# Decoding Game of Thrones

## The Effect of Gender Identity on Learning Outcomes

*Final Project Research Report*

UC Berkeley School of Information | W241 Experiments and Causal Inference | Fall 2019
Alex West, Anna Jacobson, and Apik Zorian

## Abstract

Possible causes of the persistent gender gap in science, technology, engineering, and mathematics (STEM) fields are hotly debated, as are its proposed solutions. At the college level, recruiting more female faculty to teach STEM classes has been suggested as an option to reduce the gender gap based on the potential to attract and retain female students through a role model effect. While a growing body of literature has examined the role of instructor gender at the higher education level, it typically focuses on academic outcomes and in-person education. The increasing prevalence of online learning through video presents a new facet to this discussion and one that has received less attention from researchers. This paper undertakes a unique approach by exposing students to identical content for videos (utilizing a lecture about *Game of Thrones* as a basis for understanding) recorded by both male and female instructors and measuring students' learning outcomes. In general, we find little to indicate a significant relationship between instructor gender, student gender, and student performance, but we discuss some interesting findings with regard to the order of instructors and female student outcomes. We conclude by discussing the implications of these findings and further research possibilities.

# 1.0 | Introduction

Despite great progress in college enrollment, degree attainment, and representation in some fields historically dominated by men, women are still poorly represented in many areas of science, technology, engineering, and mathematics (STEM). One proposed solution to rectify this problem is to encourage women to take part in online learning opportunities through formal educational institutions offering degree programs remotely; massive open online courses (or MOOCs) like Coursera, Udacity, or edX; or educational videos on YouTube. On one hand, these opportunities represent a tremendous democratization of knowledge: available anywhere at anytime, to anyone with access to the internet, and of particular benefit to those who have to balance their education with other responsibilities such as work and family. On the other hand, the gender gap in STEM careers and academic fields persists, and the vast majority of online video instructors in these subjects are male. This serves as a reminder that access to education is perhaps more complicated than simply opening the gates to anyone with any level of interest. How can we transform online instruction in STEM to provide true access to opportunity for women?

The gender gap in STEM fields remains a large problem with many avenues for study. Solutions that address it have implications for long-term economic mobility, wage inequality, and opportunities for innovation from other more diverse populations. This study examines one specific aspect of the online learning environment that lends itself well to field experimentation: instructor gender on video.

The intersection of gender and education presents a rich environment for scholarly research. A great deal of the existing body of research centers specifically on young children and K-12 education. In studies addressing higher education environments, the focus is often on pure academic outcomes in a classroom or in-person learning environment (e.g., Bettinger & Long, 2005; Canes & Rosen, 1995; Carrell, Page, & West, 2010; Hoffman & Oreopoulos, 2009; Neumark & Gardecki, 1998; Robst, Keil, & Russo, 1998). Other literature examines less concrete measures of success and barriers to achievement, such as motivation (Solanki & Xu, 2018), stereotype threat (Kapitanoff & Pandey, 2017, Schroeder & Adesope, 2015), and student-instructor rapport (Lammers & Byrd, 2019). The prevalence of online learning has been touted by multiple papers and articles tracking trends in enrollment and completion (Shah, 2019, Dijsselbloem, 2018), but emphasizing the effects of distance and self-regulated learning rather than gender.

This study builds on existing research and addresses an underlying gap by focusing on online video learning and instructor gender. Our approach uses identical content for recorded videos from both male and female instructors, and measures student performance on several measures of learning. As students in a STEM field currently engaged in distance learning through a major university ourselves, we chose to examine online video learning specifically because of its ubiquity, scale, and potential to accelerate more equitable representation in STEM. In addition, a student's experience in these courses may affect interest in subsequent learning, influencing major academic and career decisions. Our focus on online video learning fills an important gap in current literature that mainly focuses on in-person interactions as the basis for study.

## 1.1 | Research Hypothesis

Gender is a deep, multi-faceted concept with infinite avenues of exploration, study, and debate. We chose a specific and tightly-defined aspect of gender for our experiment, appropriate for the time and resource constraints of the project, and incorporated as many features as possible. Within those limits, we attempted to answer the following question:

***"What is the relationship between instructor and student gender on students' learning outcomes in higher education in STEM?"***

We hypothesized that if an instructor and student share the same gender identity, student performance will be improved. Our null hypothesis, therefore, was that instructor and student genders have no relationship with student performance.

# 2.0 | Experiment Design

## 2.1 | Overview

In order to test our hypothesis that shared instructor and student gender identity improves student performance, we designed an experiment in which the subjects ("students") were exposed to both a treatment and a control lecture video, then tested to measure how much they had learned from each video.

## 2.2 | Logistics

Following a small-scale pilot experiment to assess viability, the full-scale experiment took place between November 10 and November 25, 2019. Subjects were recruited from the authors' networks, including UC Berkeley School of Information, University of Michigan, professional contacts, and family and friends. Although recruitment was not limited to students, the majority of participants were current students or recent graduates. Recruitment took place online on platforms such as Slack, Twitter, and Reddit (as well as through direct email), whereby subjects were provided with a link to the Qualtrics survey platform. When they followed the link, they were taken to Phase I of the experiment, where they were asked to watch Video A, followed by eight multiple-choice questions, then Video B, followed by another eight questions. One day later, they received an email with a link to Phase II of the experiment, which was another quiz comprising ten of the sixteen questions from Phase I.

By recruiting within groups whom we believed to be particularly cerebral, specifically interested in the content of the experiment, and/or especially inclined to participate due to their personal relationships with the authors, we hoped to achieve our target sample size of 100 subjects (with the goal of achieving a minimum statistical power of 0.80 in our experiment). For additional motivation, in our recruitment efforts we offered an incentive of $100 Amazon gift cards to the five highest-scoring participants. However, we chose both our recruitment and incentivization strategies at the expense of random sampling of our target population, which we discuss further in Section 3.1.1.

## 2.3 | Treatment

The experiment utilized a lecture about Natural Language Processing (NLP), using George R. R. Martin's series of fantasy novels *A Song of Ice and Fire* as a basis for understanding; see Appendix C. The lecture was divided into two parts (A and B) which were structured similarly and approximately equal in length and complexity. Each part discussed a different aspect of lexical diversity (variability in Part A and density in Part B).

The content of the lecture was selected to meet several objectives. *Game of Thrones*, the television series based on *A Song of Ice and Fire*, is enormously popular; by using it as a theme, we hoped to interest and amuse potential subjects. However, we also wanted to use material about which the students' baseline knowledge could be expected to be relatively low; the NLP concepts introduced in the lectures are fairly obscure and even subjects familiar with NLP might not be conversant in the precise mechanisms and terminology discussed in the lectures. In addition, the content included various terms, facts, and figures that were well-suited to the types of questions we intended to use for outcome measurement; see Section 2.4.

Each subject watched both parts of the lecture, one with a female instructor and the other with a male instructor. After blocking by the subject's gender, the order of the instructors was randomly assigned using Qualtrics's "Flow" feature. Half of the subjects received Video A with a female instructor and Video B with a male instructor, and the other half received the opposite. Within those two groups, half were female subjects and half were male subjects.

## 2.4 | Outcome Measurement

Conceptually, the outcome variable we wanted to use in our experiment was student learning. However, learning outcomes are notoriously difficult to measure. First, the definition of "learning" itself is highly subjective and complex. In an academic setting, a student might be considered to have successfully "learned" if they fall anywhere on a spectrum of mastery from simply being able to remember the content, all the way up to being able to transfer the content to other contexts and build upon its concepts to generate original ideas.

In our experiment, in order to focus the definition of "learning", we selected three dimensions of learning that we felt were fundamental and meaningful:

1. **Recall:** Short-term recollection of information explicitly stated in the text of the content.
2. **Comprehension:** Understanding of concepts that were implicit but not explicitly stated in the text of the content.
3. **Retention:** Longer-term memory of both explicit and implicit aspects of the content.

A second challenge to the measurement of learning outcomes is that while "learning" is a qualitative concept, our experiment design required quantitative data. All current methodologies that are used to quantitatively measure student learning have limitations and biases, and no method is considered to be completely accurate (Breslow, 2007). In our experiment, we utilized the widely-accepted (though admittedly imperfect) methodology of direct measurement through standardized testing. Our experimental outcomes were based on performance on a series of multiple-choice questions that were modeled on the questions used in college-level standardized tests such as the GRE; see Appendix C. We designed these questions to measure recall and comprehension immediately following treatment. Retention was measured through a combination of recall and comprehension questions administered one day after treatment. We gathered the scores for each set of questions, as well as the total scores, in which each question was weighted equally and the total was scaled to 100.

Using these performance scores alone would make the experiment susceptible to bias from individual effects (i.e. any individual student's scores could vary due to their own specific aptitude, baseline knowledge of the content, test-taking skill, or any other of an infinite list of possible individual-level confounds). This is a particular concern given the relatively small scale of our experiment, in which outliers could unduly influence the results. In order to mitigate these individual effects, we considered an additional individual fixed effects variable (IFE_Scaled) that was calculated from the mean of each subject's total Phase I scores on Videos A and B. This variable represents the probable outcome for any individual student regardless of instructor or video.

## 2.5 | Within-Subject Design

Our experiment utilized a within-subject design in which all of the subjects were in both treatment and control (i.e. every subject viewed one lecture with a female instructor and one lecture with a male instructor). This experiment design type was selected in order to leverage each participant's involvement to capture as much data per person as possible and maximize the statistical power of the experiment. It also had the advantages of eliminating bias due to between-group individual effects and allowing for paired analysis at the individual level.

However, the disadvantage of the within-subject design is that it necessitated that we use different content for treatment and control, since repeating the content would have biased the learning outcomes toward the subjects'

second exposure to the material. Having two different lectures introduced potential “specific lecture effects”, in which differences in the unique attributes of the lectures - including the content and assessment questions - might affect outcomes. We addressed this concern in several ways. First, we used content and questions for each lecture that were as similar as possible in length, structure, style, and complexity. Second, both instructors recorded both lectures, to allow for comparability between outcomes on the two parts of the lecture independent of the instructor. Finally, we made the videos as visually similar as possible, with the instructors recorded in the same aspect ratio wearing similar clothing against similar backgrounds.

