PLSC 341 Project: The Difficulty Of Wordles

Jason Wang

May 9th, 2022

1 Introduction

The hit video game Wordle has hit the internet by storm, and its popularity shows no sign of waning. The objective of the game is to guess a five-letter word in six guesses, with *Mastermind*-like clues given after each guess as to the letter composition of the word. A green letter means the letter was guessed correctly and is in the right place, a yellow letter means a letter is in the word but is in the incorrect place, and an uncolored letter means the letter is not in the word.

Figure 1: A screenshot of Wordle.

Many variants of Wordle have popped up due to the game's popularity; one such example is Hello Wordl, which is nearly identical to the original game, save for the fact that players can adjust the length of the words from four to eleven letters. This in turn will adjust the difficulty of the game; but by how much and in what direction (easier/harder) does it do so? Through a blocked random experiment, we explore the changes in difficulty these modifications have on the game.

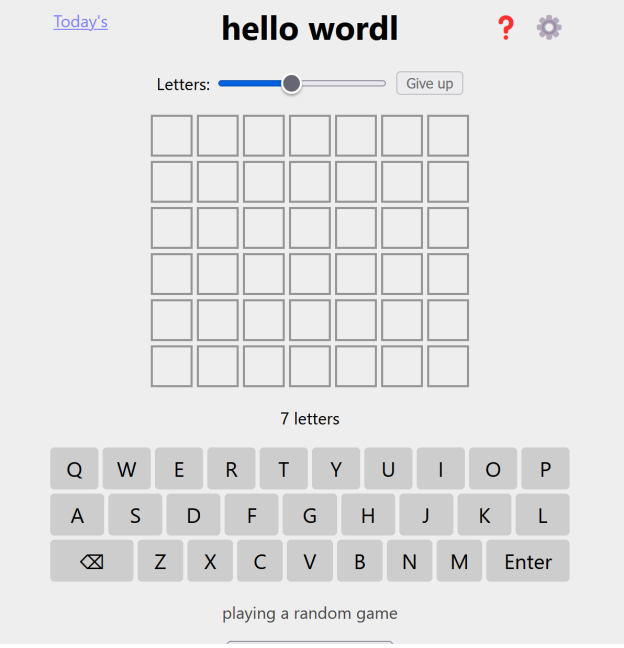


Figure 2: A screenshot of the Wordle variant Hello Wordl, with the word length adjusted to 7.

2 Theory

To measure the difficulty of a Wordle format, we observe two statistics: the average time taken and the average number of guesses for each player to solve a Wordle of that particular format. We must take into account both, as there is a number of different strategies players may take in solving a Wordle; some may prioritize the number of guesses taken, but take longer to make more strategic guesses (we shall call this type of player "Strategists"), while others may race to the finish line in order to get the answer quickly (dubbed "Speedrunners").

For longer Wordles, we expect the average amount of time taken to increase while the average number of guesses taken to solve decreases; this is because as more letters are introduced, although the number of possible words increases (up to a length of 8[2]), more information is given per word, helping to narrow down the number of guesses. However, as it is also harder to think of longer words of a certain length (as it also takes more time to internally count the number of letters in the word).

For shorter Wordles (which, in the case of this study, specifically refers to Wordles with 4 letters), it may also make sense for the average number of guesses to increase, since less information is provided per word. However, we may expect players to take less time per Wordle, since four-letter words are more commonly used than five-letter words, so players will be more familiar with them, even if there are objectively fewer distinct four-letter words in the English language.[1]

3 Experimental Design

The original plan as detailed in the pre-analysis consisted of pre-assigning Wordles of various lengths on Speedle, a fork of Hello Wordl. The original plan also called for a much wider range of treatments, including testing the effect of turning on "Hard Mode" (which restricts future guesses

based on the results of previous ones) and 8/9 letter Wordles. These additional treatments were eventually dropped due to time constraints. A pilot study was first conducted to test out the plan and any potential issues that may arise.

