Experiments and Casuality: Final Project

Danielle Adler, Osmar Coronel, Dan Kent

April 24, 2019

Abstract

The goal of our experiment is to assess the impact of music on cognitive function through the performance of activities that require logical reasoning. In our experiment, subjects were exposed to terrible music (the I Love You song by Barney and Friends) and white noise music during two Sodoku puzzle challenges back-to-back. Subjects were recruited through Mechnical Turk from different states / countries, with different levels of education, different levels expertise in logical games, and different ages ranging from 24 to 67 years old. The outcome measures we analyzed related to completion rate of the game as well as number of incomplete (white) spaces left after a given time limit. The results of our experiment did not find a significant casual relationship between different types of music and the ability to perform logic-related tasks.

Experiment Overview

Research Question

What makes individuals productive? Does different noise affect this productivity? Some individuals seem to work better with complete silence and no interruptions, while others need music or some level of stimulation. Others may need to read in silence, but code with loud music. Thus, stimulation needed seems not only different from person to person, but differ within each task a person performs. As we thought about a research question, we wondered if there were any generalizations that can be applied to the role of stimulation, cognitive ability, and focus. What helps us focus? What makes it harder for us to focus? What makes humans more productive on tasks at hand? The research question, holistically, is: **do different sensory stimuli impact cognitive ability?**

Why is This Interesting?

We thought about different ways to increase or decrease cognitive ability with music. While, many times, music choice depends on personality traits and environment, the aim of this experiment will be to see if any generalizations exist with music's impact on cognitive function. The music used to decrease cognitive function is motivated by the U.S. government, which makes it quite interesting to us. In wars, the U.S. government and many other governments have used music on repeat as a form of torture, elaborated in two BBC articles: one and two. Several songs have been very effective at allowing governments to gain access to valuable information. While the U.S. government has these songs on repeat for extended periods of time, our experiment only lasts a matter of minutes. Therefore, song use would not rise to the level of torture or unethical treatment of subjects. One of the key songs used by the gonverment that also seemed appropriate for our experiment was the children's song, I Love You by Barney and Friends.

For the cognitive function piece of the experiment, or what we will actually expect subjects to *do*, we chose Sudoku. This game is a key logic game that is relatively familiar to the general public and knowledge of a specific subject or language is not critical. We also anticipate that Sudoku is the type of game where stimuli changes could affect cognitive function.

What Observational Work has Been Done?

Our review of other observation work indicates that no studies have directly proven a causal relationship between bad or distracting music and logical reasoning performance; however, a number of studies address adjacent topics and lead us to a mix of conclusions:

- A study by PsycNet detailed working memory is susceptible to disruptions in speech and music, though familiar music had little effects on performance.
- One study by BioMed Central did not find "substantial and consistent" influence of background music on verbal learning.
- A study published in the Journal of Consumer Research indicated that a high level of noise hurts creativity.
- One review of existing studies by a team at McGill University identified that listening to music could reduce a patient's stress by reducing the production of the stress hormone cortisol in the body.
- Another study published in Research Gate indicated that unpleasant music disrupted performance on a memory task.
- A small study by PsycNet indicated that subjects with background music achieved greater productivity when background music was in "major mode".
- Another small study published in Wiley showed that participants psychological and physiological relaxation was greater for individuals who rested and listened to music versus those who just rested.
- Another study by the National Center for Biotechnology Information identified that thinking-intensive tasks such as proofreading was impacted by environments, which included speech.
- A study published in The Journal of the Acoustical Society of America indicated that ambient natural sounds could improve employee productivity and moods.
- Another small study published in Sage Journals shows that individuals who did not listen to music (versus listening to music) had greater time-on-task but lower quality-of-work.
- One of the first studies published in Science Direct on this topic showed that music is "effective at increasing efficiency in [repetitive work]."
- A further study published in Research Gate demonstrated that exposure to different types of music can enhance performance on a variety of cognitive tests.
- A study published in Wiley probing the "Mozart Effect" demonstrated that there was no impact on children who listened to a specific Mozart sonata with respect to spatial ability.
- Another study published in Wiley generally confirmed its hypothesis that performance of introverts on complex cognitive tasks was worse in the presence of music.

These studies and reviews demonstrate that there has been little research conducted regarding "bad" or "annoying" music and its impact on logical reasoning performance; rather, most of the literature focuses on "good" music or background music to aid reasoning, cognitive and physical tasks.