Within-subject design also necessarily introduces a potential temporal or ordered effect, as a subject cannot be in treatment and control simultaneously. In our experiment, this meant that the subjects might have different outcomes for the second lecture simply because it occurred later in the experiment. Reasons for this could include fatigue, loss of focus, and boredom. A potential mitigation might have been to introduce differentiation between the two videos, in order to try to maintain the subjects’ level of interest and engagement from Part A through Part B. Unfortunately, however, this would be in direct conflict with the “specific lecture effects” mitigations described in the preceding paragraph. Therefore, we chose to address this concern by limiting the length of both the lectures to approximately 2.5 minutes and limiting the number of questions after each video to eight. With an estimated completion time for each video/quiz segment quite short at about 5 minutes, we hoped to alleviate any temporal effects.

## 2.6 | Covariates

Before each subject started Phase I of the experiment, we gathered key personal information that we thought might affect their outcomes, as follows:

- **Age:** We hypothesized that younger people might be more comfortable with the online delivery method of the lectures, whereas older people may not have ever encountered this format in their own educations.
- **Ethnicity:** As there can be cultural differences in attitudes both toward education and toward gender, we thought that ethnicity might be an informative covariate.
- **English Language Acquisition:** We thought that there might be differences between native and non-native English speakers. For instance, we thought that non-native speakers might perform better on questions related to recall, as recall is an essential skill in learning a new language.
- **Student Status:** Similar to the age covariate, we thought that current students might be more accustomed to online learning that former students, and therefore they might have better outcomes with that delivery method.
- **Level of Education Completed:** We hypothesized that higher levels of education might be correlated with performance.

Because within-subject design guarantees perfect covariate balance between treatment and control (since every subject is in both treatment and control), these covariates were not useful for regression modeling. However, the information was informative for exploratory data analysis about our test population, as well as allowing for subgroup analysis; see Exhibit B.

# 3.0 | Experimental Results & Analysis

A total of 112 subjects completed Phase I of the experiment, with 54 female and 58 male participants (though we had hoped to also assess outcomes for non-binary subjects, only one participant identified as such and therefore could not be used in our analysis). 94 of these subjects completed Phase II, with 47 female and 47 male participants. On average, the subjects in Phase I were in their twenties or thirties (64%), white (63%), native English speakers (74%), current students (54%), and highly educated (66% with graduate, professional, or postgraduate degrees); see Exploratory Data Analysis in Appendix B. The mean subject location was in the central United States ((38.43, -87.06), which is Madison Township, IN). The population we intend our experiment to generalize to is American college students, which is a younger, more ethnically diverse, and less educated population than our experiment subjects. Moreover, our subjects were not randomly selected; the authors selected participant groups and the individual participants from those groups self-selected into the experiment. Nonetheless, we feel that there is enough variation within the experiment population to provide interesting results, even if not conclusive or completely generalizable.

Table B.1 in Appendix B shows the results of our recall, comprehension, and retention scores for each video, as well as the total Phase I scores. We can see that all students performed better on Video A than Video B, with male subjects scoring higher than female subjects on Video B, and female subjects scoring better than male on Video A. This could have happened for a number of reasons, such as the script for Video A being easier to comprehend than that of Video B, or the questions for Video B being more difficult. We look further into the video-level performances later in this section.

We observed a 16% attrition rate for subjects from Phase I to Phase II. Our initial thought was that these subjects may have been overwhelmed by the dense subject matter or had found the questions too difficult in Phase I, and had opted to not take part in Phase II. Upon reviewing the scores on Phase I, we saw that the average overall score of attriters in Phase I was 76%, while those who did return for Phase II scored 81% on Phase I questions. This supports our theory that because the attriters did not perform well on Phase I questions, they may have not felt confident in performing well enough in Phase II to be in the top scores and qualify for the prize. We also saw that a whopping 2/3 of the attriter group reported themselves as current students. In comparison, only 42% of the retainers were students. This may suggest that students have less free time to dedicate to our study. Perhaps because students are regularly bogged down with taking tests, our study may have felt like less of a fun video learning activity and more like a burden on top of their academic schedule.

## 3.1 | Primary Analyses

### 3.1.1 | *Overview*

We used three different subject groupings for our primary analyses. In the first design, a two-group design, the treatment is defined as an instructor who shares the student's gender identity, while control is an instructor with the opposite gender identity from the student; see Table 3.1. In this analysis, we did not differentiate between female and male students.

| Experiment Design 1 | | Instructor Gender ($G_i$) | |
|---|---|---|---|
| | | Female | Male |
| Student Gender ($G_s$) | Female | Same_Gen | Mixed_Gen |
| | Male | Mixed_Gen | Same_Gen |

**TABLE 3.1**
Experiment Design 1 - Two-Group Design

In our second design, a 2x2 design, there are four different treatment groups; see Table 3.2. Each of the four possible combinations of student gender and instructor gender is considered separately.

| Experiment Design 2 | | Instructor Gender ($G_i$) | |
|---|---|---|---|
| | | Female | Male |
| Student Gender ($G_s$) | Female | FF | FM |
| | Male | MF | MM |

**TABLE 3.2**
Experiment Design 2 - Four-Group (2x2) Design

In our third design, a 2x4 design, each of the eight possible combinations of student gender, instructor gender, and Video A or B is considered separately; see Table 3.3. The purpose of this design was to isolate the potential ordered effect of treatment. In other words, does it matter if a subject receives same-gender treatment before or after opposite-gender treatment?

| Experiment Design 3 | | | | | |
|---|---|---|---|---|---|
| Video | | A | B | A | B |
| Instructor Gender ($G_i$) | | Female | Male | Male | Female |
| Student Gender ($G_s$) | Female | FFA | FMB | FMA | FFB |
| | Male | MFA | MMB | MMA | MFB |

**TABLE 3.3**
Experiment Design 3 - Eight-Group (2x4) Design

For each of our three experiment designs, we plotted the individual outcomes for each group, performed a t-test, measured the statistical power, and ran linear regression models. The results of the t-test and power estimates can be seen in Table 3.4, and we will delve more deeply into the designs in Sections 3.1.4 through 3.1.6.

### 3.1.2 | Notation

Throughout our analyses, we use indicator variables for each subject group. We use a specific notation for these variables in Experiment Designs 2 and 3. For Experiment Design 2, our group variables are named as *SubjectGender_InstructorGender* and for Experiment Design 3, our group variables are named as *SubjectGender_InstructorGender_VideoWatched*. Therefore, Group FMA describes the subset of female subjects that watched the version of Video A that was recorded by the male instructor, while Group MFB is the subset of male students who watched Video B that was recorded by the female instructor (Experiment Design 2 is not video-specific, so the variables do not include the last letter).

### 3.1.3 | Analysis Results Overview

We found no significant treatment effects in our analyses, with the exception of Group FFA in Experiment Design 3. The analyses of this group were also the only hypothesis tests that met our target statistical power threshold of 0.80.

| Specifications | | | Paired t-test | | Paired t-test Power Calculation | | |
|---|---|---|---|---|---|---|---|
| **Experiment Design** | **Group 1** | **Group 2** | **t-value** | **p-value** | **n** | **d** | **power** |
| **1** | **Same Gender** | **Mixed Gender** | -0.101 | 0.919 | 112 | 0.014 | 0.05 |
| **2** | **FF** | **FM** | 0.857 | 0.395 | 54 | 0.179 | 0.200 |
| | **MF** | **MM** | 0.976 | 0.333 | 58 | 0.153 | 0.200 |
| **3** | **FFA** | **FMB** | **3.20** | **0.004** | **26** | **0.697** | **0.920** |
| | **FMA** | **FFB** | 1.12 | 0.272 | 28 | 0.215 | 0.196 |
| | **MFA** | **MMB** | 1.397 | 0.173 | 31 | 0.310 | 0.387 |
| | **MMA** | **MFB** | 0.138 | 0.892 | 27 | 0.030 | 0.053 |
| | **FFA** | **All Others** | **3.481** | **0.001** | **26/198** | **0.591** | **0.081** |

**TABLE 3.4**
Hypothesis Test Results
*Note: All hypothesis tests in this tables are paired t-tests, which the exception of the final test in Experiment Design 3 (FFA vs. All Others), which is a two-sample t-test.*

### 3.1.4 | Experiment Design 1: Two-Group Model

In our first experiment design, we looked at the results for only treatment (same-gender instructor) and control (opposite-gender instructor) groups. Figure 3.1 shows an individual values plot that compares these two groups; we can see the results are very similar, as the mean value and the first and third quartile are all almost identical. Judging by these results, we would not expect to see any treatment effect. Upon running a t-test on our experiment, our assumption proved to be correct, as we did not see any significant treatment effects; see Table 3.4. We also failed to achieve high statistical power. We ran a linear regression on this design and our analysis again as expected showed no significant treatment effects; see Table B.2 in Appendix B.


**FIGURE 3.1**
Experiment Design 1 - Individual Values Plot

### 3.1.5 | *Experiment Design 2: Four-Group Model*

Our second experiment design looked at our 4-group model. Figure 3.2 shows the individual values plot for this experiment, and we do see some differences in the groups. Observing particularly the FF and MM groups, we notice the means are different and the boxes are shaped and positioned differently. This led us to hypothesize that we might find a significant treatment effect for either or both of these groups.