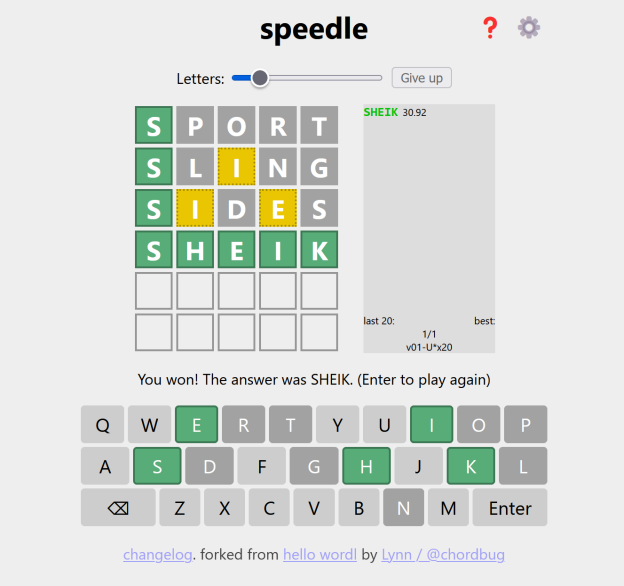


Figure 3: A screenshot of the Wordle variant Speedle.

3.1 Pilot Study

Over the course of a two-hour-long Wordle session (refreshments provided), five players were given Wordles to solve. The Wordles were block-assigned by player, with the experimental unit being the Wordles, and the randomly-assigned treatment being the length of the Wordle. Between solves, players were asked to submit the length of the Wordle, the time it took to solve the Wordle, and the amount of guesses taken to solve the Wordle using a Google Form. If the player did not solve a certain Wordle, they were asked to mark the score with "X" instead of a numeric value. Furthermore, players were notified that *both* time and score data were being collected per Wordle. This was to discourage Speedrunners from adjusting their strategy to that of the Strategists, as simply assigning Wordles does not make it immediately clear that time is of importance. Players were *not* specifically asked to refrain from collaboration, as the answers to the Wordles were not shared between players anyway, and the participants were known to collaborate for regular Wordles in the past.

A few issues came about. Firstly, players reported feeling pressured to finish quickly by the website due to past finish times being visible. While this may make sense for the original purpose of the website, speed-solving Wordles (hence the name Speedle), this is an issue for the study, as we are trying not to affect the strategy of the players.

Another issue is my fault; to save time, I asked the participants to randomize themselves instead of assigning them in a bid to save time. While this worked initially, players soon discovered that shorter Wordles were faster (and therefore perceived by them to be "easier"), so there was a large bias towards 4- and 5-letter Wordles being solved.

Additionally, there were issues with the timer; it seemed to start tracking the time for a Wordle

as soon as the previous Wordle was solved. While this might seem like a negligible difference, this may not be the case when players had to both copy over their results and adjust the length for the next wordle. This leads us to another issue:

The website did not provide a "share" button to copy the results into the clipboard, meaning that players had to manually copy over their results. Additionally, the site did not save the number of guesses it took to solve the Wordle, meaning that if a mis-click happened, there would be no way to recover the info and the data would be lost. On this note, it was very easy to accidentally skip past the results screen, since a single additional press of the "Enter" button after submitting a final guess would immediately send the player to the next Wordle.

An attempt was made to fix these issues in the revised study.

3.2 Revised Study

For the revised study, I forked Speedle on GitHub and modified the site to create my own version, *Researchdle*, to fit the needs of the study. To fix randomization issues, I used JavaScript's `Math.random()` function to randomly assign the Wordle length for each Wordle played. Additionally, I stored the most recent Wordle length generated by JavaScript in local storage to prevent players from being able to refresh the browser to obtain a new Wordle.

