What is the Most Effective Way to Learn: Spaced or Mass Studying?

Keane Johnson, Dan Lee, Anu Sankar

School of Information, University of California, Berkeley
W241: Experiments and Causality

August 12, 2019

Abstract

We have experimented with two different study methods “spaced” learning vs. “mass” learning (cramming) to learn vocabulary from a fictional language. We found that spaced learners score higher and our results concur with some of the previous studies that have compared these two approaches. Though the spaced learners scored higher, our study interestingly also found that their speed of recall was slightly slower, suggesting that spaced learners may also be recalling information from long-term memory.

Background

One of the most persistent activities through life is learning. Although it is most associated with adolescence and school, learning is necessary as one moves through their career and takes on different roles, as well as in one’s leisure time in the pursuit of new hobbies or interests. Its constant presence has led to debate on what is the best, most time-effective method to learning. The two predominant schools of thought are: studying in a single longer session, and studying in shorter sessions spaced over multiple days.

Prior research has suggested that multiple spaced learning sessions are more effective than a single massed studying session. Proposed reasons for this range from the positive effects of sleep before and after learning a new task (Division of Sleep Medicine at Harvard Medical School, 2007), to a stronger memory trace built over time through reinforced cellular and molecular pathways (Smolle, Zhang, and Byrne, 2016).

Past experiments have explored this domain by having participants learn words using a web-based application (Kornell, 2009). Kornell found that spaced studying was more effective for 90% of subjects. The aim of this paper is to replicate the results produced by the current research and demonstrate through a randomized experiment that spaced studying is more effective than massed studying.

Research Question

Do multiple shorter study sessions - spaced studying - produce better learning outcomes in comparison to one longer study session - massed studying?

The prevailing theory is that shorter, spaced study sessions are more effective. Our hypothesis is in line with this prior research and we expect participants assigned to spaced study sessions to have better recall than participants assigned to the massed studying session.

Experiment Overview / Study Design

We implemented a crossover study design with repeated measures to test this hypothesis. A crossover study is a randomized, controlled experiment that administers a series of different treatments to participants. The repeated measures enables researchers to collect data on participants after each treatment.

In this experiment, the different treatments were the two different studying styles: spaced and massed. The repeated measures were assessments to test participants' knowledge after receiving the prescribed treatment. We measured their knowledge in two different ways: accuracy and speed. Accuracy was calculated as the percentage of correct answers for each assessment, out of a total of fifteen questions. Speed was calculated as the average time, in seconds, it took to answer each assessment question

In order to test the effectiveness of the learning styles, it was crucial to have participants learn a subject of which they had no, or very limited, prior experience. We selected High Valyrian because it is a fictional language, minimizing participants' previous knowledge. We also believed its use in the popular Game of Thrones television series would entice individuals to participate in the study.

Figure 1 illustrates the experimental design and setup. The experiment was divided into two four-day phases, with differing treatment depending on the phase. In the first phase, both groups studied from the "Basics 1" lesson. One group was instructed to study a specific number of High Valyrian sub-modules from Basics 1 every day for three days, while the other group was instructed to study all the sub-modules on the third day. In the second phase, a new lesson was introduced and the studying treatment was flipped. The first group studied all modules from the "Phrases 1" lesson on the third day and the second group studied a specific number of sub-modules from "Phrases 1" each day over the three days.

Grp.	Assign.	Pre-Treat	July 16	July 17	July 18	July 19	July 20	July 21	July 22	July 23
1	R	O	X_{1S}	X_{1S}	X_{1S}	O			X_{1M}	O
2	R	O			X_{2M}	O	X_{2S}	X_{2S}	X_{2S}	O

Figure 1. Experimental Design

Project Timeline

Although the experimental phase of this project was conducted over the course of eight days, planning and recruiting occurred over multiple weeks before the experiment began. On July 1st, we sent the first of a series of emails and slack messages recruiting individuals to our experiment. Two weeks later, on July 16th, we began the experiment. The majority of our analysis was conducted between July 25th and August 4th, with the final paper written between August 5th and August 12th.

