

## **Exploring the relationship between my English words learning process and the language I choose for searching information**

### **Overview**

I used to take notes of new English words for exams back in high school but gradually dropped this habit after I graduated. Recently, I noticed that I was often searching for the meanings of words that were searched before. This meant that I was not actually learning the words. Thus, I was trying to see if I can improve my English via the process of working on this final project.

Additionally, I found out that when I'm searching for information on the Internet or videos on YouTube, I'm more likely to use Chinese first instead of English. However, in many cases English contents are more accurate and useful. Therefore, besides improving my English, I'm also encouraging myself to read and watch more English contents instead of passively learning by just collecting data when needed. I hope to discover if I will be more willing to search in English while I am familiar with more English words.

For this final project, I was exploring myself by collecting data according to the topic I decided to work on which is about my learning process and the language I choose for searching information. Through the whole process of the project, I first made hypotheses about the relationships between the data I was collecting. Then I tried to sketch different ideas for the final visualization. Finally, I was able to explore the relationship between my English words learning process and the language I choose for searching information according to my final visualization.

### **Collecting data**

Following are the data I was collecting:

- **The date of every observation(YYYY-MM-DD):** I started to collect my data

every day since October 18th until December 15th.

- **Amount of New English Words I noted every day:** When I read or heard an English word that I didn't know, I noted it down. At the end of the day, I recorded the total number of the words I noted.
- **Average difficulty of the noted English words each day:** For every noted word I gave a difficulty (1-5) based on how hard I thought it was to learn the word. The words with more letters or special terms from different profession were more likely to have higher difficulty. At the end of the day, I calculated the average difficulty of the words.
- **Amount of words that I have collected before but still noted down the other day after**
- **The amount of searches I made every day and how many are them is in English:** I recorded every searches I made on different search platform (Google, YouTube, Amazon, etc.) so I can calculate the percentage of English searched I made each day.
- **Two test scores (%) I get on Quizlet (a learning app):** Quizlet is a learning app (See official website of Quizlet: <https://quizlet.com>) that provide a function to create tests based on the words noted by the users. The scores I got on the tests could indicate how well I was learning new English words. At the end of each day, I took a test based on the words noted on that day. In addition, to see if I was really memorizing the words, I also took another test based on the same list of words on the next day.
- **How many tries I take to get 100% on the tests I took on Quizlet**

Table of data without second test scores every day

Date	Noted New Words each Day	Average Difficulty of Noted New Words each Day (0-5)	Amount of Words Noted Before each Day	Total Searches each Day	English Searches each Day	Test performance on Quizlet		
						How Many Words tested	Score (%)	How many tries to get 100%
2021-10-18	11	3.1	0	5	3	11	100	1
2021-10-19	19	3.5	0	7	4	19	84	3
2021-10-20	15	2.9	0	16	7	15	100	1
2021-10-21	9	2.8	0	9	2	9	100	1
2021-10-22	13	3.7	0	7	5	13	92	2
2021-10-23	18	3.3	0	10	7	18	72	3

Table of data after I started collecting second test scores every day

Date	Noted New Words each Day	Average Difficulty of Noted New Words each Day (0-5)	Amount of Words Noted Before each Day	Total Searches each Day	English Searches each Day	Test performance on Quizlet			
						Which Day's words tested	How Many Words tested	Score (%)	How many tries to get 100%
2021-10-24	9	2.9	0	13	10	Same day	9	89	2
						The day before	18	67	3
2021-10-25	5	3.4	0	3	2	Same day	5	80	2
						The day before	9	67	2
2021-10-26	9	3.0	0	2	1	Same day	9	100	1
						The day before	5	80	2
2021-10-27	14	3.4	0	6	3	Same day	14	93	2
						The day before	9	89	2
2021-10-28	6	2.8	0	5	3	Same day	6	100	1
						The day before	14	100	1
2021-10-29	12	3.5	0	4	2	Same day	12	100	1
						The day before	6	83	2
2021-10-30	2	3.5	0	6	4	Same day	2	100	1
						The day before	12	92	3
2021-10-31	9	3.1	0	7	2	Same day	9	89	2
						The day before	2	100	1
2021-11-01	17	3.8	1	7	4	Same day	16	81	2
						The day before	9	89	2
2021-11-02	2	3	0	5	3	Same day	2	100	1
						The day before	16	94	2
2021-11-03	9	3.2	0	2	2	Same day	9	100	1
						The day before	2	100	1
2021-11-04	7	3.4	0	4	4	Same day	7	86	2
						The day before	9	67	2
2021-11-05	2	3	0	0	0	Same day	2	100	1
						The day before	7	57	2
2021-11-06	6	3.2	1	4	1	Same day	6	100	1
						The day before	2	100	1
2021-11-07	7	3.4	0	3	3	Same day	7	67	2
						The day before	6	100	1