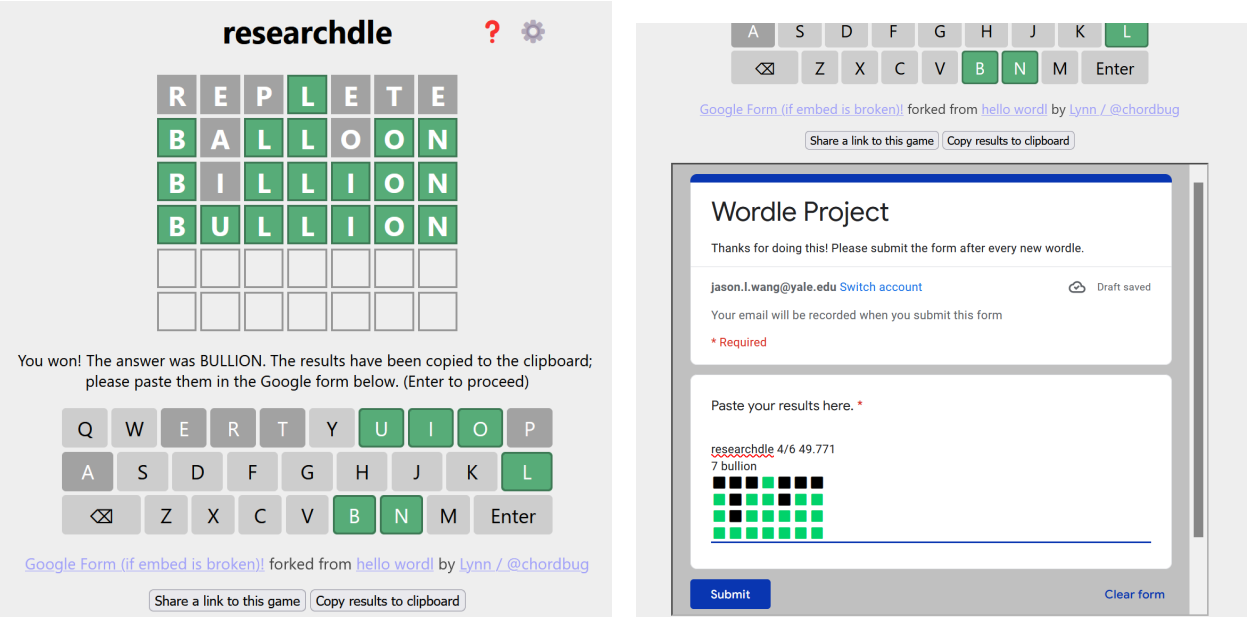


Figure 4: Two screenshots of Researchdle and the embedded Google Form used to submit data.

To help prevent mis-clicks, I added an extra step so that players would have to press "Enter" twice in order to move on to the next Wordle. I added a "share" functionality and embedded the Google form into the website, allowing for players to copy their results and submit without having to switch tabs. This had the additional benefit of keeping track of more data; whereas before it was only feasible to collect the number of guesses and the amount of time taken, now, it was possible to get detailed information about each guess the player made. This data falls beyond the scope of this paper and was not analyzed, but it is nonetheless provided in the replication archive. The source code for Researchdle is available in the GitHub repository linked in the Appendix.

This new website allowed for a smaller time commitment for the players as well. Instead of asking players to commit to a single session, players were asked to gradually complete as many Wordles as they desired over the course of about a week. This had the added side benefit of further reducing time pressure.

3.3 A note on failed Wordles

It is possible to fail to solve a Wordle, or in other words, fail to guess the Wordle within six guesses. This feature was kept in, as player strategies may change if infinite guesses were allowed. However, this means that we must impute a value for these failed Wordles, as we do not know the true number of guesses it would have taken the player to correctly solve it. We use a value of 7, because it seemed to be a fair assumption for players to guess the word given more attempt, but analysis was also done with a failure value of 10. This value can be adjusted by changing the `failureValue` variable in the provided R analysis script in the replication archive.

4 Results & Discussion

4.1 Results

From the revised study, 138 Wordle results were collected across 6 players. The mean score for 5-letter (control) Wordles was 4.917, as compared to mean scores for 4-, 6-, and 7-letter (treatment) Wordles of 5.317, 5.000, and 5.485, respectively. The difference in means estimates are 0.400 ($p = 0.201$), 0.083 ($p = 0.822$), and 0.568 ($p = 0.075$), respectively. Although none of the p-values are less than 0.05, so we fail to reject the null hypothesis that the length of the Wordle has an effect on its difficulty, as quantified by score, we note that the data for 7-letter Wordles, with $p < 0.10$, trends towards statistical significance. See Table 1 and Figure 5.