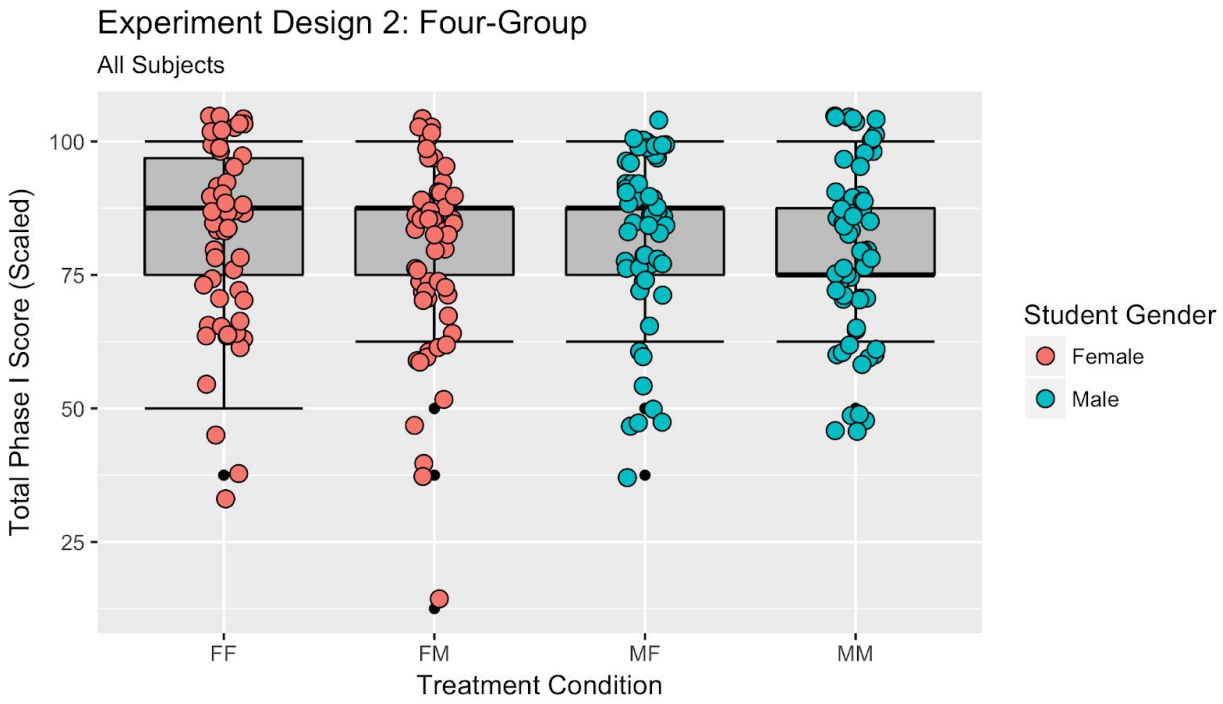



**FIGURE 3.2**
Experiment Design 2 - Individual Values Plot

We tested this hypothesis by running separate t-tests comparing the female student groups with different instructor genders (FF vs FM), as well as the male student groups with different instructor genders (MF vs MM). As the results show in Table 3.4, we did not find a significant treatment effect and failed to reach our threshold for statistical power. We also ran linear regressions on this design and our analysis again showed no significant treatment effects; see Table B.3 in Appendix B**.** This shows that the visibly perceptible differences observed in the individual values plots are likely to have occurred by chance.

### 3.1.6 | *Experiment Design 3: Eight-Group Model*

Our third and final experiment design looked at each of the eight possible combinations of video (A/B), instructor gender (male/female), and subject gender (male/female). Figure 3.3 shows an individual values plot that compares these eight groups. We do see a noteworthy difference in Video A scores for female subjects based on if they had a male or female instructor; female subjects who had a female instructor for Video A had much higher scores than female subjects who had a male instructor for the same video. We also see that in the box plot for the female instructor group, the majority of the scores appear to be above the mean, while the male instructor group did not perform as well, with some subjects scoring below 50%.

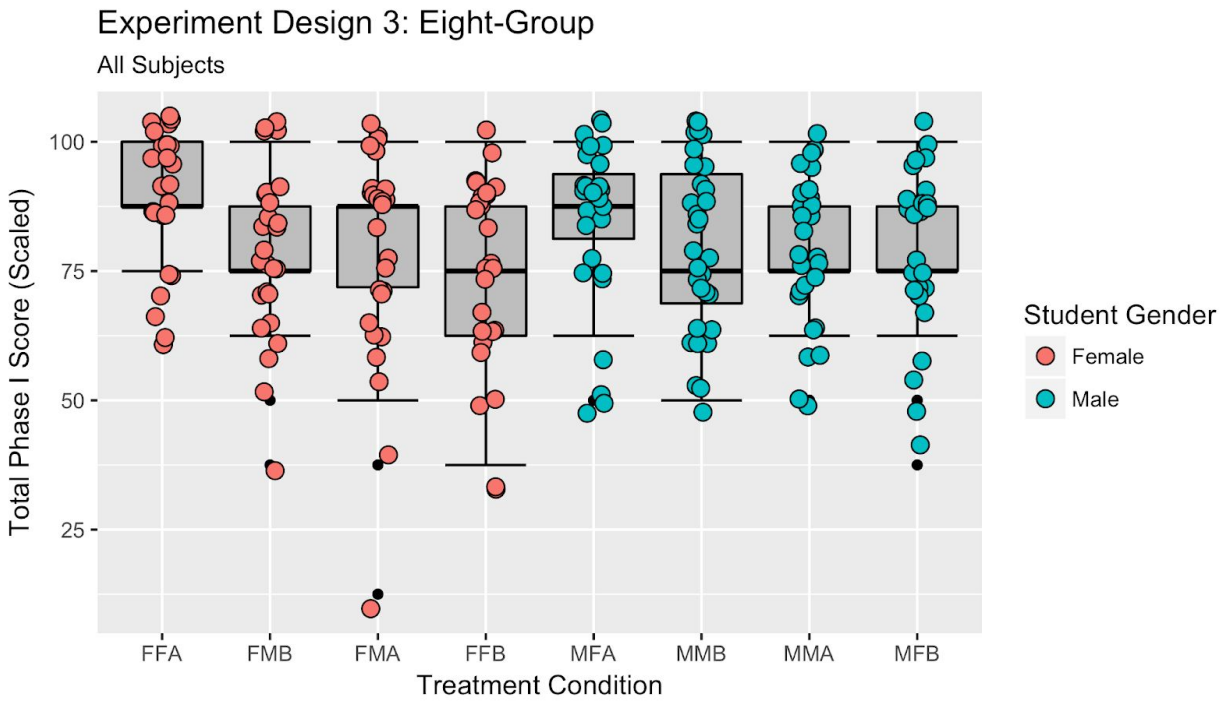



**FIGURE 3.3**
Experiment Design 3 - Individual Values Plot

Motivated by these findings, we ran t-tests on the different combinations of groups where the subjects were the same and the gender and videos changed. For example, the same group of men who watched Video A with a male instructor and Video B with a female instructor or the same female subjects who watch Video A with a female instructor and Video B with a male instructor. As discussed earlier, in our experiment, it was not possible to have the same group of subjects watch Video A with both a male and female instructor, as the scores would likely be higher for the second viewing. However, as the content and difficulty of both videos was designed to be equivalent, we did not see an issue with the videos changing, as long as we were comparing scores of the same group of subjects with varying instructor genders.

The results of these t-tests can be seen in Table 3.4, which shows that we did find a significant treatment effect between the female subjects with different gendered instructors (FFA vs FMB). Our power test also showed statistical power of 0.92. These results were very promising, and we followed this up by comparing FFA to the rest of the groups combined (FFA vs All Others in Table 3.4). For this, we ran a two sample t-test and power calculation, as the two groups do not include the same subjects. We again saw significant results for this t-test, and we achieved 0.805 statistical power, which was just above the threshold at which we could declare the experiment to be well-powered.

Next, we ran a linear regression with all eight variables from Experiment Design 3 to confirm our hypothesis that FFA was significant and to ascertain the magnitude of the treatment effect. These results are shown in Table 3.5.

| | *Dependent variable:* | |
|---|---|---|
| | Total_Scaled | |
| | (1) | (2) |
| FFA | 9.776** | 6.035** |
| | (4.103) | (2.410) |
| FMB | −0.801 | −4.542* |
| | (4.466) | (2.420) |
| FMA | −1.042 | 3.093 |
| | (5.044) | (2.519) |
| FFB | −5.060 | −0.925 |
| | (4.567) | (2.537) |
| MFA | 5.511 | 3.147 |
| | (4.167) | (2.572) |
| MMB | 0.672 | −1.692 |
| | (4.369) | (2.506) |
| MMA | 0.463 | 0.463 |
| | (4.308) | (2.530) |
| IFE_Scaled | | 0.986*** |
| | | (0.060) |
| Constant | 79.167*** | 0.396 |
| | (3.189) | (5.283) |
| Type of Std. Error | Robust | Robust |
| Observations | 224 | 224 |
| $R^2$ | 0.065 | 0.681 |
| Adjusted $R^2$ | 0.034 | 0.669 |
| Residual Std. Error | 16.136 (df = 216) | 9.444 (df = 215) |
| F Statistic | 2.135** (df = 7; 216) | 57.411*** (df = 8; 215) |
| *Note:* | | *p<0.1; **p<0.05; ***p<0.01 |

**TABLE 3.5**
Experiment Design 3 - Eight-Group Regression Table with Robust Standard Errors
*Note: In the models in this table, MFB is considered the control group and has a coefficient of 0.*

In these regression models, we find that FFA is significant and has a large coefficient compared to the other variables. This finding is consistent in both Model (1), which includes only the group indicator variables, and in Model (2), which also includes the Individual Fixed Effects variable.

These findings supported our hypothesis, so we ran a regression using *just* FFA to predict performance on Phase I. The results for this regression are shown in Table 3.6.

| | *Dependent variable:* | |
|---|---|---|
| | Total_Scaled | |
| | (1) | (2) |
| FFA | 9.713*** | 6.089*** |
| | (2.837) | (1.800) |
| IFE_Scaled | | 0.971*** |
| | | (0.059) |
| Constant | 79.230*** | 1.590 |
| | (1.177) | (5.103) |
| Type of Std. Error | Robust | Robust |
| Observations | 224 | 224 |
| $R^2$ | 0.036 | 0.661 |
| Adjusted $R^2$ | 0.032 | 0.658 |
| Residual Std. Error | 16.159 (df = 222) | 9.606 (df = 221) |
| F Statistic | 8.303*** (df = 1; 222) | 215.308*** (df = 2; 221) |
| *Note:* | | *p<0.1; **p<0.05; ***p<0.01 |

**TABLE 3.6**
Experiment Design 3 - FFA Regression Table with Robust Standard Errors

Again, we find that FFA is significant and has a large coefficient, even in Model (2) which controls for individual fixed effects.

## 3.2 | Subgroup Analyses

We also performed subgroup analyses based on the covariate data we collected; see Appendix B. These included the following subgroups:

1. **Current Students (N = 60 Pairs)**
2. **Subjects Under 30 Years Old (N = 45 Pairs)**
3. **Subjects 30 Years and Older (N = 67 Pairs)**
4. **Highly Educated Subjects (Masters and professional degrees and PhDs) (N = 74 Pairs)**
5. **Native English Speakers (N = 82 Pairs)**

We had intended to analyze subgroups based on ethnicity as well, but because our study population was overwhelmingly white, we did not feel that this would be a fruitful avenue of exploration.