2021-11-08	3	4	0	3	3	Same day	3	100	1
						The day before	7	100	1
2021-11-09	2	2	0	4	3	Same day	2	100	1
						The day before	3	100	1
2021-11-10	9	3.2	0	6	3	Same day	9	89	2
						The day before	2	100	1
2021-11-11	11	3.8	1	3	0	Same day	11	64	2
						The day before	9	67	2
2021-11-12	5	3.2	0	3	1	Same day	5	100	1
						The day before	11	73	2
2021-11-13	8	2.9	0	9	4	Same day	8	100	1
						The day before	5	60	2
2021-11-14	4	3	0	7	3	Same day	4	100	1
						The day before	8	75	2
2021-11-15	3	3.3	0	2	0	Same day	3	100	1
						The day before	4	50	2
2021-11-16	7	3.4	0	4	2	Same day	7	100	1
						The day before	3	100	1
2021-11-17	8	3.8	0	3	0	Same day	8	88	2
						The day before	7	71	2
2021-11-18	3	4.3	0	4	1	Same day	3	66	2
						The day before	8	75	2
2021-11-19	3	3.3	0	2	0	Same day	3	100	1
						The day before	3	100	1
2021-11-20	12	3.3	0	7	5	Same day	12	92	2
						The day before	3	100	1
2021-11-21	7	2.9	1	7	3	Same day	7	71	2
						The day before	12	58	4
2021-11-22	4	3.8	0	5	2	Same day	4	75	2
						The day before	7	57	2
2021-11-23	1	3	0	3	1	Same day	1	100	1
						The day before	4	100	1
2021-11-24	8	3.4	0	3	0	Same day	8	100	1
						The day before	1	100	1
2021-11-25	4	3.5	0	6	5	Same day	4	100	1
						The day before	8	100	1
2021-	13	2.9	1	2	0	Same day	13	75	2