Research Design

Environment

As we anticipated including over 130 subjects (see power calculations below), we conducted some brief math to understand how best to move forwards with the research design. We began with estimating that treatment (discussed below), will take a minimum of approximately 10 minutes, plus 3 minutes of instruction and setup and 2 minutes debrief and payment. Thus, 15 minutes times 130 subjects yields over 32 hours, exclusive of the time it takes to acquire the subjects, coordinate them, test them, and analyze their results. As a result of these back-of-the-envelope calculations on time, the team decided to prioritize efficiency and acquire subjects through Amazon's Mechanical Turk service and conduct the experiment online.

The benefits for using Mechanical Turk included speed, consistency, secure payment, and a diversity of subjects. With respect to speed, the primary benefit of leveraging the Mechanical Turk platform was that, similar to a MapReduce process, we could map a job (experiment) across multiple subjects simultaneously, then the results are reduced by technology and verified by the human-in-the-loop.

Instead of over 32 hours of continuous human supervision, we were able to reduce the time to a handful of hours of setup, execution, and verification of the data. Mechanical Turk also provided a more consistent experience as it was agnostic to the experimenter (as we would suspect that Danielle, Osmar, or Dan, even if following a script, would vary their instructions or environment slightly, which would require another blocked variable and analysis). Further, as the data was automatically collected, errors and omissions were reduced yielding more consistent data.

Additionally, remuneration for subjects was handled automatically via the Mechanical Turk system. This obviated the need to hold money in cash, which presents additional risk for the experimenters. Further, there is a more robust paper-trail in case of auditing in relation to the experiment.

Finally, Mechanical Turk presented the opportunity to recruit subjects from a diversity of backgrounds. Whereas we would previously have been limited to engaging with local subjects in the Berkeley, Miami, and New York areas, Mechanical Turk facilitated the reach into new geographies as the platform is location agnostic. If we were all recruiting among our friend or work groups in specific locations, the changes of spillover would be higher. While Mechnical Turk subjects may talk to each other about HITs, we still believed the opportunity to mitigate spillover would be better on Mechnical Turk

The primary drawback for using Mechanical Turk regarded compliance monitoring. As we were unable to visually confirm complete compliance of our experiment protocols, we compromised and instituted several compliance checks, described below. Again, our research team felt that this was an acceptable tradeoff and we would work to ensure compliance with a few procedures also described below.

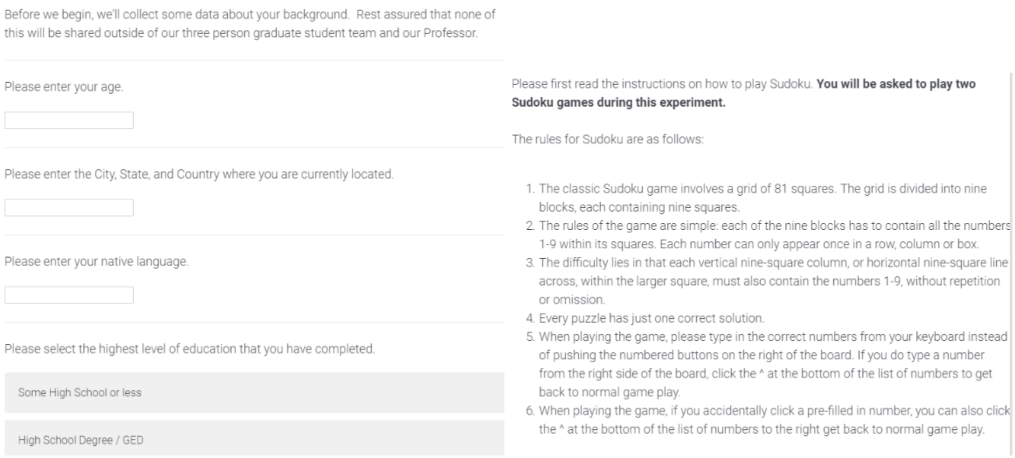
Experimental Design Overview

We developed a three-part experiment leveraging two platforms to acquire, treat, and analyse subjects.

First, we found subjects on Mechanical Turk, recruiting all those that were Mechnical Turk Masters and had clearly done many Human Intelligence Tasks (HITs). After participants opened the Mechnical Turk survey, they were then directed to open a Qualtrics survey link. We set up the formatting of the survey so that the Berkeley Logo displayed as a header on every Qualtrics page to identify the survey as an academic research initiative.

The first component of the Qualtrics survey asked for demographic and behavioral information that we could later control for in our analysis described below. We mandated that all respondents complete all fields. We collected demographic information that included age, current city, state, and country, native language, a 6-option scale of highest education encompassing some high school or less to doctoral degree, and behavioral information that asked about respondents self-identification of being a “morning” or “night” person, expertise of logic games on a 1 to 5 scale, how many minutes does the respondent play logic games in an average week, and knowledge of sudoku on a 1 to 5 scale. All these questions were displayed on one page pictured below.