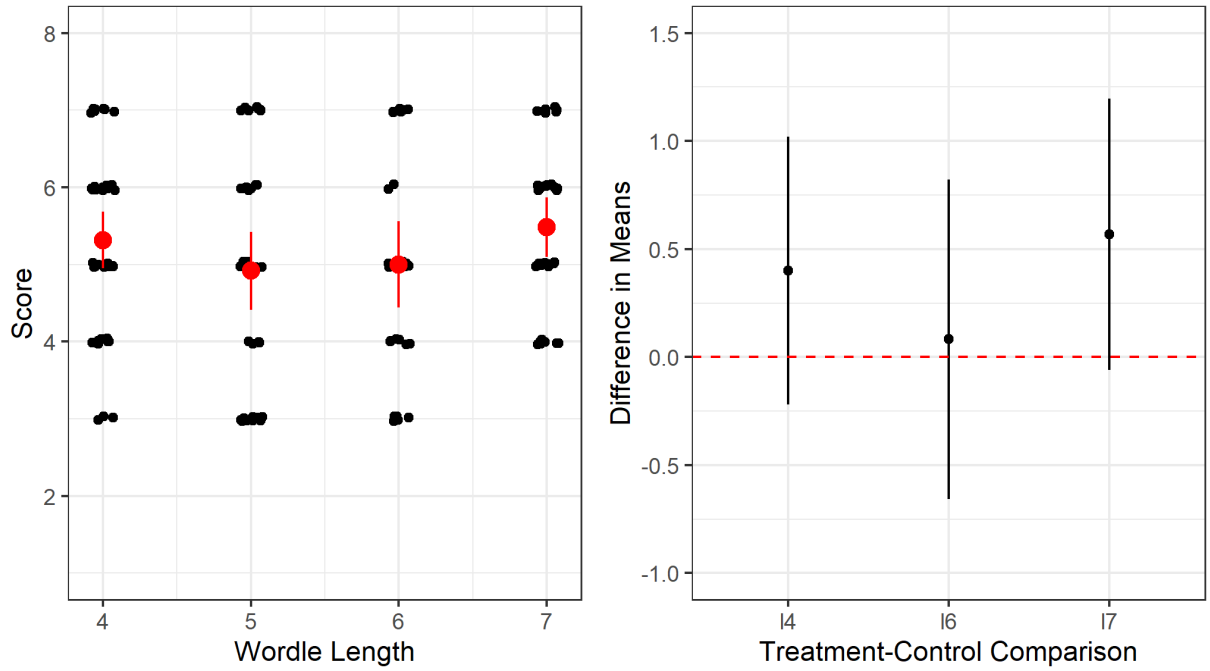


Figure 5: A plot of the score data collected, with failures being denoted as a 7.

Table 1: Difference in Means (score)

	6 v. 5	7 v. 5	4 v. 5	7 v. 4	6 v. 4	7 v. 6
(Intercept)	4.917***	4.917***	4.917***	5.317***	5.317***	5.000***
	(0.250)	(0.250)	(0.250)	(0.183)	(0.183)	(0.272)
l6	0.083				−0.317	
	(0.370)				(0.328)	
l7		0.568		0.168		0.485
		(0.314)		(0.264)		(0.332)
l4			0.400			
			(0.310)			
R^2	0.001	0.045	0.022	0.005	0.015	0.036
Adj. R^2	−0.015	0.031	0.009	−0.008	0.000	0.020
Num. obs.	64	69	77	74	69	61
RMSE	1.474	1.321	1.335	1.137	1.286	1.264

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

If we impute a score of 10 rather than 7 for failed guesses, then all the means rise, but the confidence intervals widen. The mean score for 5-letter Wordles was 5.500, as compared to mean scores for 4-, 6-, and 7-letter Wordles of 5.829, 5.750, and 6.121, respectively. The difference in means estimates are 0.329 ($p = 0.536$), 0.250 ($p = 0.536$), and 0.621 ($p = 0.273$), respectively. In this case, none of the observed p-values are less than 0.05, so we fail to reject the null hypothesis that the length of the Wordle has an effect on its difficulty, as quantified by score. See Table 2 and Figure 6.

