Quiz (Ms)	Elapsed Time (Ms)
1	User Preference	5	1297938	1297938
2	User Preference	10	347591	347591
3	User Preference	8	1150233	1150233
4	User Preference	7	472918	472918
5	User Preference	10	301286	301286
6	User Preference	5	586344	586344
7	User Preference	4	1134848	1134848
8	User Preference	9	769306	769306
9	User Preference	7	1699347	1699347
10	User Preference	8	517271	517271
11	User Preference	7	755700	755700
12	User Preference	9	933690	933690
13	User Preference	7	1211404	1211404
14	User Preference	5	758319	758319
15	User Preference	6	791021	791021
16	User Preference	11	558903	558903

Table 5. Average User-Preferred Settings

In summary, the results were statistically significant when just time was considered, with significant differences between user-selected and mixed settings as well as between the optimal and mixed settings. In this case, user-selected options outperform optimal and mixed settings with 400–1200% quicker response times, respectively. The results were statistically significant once again when accuracy alone was taken into consideration, with substantial differences between user settings and mixed settings along with user settings and optimal settings. Here, when compared to optimized and mixed settings, the accuracy of the mixed settings performed 24–39% better. User settings performed best when accuracy and speed were considered collectively, with a 24% decrease in accuracy from mixed settings but a 1200% increase in overall recall time.

5 DISCUSSION

Based on the previous analysis of our results, several discussion topics of note were brought up regarding the best-performing settings type for each of the reviewed metrics (speed, accuracy, and both). One key question is why the mixed settings set performed remarkably better in accuracy but incomprehensibly worse in time. While the mixed settings increased in accuracy by 24% over user-selected settings, speed decreased by nearly 1,200%. One possible answer for this lies in the way in which each setting set was selected. The user-selected settings obviously had a lot of variance due to different users selecting different text settings, whereas the optimized settings are static and have very little variance between users. However the mixed settings were generated randomly by swapping ½ of the features selected in user settings with optimized settings. This overall could have generated a larger variance and contributed to the speed deficit. Another possible reason for this may be more simply explained as mixed settings having no cohesion. This may have made the mixed settings distracting and the lack of cohesion could cause a longer time to recall the information. Similarly, this phenomenon may have been what aided the recall accuracy for the mixed settings due to the suspected memory benefits of obscurification. Another point of interest was our selected number of features. It is possible that some of the experiment's results could be the cause of the number of text options and features being too high for the number of participants whose data was collected. In total, there are 972 possible combinations for user settings and after combining these into mixed settings there are 19,440 possible combinations for user settings and mixed settings. In our experiment, 16 participants was not enough to accurately address all of the possible combinations.

5.1 Limitations of the Study

Various limitations arose throughout the study. Considering our music being played was using the Wizard-of-Oz effect, and our optimal music settings was 'Classical', we needed to ensure that the mixed settings was in fact a mixture. The issue with this is that there was no way to tell what type of music should be currently playing. We knew that for the first set that it should be the user's choice. We would need to pay attention to what type of music they chose, change it to classical for the optimized sections, and determine whether we would continue with either their classical or preferred type of music. But after that our data would assign the variable to either classical or whatever the user had chosen, and without any given indicator of what should be playing, we had to choose what the user was experiencing. For the most part, we kept the audio to classical once the optimized settings had run through, as otherwise we would have to go in and manually change all the data entry points as they may or may not match whatever we as humans had chose.

Another limitation of the study was our environment. While we set out to find how a virtual world would effect the subjects' performance, we only had enough time to create a single, peaceful environment. Our research showed that an engaging virtual environment was the best for recall[12]. While this was a good finding, the term 'engaging' is very abstract and the meaning can change from person to person. So we decided upon something familiar to us in Colorado, a peaceful mountain valley with trees surrounding that draws your attention; keeping the focus on the flash cards directly in front of the user. While this is the ideal virtual environment, time was not allotted to go more in-depth with different worlds that the user may find more distracting or distraught. For instance, some studies showed that students' conceptual and spatial learning benefited from fewer distractions in the virtual world [35, 39]. It would have been advantageous to introduce this variation into the experiment to hopefully observe similar findings. We also wanted to keep the variables limited as our focus was on user preferences versus optimization, and this addition could have caused unnecessary noise in our study.