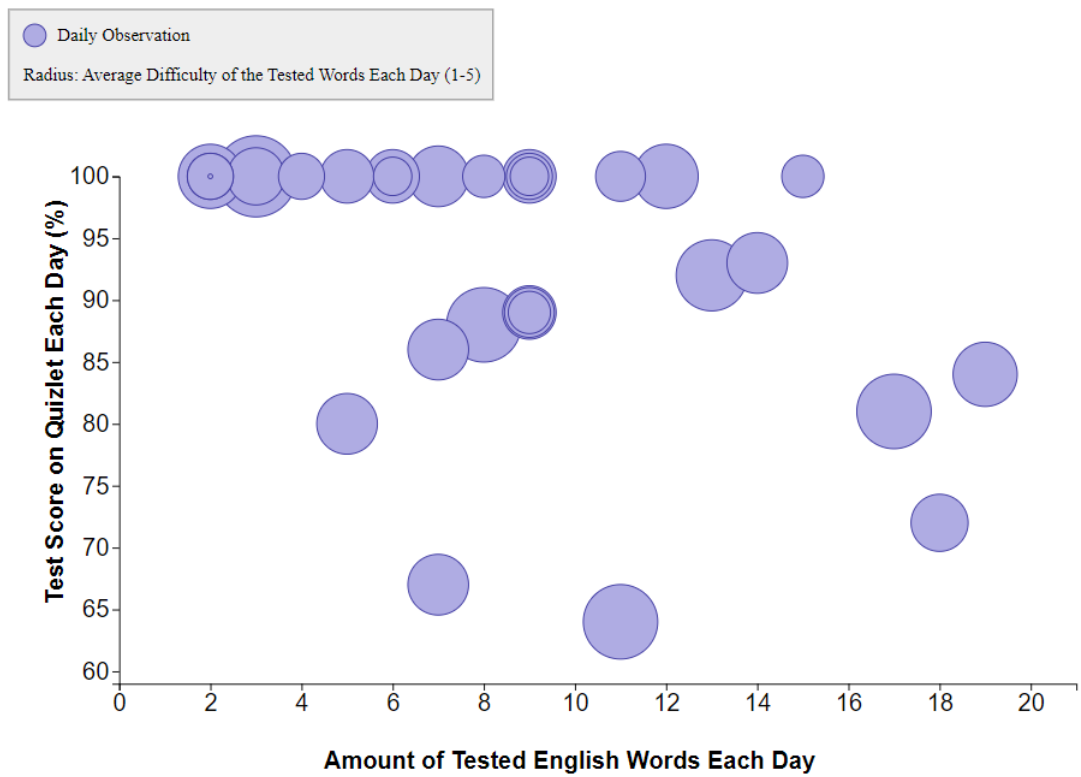
11-26						The day before	4	77	3
2021-11-27	9	2.9	0	4	3	Same day	9	78	2
						The day before	13	62	2
2021-11-28	3	4	0	6	4	Same day	3	100	1
						The day before	9	89	2
2021-11-29	4	3.3	0	2	1	Same day	4	100	1
						The day before	3	100	1
2021-11-30	2	3.5	0	5	4	Same day	2	100	1
						The day before	4	100	1
2021-12-01	6	3.2	0	13	6	Same day	6	83	2
						The day before	2	100	1
2021-12-02	16	3.1	2	5	4	Same day	16	81	2
						The day before	6	100	1
2021-12-03	4	3.5	0	3	2	Same day	4	100	1
						The day before	16	50	3
2021-12-04	5	2.8	0	14	4	Same day	5	60	2
						The day before	4	100	1
2021-12-05	2	2	0	3	2	Same day	2	100	1
						The day before	4	80	2
2021-12-06	8	3	0	4	0	Same day	8	88	2
						The day before	2	100	1
2021-12-07	5	2.8	0	7	2	Same day	5	100	1
						The day before	8	100	1
2021-12-08	3	3.3	0	2	1	Same day	3	100	1
						The day before	5	80	2
2021-12-09	2	2	1	6	5	Same day	2	100	1
						The day before	3	100	1
2021-12-10	10	3.7	0	9	5	Same day	10	70	2
						The day before	2	100	1
2021-12-11	4	3.5	0	2	2	Same day	4	100	1
						The day before	10	90	2
2021-12-12	3	3.7	0	6	1	Same day	3	100	1
						The day before	4	100	1
2021-12-13	6	3.7	0	4	3	Same day	6	100	1
						The day before	3	100	1
2021-12-14	5	3.4	0	3	1	Same day	5	100	1
						The day before	6	100	1