Using Experiment Design 3 with individual fixed effects (IFE_Scaled), we found that as in our general population model, FFA continues to be significant for Subgroups 1 (Current Students), 3 (Subjects 30 Years and Older), and 5 (Native English Speakers). However, FFA is not significant for Subgroups B (Subjects Under 30 Years Old) and D (Highly Educated Subjects). And unlike our general population model, we found significant *negative* treatment effects for FMB for Subgroups 1 (Current Students) and 2 (Subjects Under 30 Years Old).

These results may simply be a result of the eccentricities of the relatively small pool of subjects in the experiment. However, because the models are controlled for individual fixed effects, we do not feel that they should be dismissed out of hand. Of particular interest is Subgroup 1 (Current Students), which is the population to which we intended our experiment to generalize. The female students in this subgroup showed large, significant treatment effects when exposed to Video A with a female instructor and Video B with a male instructor (positive and negative effects, respectively). Further study to determine if these results are replicable is warranted.

## 3.3 | Concerns

An unanticipated finding from our experimental results is that subjects' outcomes on Video B are worse than on Video A, irrespective of instructor or student gender. In a paired t-test, the mean of the differences in the results from Videos A and B (4.91 points) is significant, $t(111) = 2.84$, $p = 0.005$, power = 0.88. This suggests that we were not successful in fully mitigating either the specific video effects or the temporal effects - or both - as intended by our experiment design.

Additionally, if there is a positive treatment effect for female students to be exposed to a female instructor first, we might logically expect to see a negative treatment effect for female students to be exposed to a male instructor first. In Model (2), which includes the Individual Fixed Effects variable, we actually observe the opposite; the coefficient for variable FMA is positive and relatively large - and not significant. Although the presence of a positive treatment effect in one condition does not necessitate the presence of a negative treatment effect in the opposite condition, the conclusion might be stronger if it were the case. It is possible that we do not see this effect because of the difference in outcomes for Videos A and B described in the preceding paragraph. However, it is also possible that it is due to the complexity of gender itself; although for the purposes of this analysis we treat it as two discrete, mutually exclusive conditions, gender is in fact embodied in many different attributes that can be uncorrelated and widely varied in any one individual, regardless if they identify as male or female.

Another potential criticism of this experiment, given the three different primary analyses, is the problem of multiple comparisons; i.e., as the number of comparisons increases, it becomes more likely that the groups being compared will appear to differ in terms of at least one attribute. However, even if we use the Bonferroni Correction to adjust our significance threshold to 0.0167 (derived from the typical value of 0.05, divided by 3 comparisons), we find that the results described in Experiment Design 3 remain significant.

## 4.0 | Conclusions and Further Research

Through our experimental results and analysis, we found a causal relationship in one specific case: when a female student watched the female instructor video first, she was more likely to perform better on the quiz for that video (termed FFA in our results). These women performed significantly better than all other groups, including men who watched the male instructor video first, and women who saw the female instructor video second. Conceptually, the temporal aspect of instructor gender was found to be significant for improving female student outcomes. In practical terms, when a female student begins instruction with a female instructor, her performance on metrics related to her ability to learn improves significantly.

This FFA finding merits further discussion and scrutiny, due to the relatively small sample size of our experiment and multiple comparisons of our analysis. If valid, there are compelling implications to this discovery, as it presents novel ideas for addressing the gender gap in STEM. What we found could be termed an “anchoring effect” of seeing a female instructor first, and could be utilized in online education to offset some of the heavily male-dominated subjects without needing to re-record all videos. Institutions could simply place female instructor videos strategically as the first video seen in a section. Outside the online education environment, this finding could help expose young women to STEM subjects through initial introductions by female instructors or practitioners, or even in professional environments to onboard new female employees. The potential applications are endless and could have deep impacts on education and beyond.

Apart from this finding, however, we failed to reject the null hypothesis; we did not prove a broad causal relationship between instructor gender and student gender and student performance. There are many reasons this could be the case, both as a result of our research design and our subject population. On the positive side, our results could indicate that student performance is within the individual’s own control and not dependent on external conditions such as instructor gender. Or perhaps students are accustomed to having both male and female instructors and have adapted to mitigate any effects. The subjects recruited for this experiment were a highly educated and motivated group who might be less impacted by the treatment than the general student population. Finally, our research design had only two instructors, two videos, and 8 questions; this lack of variation may have influenced our ability to measure more subtle effects.

An exciting element of this research is the abundance of possible future studies that can build upon these results. First and foremost, to confirm our FFA finding, the study could be repeated immediately with the exact same content and questions with a relatively small sample size of women to see if we can replicate the treatment effect we found for Video A. Beyond the results of this study, the design lends itself well to expansion in the scale of both subjects and substance. Further research could increase the sample size to a wider subset of the population, augment the number of instructors to mitigate instructor-specific effects, maximize the number of questions after each video for more precise measurement and variance in scores, record more videos and write more questions to alleviate content-specific effects, and finally, run the study over a longer period of time.

The understanding of gender is moving away from a strict binary interpretation, making the task of asking gender-related research questions both more complex and increasingly necessary. Furthermore, the problem of lack of diversity in STEM encompasses more than just gender; it includes race, ethnicity, sexual orientation, and socioeconomic status, among other aspects. In electing to focus on just one sliver of this problem, we hope to lay the groundwork for future field experiments addressing all types of representation in online education and distance learning. In addition, our research involved only American institutions and subjects but has implications for how global bodies approach issues of representation, whether by addressing with action, discussing with inaction, ignoring, or rejecting them. Continuing to ask these types of research questions not only has the benefit of revealing

discrepancies and biases but can also be a bridge to tangible real-world action and change. Finally, institutional change in a country like the US whose education system attempts to offer equal opportunities to women and men can, in turn, contribute to generating solutions to combat structural gender inequality throughout the world.

## Appendix A - List of References


Bersin, J. (2016, January 1). Use Of MOOCs And Online Education Is Exploding: Here's Why. Retrieved 5 December 2019, from https://www.forbes.com/sites/joshbersin/2016/01/05/use-of-moocs-and-online-education-is-exploding-heres-why/#1e1a99b7649f

Bettinger, E., & Long, B. (2005). Do faculty serve as role models? The impact of instructor gender on female students. American Economic Review, 95, 152–157.

Breslow, L. (2007). Methods of Measuring Learning Outcomes and Value Added. Massachusetts Institute of Technology. https://tll.mit.edu/sites/default/files/guidelines/a-e-tools-methods-of-measuring-learning-outcomes-grid-2.pdf

Canes, B., & Rosen, H. (1995). Following in her footsteps? Women's choices of college majors and faculty gender composition. Industrial and Labor Relations Review, 48, 486–504.

Carrell, S., Page, M., & West, J. (2010). Sex and science: How professor gender perpetuates the gender gap. The Quarterly Journal of Economics, 125, 1101–1144.

Dijsselbloem, J. (2018, April 11). The Rise of MOOCs: Can Online Distance Learning Replace Traditional Education? Retrieved December 5, 2019, from https://www.diggitmagazine.com/papers/rise-moocs-can-online-distance-learning-replace-traditional-education.

Hoffman, F., & Oreopoulos, P. (2009). A professor like me: The influence of instructor gender on university achievement. Journal of Human Resources, 44, 479–494.

Kapitanoff, S., & Pandey, C. (2017). Stereotype threat, anxiety, instructor gender, and underperformance in women. Active Learning In Higher Education, 18(3), 213-229. doi: 10.1177/1469787417715202

Lammers, W. J., & Byrd, A. A. (2019). Student Gender and Instructor Gender as Predictors of Student–Instructor Rapport. Teaching of Psychology, 46(2), 127–134. doi: 10.1177/0098628319834183

Neumark, D., & Gardecki, R. (1998). Women helping women? Role model and mentoring effects on female Ph.D. students in economics. Journal of Human Resources, 33, 220–246.

Paul, A. (2014, September). Even Online Classes Have a Gender Gap. Here's How to Close It. Retrieved 5 December 2019, from https://slate.com/technology/2014/09/mooc-gender-gap-how-to-get-more-women-into-online-stem-classes.html

Robst, J., Kell, J., & Russo, D. (1998). The effect of gender composition of faculty on student retention. Economics of Education Review, 29, 429–439.

Schroeder, N. L., & Adesope, O. O. (2015). Impacts of Pedagogical Agent Gender in an Accessible Learning Environment. Educational Technology & Society, 18 (4), 401–411.

Shah, D. (2019, January 2). Year of MOOC-based Degrees: A Review of MOOC Stats and Trends in 2018 - EdSurge News. Retrieved 5 December 2019, from https://www.edsurge.com/news/2019-01-02-year-of-mooc-based-degrees-a-review-of-mooc-stats-and-trends-in-2018

Solanki, S., & Xu, D. (2018, August). Looking Beyond Academic Performance: The Influence of Instructor Gender on Student Motivation in STEM Fields. American Educational Research Journal, 55(4), 801-835.