The next page alerted the subject that there would be two Sudoku games in the experiment and invited the subject to read a description about the rules of the game, along with viewing an example Sudoku game board.



Before we begin, we'll collect some data about your background. Rest assured that none of this will be shared outside of our three person graduate student team and our Professor.

Please enter your age.

Please enter the City, State, and Country where you are currently located.

Please enter your native language.

Please select the highest level of education that you have completed.

Some High School or less

High School Degree / GED

Please first read the instructions on how to play Sudoku. **You will be asked to play two Sudoku games during this experiment.**

The rules for Sudoku are as follows:

1. The classic Sudoku game involves a grid of 81 squares. The grid is divided into nine blocks, each containing nine squares.
2. The rules of the game are simple: each of the nine blocks has to contain all the numbers 1-9 within its squares. Each number can only appear once in a row, column or box.
3. The difficulty lies in that each vertical nine-square column, or horizontal nine-square line across, within the larger square, must also contain the numbers 1-9, without repetition or omission.
4. Every puzzle has just one correct solution.
5. When playing the game, please type in the correct numbers from your keyboard instead of pushing the numbered buttons on the right of the board. If you do type a number from the right side of the board, click the ^ at the bottom of the list of numbers to get back to normal game play.
6. When playing the game, if you accidentally click a pre-filled in number, you can also click the ^ at the bottom of the list of numbers to the right get back to normal game play.

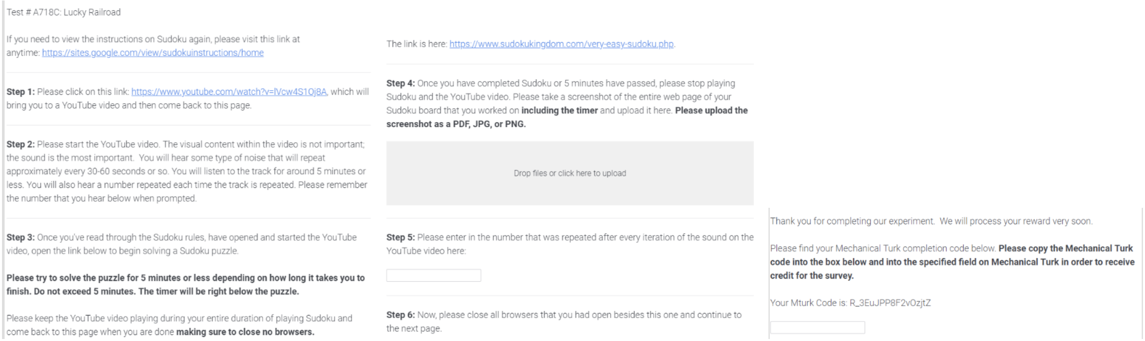
Figure 1: Survey Questions and Sudoku Rules

The subjects were then presented the first test page and then the second subsequent test based upon randomization described below. On each test page, the subjects were presented with a title and serial number of the test, for example, Test # A718C: Lucky Railroad or Test # A122G: Blue Waterfall. These were dummy names and test serial numbers used as a secondary measure for indicating that there were two different tests running sequentially, to reduce the potential frustration of users feeling that they just filled out this test before.

The page included a link to an external webpage we created with the same Sudoku instructions should a participant need to refer back to them while playing the game.

The Qualtrics survey then broke down the current experiment into six steps. The first step directed the subject to open a Youtube video with the control or treatment conditions. The second step instructed the user to play the Youtube video at 50% computer volume and alerted the subject of a confirmation condition where a number will be repeated every 30-60 seconds in the treatment. Step three prompts the subject to solve the randomly generated sudoku puzzle for five minutes, or less if completed sooner, and provides a link to the Sudoku game and instructions to continue to listen to the treatment or control music. Step four instructs the user to stop the Youtube video, take a screenshot of the Sudoku board at the completion of the time or their completion of the game along with the timer that is displayed on the board and upload it as a PDF, JPG, or PNG. Then, the fifth step, as described above, asks the user to input the compliance check number. Finally, the user is prompted to close all browsers except the current survey and proceed to the next page. Again, this protocol repeats twice.

After completing the two instances of the experiment, on the last page in the Qualtrics survey, the platform displays the user's unique Mechanical Turk code, which is the unique identifier that Qualtrics uses as the user ID. The survey prompts the subject to enter the ID into both the box below the question in Qualtrics as well as on Mechanical Turk to facilitate cross-validation of the user. Finally, there is a text box enabling subjects to input their email if they would like to be updated about the results of the survey.