6 ACKNOWLEDGMENTS

This experiment would not have been possible without support from Professor Ortega who allowed us to borrow a VR headset and guided our research and development to help us succeed. We are grateful to all of our participants for taking part in the experiment and for their willingness to volunteer their time so we could gather valuable information from them.

REFERENCES

[1] Christina Bachmann and Lauro Mengheri. 2018. Dyslexia and Fonts: Is a Specific Font Useful? *Brain Sciences* 8, 89 (2018). https://doi.org/10.3390/brainsci8050089

[2] William R. Balch, Kelly Bowman, and Lauri A. Mohler. 1992. Music-dependent memory in immediate and delayed word recall. *Memory Cognition* 20, 1 (1992). https://doi.org/10.3758/BF03208250

[3] James Bennett-Levy and Graham E. Powell. 1980. The Subjective Memory Questionnaire (SMQ). An investigation into the self-reporting of 'real-life' memory skills. *The British journal of social and clinical psychology* 19 (1980), 177–188.

[4] Jeffrey Beyon and Chante Cox-Boyd. 2020. The Influence of Font Type and Color on Online Information Recall. *North American Journal of Psychology* 22, 1 (2020). link.gale.com/apps/doc/A618129321/HRCA?u=anon~9bb4cd26&sid=googleScholar&xid=47d64da9

[5] Nancy L. Bowles and Leonard W. Poon. 1985. Aging and Retrieval of Words in Semantic Memory. *Journal of Gerontology* 40, 1 (01 1985), 71–77. https://doi.org/10.1093/geronj/40.1.71 arXiv:https://academic.oup.com/geronj/article-pdf/40/1/71/1927134/40-1-71.pdf

[6] Marc Brysbaert, Michaël Stevens, Paweł Mandera, and Emmanuel Keuleers. 2016. How Many Words Do We Know? Practical Estimates of Vocabulary Size Dependent on Word Definition, the Degree of Language Input and the Participant's Age. *Frontiers in Psychology* 7 (2016). https://doi.org/10.3389/fpsyg.2016.01116

[7] Kyle Jeshua Casumbal, Matthew So, Anneliese Ng, Francesca Guzman, Nina Fernandez, and Chloe Chan. 2019. The Effects of Low-Fidelity Music and Font Style on Recall. (12 2019). https://doi.org/10.13140/RG.2.2.31182.41286

[8] Michael S. Cohen, Jesse Rissman, Mariam Hovhannisyan, Alan D. Castel, and Barbara J. Knowlton. 2017. Free Recall Test Experience Potentiates Strategy-Driven Effects of Value on Memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 43 (2017), 1581–1601.

[9] Fergus I. M. Craik and Joan M. Mcdowd. 1987. Age differences in recall and recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition* 13 (1987), 474–479.

[10] Meredyth Daneman and Patricia A. Carpenter. 1980. Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior* 19, 4 (1980), 450–466. https://doi.org/10.1016/S0022-5371(80)90312-6

[11] Amitabh Varshney eric Krokos, Catherine Plaisant. 2019. Virtual memory palaces: immersion aids recall. *Virtual Reality* 23, 15 (2019), 1–15. https://doi.org/10.1007/s10055-018-0346-3

[12] Katherine Franceschi, Ronald M. Lee, Stelios H. Zanakis, and David Hinds. 2009. Engaging Group E-Learning in Virtual Worlds. *Journal of Management Information Systems* 26, 1 (2009), 73–100. https://doi.org/10.2753/MIS0742-1222260104 arXiv:https://doi.org/10.2753/MIS0742-1222260104

[13] Vered Halamish, Hila Nachman, and Tami Katzir. 2018. The Effect of Font Size on Children's Memory and Metamemory. *Frontiers in Psychology* 9 (2018). https://doi.org/10.3389/fpsyg.2018.01577

[14] Carlos M. Hamamé, Juan R. Vidal, Tomás Ossandón, Karim Jerbi, Sarang S. Dalal, Lorella Minotti, Olivier Bertrand, Philippe Kahane, and Jean-Philippe Lachaux. 2012. Reading the mind's eye: Online detection of visuo-spatial working memory and visual imagery in the inferior temporal lobe. *NeuroImage* 59, 1 (2012), 872–879. https://doi.org/10.1016/j.neuroimage.2011.07.087 Neuroergonomics: The human brain in action and at work.