Start Recruitment	Welcome and Instructions	Phase 1 Studying	Phase 1 Assessment	Phase 2 Studying	Phase 2 Assessment	Analysis, Awards
July 1-12	July 16	July 16-18	July 19	July 20-22	July 23	July 24 - August 12

Figure 2. Project Timeline

Discussion of Tools Used

With email as a communication tool, we also used Duolingo for studying and Qualtrics to present instructions, record participation, and test learning. Duolingo is an online learning platform that enables students to learn new languages. We selected Duolingo over other language-learning platforms because it is free, the lessons are modular and can be broken down into short segments, and it already had a program for High Valyrian.

Figure 3 presents the Duolingo homepage. Clicking "START" takes the participant into the next learning module to be studied. As seen in **Figure 4,** the module presents words and phrases in High Valyrian and a set of words in English to the participant. The words can be delivered out loud by clicking on the High Valyrian word or the speaker icon. In order to progress through the module, the student must match the High Valyrian words to their correct English meaning in the correct order.

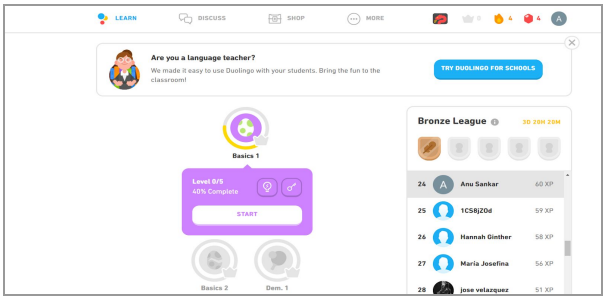

Figure 3. Screenshot of the Duolingo homepage.

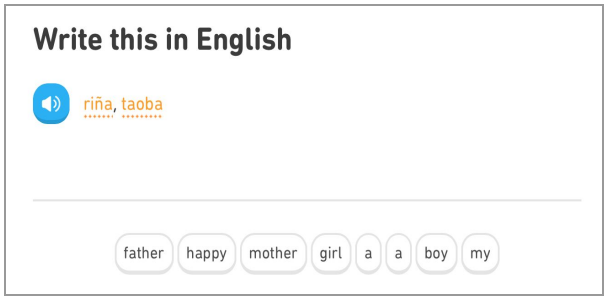

Figure 4. Example of a Duolingo learning exercise.

We used Qualtrics to deliver instructions, record participation and test learning. We used survey-type forms to deliver instructions because it enabled us to record who was reading

and, presumably, following our instructions. This allowed us to collect data on our participants' compliance. Qualtrics was an optimal choice for testing our participants' learning because it provided flexibility in crafting our assessments, and permitted us to record how long each participant took to answer each question of their assessment. **Figure 5** presents an example set of study instructions delivered via Qualtrics and **Figure 6** presents an example assessment question.

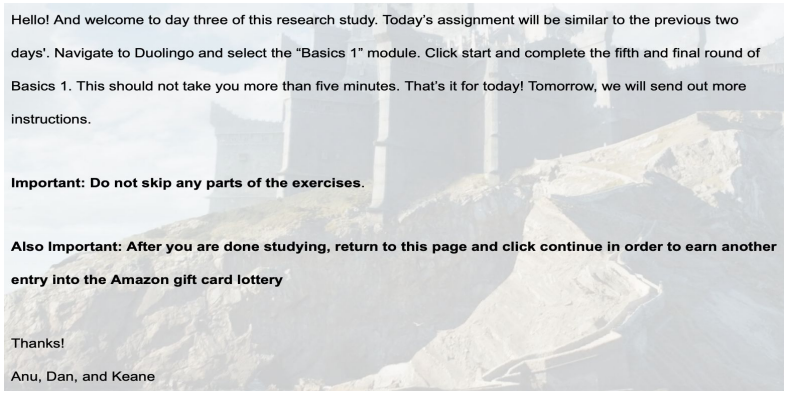

Figure 5. Example of study instructions delivered via Qualtrics.

Figure 6. Example of assessment question delivered via Qualtrics.

Recruitment Process

We recruited friends and family as the subjects for this experiment through our personal network and also through the iSchool Slack channels. Our invitation to the participants included a brief message and a link to a qualtrics survey to register for the experiment. The registration was open for twelve days and the experiment ran for a duration of eight days with each group participating only on six of those days. In order to attract participants, we announced that we would raffle ten $25 Amazon gift cards among the participants. Personalized email messages were sent to prospective participants from our personal network. Here is the screenshot of a more generic recruitment message posted on one of the Slack channels.