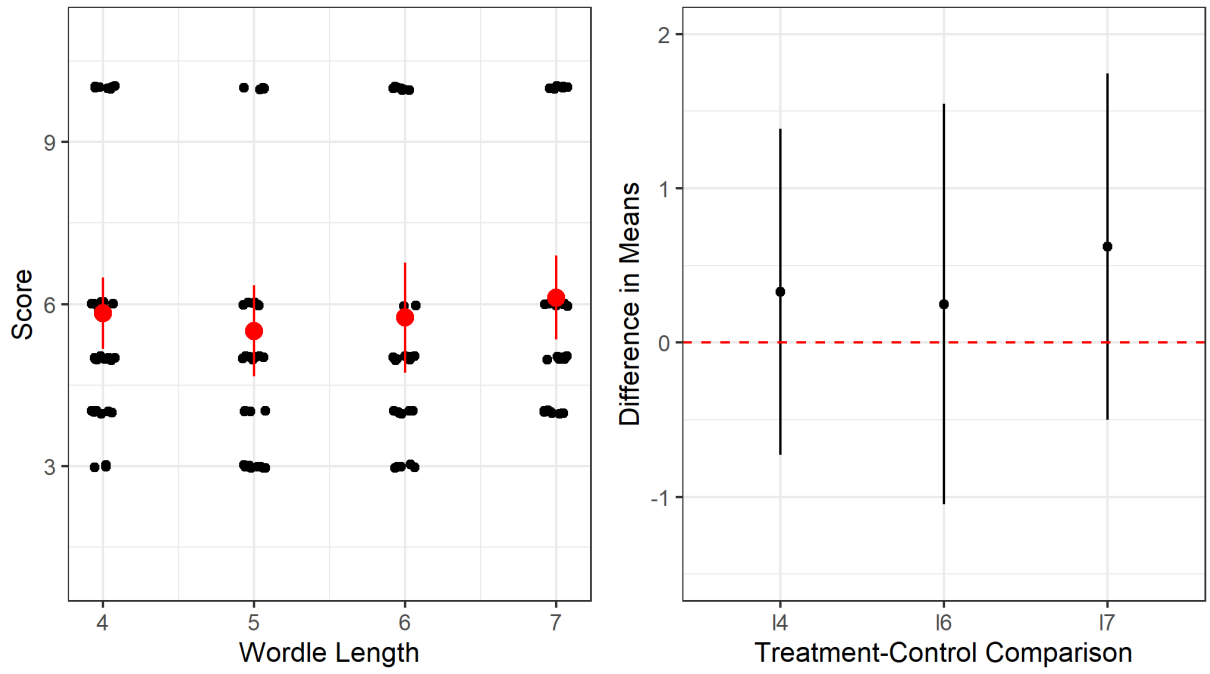


Figure 6: A plot of the score data collected, with failures being denoted as a 10.

Table 2: Difference in Means

	6 v. 5	7 v. 5	4 v. 5	4 v. 7	4 v. 6	6 v. 7
(Intercept)	5.500***	5.500***	5.500***	5.829***	5.829***	5.750***
	(0.415)	(0.415)	(0.415)	(0.329)	(0.329)	(0.498)
l6	0.250				−0.079	
	(0.648)				(0.597)	
l7		0.621		0.292		0.371
		(0.562)		(0.502)		(0.625)
l4			0.329			
			(0.530)			
R^2	0.002	0.018	0.005	0.005	0.000	0.006
Adj. R^2	−0.014	0.003	−0.008	−0.009	−0.015	−0.011
Num. obs.	64	69	77	74	69	61
RMSE	2.553	2.345	2.294	2.139	2.334	2.396

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

The mean time to solve a 5-letter control Wordle was 130.830s, while the mean times to solve 4-, 6-, and 7-letter Wordles were 97.234s, 261.190s, and 388.774s, respectively. The difference in means estimates are −33.596s ($p = 0.354$), 130.350s ($p = 0.015$), and 257.944s ($p = 0.0029$), respectively. Definitively, we reject the null hypothesis that the length of the Wordle has an effect on its difficulty, as quantified by *time*, for 6 and 7-letter Wordles, while we fail to reject the null hypothesis for 4-letter Wordles. See Table 3 and Figure 7.