2021-12-15	2	4	0	2	0	Same day	2	100	1
						The day before	5	80	1

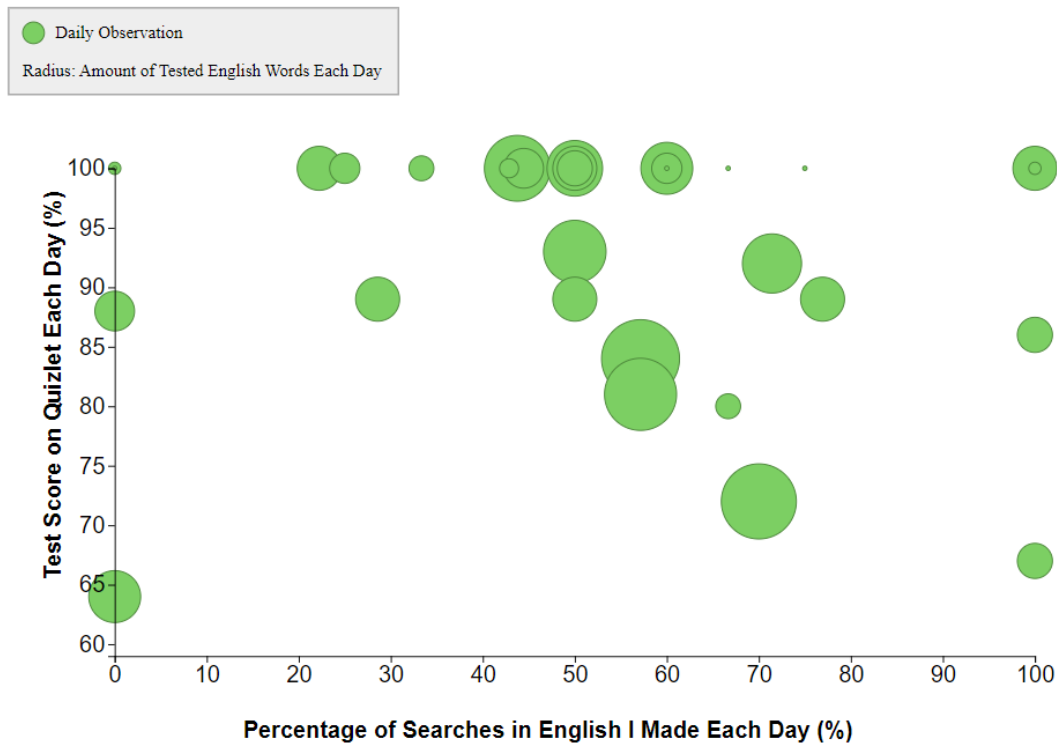
### Initial Hypothesis

At the beginning of this project, I made two hypotheses based on the data I was planning to collect. First was *as the amount of English words I'm unfamiliar with decreases, the frequency of searching information in English increases*. To see if my data support my hypotheses, I created scatterplots to observe the relationship between the factors. Second was *as the amount of noted English words increases, my test scores on Quizlet decreases*.

As shown in Scatter Plot 1 below, I was able to my scores decreasing when the amount of noted English words is greater than 10. So the plot was able to prove my hypothesis under a certain condition. On the other hand, Scatter Plot 2 shows the lack of relationship between the percentage of my English searches and my test performance.



Scatter Plot 1



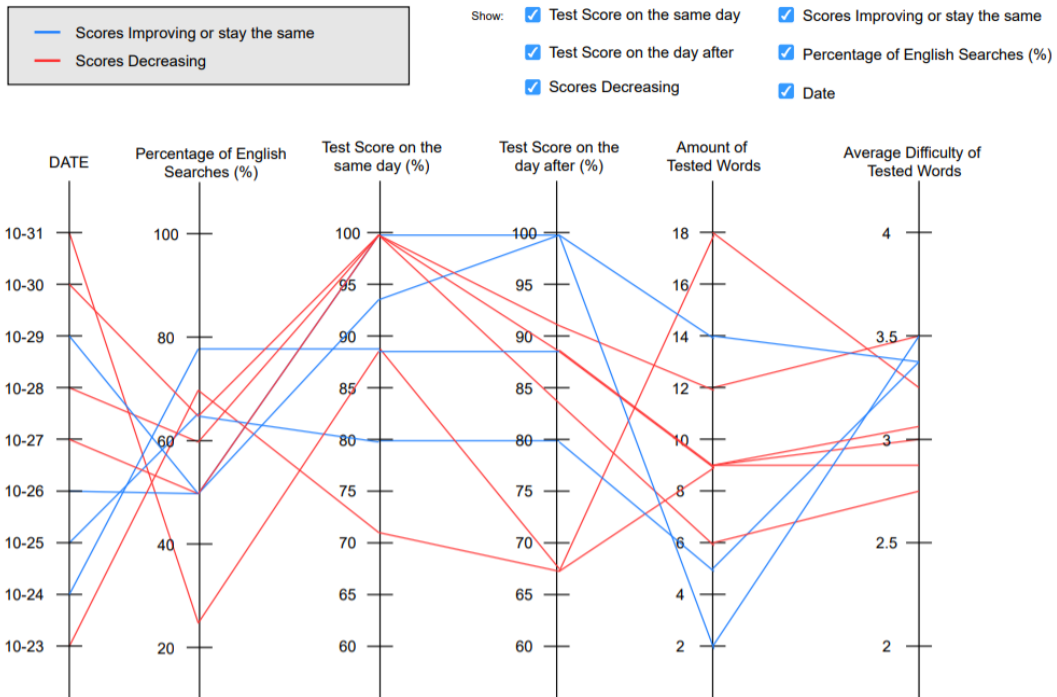
Scatter Plot 2

## Sketching Visualization

After creating the scatter plots above, I found that the way I plotted my data was not able to see how my test performance and searching habit changing while I was familiar with more English words. Thus I started to consider how could I rearrange the value on x and y axis, the radius of the circles, and the color of the circles to represent the data I wanted to show on my plot. Also, this is the point I started to add my second test scores into my visualization so I can see if there would be anything new to explore besides my initial hypotheses.