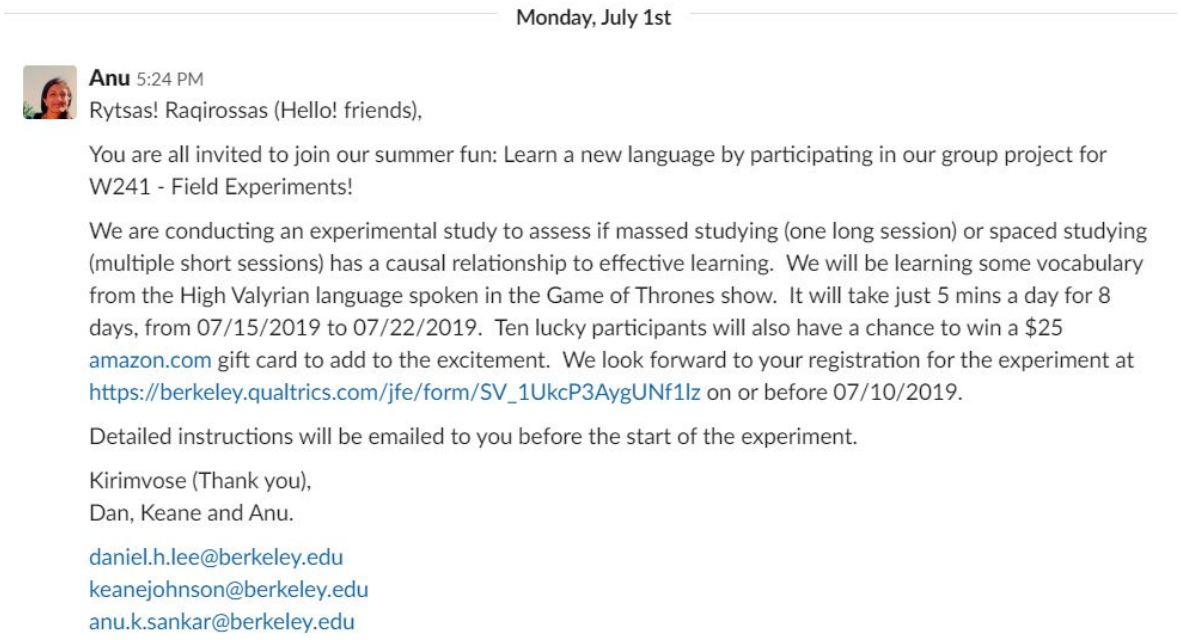

Figure 7. Example of recruitment message.

As we ran the experiment for eight days, it was crucial to keep the subjects motivated over the entire duration of the study. We encouraged their participation by reminding them to submit their name for an additional entry into the lottery after completing their daily learning task (treatment).

Enrollment

Inclusion Criteria

We had not specified an inclusion criteria in our recruitment message, but the participants in our study were all adults, 18 years or older. This occurred naturally because we had tapped into only our friends and family circle. We did not impose any geographical limitations on the subjects' location. The location information was collected as part of the metadata in our recruitment survey. This was useful just to allow for extra time to those who lived outside the US.

Exclusion Criteria

We had not specified an exclusion criteria either. We welcomed participation from anyone interested in learning a new language and could spare about five minutes a day for six days towards that activity.

Pre-treatment Covariates

Early in the project, we decided to collect some basic information about the participants in the recruitment survey so that we could use some of that data later on as covariates in our regression models. In addition to name and email address, the participants were asked to

submit their age group, highest educational degree earned, the number of languages they have basic proficiency in, a list of those languages, familiarity with the Game of Thrones TV show and self-assessed knowledge level in High Valyrian. The answers to these questions might correlate highly with the ability to learn and recall.