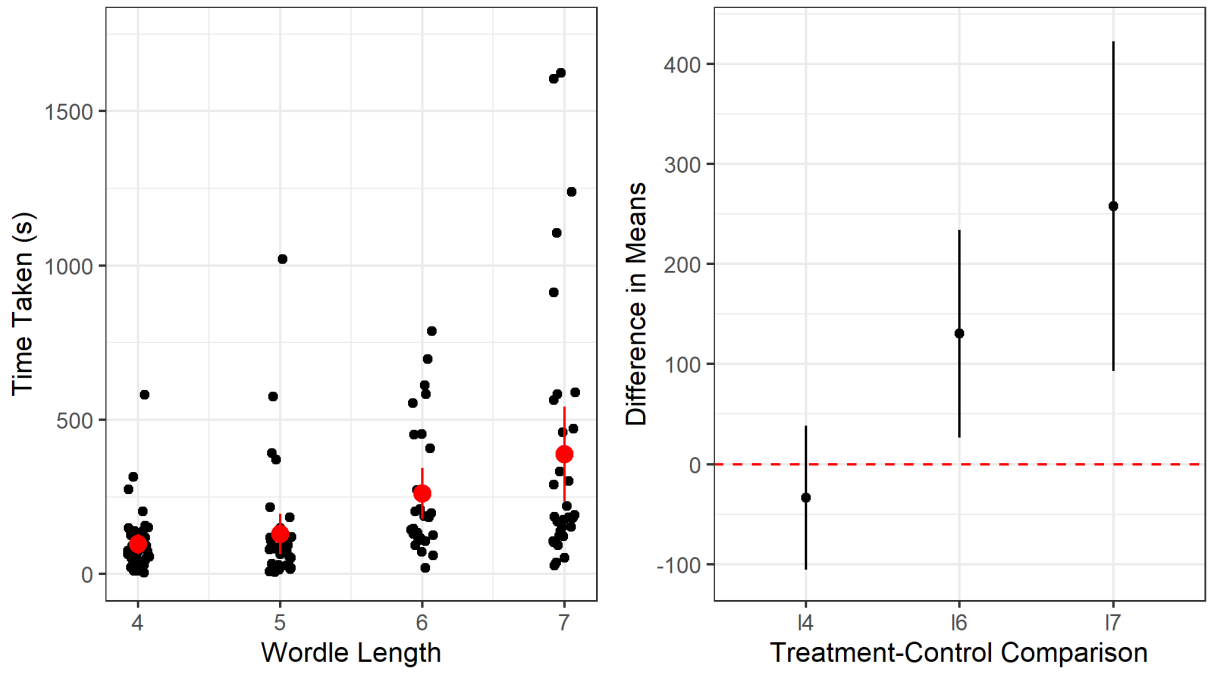


Figure 7: A plot of the time data collected.

Table 3: Difference in Means (time)

	6 v. 5	7 v. 5	4 v. 5	7 v. 4	6 v. 4	7 v. 6
(Intercept)	130.830***	130.830***	130.830***	97.234***	97.234***	261.180***
	(32.197)	(32.197)	(32.197)	(15.935)	(15.935)	(40.534)
l6	130.350*				163.946***	
	(51.765)				(43.554)	
l7		257.944**		291.540***		127.594
		(81.807)		(76.875)		(85.433)
l4			−33.596			
			(35.924)			
R^2	0.095	0.136	0.012	0.196	0.212	0.033
Adj. R^2	0.080	0.123	−0.001	0.184	0.201	0.017
Num. obs.	64	69	77	74	69	61
RMSE	202.734	329.602	151.552	297.886	157.337	349.690

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Additional plots showing the average treatment effects by player are available in the Appendix.