Vesterberg, P. (2019, April 14). Decoding 'Game of Thrones' by way of data science, Part 1 - A numerical exploration of 'A Song of Ice and Fire'. Noteworthy - The Journal Blog. https://blog.usejournal.com/decoding-a-game-of-thrones-by-way-of-data-science-fd81e66d1255

# Appendix B - Tables & Figures

## B.1 | Exploratory Data Analysis: Scores

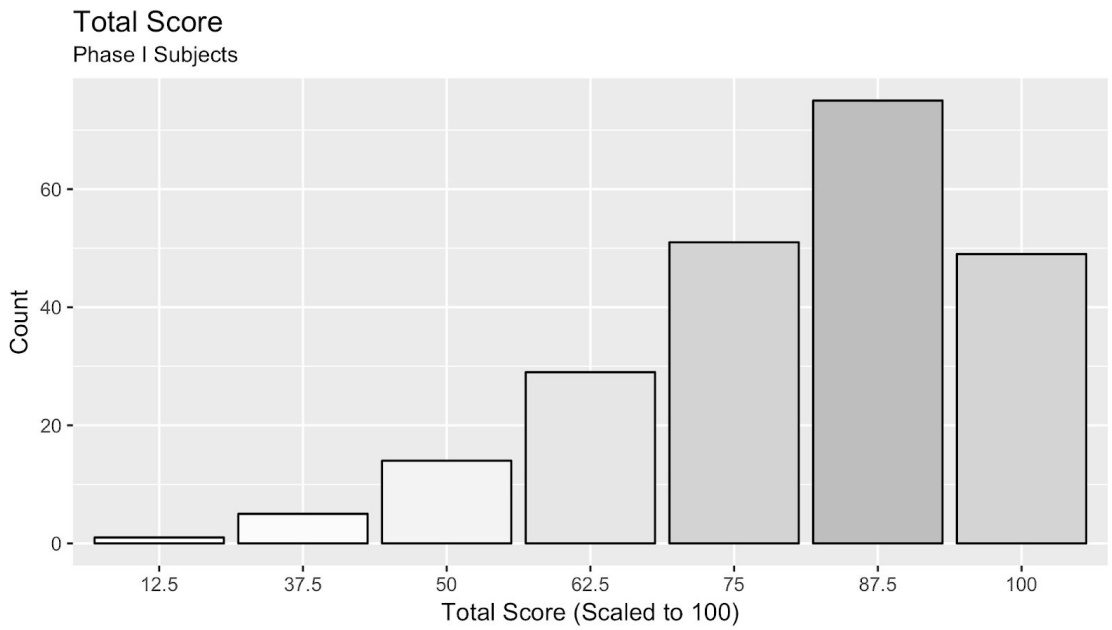



**FIGURE B.1**
Total Score Distribution

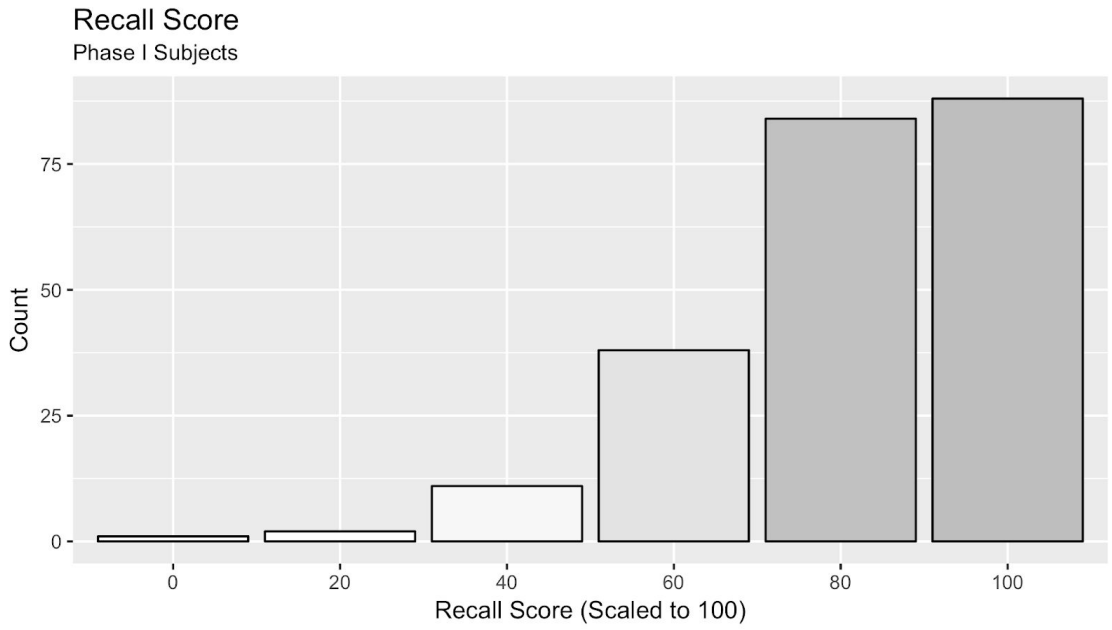



**FIGURE B.2**
Recall Score Distribution

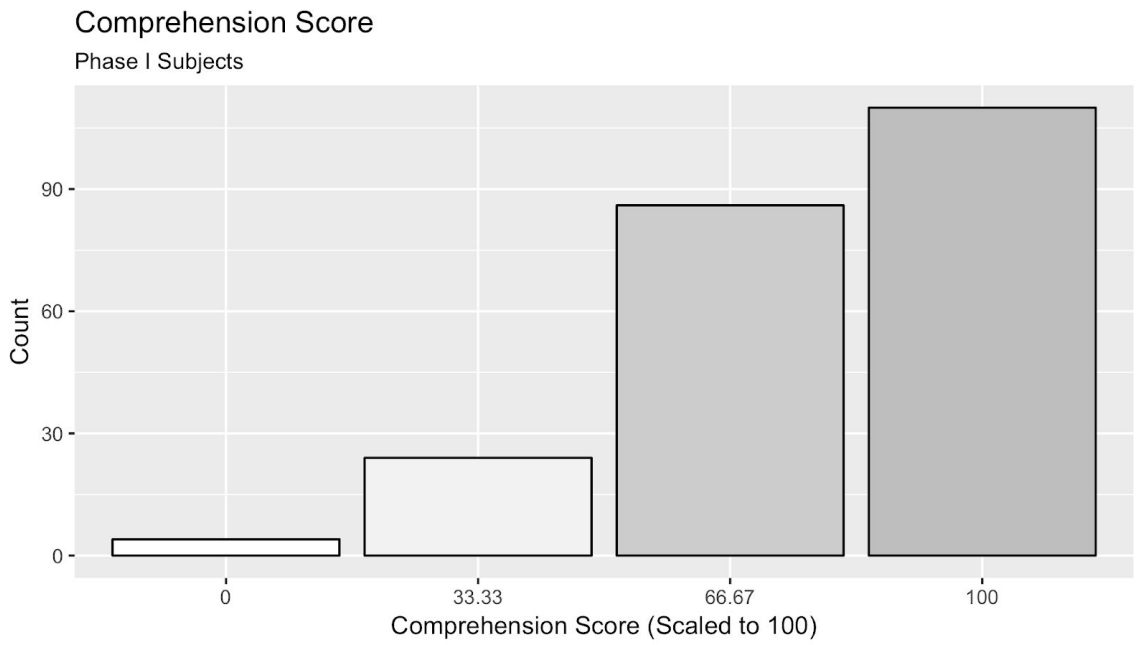



**FIGURE B.3**
Comprehension Score Distribution

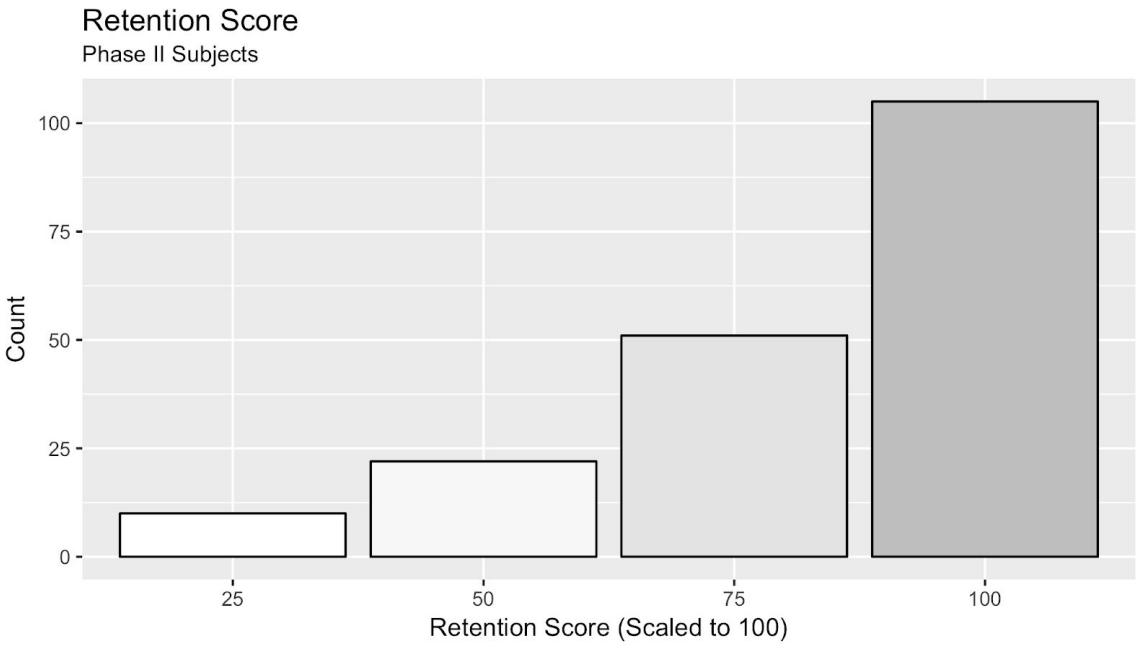



**FIGURE B.4**
Retention Score Distribution

| Video | Score | All Students | Female Students | Male Students |
|---|---|---|---|---|
| A | Total Phase I | 82.81 | 83.33 | 82.33 |
| | Recall | 81.07 | 80.74 | 81.38 |
| | Comprehension | 85.71 | 87.65 | 83.91 |
| | Retention | 90.16 | 88.30 | 92.02 |
| B | Total Phase I | 77.90 | 76.16 | 79.53 |
| | Recall | 82.14 | 80.74 | 83.45 |
| | Comprehension | 70.93 | 68.52 | 72.99 |
| | Retention | 76.60 | 78.19 | 75.00 |