Randomization and Blocking

Forty-nine persons had registered to participate in our field experimental study. Based on the pre-treatment covariates that we collected from the recruitment survey, we decided to block only on prior High Valyrian knowledge and then randomize the treatment group assignment. Five participants had marked their knowledge level of High Valyrian as "Intermediate" (can recognize a few words and phrases). The remaining forty four had identified themselves as a "Beginner" (couldn't recognize it outside the Game of Thrones show) or as having no knowledge of High Valyrian. None of the participants had a knowledge level of "Advanced" (recognize over 10 words and phrases, speak some) or "Proficient". **Figure 8** shows the code used to generate the blocked random assignment of treatment which resulted in two groups.

````
## Data Import

```{r import, echo=FALSE}
# Import participant data
d <- fread('../data/interim/participants_cl.csv', header = TRUE) # Participant pre-treatment survey
head(d)
```

## Blocked Randomization

```{r randomize, echo=FALSE}
d <- d[order(-pre_hv)]
pre_hv_obs <- sum(d[ ,pre_hv])
total_obs <- d[, .N]

randomize_blocked <- function(){
  c(sample(c(rep(0, round(pre_hv_obs/2)), rep(1, pre_hv_obs - round(pre_hv_obs/2)))), #group A
    sample(c(rep(0, round((total_obs- pre_hv_obs)/2)), rep(1, round((total_obs- pre_hv_obs)/2))))) #group B
}

# Assignment treatment
d <- d[ , treatment := randomize_blocked()]

# Check random assignment
table(d[ ,pre_hv], d[ ,treatment])
```

### Export cleansed data to interim folder

```{r export, echo=FALSE}
fwrite(d, "../data/interim/participants_cl_blkrand.csv")
```
````

Figure 8. Block Randomization Code.

The first group with twenty four members were assigned the treatment of spaced and then massed studying. The second group with twenty five members were assigned to massed and then spaced studying treatment.

Randomization Check: Covariate Balance

After blocked random assignment to treatment, we performed a covariate balance check using linear regression models. The base model has the outcome variable "treatment" regressed on a numeric constant value of 1. The second model has the outcome variable "treatment" regressed on the previous knowledge of High Valyrian (pre_hv), age groups, education and number of languages known currently. The stargazer report below shows that none of the covariates have a statistically significant coefficient.

	Dependent variable:	
	treatment	
	(1)	(2)
pre_hv		0.227
		(0.262)
age25 - 34		-0.624
		(0.550)
age35 - 44		-0.433
		(0.569)
age45 - 54		-0.839
		(0.631)
age65 - 74		-0.922
		(1.078)
education4 year degree		0.161
		(0.631)
educationDoctorate		0.636
		(0.834)
educationGraduate degree		0.147
		(0.637)
num_languages2		-0.149
		(0.199)
num_languages3		-0.197
		(0.231)
num_languages4		0.154
		(0.367)
num_languages5 or more		0.299
		(0.569)
Constant	0.510***	0.987
	(0.072)	(0.849)
Observations	49	49
R2	0.000	0.182
Adjusted R2	0.000	-0.090
Residual Std. Error	0.505 (df = 48)	0.527 (df = 36)
F Statistic		0.668 (df = 12; 36)
Note:		*p<0.1; **p<0.05; ***p<0.01
Analysis of Variance Table		

Figure 9. Table for Randomization / Covariate Balance Check.

An Anova test in R to compare the two models did not have a statistically significant p-value and therefore we can conclude that the covariates are balanced and that the randomization of treatment assignment was successful.

```
Model 1: treatment ~ 1
Model 2: treatment ~ pre_hv + age + education + num_languages
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     48 12.245
2     36 10.014 12    2.2306 0.6682 0.7694
```

Figure 10. Anova test for Randomization / Covariate Balance Check.

Data Collection and Processing

Upon completion of our experimental study, we were able to compile descriptive and demographic data, treatment compliance, and assessment performance from the Qualtrics platform for each participant. Descriptive and demographic data included age, level of education, number of languages spoken, pre-familiarity with Game of Thrones and the High Valyrian language, and metadata on their experimental experience, such as the browser they used for studying and whether they used a mobile device or desktop device. Many of these covariates were used to control for fixed effects. Treatment compliance was approximated using the completion survey captured at the end of each day's activity. Finally, to capture assessment performance, we generated two main outcome variables -- accuracy and speed -- for each participant and assessment. These served as the basis for two OLS regression models discussed in detail below.

Data Completeness

As seen in **Figure 11**, our initial experimental field of 49 participants was reduced to 24 participants who completed both Assessment 1 and Assessment 2. We concluded that the most significant attrition (20 participants) occurred during the week prior to administering Assessment 1, indicating that these participants never intended to participate in the study and could thus be excluded from the experiment altogether as "non-participants".