4.2 Discussion

The observed data seemed to fit the predictions as listed in the Theory section. Increasing the length of a Wordle appears to contribute noticeably to the difficulty in terms of time taken, and decreasing the length appears to potentially decrease the difficulty in terms of time taken. Although conclusions can not be drawn as definitively for the score data than the time data, the trends can still be seen in the coefficient plots; all variants in the length of the Wordle seemed to contribute a slight increase in the score.

One potential justification for the relative inconclusivity of the score data as compared to the time data is a outcome range issue; the possible outcomes for the score are 1 through 7, the failure value, whereas the possible outcomes for the time are any positive real number. Changing the failure value does not help, as it ends up increasing the variance in the data more than it affects the differences of means.

For future studies, more data should be collected so that data can be conclusively drawn for both metrics, score and time. Alternatively, it may be worthwhile to consider incorporating score and time data together into a combined difficulty index, or to consider metrics other than time or difficulty, such as an average number of correct letters for the first n guesses for some fixed n. It may also be helpful to consider a time penalty (and a penalty for any other difficulty metrics) for failed Wordles as well.

References

[1] Peter Norvig. English letter frequency counts:mayzner revisited or etaoin srhldcu, 2012. URL: `https://norvig.com/mayzner.html`.

[2] Ravi Parikh. Distribution of word lengths in various languages, 2013. URL: `http://www.ravi.io/language-word-lengths`.