**TABLE B.1**
Summary of Scores for All, Female, and Male Students

## B.2 | Exploratory Data Analysis: Covariates

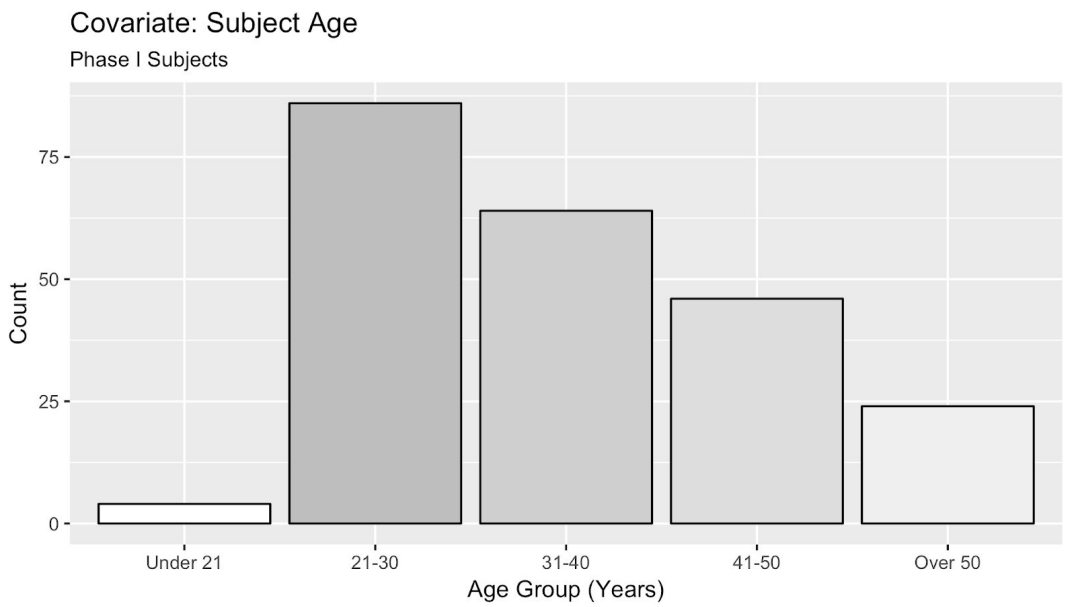



**FIGURE B.5**
Subject Age

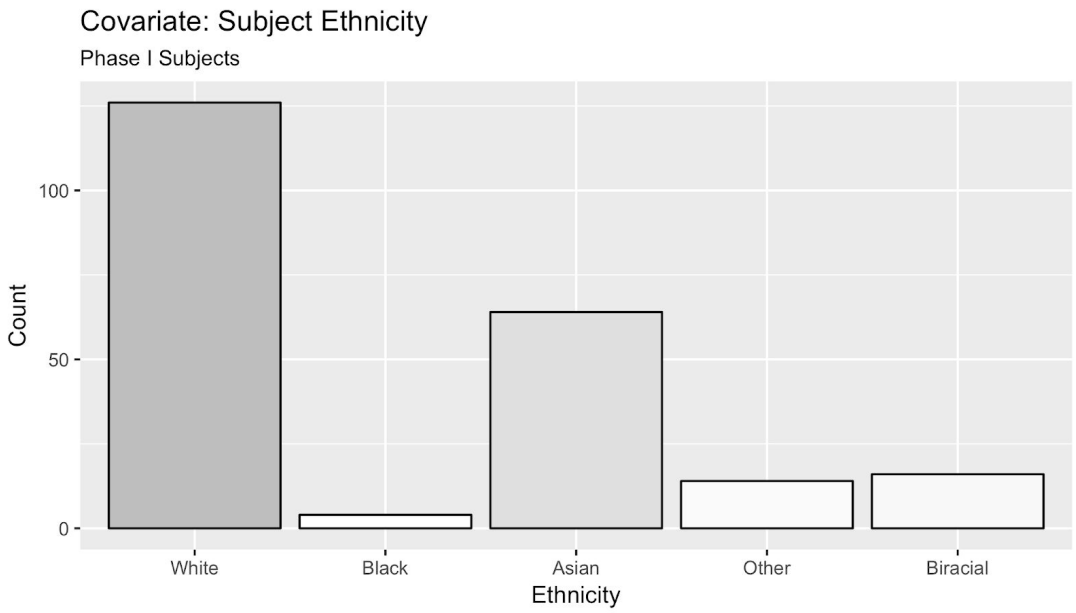



**FIGURE B.6**
Subject Ethnicity

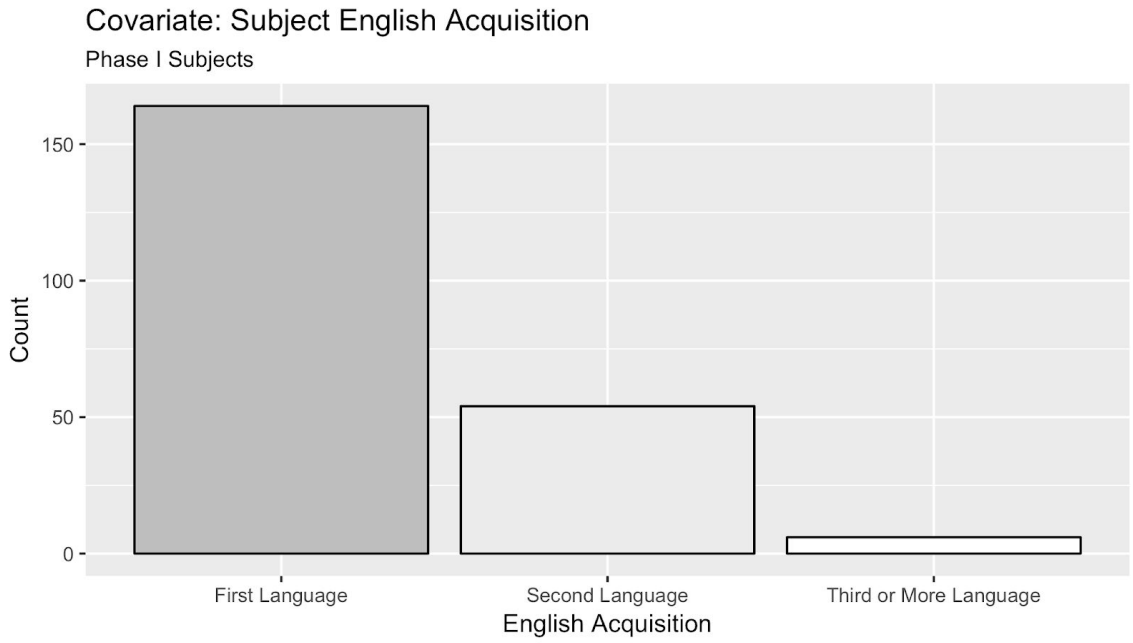



**FIGURE B.7**
Subject English Language Acquisition

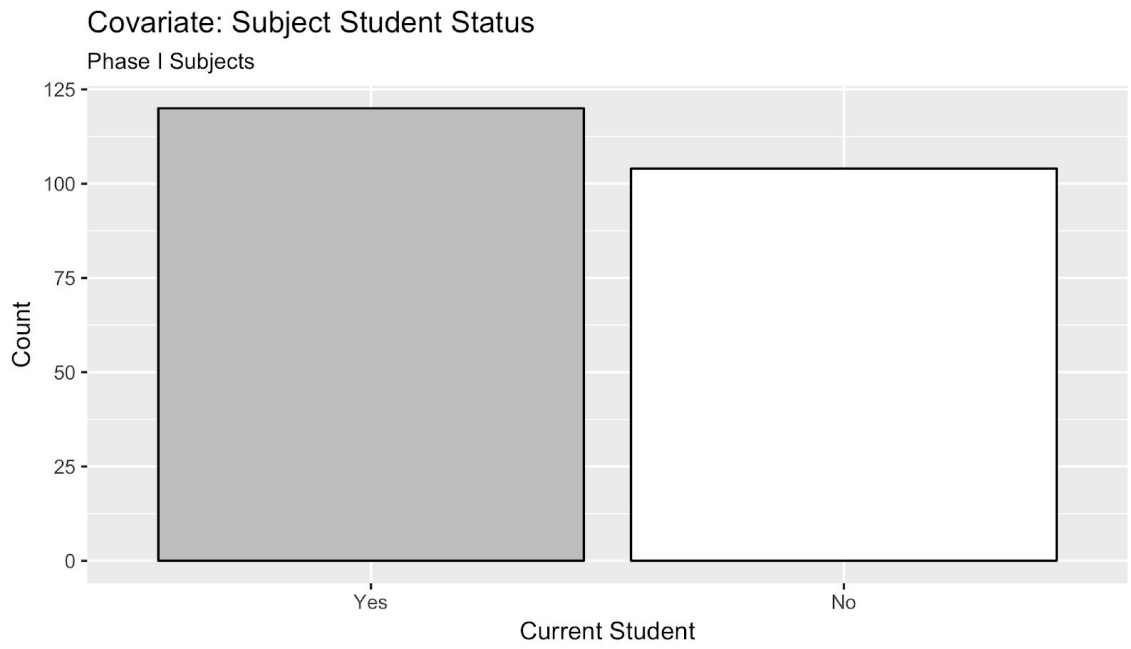



**FIGURE B.8**
Subject Student Status

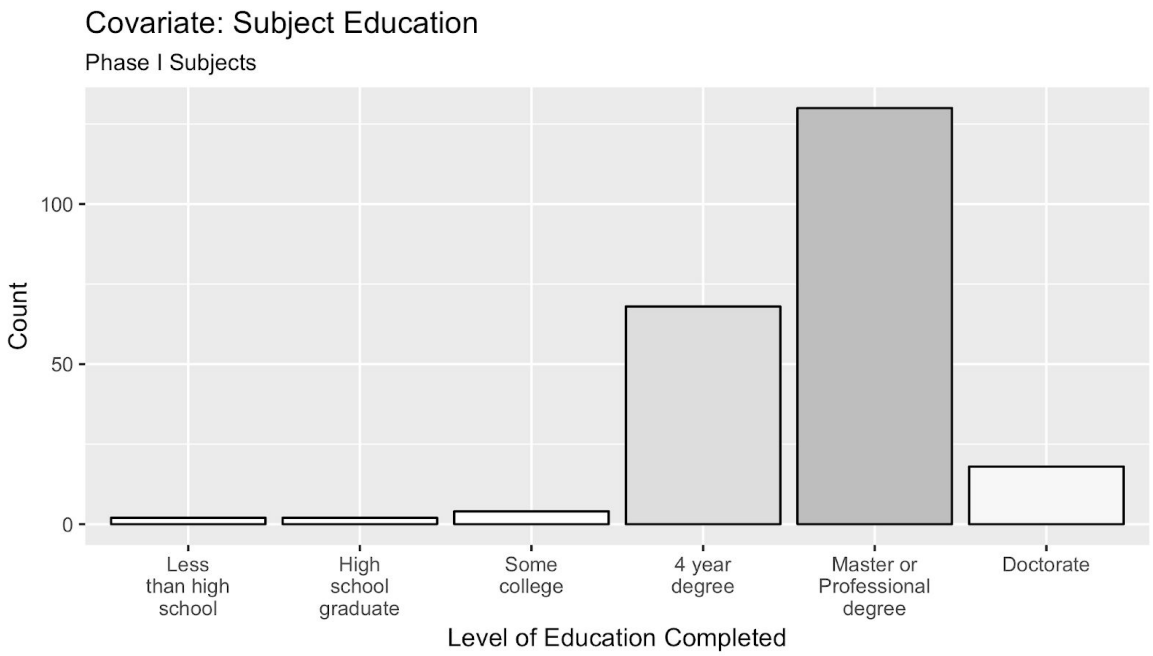



**FIGURE B.9**
Subject Education

## B.3 | Experiment Design 1: Two-Group Model

| | *Dependent variable:* | |
|---|---|---|
| | Total_Scaled | |
| | (1) | (2) |
| Same_Gen | 0.223 | 0.223 |
| | (2.209) | (1.322) |
| IFE_Scaled | | 0.984*** |
| | | (0.060) |
| Constant | 80.246*** | 1.140 |
| | (1.598) | (5.236) |
| Type of Std. Error | Robust | Robust |
| Observations | 224 | 224 |
| $R^2$ | 0.00005 | 0.647 |
| Adjusted $R^2$ | −0.004 | 0.644 |
| Residual Std. Error | 16.458 (df = 222) | 9.803 (df = 221) |
| F Statistic | 0.010 (df = 1; 222) | 202.381*** (df = 2; 221) |
| *Note:* | | *p<0.1; **p<0.05; ***p<0.01 |