Of the 29 participants who completed Assessment 1, only 24 also completed Assessment 2. Because we did not have a complete set of matched observations for 5 individuals in our remaining group, we have defined attrition accordingly as those who did not complete both assessments. Ultimately, this resulted in 27% attrition for treatment group 1 and 7% attrition for treatment group 2.

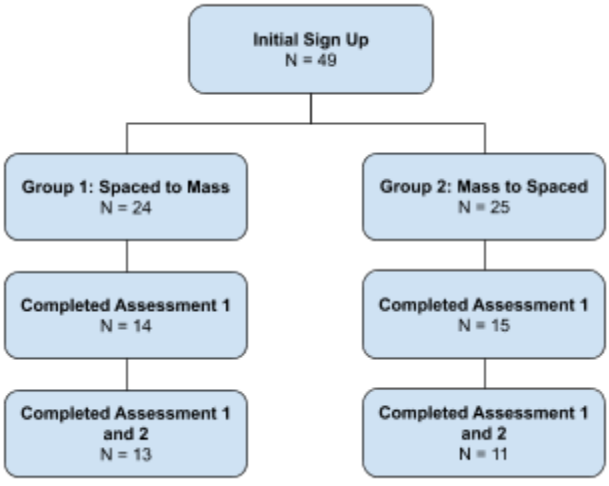

Figure 11. Treatment vs. Control Flow Diagram.

Results

Initial Exploratory Data Analysis

Initial EDA of Assessment 1 and Assessment 2 outcomes revealed slight differences in the distribution of scores for participants assigned to spaced studying vs. mass studying. As seen in **Figure 12,** the distribution of Assessment 1 scores suggest that mass studying is more effective for a subset of participants, but on average produced lower scores with higher variability across the entire mass studying group, as compared to the spaced studying group. Interestingly, the effects are reversed for Assessment 2, where spaced studying appears to produce lower scores on average with higher variability in scores.

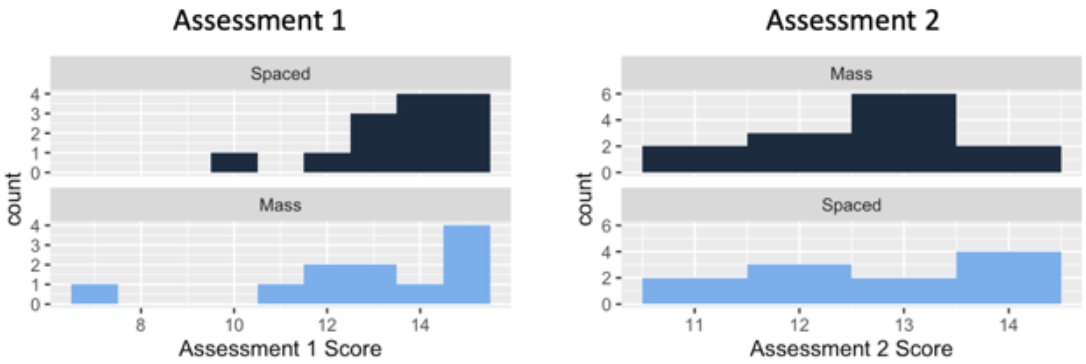

Figure 12. Distribution of Scores across Assessment 1 and Assessment 2.

Figure 13 illustrates the distribution of the outcome variable speed across Assessment 1 and Assessment 2. These charts indicate significant differences in the distribution of answer times for participants assigned to spaced studying vs. mass studying, and are consistent with our hypothesis in that they suggest that mass studying produces significantly lower answer times than spaced studying.