One of the way I thought of was to use the **parallel coordinates plot** I saw from [https://datavizcatalogue.com/methods/parallel\\_coordinates.html](https://datavizcatalogue.com/methods/parallel_coordinates.html) because this is a graph that can include multiple factors and observe the relationships between them. I was mainly adding a date axis to observe the date and the percentage of English searches each day in this graph. Since as the day passing, I should be familiar with more English words, so the date could help me with exploring my initial hypothesis about the relationship between how am I familiar with English words and the language I choose for searching information.

### Relationships Between the Factors That Affect My Test Score On Quizlet



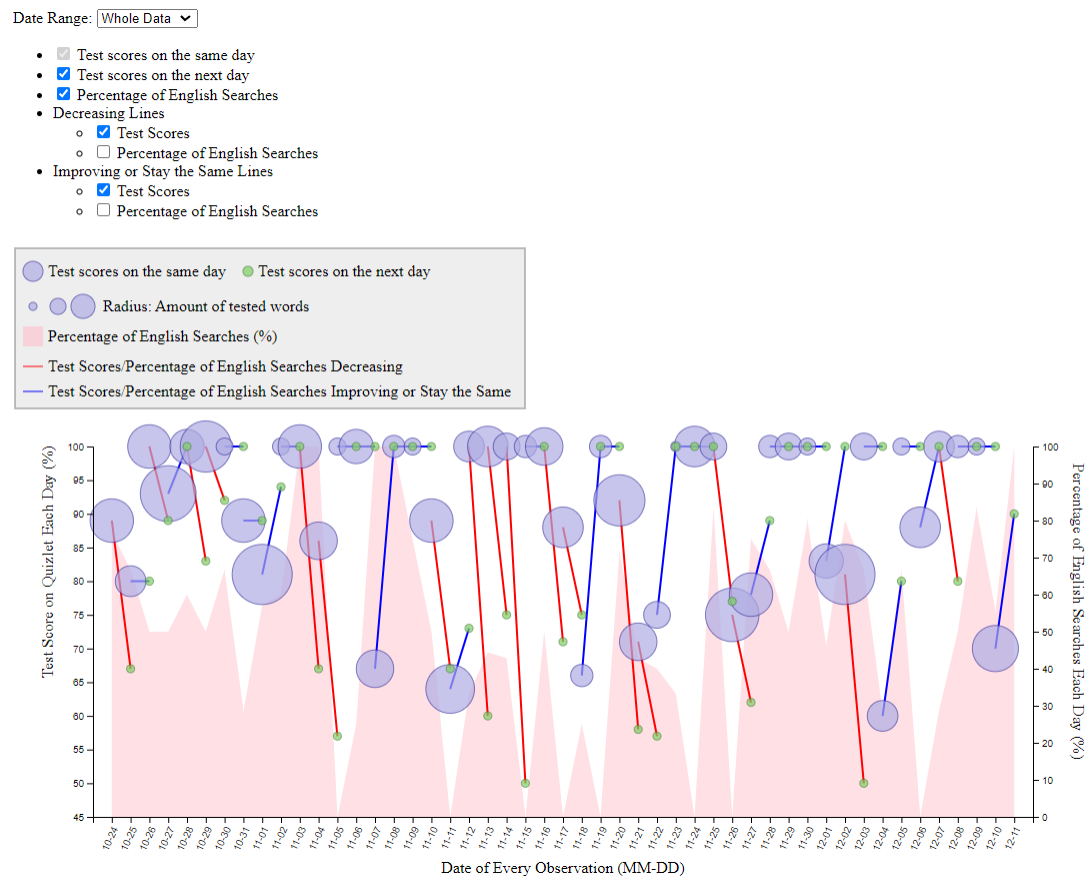
Sketch of Parallel Coordinates Plot Based on Part of My Data

I didn't end up using this sketch for my final visualization because the complexity of it. Though it may be able to contain multiple factors, the overlapping lines would hide a lot of my data. Furthermore, the order of the axis was also letting the relationships harder to be explored. Even when I tried to add options for the readers to help understanding this plot, I thought it would still be confusing.