Test # A718C: Lucky Railroad

If you need to view the instructions on Sudoku again, please visit this link at anytime: https://sites.google.com/view/sudokuinstructions/home

Step 1: Please click on this link: https://www.youtube.com/watch?v=lVcw4S1Oj8A, which will bring you to a YouTube video and then come back to this page.

Step 2: Please start the YouTube video. The visual content within the video is not important; the sound is the most important. You will hear some type of noise that will repeat approximately every 30-60 seconds or so. You will listen to the track for around 5 minutes or less. You will also hear a number repeated each time the track is repeated. Please remember the number that you hear below when prompted.

Step 3: Once you've read through the Sudoku rules, have opened and started the YouTube video, open the link below to begin solving a Sudoku puzzle.

Please try to solve the puzzle for 5 minutes or less depending on how long it takes you to finish. Do not exceed 5 minutes. The timer will be right below the puzzle.

Please keep the YouTube video playing during your entire duration of playing Sudoku and come back to this page when you are done **making sure to close no browsers.**

The link is here: https://www.sudokukingdom.com/very-easy-sudoku.php.

Step 4: Once you have completed Sudoku or 5 minutes have passed, please stop playing Sudoku and the YouTube video. Please take a screenshot of the entire web page of your Sudoku board that you worked on **including the timer** and upload it here. **Please upload the screenshot as a PDF, JPG, or PNG.**

Step 5: Please enter in the number that was repeated after every iteration of the sound on the YouTube video here:

Step 6: Now, please close all browsers that you had open besides this one and continue to the next page.

Thank you for completing our experiment. We will process your reward very soon.

Please find your Mechanical Turk completion code below. **Please copy the Mechanical Turk code into the box below and into the specified field on Mechanical Turk in order to receive credit for the survey.**

Your Mturk Code is: R_3EuJPP8F2vOzjtZ

Figure 2: Survey Experiment and Confirmation

Renumeration

All subjects were incentivized to participate in the experiment through the use of a monetary reward. For our pilot test, we set the price for successful completion of the protocol at \$2.50. The price was set higher than other comparable tests because we had to set up a new account with no history and thereby leverage an increased price to mitigate the increased risk of a new requester profile. This pilot HIT attracted many respondents and provided quick results.

As we discovered we would need to increase the sampled subjects, we reduced the price to \$2.15 assuming that the project would be completed perhaps more slowly, but within acceptable time parameters. The value we set was based upon the success we had in our pilot, the constraint of needing more sampled subjects, and mitigation of our new account. We also set the maximum completion of the survey to 35 minutes so that users would respond and complete the experiment in a reasonable time. We observed that we priced appropriately and ethically by comparison to other surveys because we quickly received enough respondents within a handful of hours. All remuneration was issued by Mechanical Turk and deducted from an account set up by the researchers.

Pilot Experiments

The team conducted several pilot experiments that informed our direction of our final experiment.

Our first battery of experiments involved asking users to attempt to complete the Sudoku puzzles without any control or treatment noise. We were testing to see how feasible it was to expect the average user to complete the entire puzzle without any type of time limit. We observed from this pilot that it was unrealistic to expect the majority of individuals to complete the puzzle without giving up. Most individuals who gave up cited frustration or too much time spent.

One interesting observation was that for subjects who were familiar with Sudoku, but did not regularly play the puzzle game, there was a marked learning curve where the first game took a relatively longer time to complete than subsequent games. As a result of this outcome, we restructured our experiment to test two games of Sudoku, where one game would facilitate the learning process and the second Sudoku game would allow for the user to demonstrate their skill.

As a function of our findings of a learning curve, the research team decided to move forward with a test that randomly applied a treatment (of the test music) and a control (of white noise) to measure a difference-in-differences effect of ability in each game.

Our final pilot experiment was executed on Mechanical Turk with 10 subjects. The pilot experiment was generally successful and helped us tweak instructions and protocol for the subjects. Following this pilot, we launched our full experiment.

Compliance Checks

Compliance for the experiment consisted of three checks:

The first check was the capture of the time in the screenshot that the subjects would take and upload to the website. The protocol instructed subjects to take five minutes to solve the puzzle. The research team was able to validate compliance by individually viewing and recording each screenshot of the Sudoku board that included a stopwatch that indicated how long the subject had cumulatively spent on the puzzle. The stopwatch only begins when a user enters in a number, which protects against individuals loading the page and then waiting to begin the puzzle moments later; however, it does not guard against individuals solving the puzzle first and then inputting in the numbers. From our pilot survey we observed that this concern was not material. This compliance check served to ensure that users spent the correct amount of time completing the puzzle and that we would be able to make an apples-to-apples comparison. Beyond editing the photo or solving the puzzle before inputting the numbers, we did not imagine a way of circumventing this concern.