[15] Joel Harman, Ross Brown, Daniel Johnson, and Selen Turkay. 2020. The Role of Visual Detail during Situated Memory Recall within a Virtual Reality Environment. In *Proceedings of the 31st Australian Conference on Human-Computer-Interaction* (Fremantle, WA, Australia) *(OzCHI '19)*. Association for Computing Machinery, New York, NY, USA, 138–148. https://doi.org/10.1145/3369457.3369469

[16] III Henry L. Roediger and Jeffrey D. Karpicke. 2006. Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention. *Psychological Science* 17, 3 (2006), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x arXiv:https://doi.org/10.1111/j.1467-9280.2006.01693.x PMID: 16507066.

[17] Sarah Judde and Nikki Rickard. 2010. The effect of post-learning presentation of music on long-term word-list retention. *Neurobiology of Learning and Memory* 94, 1 (2010), 13–20. https://doi.org/10.1016/j.nlm.2010.03.002

[18] Jeffrey D. Karpicke and Henry L. Roediger. 2008. The Critical Importance of Retrieval for Learning. *Science* 319, 5865 (2008), 966–968. https://doi.org/10.1126/science.1152408 arXiv:https://www.science.org/doi/pdf/10.1126/science.1152408

[19] Liu J. Khan. 2020. The impact of colors on human memory in learning English collocations: evidence from south Asian tertiary ESL students. *Asian-Pacific Journal of Second and Foreign Language Education* 5, 17 (2020). https://doi.org/10.1186/s40862-020-00098-8

[20] Maarten H. Lamers and Maik Lanen. 2021. Changing Between Virtual Reality and Real-World Adversely Affects Memory Recall Accuracy. *Frontiers in Virtual Reality* 2 (2021). https://doi.org/10.3389/frvir.2021.602087

[21] Kung Wong Lau and Pui Yuen Lee. 2015. The use of virtual reality for creating unusual environmental stimulation to motivate students to explore creative ideas. *Interactive Learning Environments* 23, 1 (2015), 3–18. https://doi.org/10.1080/10494820.2012.745426 arXiv:https://doi.org/10.1080/10494820.2012.745426

[22] Chen Audrey H. Lorch, Robert F. 1986. Effects of number signals on reading and recall. *Journal of Educational Psychology* 78, 4 (1986), 263–270. https://doi.org/10.1037/0022-0663.78.4.263

[23] Nina Mainz, Zeshu Shao, Marc Brysbaert, and Antje S. Meyer. 2017. Vocabulary Knowledge Predicts Lexical Processing: Evidence from a Group of Participants with Diverse Educational Backgrounds. *Frontiers in Psychology* 8 (2017). https://doi.org/10.3389/fpsyg.2017.01164

[24] Tasnim Makada, Danyal Ozair, Mudassir Mohammed, and Cheryl Abellanoza. 2016. Enhancing Memory Retention by Increasing Alpha and Decreasing Beta Brainwaves Using Music. In *Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments* (Corfu, Island, Greece) *(PETRA '16)*. Association for Computing Machinery, New York, NY, USA, Article 60, 4 pages. https://doi.org/10.1145/2910674.2935851

[25] Robert J. Matheis, Maria T. Schultheis, Lana A. Tiersky, John DeLuca, Scott R. Millis, and Albert Rizzo. 2007. Is Learning and Memory Different in a Virtual Environment? *The Clinical Neuropsychologist* 21, 1 (2007), 146–161. https://doi.org/10.1080/13854040601100668 arXiv:https://doi.org/10.1080/13854040601100668 PMID: 17366282.