Therefore, I chose another sketch for my final visualization and started work on creating the plot.



## Final Visualization: Test Scores on Quizlet Variation Plot



### Question I was exploring

After creating different scatter plots to support my hypothesis and sketching idea for my final visualization, my hypothesis turned into a further question which was: *As I am familiar with more English words, how will my searching habit affect my test scores on Quizlet?*

### Data I used for this visualization

For this final visualization, I didn't use every factor I was collecting for this plot. Here are the ones I was adding to the plot and how they were shown in the plot.

- **Date of every observation:** The X axis value of this plot. I only started taking two tests every day after October 23rd so I didn't show the data before on this final visualization.
- **Amount of tested new English words each day:** This is shown by the radius of the blue circle in this plot.
- **Amount of searches I made each day and how many of them are in English:** To calculate the percentage of English searched I made each day. This is shown by the pink area in this graph

- **Test scores on Quizlet:** This is shown by the Y axis value of the circles and the colors represent which day of words I was taking tests on.

### **Choosing the type of visualization**

For this visualization, I'm trying to create a graph that's **similar to a cleveland dot plot** that draws pairs of circles linked by lines. (Cleveland dot plot was mentioned in previous classes and there was an example provided in the `example_cleveland_dot_plot` file from class 9 materials.) However, to fit the data I am collecting, I made some adjustment in my plot. Since I wanted to see how well I am learning the new English words I noted every day, I'm not only taking tests at the end of everyday but also taking tests about the same list of words on the next day. By using the concept of cleveland dot plot, I can observe the changes between the scores of two tests I take for the same words on different days. I think cleveland dot plot suits well for my data when reviewing it so I tried to implement it to this visualization.

The other kind of graph I used in my visualization was **area graph** I learned from: [https://datavizcatalogue.com/methods/area\\_graph.html](https://datavizcatalogue.com/methods/area_graph.html). It's a simpler version of stacked area graph mentioned in class but I only used area graph just to see how the percentage of English searches I made changes day by day. I wanted to see the relationship between my searching habit and my test scores on Quizlet, so I add the area graph about the percentage of English searches I made every day as a background of this visualization.

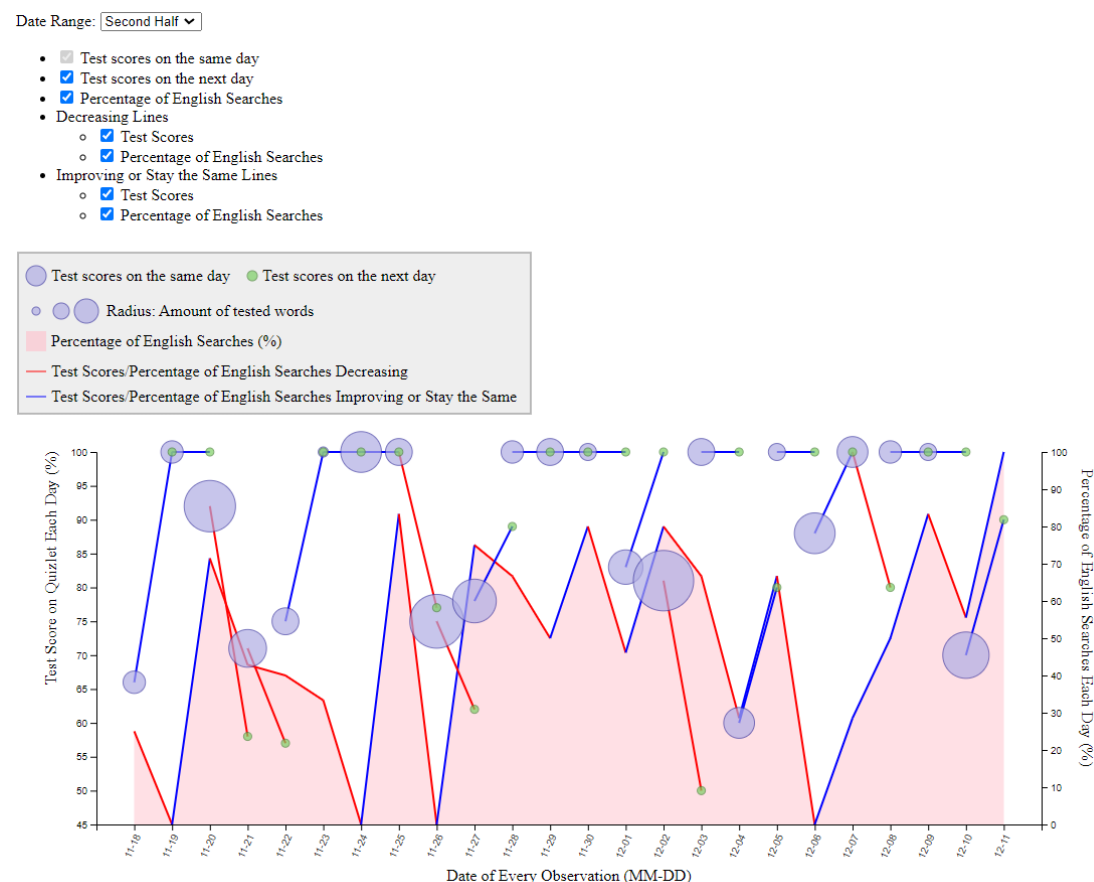
### **Adding Interactive Design to My Visualization**