**TABLE B.2**
Experiment Design 1 - Two-Group Regression Table with Robust Standard Errors
*Same_Gen = indicator variable for students and instructors of the same gender*
*IFE_Scaled = individual fixed effects*

## B.4 | Experiment Design 2: Four-Group Model

| | *Dependent variable:* | |
|---|---|---|
| | Total_Scaled | |
| | (1) | (2) |
| FF | 1.509 | 3.113* |
| | (3.087) | (1.852) |
| FM | −1.501 | 0.104 |
| | (3.218) | (1.862) |
| MF | 2.371 | 2.371 |
| | (2.908) | (1.819) |
| IFE_Scaled | | 0.985*** |
| | | (0.060) |
| Constant | 79.741*** | −0.215 |
| | (2.053) | (5.188) |
| Type of Std. Error | Robust | Robust |
| Observations | 224 | 224 |
| $R^2$ | 0.008 | 0.654 |
| Adjusted $R^2$ | −0.005 | 0.647 |
| Residual Std. Error | 16.466 (df = 220) | 9.751 (df = 219) |
| F Statistic | 0.597 (df = 3; 220) | 103.362*** (df = 4; 219) |
| *Note:* | | *p<0.1; **p<0.05; ***p<0.01 |

**TABLE B.3**
Experiment Design 2 - Four-Group Regression Table with Robust Standard Errors
*FF = indicator variable for female students and female instructors*
*FM = indicator variable for female students and male instructors*
*MF = indicator variable for male students and female instructors*
*IFE_Scaled = individual fixed effects*

## B.5 | Subgroup Analyses

| | *Dependent variable:* | | | | |
|---|---|---|---|---|---|
| | Total_Scaled | | | | |
| | (1) | (2) | (3) | (4) | (5) |
| FFA | 7.103** | 3.750 | 7.310** | 4.924 | 6.703** |
| | (3.033) | (4.145) | (3.098) | (3.080) | (3.046) |
| FMB | −6.290** | −8.750** | −1.857 | −1.326 | −2.508 |
| | (3.040) | (4.145) | (3.137) | (3.084) | (3.050) |
| FMA | 1.116 | −3.750 | 7.024*** | 1.395 | 4.390 |
| | (3.758) | (5.397) | (2.583) | (3.212) | (3.121) |
| FFB | −0.446 | −1.250 | −0.615 | 2.089 | 0.818 |
| | (3.740) | (5.397) | (2.703) | (3.200) | (3.156) |
| MFA | 0.830 | −0.714 | 5.670 | 1.530 | 3.901 |
| | (3.484) | (3.958) | (3.731) | (2.901) | (3.031) |
| MMB | −1.849 | −4.286 | −0.212 | 1.009 | 1.297 |
| | (3.424) | (3.958) | (3.553) | (2.898) | (2.956) |
| MMA | 0.781 | −5.000 | 3.676 | 1.786 | 2.778 |
| | (3.405) | (4.497) | (3.122) | (3.328) | (3.347) |
| IFE_Scaled | 0.992*** | 1.000*** | 0.972*** | 0.995*** | 0.991*** |
| | (0.096) | (0.111) | (0.068) | (0.075) | (0.069) |
| Constant | 0.295 | 2.500 | −0.369 | −1.339 | −1.337 |
| | (8.438) | (10.031) | (5.926) | (6.694) | (6.376) |
| Type of Std. Error | Robust | Robust | Robust | Robust | Robust |
| Observations | 120 | 90 | 134 | 148 | 164 |
| $R^2$ | 0.693 | 0.708 | 0.678 | 0.688 | 0.701 |
| Adjusted $R^2$ | 0.671 | 0.679 | 0.658 | 0.671 | 0.686 |
| *Note:* | | | | | *p<0.1; **p<0.05; ***p<0.01 |

**TABLE B.4**

Experiment Design 3 (Subgroups) - Eight-Group Regression Table with Robust Standard Errors

*(1)* *Current Students*

*(2)* *Subjects Under 30 Years Old*

*(3)* *Subjects 30 Years and Older*

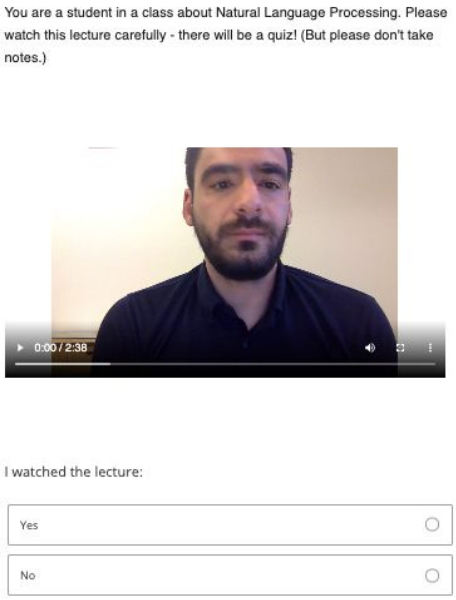
*(4)* *Highly Educated Subjects (Masters and professional degrees and PhDs)*

*(5)* *Native English Speakers*

*IFE_Scaled = individual fixed effects*

# Appendix C - Lecture Scripts and Quiz Questions

## C.1 | Video A Script




This week will be focused on lexical diversity. Similar to biodiversity in an environmental ecosystem, lexical diversity tells us how rich and robust a corpus of text is. To measure lexical diversity, we assess metrics that allow us to frame qualitative assessments in numerical terms. These metrics include the following:

- Volume: length of the text in number of words
- Variability: ratio of the number of unique words to the total number of words
- Density: estimated measure of information density

George R. R. Martin's series of fantasy novels A Song of Ice and Fire' has a massive volume, close to 1.8 million words. Compare this to William Shakespeare's complete works, comprising around 800,000 words. Martin used 22,000 different distinct words to tell his story, with a variability of around 1% (variability is in this case calculated as the distinct words divided by the total words in the text, also known as Type-Text Ratio (TTR)).
Variability is one of many different measures of the complexity of the text, the richness of the vocabulary, and to what degree repetition is avoided. We use a technique called lemmatization to avoid counting the same word in different forms more than once. For example, the two verbs "run" and "ran" should only be counted as one distinct word; lemmatization helps us to reduce words like these to their common root-word. Shakespeare is said to have one of the richest vocabularies in print, with Hamlet's 30,000 words written using 4,200 distinct words and an impressive variability of 13%! However, comparing texts with such a large difference in volume is an unfair comparison, since one's ability to use unique words decreases as the volume of a text increases. Looking at Shakespeare's complete works we find a variability of around 3%. Going further still we can take the moving average of the variability (looking at a window of 1000 words at a time) and we end up with a TTR-score for Martin of 43% compared to Shakespeare's 41%. It should be noted that Martin introduces a host of invented names for all the fantastical characters, objects, and places, making up a large part of the variability.

## C.2 | Video A Quiz - Recall

Based only on what you remember from the lecture you just watched, please answer the following questions. (No Googling please!)

Lemmatization is a technique to:

- Reduce different forms of words to their common root-word.
- Count the total number of words in a text.
- Calculate how many invented words are used.
- Determine the difference in volume between two texts.

In lexical diversity, variability is:

- The length of the text in number of words.
- The ratio of the number of unique words to the total number of words.
- The estimated measure of information density.
- The measurement of how different the words in a text are from each other.

Hamlet has a variability of:

- 1%
- 13%
- 41%
- 43%

The volume of *A Song of Ice and Fire* is:

- 1.9 million words
- 800,000 words
- 1 million words
- 1.8 million words

Using a moving average of variability based on a 1000-word window, Martin's Type-Text Ratio score is:

- Higher than Shakespeare's.
- Lower than Shakespeare's.
- The same as Shakespeare's.
- Not possible to calculate.

→

## C.3 | Video A Quiz - Comprehension

The passage suggests that Martin's work displays which of the following qualities?

- Use of repetition to emphasize key themes and ideas.
- An attempt to elevate fantasy to an artistic status more closely approximating that of other genres of literature.
- A tendency to incorporate elements from the works of Shakespeare.
- Use of language that gives the text richness and complexity.

What is a true statement about lexical variability?

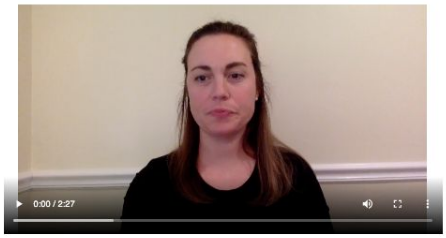
- High variability indicates a lower number of unique words relative to the volume of a text.
- In general, variability decreases as the volume of a text increases.
- Variability is a measurement of the amount of information in a text.
- Low variability is a sign that a text is highly complex.

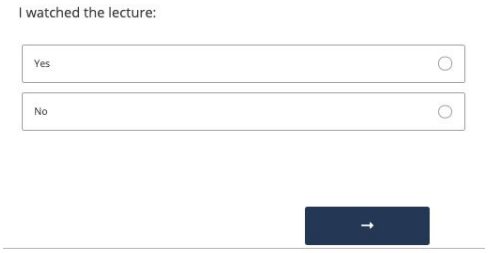
What is one reason that *A Song of Fire and Ice* has high variability based on a 1000-word window?

- Martin uses many nouns, adjectives, verbs, and adverbs.
- Martin uses a large number of words overall.
- Martin uses many invented words.
- Martin uses many forms of the same root-word.