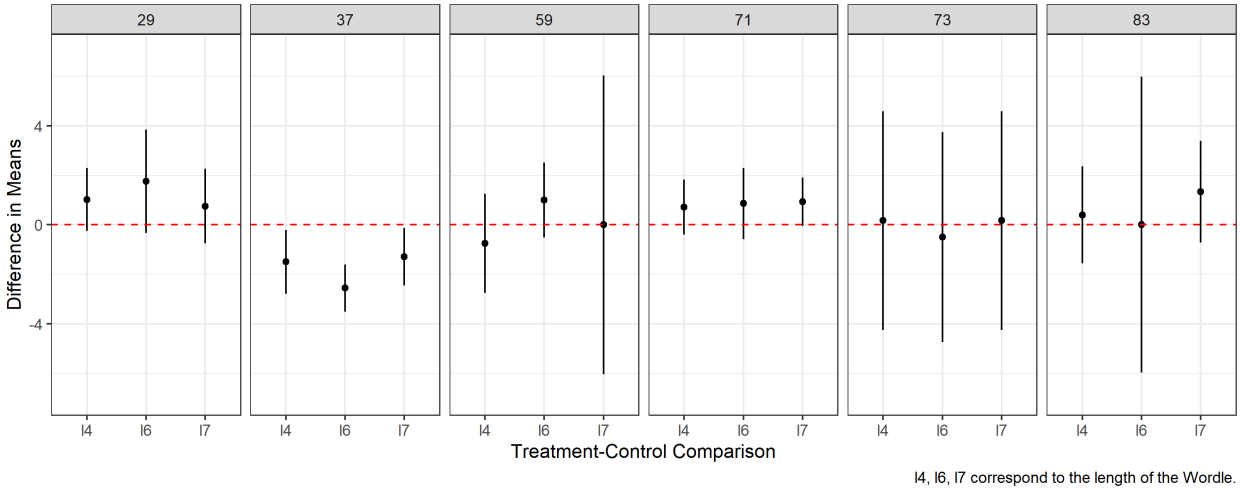
Appendix

Links

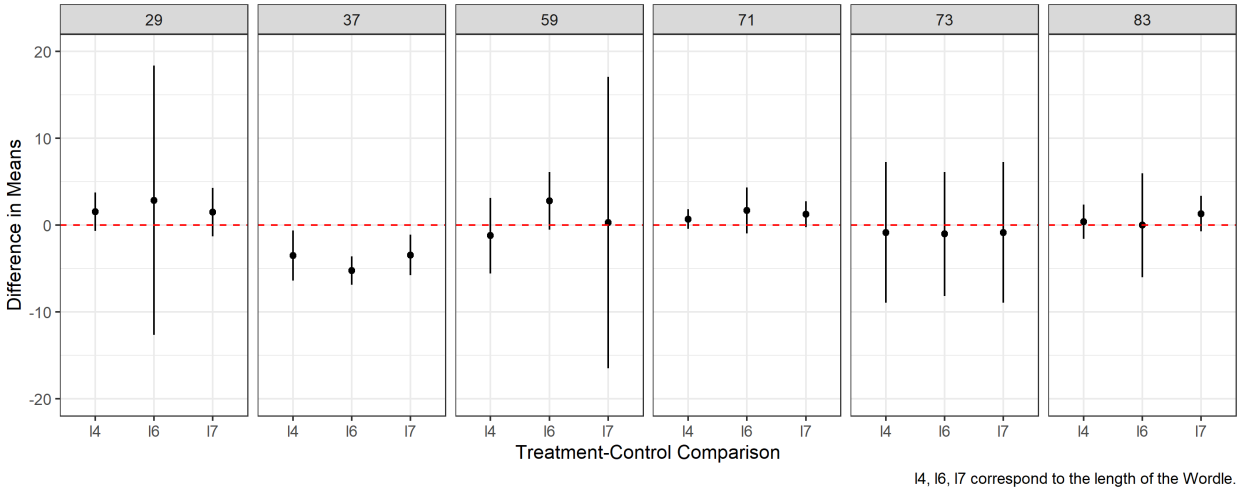
- Wordle: `https://www.nytimes.com/games/wordle/index.html`
- Hello Wordl: `https://hellowordl.net/`
- Speedle: `https://tck.mn/speedle`
- Researchdle: `https://jas0n.net/researchdle`
  - Github Repo: `https://github.com/piis31415/researchdle`

Plots

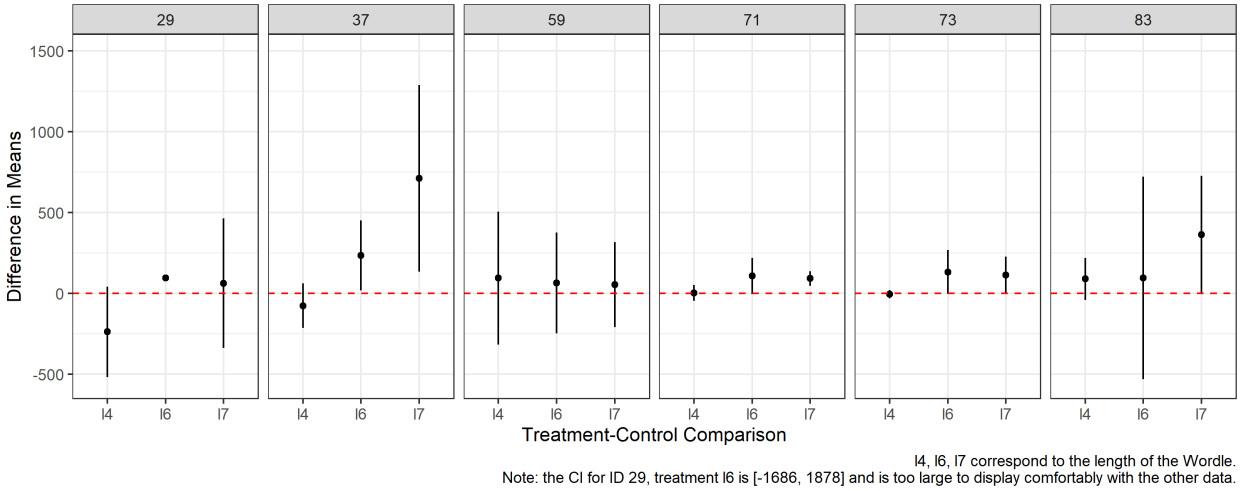
ATE Plots by Player



A ATE coefficient plot of the score data collected, with failures being denoted as a 7, by player ID.



A ATE coefficient plot of the score data collected, with failures being denoted as a 10, by player ID.



A ATE coefficient plot of the time data collected, by player ID.