The second compliance check occured after the user finished the puzzle or five minutes had elapsed, and then proceeded to ask the user to input a number that was repeated in the treatment and control audio. For control, the number was 83 and for treatment the number was 67 (although we accepted 57 as well). This compliance check measured whether the user had listened to the audio recordings while playing the game. One means of circumventing this compliance check was to listen to the recording until the subject heard the number and then mute the audio. While we instructed the subjects to set their computer volume at 50%, another means of non-compliance would be to lower the volume level after they hear the two-digit number.

The final compliance check was on the last page of the experiment and prompted the user to enter in a code displayed above the text field box. This was a code that was dynamically displayed based upon the randomly generated user ID associated with their Qualtrics session. This compliance check served as extra assurance that users completed both the treatment and control. We could not imagine a means to circumvent this compliance check.

Randomization

ROXO Grammar:

Experimental Group 1: E O R – X O Experimental Group 2: E O R X – O

- E = randomization allocation to **e**quivalent groups
- O = **o**bservational information (demographics, initial questions, response back into MTurk)
- R = **r**andomized order of treatment and control
- X = treatment
- – = control

One level of randomization was used in this experiment. We leveraged Qualtrics's built in randomization functionality to randomize the order in which respondents saw treatment and control. This procedure allowed us to measure an effect whereby the music could impact both the learning (treatment upon first test) as well as the impact of music on subsequent plays of the game. Part of the rationale for this was due to the fact that we observed through pilot studies that individuals frequently performed significantly better with respect to percentage completion of the puzzle in addition to the time of completion after their first game.

A further level of randomization that was not measured or controlled in this experiment was the fact that respondents were sent to a website that randomly generated Sudoku puzzles. The research team did not investigate the random chance that the subject could receive the same puzzle twice in a row, but made the assumption that this was very improbable. The website that generated the puzzles always made sure that within levels of expertise, the same amount of numbers were filled in at the beginning of the game.

Experimental Materials

Subjects were exposed to a randomized treatment and control protocol, described above. All valid, non-attriting, compliant subjects experienced the treatment and control protocols.

The control protocol involved listening to a Youtube video where the visual content was not important; rather, as described to the subject, the audio was only what the subject needed to focus on while completing the exercise. The control audio consisted of an approximately 30-minute track of one minute of randomly generated white noise and then a repeating two-digit number, 67.

The treatment protocol consisted of a Youtube video similar to the control video where the visual content was not important but consisted of a song associated with the children's television show Barney & Friends. This song was selected due to its millennial pop-culture association with being a hypnotizingly "bad" song. This song along with another two-digit number, 83, composed the approximately 30 minute-long treatment track.

The logical reasoning component of the experiment that measured the degree of progress was the Sudoku website SudokuKingdom. The experiment leveraged SudokuKingdom's randomly generated Sudoku game boards and the researchers selected the "very easy" mode, in which 45 of the 81 squares are filled in. The puzzle displayed that these boards were categorized as "very easy" along with the functionality to undo a move, restart, print, and monitor how much time has elapsed. Each group of nine squares subsequently had five numbers filled in. We selected SudokuKingdom as our platform because of its intuitive means of filling in the number in the Sudoku boards, its ability to automatically load a "very easy" level board, and its clearly visible timer.

The researching team decided to use this platform for its simplicity though realized that there were drawbacks. The website also had a fair amount of superfluous content, such as banner advertisements on both the left and right side of the Sudoku board, a player rankings board below the sudoku board, and a picture that filled in as the user completed the board to the left of the board. The website also allowed subjects to reload the page and get a new puzzle if a subject wanted to switch for some reason.

Further, in hindsight, it would have behooved the experiment if the research team selected two pre-set boards so that all users could complete the same puzzles. While we are confident that the "very easy" selection is fairly consistent across all randomly designed boards, being able to more easily compare apples-to-apples occured to the research team after the fact. This being the case, however; there might not have been a perfect solution because SudokuKingdom does not have a means of serializing the individual games and we would have needed to find an alternate platform to conduct the experiment as well as ensure no spillover among subjects.