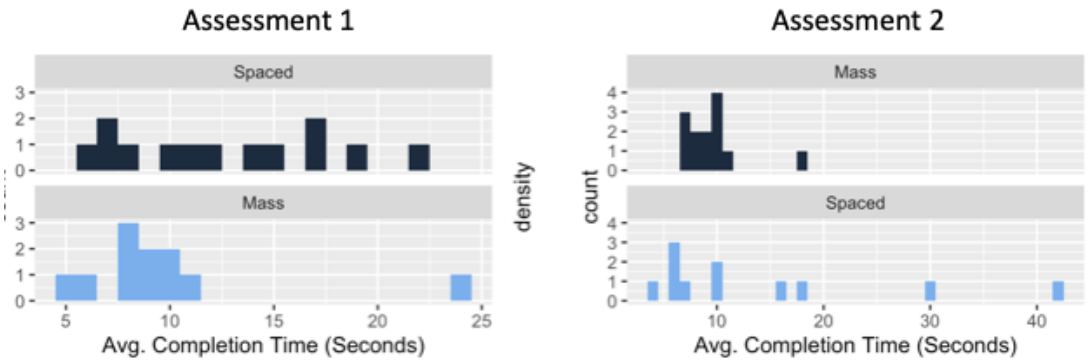

Figure 13. Distribution of Speed across Assessment 1 and Assessment 2.

Analysis of Point Estimates

Our crossover study design allowed us to measure treatment effects for the same participant across two different treatment conditions, and for our treatment groups within the same assessment wave. This provided multiple points of comparison from which to derive conclusions on the treatment effect. The crossover design provided the additional benefit of controlling for confounders, since each participant provided an observation under each treatment condition.

Figure 14 illustrates the analysis setup, with each τ representing an effect size between our spaced and mass studying groups. The treatment groups are measured against each other in τ_1 and τ_2, and participants are measured against each other across treatments in τ_3 and τ_4. All effects are measured as spaced minus mass to retain consistency in the direction of the effect.

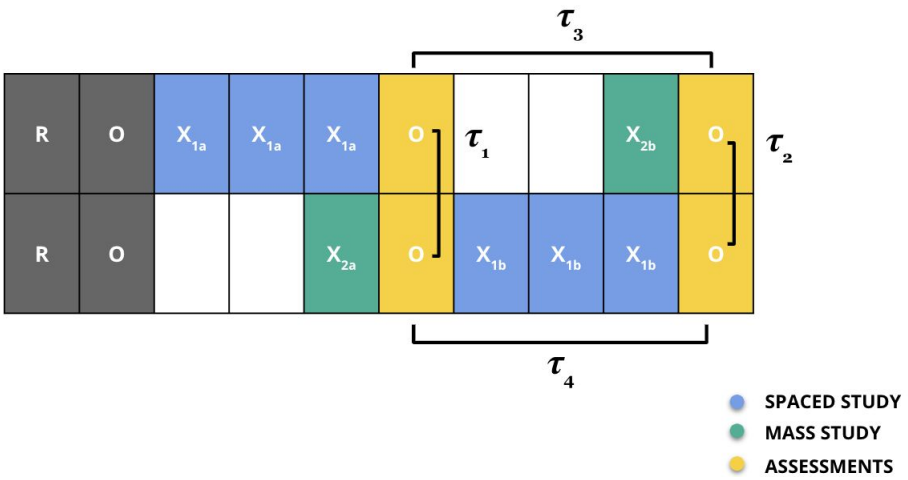

Figure 14. Setup of Point Estimate Analysis

The full results of this analysis are reflected in **Figure 15.** Average treatment compliance for each day of studying (α) was captured for each treatment group and applied to derive a Complier Average Causal Effect (CACE) in addition to the Intent to Treat (ITT) effect. The average compliance rate for Spaced to Mass learners was 0.82, while Mass to Spaced learners was 0.87. For the outcome variable of scores, these point estimates suggest an average ITT (CACE) of 0.41 (0.46)**,** indicating spaced learners score 0.41 points higher than their mass studying peers. For the outcome variable of speed, they show an ITT (CACE) of 3.71 (4.53), indicating spaced learners take almost 4 seconds longer to answer questions, on average, compared to their mass studying peers.

Figure 15 also includes a row comparing the effects between τ_3 and τ_4, which is useful to understand the composite effects of the ordering of treatments. The observed difference in ITT and CACE for scores and times between these two point estimates suggest that there is some significance in the ordering of treatment, i.e. whether mass studying or spaced

studying is conducted first. These results suggest that spaced to mass studying results in higher accuracy and lower speeds than mass to spaced studying. Further discussion of the potential drivers of these outcomes are explored in detail in the sections below.