Following the suggestions that I got from Prof. Jay and my classmates in class, I tried to add options to this visualization.

- By selecting the date range, the users can choose to see which part of my data so the circles won't be clustered as the plot for every day.
- By checking the box for the scores on the next day, the users can see the green circles and the lines that connecting every pair of circles.
- By checking the box for percentage of English searches, the users can see the area graph showing the change of my searching habit.
- By checking the box for different lines, the users can see the blue and red lines then see the relationships between them.

Also, I tried to add vertical lines to help users to see which date the circles are refer to when they move the mouse onto the circles, but there are already a lot of lines in my plot so I didn't add the vertical lines at last.

## My observation of this visualization



At first, I couldn't really find a clear relationship between the percentage of my English searches and my test scores on Quizlet. I was expected the scores would fall on lower positions when the percentage is also low but it didn't appear to be that way.

However, by observing the changes between the scores on different days, I could see that most of the red lines connecting the circles (represent decreasing scores) are matching the red lines on the area graph (represent decreasing percentage of English searches) and most of the blue lines are matching as well. Thus I found out that this plot was giving me a new idea of: *when the percentage of English searches I made increases on the next day, my scores on Quizlet are more likely to improve or remain the same on the next day*. That might be telling that searching in English was helping me memorize the words I learned on the previous day better so my test scores improved. By choosing which kind of lines to show on the plot using the provided checkboxes, the relationships can be easier to see.

In addition, I assume the amount of tested words is also a factor that affects my test scores, so the larger circles should fall on lower positions. According to this plot it quite support my expectation, I can see most of the smaller circles fall on the top of

this plot. That means I can learn better when the amount of noted words is under a certain number. By hiding the area graph of percentage of English searches and the test scores on the next day using the check box, I can focus more on the difference of the blue circles to see this relationship.

## **Conclusion**

At the end of this project, I was quite surprised with the results. I wasn't able to clearly prove my initial hypotheses when creating the scatter plots at the beginning. Even when I was trying different sketches for the final visualization, it seems to be harder to explore relationships between my data since I tried to add the most data as possible to them. Until when I shared my draft in class and got feed backs from Prof. Jay and my classmates, I started to see some ideas from the draft which became my final visualization. The way my plot told me that how the language I chose for searching information affects my test performance was not expected before.

Though the process of collecting the data every day kind of took me a lot of work but the result came out surprisingly interesting. Now that I have completed this project, I'm more curious about what's affecting the language I choose for searching besides my English words learning process. I think the topic I am searching for or the search engine I'm using will be the main factors that affect my searching habit.

## **Works Cited**

“Area Graphs - Learn about this chart and tools to create it.” *The Data Visualisation Catalogue*, [https://datavizcatalogue.com/methods/area\\_graph.html](https://datavizcatalogue.com/methods/area_graph.html).

“Parallel Coordinates Plot - Learn about this chart and tools.” *The Data Visualisation Catalogue*, [https://datavizcatalogue.com/methods/parallel\\_coordinates.html](https://datavizcatalogue.com/methods/parallel_coordinates.html).