Experiment Potential Outcomes

Our experiment had two groups of observed subjects. The structure of our outcomes is as follows for our pair-wise study:

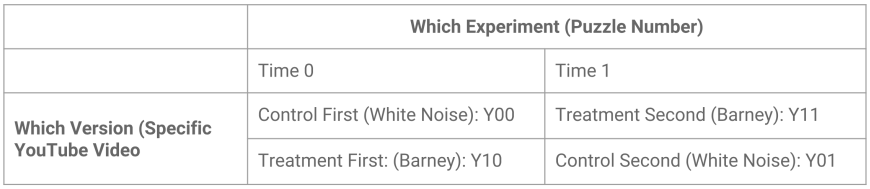
Group A

- y00: y0, Control (white noise) in Time 0
- y11: y1, Treatment (Barney I Love You song) in Time 1

Group B

- y10: y1, Treatment (Barney I Love You song) in Time 0
- y01: y0, Control (white noise) in Time 1

The figure below describes the experimental design further:



	Which Experiment (Puzzle Number)	
	Time 0	Time 1
Which Version (Specific YouTube Video	Control First (White Noise): Y00	Treatment Second (Barney): Y11
	Treatment First: (Barney): Y10	Control Second (White Noise): Y01

Figure 3: Experimental Design

Measurement of Variables

We chose three key ways to measure the degrees of success in this experiment. Two of these outcomes were within subject experiments and one was between subject experiments:

- Within-Subject Outcomes:
  - *Completion* - 1 if the player finished the game within five minutes; 0 otherwise. This outcome applied to all respondents who finished the game in one puzzle and did not finish the game in another
  - *Number of Boxes Left Empty* - The data in this column represents the number of white spaces or boxes that are not filled left on the Sudoku board in an individual game. If a subject completed the entire game, no white spaces would be left and this value would be zero. As described above, we assumed that all cells filled within the Sudoku board were correct. These values were counted manually by the research team and are subject to some counting and recording errors. The possibility does exist for incorrect numbers to be entered, although if the number already existed in the box, row or column, the puzzle would not let it be entered. The only way to ensure true accuracy of the puzzles would be to solve each one, which would be a trenmendous undertaking. The research team decided to forgo this exercide and assume that the squares were correct. Even if the squares were not correct, the incorrectness would be randomly assigned.
- Between Subject Outcomes:
  - *Completion of the First Game* - 1 if the player finished the first game, 0 otherwise. In this situation, we are comparing outcomes between subjects. We will be blocking for the treatment vs. control represented in the first game (i.e. the Barney song vs. white noise). We will also be controlling for various covariates such as level of education, knowledge of Sudoku, knowledge of logic games in general, time spent playing games, morning vs. night person, and all other variables that we asked about in the beginning of the Qualtrics survey.

Pre-Experiment Power Calculation

We estimated the power of our experiment for two out of three outcomes using Cohen's D distance. More specifically, we estimated the power outcomes on both of our within-subject measures, namely number of white spaces left on the board at five minutes of completion, and completion or not of the game at five minutes of doing the puzzle. First we decided that our practical significance would be to detect a difference of at least one white space and assumed a standard deviation twice the detection value giving us a Cohen's distance of 0.5. Similarly, for the Completion outcome, the minimum completion percentage of practical significance that we wanted to detect was 10% or 0.01, and assumed a standard deviation of 0.02, which gave us a Cohen's distance of 0.5.

```
##
##      Paired t test power calculation
##
##               n = 100
##               d = 0.5
##       sig.level = 0.05
##           power = 0.9986097
##     alternative = two.sided
##
## NOTE: n is number of *pairs*
```

Outcome	N	Cohen's D effect size	Test Type	Power
White spaces	100	0.50	Paired	0.99
Completion	100	0.50	Paired	0.99

Result Overview

We exported the data from Qualtrics to Google Sheets to do our manual quality assurance, and then to a CSV to enable analysis and processing in RStudio. A preview of our data is below.

ID	EndDate	ResponseId	Age	Degree	Expertise_logic_games	Time_playing_logic_games	Expertise_logic_sodoku	Duration	White_spaces	Video_number	Treat	Puzzle_Num	Morning	High_School
1	43558.73	R_1gBCzyAiutSEpfw	25	Bachelor	3	30	3	246	0	67	1	0	1	0
2	43558.73	R_2rqmMgzIkH34n3m	32	Bachelor	3	0	2	220	0	67	1	0	0	0
3	43558.73	R_3NOs1WeiyFcCGbu	54	Master	2	35	2	259	0	67	0	0	0	0
4	43558.73	R_1gHfdbCG0HGWyaK	34	Bachelor	2	15	2	304	13	67	0	0	0	0
5	43558.73	R_2WPeoqEnJG8XksN	50	Bachelor	3	0	1	303	3	67	0	0	1	0
6	43558.73	R_2bOfjaoW54ZoydI	30	High School	2	0	2	301	16	67	1	0	0	1

Results Description

Key Outcome-Related Variables:

- *Treat* - This column represents the treatment randomization. A `1` within this column represents that a subject was exposed to treatment (i.e. the Barney song) and a `0` represents that a subject heard the white noise control.
- *Puzzle_num* - This column represents the puzzle number of the respondent, keeping in mind that each respondent completed two puzzles. A `0` means that a respondent is playing their first game and `1` means that a respondent is in their second game. As we are doing a pair-wise experiment, it's very important to ensure that we are controlling for whether someone is playing their first or second game. Due to the learning curve described above, respondents could improve in their Sudoku playing on their second game, which contributes to our decision to measure difference in differences in some of our outcomes.
- *ID* - As described above, we conducted a pair-wise study, so each subject ID is listed in here twice to represent the two games that each survey respondent played.