[26] Michael L. Mueller, John Dunlosky, Sarah K. Tauber, and Matthew G. Rhodes. 2014. The font-size effect on judgments of learning: Does it exemplify fluency effects or reflect people's beliefs about memory? *Journal of Memory and Language* 70 (2014), 1–12. https://doi.org/10.1016/j.jml.2013.09.007

[27] Adin Nelson and Kinga L. Eliasz. 2023. Desirable Difficulty: Theory and application of intentionally challenging learning. *Medical Education* 57, 2 (2023), 123–130. https://doi.org/10.1111/medu.14916 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/medu.14916

[28] Steven Pan, Inez Zung, Megan Imundo, Xuxin Zhang, and Yunning Qiu. 2022. User-generated digital flashcards yield better learning than premade flashcards. *Journal of Applied Research in Memory and Cognition* (12 2022). https://doi.org/10.1037/mac0000083

[29] Thomas D. Parsons and Albert A. Rizzo. 2008. Initial Validation of a Virtual Environment for Assessment of Memory Functioning: Virtual Reality Cognitive Performance Assessment Test. *CyberPsychology & Behavior* 11, 1 (2008), 17–25. https://doi.org/10.1089/cpb.2007.9934 arXiv:https://doi.org/10.1089/cpb.2007.9934 PMID: 18275308.

[30] Gaën Plancher, Julien Barra, Eric Orriols, and Pascale Piolino. 2013. The influence of action on episodic memory: A virtual reality study. *The Quarterly Journal of Experimental Psychology* 66, 5 (2013), 895–909. https://doi.org/10.1080/17470218.2012.722657 arXiv:https://doi.org/10.1080/17470218.2012.722657 PMID: 23025821.

[31] Tamara M. Rosner, Hanae Davis, and Bruce Milliken. 2015. Perceptual blurring and recognition memory: A desirable difficulty effect revealed. *Acta Psychologica* 160 (2015), 11–22. https://doi.org/10.1016/j.actpsy.2015.06.006

[32] Steven Smith and Edward Vela. 2001. Environmental context-dependent memory: A review and meta-analysis. *Psychonomic Bulletin Review* 8 (2001). Issue 2. https://doi.org/10.3758/BF03196157

[33] Melanie C. Steffens, Rul von Stülpnagel, and Janette C. Schult. 2015. Memory Recall After "Learning by Doing" and "Learning by Viewing": Boundary Conditions of an Enactment Benefit. *Frontiers in Psychology* 6 (2015). https://doi.org/10.3389/fpsyg.2015.01907

[34] Behzad Tabibian, Utkarsh Upadhyay, Abir De, Ali Zarezade, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. 2019. Enhancing human learning via spaced repetition optimization. *Proceedings of the National Academy of Sciences* 116, 10 (2019), 3988–3993. https://doi.org/10.1073/pnas.1815156116 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1815156116

[35] Hakan Tüzün and Fatih Özdinç. 2016. The effects of 3D multi-user virtual environments on freshmen university students' conceptual and spatial learning and presence in departmental orientation. *Computers Education* 94 (2016), 228–240. https://doi.org/10.1016/j.compedu.2015.12.005

[36] Monika Undorf and Malte Zimdahl. 2018. Metamemory and Memory for a Wide Range of Font Sizes: What is the Contribution of Perceptual Fluency? *Journal of experimental psychology. Learning, memory, and cognition* 45 (04 2018). https://doi.org/10.1037/xlm0000571

[37] Courtney Upchurch. 2017. The Effect of Font Type on Memory for Instruction. *Honors Theses* (2017), 1–20. https://egrove.olemiss.edu/hon_thesis/232

[38] Shaun Wallace, Zoya Bylinskii, Jonathan Dobres, Bernard Kerr, Sam Berlow, Rick Treitman, Nirmal Kumawat, Kathleen Arpin, Dave B. Miller, Jeff Huang, and Ben D. Sawyer. 2022. Towards Individuated Reading Experiences: Different Fonts Increase Reading Speed for Different Individuals. *ACM Trans. Comput.-Hum. Interact.* 29, 4, Article 38 (mar 2022), 56 pages. https://doi.org/10.1145/3502222

[39] Wenxue Wang. 2022. Research on the Influence of High and Low Spatial Ability on VR Panoramic Video Learning Task. In *2022 13th International Conference on E-Education, E-Business, E-Management, and E-Learning (IC4E)* (Tokyo, Japan) *(IC4E 2022)*. Association for Computing Machinery, New York, NY, USA, 9–15. https://doi.org/10.1145/3514262.3514321

[40] Chunliang Yang. 2018. Enhancing learning and retrieval of new information: a review of the forward testing effect. *Science Learn.* https://doi.org/10.1038/s41539-018-0024-y

Received 27 March 2023