τ	ITT (Scores)	CACE (Scores)	ITT (Speed)	CACE (Speed)
1	0.71	-0.04	2.95	3.16
2	-0.18	-0.22	4.2	5.13
3	1	1.16	3.21	3.71
4	0.11	0.97	4.46	6.1
Avg: 1 through 4	**0.41**	**0.46**	**3.71**	**4.53**
3 minus 4	**.89**	**0.19**	**-1.25**	**-2.39**

Figure 15. Summary of Point Estimate Analysis.

Regression Analysis and Results

To validate the experimental outcomes and incorporate the effects of additional covariates, we transformed the dataset into a tall representation to enable analysis via OLS regression. Figure 16 illustrates the distributions of this complete dataset representation in box and whisker plots to validate consistency with the original dataset and explore the presence of outliers.

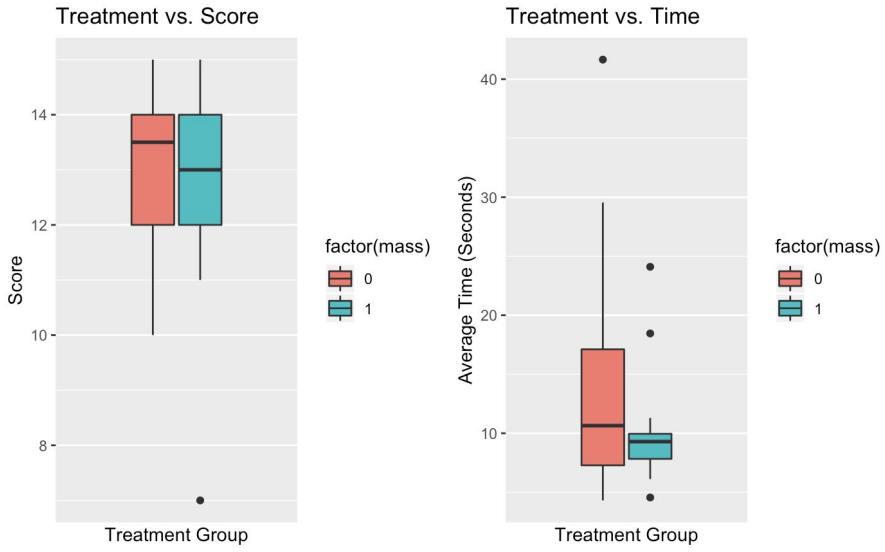

Figure 16. Box and Whisker Plot of Treatment Effect on Score and Time.

Our OLS regression models incorporated the fixed effects described above (including age, education level, languages known, and High Valyrian knowledge). Because we hypothesized that there may be differences in effects driven by the increased difficulty of Assessment 2, we created a new indicator variable "r2" to signal whether the observation came from Assessment 1 or Assessment 2 and included an interaction term in regression to explore its relationship to the treatment condition and the presence of any heterogeneous treatment effects. Results are shown in Figure 17 with model parameters displayed below.

	score		time	
	(1)	(2)	(3)	(4)
mass	−0.458	−0.409	−3.664*	−3.705**
	(0.455)	(0.385)	(1.946)	(1.558)
r2		−0.591		0.494
		(0.385)		(1.558)
mass:r2				
Constant	13.208***	15.500***	13.382***	20.329***
	(0.322)	(1.648)	(1.376)	(6.673)
Fixed Effects: Languages	No	Yes	No	Yes
Fixed Effects: Education	No	Yes	No	Yes
Fixed Effects: HV Experience	No	Yes	No	Yes
Fixed Effects: Self	No	Yes	No	Yes
Observations	48	48	48	48
R2	0.022	0.668	0.072	0.717
Adjusted R2	0.0003	0.291	0.051	0.396
Residual Std. Error	1.577 (df = 46)	1.328 (df = 22)	6.742 (df = 46)	5.380 (df = 22)
F Statistic	1.013 (df = 1; 46)	1.772* (df = 25; 22)	3.545* (df = 1; 46)	2.232** (df = 25; 22)
Note:				*p<0.1; **p<0.05; ***p<0.01

Figure 17. Summary of OLS Regression Outputs.