Pre-test Measurements:

We converted our pre-test demographic and behavioral metrics into ordinal ranks and one-hot-encoded education atainment.

- *Expertise_logic_games* - Self ranking expertise in logic games from 1(low) to 5 (high)
- *Time_playing_logic_games* - How many minutes a respondent plays logic games in an average week
- *Expertise_logic_sodoku* - Self ranking expertise in playing Sodoku from 1(low) to 5 (high)
- *Morning* - If someone is a morning person `1`, otherwise `0`
- *Master* - `1` if education level is "Master Degree"; `0` otherwise
- *Bachelor* - `1` if education level is "Bachelor Degree"; `0` otherwise
- *Associate* - `1` if education level is "Associate"; `0` otherwise
- *High_School* - `1` if education level is "High School"; `0` otherwise

Cleaning and Exploratory Data Analysis

Our exploratory data analysis begins with subsetting the data and evaluating attrition. We will subsequently conduct an exploratory data analysis on our control and treatment variables to see examine the distribution of our data.

Attrition

First, we will begin with looking at our attrition and removing it from our exploratory data analysis as we know that the results are not accurate

1) Subjects with at least one or more missing outcomes:

ID	White_spaces	Completion	ResponseId
51	NA	NA	R_aghI4Cm7ltKAGAx
51	NA	NA	R_aghI4Cm7ltKAGAx
52	NA	NA	R_2YmQeKAjKPPATgl
52	NA	NA	R_2YmQeKAjKPPATgl
61	NA	NA	R_3fZT8FcydOjENba
109	NA	NA	R_3MLlUKECgeTBgh6
109	NA	NA	R_3MLlUKECgeTBgh6
There	are 4 subjects	with at least	one missing outcome.

2) Subjects where the second screenshot uploaded is a copy of the first game:

These IDs have the same video number and exactly the same time to finish the game.

ID	video1	video0	Duration	diff_duration
36	83	83	361	0
57	67	67	455	0
75	3	3	213	0
86	4	4	318	0
98	67	67	300	0
108	57	57	370	0
112	83	83	342	0

ID	video1	video0	Duration	diff_duration
117	67	67	337	0

There are 8 additional subjects with one missing outcome. In total we have an attrition of 12 subjects or 8.6% attrition, 91% of the outcomes are observables or 127 subject outcomes are observables. Please see the figure below to understand the dropoffs and where they occured in the experiment flow.

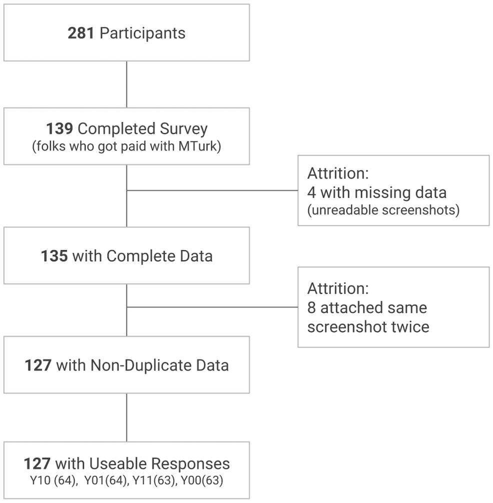

281 Participants
139 Completed Survey
(folks who got paid with MTurk)
Attrition:
4 with missing data
(unreadable screenshots)
135 with Complete Data
Attrition:
8 attached same
screenshot twice
127 with Non-Duplicate Data
127 with Useable Responses
Y10 (64), Y01(64), Y11(63), Y00(63)

Figure 4: Experimental Design Attrition Understanding

From a compliance perspective, we decided to keep all of the observable outcomes within our experiment. This eliminated any element of "picking and choosing" from our side, and prohibited us from succumbing to any fishing expeditions. Even though we told subjects to limit their time to five minutes we decided to still keep all respondents in our analysis including those who took signficantly longer than the deliniated five minutes. Our rationale is that the main focus of our experiment and our outcome measures relates to within-subject outcomes, which should control for timing variability. For our between-subject outcome, we wanted to make sure that we were comparing similar populations, and still chose not to eliminate any subjects. Our goal was to keep as many subjects in our analysis as possible to try our best to make casual and generalizable claims about our findings.