→

## C.4 | Video B Script




We calculate density by taking the number of lexical words divided by the total number of words. Lexical words are nouns, adjectives, verbs, and adverbs. These types of words are considered the key bearers of information; they give a text its meaning. The other words, called function words, include the articles (a, the), prepositions (on, at, in), and conjunctions (and, or but). The function words are important for the grammatical structure of the text but carry little information on their own about the meaning of the text. Density is therefore an effort to say something of how informative the text is respective to its length. On average, we have a much higher density in written text than in our speech; the majority of written text in English is estimated to have a lexical density above 40%, while speech is on average less than 40%. In order to calculate density, we need to perform part-of-speech tagging (POS), also known as grammatical tagging, which helps us to identify all the words' grammatical forms. POS-tagging constitutes a non-trivial part of natural language processing (indeed, the lemmatization described earlier relies on POS-tagging in order to function properly). The POS-algorithm needs to be able to separate ambiguous word-forms where a word such as "duck" could be a noun or a verb depending on the context. There are several different techniques we could use; algorithms such as Hidden Markov Models represent a well-proven approach to identify the correct form by taking the nearby words into account. We arrive at a lexical density of 58% for A Song of Ice and Fire compared to 61% for Shakespeare's collected works. However, density has limitations as a metric. For instance, the phrase "to be or not to be" has a lexical density of only 17%, where "to be" is considered an auxiliary verb and not part of the lexical words. However, most would agree that it is quite a memorable quote. This example shows that the interpretation of the different lexical diversity metrics must be exercised with caution, and we should not infer any differentiating characteristics without going deeper into the text.

## C.5 | Video B Quiz - Recall

Based only on what you remember from the lecture you just watched, please answer the following questions. (No Googling please!)

In lexical diversity, density is:

- The length of the text in number of words. ○
- The ratio of the number of unique words to the total number of words. ○
- The number of lexical words divided by the total number of words. ○
- The measurement of how different the words in a text are from each other. ○

Function words are:

- Key bearers of information that give a text its meaning. ○
- Words such as articles, prepositions, and conjunctions. ○
- Not important to the grammatical structure of the text. ○
- Used to calculate lexical diversity. ○

The density of *A Song of Ice and Fire* is:

- 61% ○
- 60% ○
- 59% ○
- 58% ○

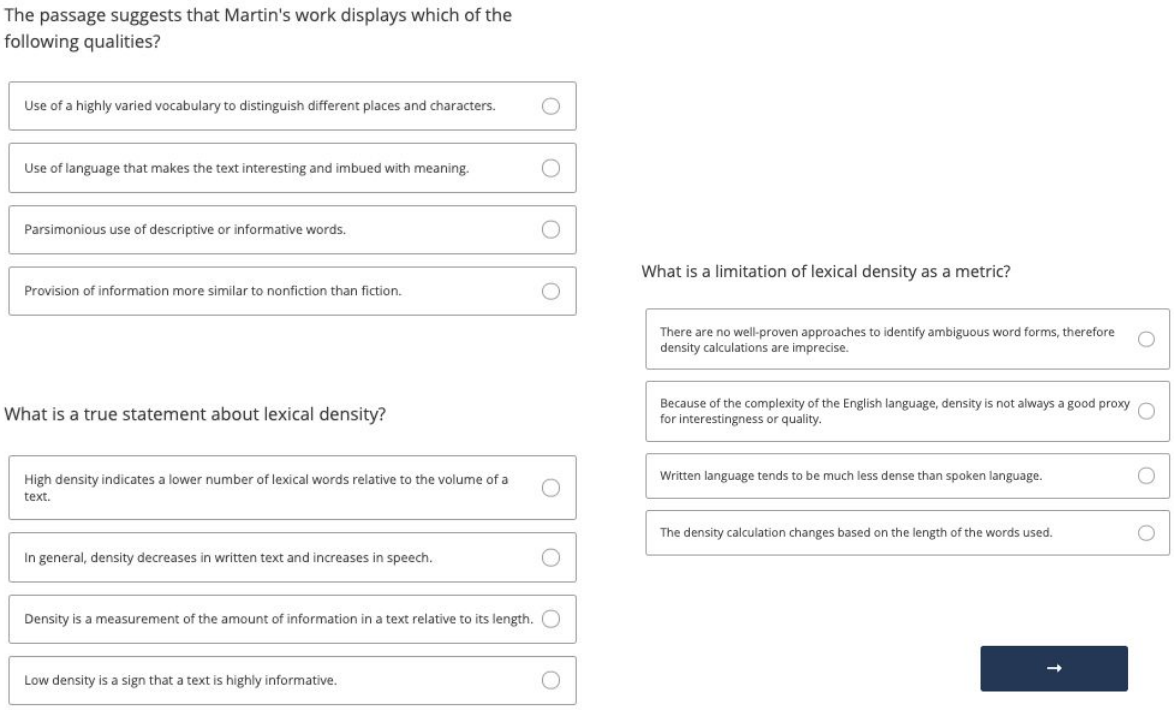
In this passage, the acronym POS stands for:

- Part of Speech ○
- Point of Sale ○
- Probability of Success ○
- Product of Sums ○

Written text typically has a lexical density:

- Higher than spoken text. ○
- Lower than spoken text. ○
- The same as spoken text. ○
- That is not possible to calculate. ○

## C.6 | Video B Quiz - Comprehension




Thank you for completing this survey. Your answers have been recorded. Please keep an eye out for the follow-up quiz which you will receive one day from now!

## C.7 | Phase II Quiz - Retention

Based on your memory of the Natural Language Processing lectures you watched yesterday, please answer the following questions. (No Googling please!)

In lexical diversity, variability is:

- The length of the text in number of words.
- The ratio of the number of unique words to the total number of words.
- The estimated measure of information density.
- The measurement of how different the words in a text are from each other.

The passage about lexical variability suggests that Martin's work displays which of the following qualities?

- Use of repetition to emphasize key themes and ideas.
- An attempt to elevate fantasy to an artistic status more closely approximating that of other genres of literature.
- A tendency to incorporate elements from the works of Shakespeare.
- Use of language that gives the text richness and complexity.

Lemmatization is a technique to:

- Reduce different forms of words to their common root-word.
- Count the total number of words in a text.
- Calculate how many invented words are used.
- Determine the difference in volume between two texts.

What is a true statement about lexical variability?

- High variability indicates a lower number of unique words relative to the volume of a text.
- In general, variability decreases as the volume of a text increases.
- Variability is a measurement of the amount of information in a text.
- Low variability is a sign that a text is highly complex.

In lexical diversity, density is:

- The length of the text in number of words.
- The ratio of the number of unique words to the total number of words.
- The number of lexical words divided by the total number of words.
- The measurement of how different the words in a text are from each other.

What is a true statement about lexical density?

- High density indicates a lower number of lexical words relative to the volume of a text.
- In general, density decreases in written text and increases in speech.
- Density is a measurement of the amount of information in a text relative to its length.
- Low density is a sign that a text is highly informative.

Function words are:

- Key bearers of information that give a text its meaning.
- Words such as articles, prepositions, and conjunctions.
- Not important to the grammatical structure of the text.
- Used to calculate lexical diversity.

What is a limitation of lexical density as a metric?

- There are no well-proven approaches to identify ambiguous word forms, therefore density calculations are imprecise.
- Because of the complexity of the English language, density is not always a good proxy for interestingness or quality.
- Written language tends to be much less dense than spoken language.
- The density calculation changes based on the length of the words used.

Thank you for contributing to science!

In appreciation of your participation, you are eligible to win one of five $100 Amazon gift cards to be awarded to the highest scores.

# Appendix D - Recruitment Messages

## Recruitment Email/Post - Phase I

Dear <name>,
You have been selected to take part in a short online assessment for a chance to win a $100 Amazon gift card. The assessment will involve watching two videos (2.5 min), each followed by a multiple-choice quiz. 1 day after completion, you will receive a new link to a short follow-up quiz. Once this is completed, your scores will be assessed, and the top 5 scores will each receive a $100 Amazon gift card (NOTE: Participants must complete the first two quizzes and the follow-up quiz to qualify).

Please start the first phase by clicking the link below. After each video, make sure to answer all the questions on the quiz, and keep an eye out for the follow-up quiz which you will receive 1 day after completion. (The assessment works best on a laptop or desktop - if on a mobile device, make sure to scroll down to access the videos.)

https://berkeley.qualtrics.com/jfe/form/SV_0wcOuLzJH7ffjMh

## Recruitment Email - Phase II (one day later)

Please use the following link for Phase II of the Decoding Game of Thrones study:

https://berkeley.qualtrics.com/jfe/form/SV_787v1m4sixfweu9

Thank you so much for your participation!

## Recruitment Email - Phase II (non-compliers)

Thank you so much for taking part in phase 1 of "Decoding Game of Thrones." Our records indicate you have not yet completed the follow-up quiz. You have 24 hours to complete it before the survey closes.

The follow-up quiz is only 10 questions and should only take a few minutes. Remember, participants must complete the follow-up quiz to qualify for the $100 Amazon gift card drawing.

https://berkeley.qualtrics.com/jfe/form/SV_787v1m4sixfweu9

Thank you so much, and best of luck!

# Appendix E - Covariate Survey Questions

The following questions were the first questions asked of all subjects.

Thank you for your participation in our study, which will take place in two phases. In Phase I of the study, we ask for your email address in order to send you a link for Phase II of the study. Your email will not be included in our research dataset, published, disseminated, or used for any form of personal identification of your data. We are also collecting other personal demographic data; this data will be used for research purposes only and will be published only in an aggregated form. We take your privacy seriously and will make every effort to respect and protect it.

- ○ I agree.
- ○ I do not agree.

My email address:

Age - I am currently:

- ○ Under 21 years old
- ○ 21-30 years old
- ○ 31-40 years old
- ○ 41-50 years old
- ○ Over 50 years old

Ethnicity - I identify as [select all that apply]:

- ☐ White
- ☐ Black or African American
- ☐ American Indian or Alaska Native
- ☐ Asian
- ☐ Native Hawaiian or Pacific Islander
- ☐ Other

Language - English is my:

- ○ First language
- ○ Second language
- ○ Third or more language

Education - I am currently a student:

- ○ Yes
- ○ No

Gender - I identify as:

- ○ Female
- ○ Male
- ○ Non-Binary