Outcome Variable: Score

Model 1:

$$Score = B_0 + B_1 * Mass$$

Model 2:

$$Score = B_0 + B_1 * Mass + B_2 * r2 + B_3 * (Mass * r2) + B_4 * factor(numlanguages) + B_5 * factor(education) + B_6 * factor(preHV) + B_7 * factor(email)$$

Outcome Variable: Time

Model 3:

$$Time = B_0 + B_1 * Mass$$

Model 4:

$$Time = B_0 + B_1 * Mass + B_2 * r2 + B_3 * (Mass * r2) + B_4 * factor(numlanguages) + B_5 * factor(education) + B_6 * factor(preHV) + B_7 * factor(email)$$

Overall, we found that the results of OLS regression were consistent with the EDA and point estimate analysis discussed above. The treatment effect of mass studying on scores in our simple model was -0.46 (1), compared with -0.41 (2) when including the “r2” indicator variable, interaction term, and fixed effects, although the treatment effects were not found to be significant. The treatment effect of mass studying on time was -3.66 (3), compared with -3.7 (4) when including the “r2” indicator variable, interaction term, and fixed effects. Both models produced significant effects for mass studying on time, but including “r2” in Model 4 produced an effect that is significant at the 95% confidence level.

These results suggest that including fixed effects in the model do not appear to affect the coefficient on treatment, but it was successful in reducing the variation of the error term. Additionally, while the coefficient for “r2” was not found to be significant, it does show that Assessment 2 in general resulted in lower scores and higher times on average, suggesting that Assessment 2 was more difficult. This intuition is further supported by the fact that there appears to be no effect driven by the interaction term between treatment and assessment number, suggesting that the treatment effect was independent of the assessments themselves. Including this covariate provides higher confidence in the true causal effects within our treatment groups.

Limitations

The sample size was a limitation to derive more power from this study. In retrospect we should have recruited many more participants through other avenues than the mere forty nine friends and family members that we started our experiment with. We encountered significant attrition (20 participants) during the week prior to administering Assessment 1. We hadn’t anticipated this level of attrition. By increasing the sample size we could have also weathered some of this attrition. Alternately, we could have increased monetary incentives by gifting a small token amount to all participants in the beginning which would have created a sense of obligation for them to participate in earnest. That in turn could have reduced the attrition and strengthened our analysis.

Conclusion

In this experiment, our objective was to learn about the effects of different studying methods on precision and recall by analyzing the outcome variables of participant scores and speed. We hypothesized that spaced learners would score higher and have a slower recall speed, which might be an indication of crystallization of the vocabulary. Conversely, we hypothesized that mass learners will have faster recall times. The spaced learners scored 0.41 points higher than the mass learners but were slower by 3.7 seconds on speed. The score is not statistically significant but the time to recall is. The results seem to validate our hypotheses.

Future enhancements would be to validate the study for a larger sample size as well as altering the phasing of the experiment to more fully mirror experimental conditions

between treatment groups. In particular, we recommend aligning the number and timing of washout days across the experiment so that both groups have the same opportunity to flush out their short-term memory on the day before the assessments, when nobody studies.

Grp.	Assign.	Pre-Treat	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8
1	R	O	X_{1S}	X_{1S}	X_{1S}	Wash out	O	X_{1M}	Wash out	O
2	R	O	X_{2M}	Wash out	O	X_{2S}	X_{2S}	X_{2S}	Wash out	O

Figure 19. Experimental Design with Proposed Wash Out Days.

Another future iteration could focus on learning logical content as opposed to vocabulary, which is largely a memorization task. The ability of spaced study participants to better learn logical concepts may produce stronger evidence to support spaced studying.

Bibliography

Division of Sleep Medicine at Harvard Medical School. "Sleep, Learning, and Memory." *Healthy Sleep*. Link.

Kornell, Nate. "Optimising Learning Using Flashcards: Spacing Is More Effective Than Cramming." *Applied Cognitive Psychology* 23: 1297–1317 (2009). Link.

Park, Denise, et al. "Theories of Memory and Aging: A Look at the Past and a Glimpse of the Future." J *Gerontol B Psychol Sci Soc Sci*. 2017 Jan; 72(1): 82–90. Link.

Settles, Burr. "How we learn how you learn." *Making Duolingo*. Link.

Smolen, Paul, et al. "The right time to learn: mechanisms and optimization of spaced learning." *Nature Reviews Neuroscience* volume 17, pages 77–88 (2016). Link.