Differential Attrition

Before we dive into the rest of our exploratory analysis and result, we wanted to look to see if we could identify patterns regarding attrition and understand its full impact. We wanted to see if we could understand *why* someone may have attrited from our experiment and if this would significantly impact our results and outcome analysis.

In order to level set first with regards to who actually attrited, we have created a table to put the amount of attrition vs. observed outcomes into context. First, we wanted to generate the necessary data on observed vs. unobserved participants:

```r
# Generating the 'Observed' variables
d[!(ID %in% duplicated_data$ID) & !(ID %in% missing_outcome$ID),
    `:=`(observed, 1)]
d[!(!(ID %in% duplicated_data$ID) & !(ID %in% missing_outcome$ID)),
    `:=`(observed, 0)]
```

```
# Limiting dataframe to unobserved outcomes
d_unobserved <- d[observed == 0]

# Limiting dataframe to observed outcomes for later on
d2 <- d[observed == 1]
```

Next, we have a table showing all that is occuring within out dataset observed and unobserved dataset.

	Treat	Control	Totals
Attrition	4	8	12
Observed	64	63	127
Initial Subjects	68	71	139

Surprisingly, attrition is a bit more commmon within the control group. The awful Barney music in the treatment group did not throw off many of our subjects. This being said, the level of attrition within treatment and control does not appear to be incredibly different. We will conduct a two proportion test below to determine if the level of difference is significant and something that we should correct for throughout our regression analysis.

Two proportion test confidence interval:

```
prop.test(c(d_unobserved[Puzzle_Num == 0 & Treat == 1, .N], d_unobserved[Puzzle_Num ==
    0 & Treat == 0, .N]), c(d[Puzzle_Num == 0 & Treat == 1, .N],
    d[Puzzle_Num == 0 & Treat == 0, .N]))$conf.int
```

```
## [1] -0.16064371  0.05293866
## attr(,"conf.level")
## [1] 0.95
```

Two proportion test p-value:

```
## [1] 0.4076749
```

From the analysis above, we can see that there is a 40.77% chance that the attrition within treatment and control happened by random chance. Therefore, we will not look to correct anything within our subsequent regressions and results analysis.

For another check to see if / how attrition could effect our results, we used a logistic regression to check the impact of differential attrition. More specifically, we used several of our main pre-treatment covariates to predict the probabilities of having an outcome in the first game or not.

```
probobs <- d[Puzzle_Num == 0, glm(observed ~ (Treat * Age) +
    (Treat * Expertise_logic_games) + (Treat * Expertise_logic_sodoku),
    family = binomial(link = "logit"))$fitted]
```

Then, we looked at the summary statistics of the probabilities for treatment and control to see the distribution or predicted probabilities of attrition among these two groups.

Control Summary Statistics:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2514  0.8721  0.9509  0.8873  0.9741  0.9982
```

Treatment Summary Statistics:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.7113  0.9355  0.9545  0.9412  0.9673  0.9768
```

From the summary statistics above, we can see that the rate of attriting is a bit more common in the control group, which goes against our intuition. However, as the average rates of attriting are not that high in either situation, and our two proportion test was insignificant, we will stay on the course of not correcting in our regressions.

Now that we have determined that the level of attrition will not signifcantly affect our analysis, we will conduct further exploratory analysis on our attriters. We will explore what a sample of our pre-treatment variables look like for those that winded up attriting out of the treatment and control groups.

Treat	Treat_Num	Mean_Age	Mean_Morning	Mean_Sudoku_skills	Mean_Logic_Game_Skills
0	8	40	0.375	2.375	3.75
1	4	42	0.500	2.500	3.00

These variables do not point us in any real general direction. The treatment group that stopped has a slight skew towards morning people compared to the control group, but overall this skew is quite small and many of the covariates are fairly equal.

General Exploration

Now, we can evaluate some of our key outcome variables with all data that we will use within our regression analysis.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.000   0.000   5.000   8.728  17.000  34.000
```

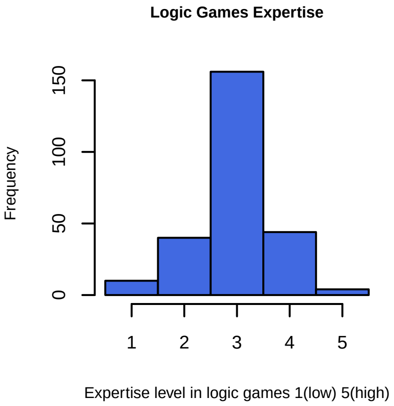
For the white spaces feature, we observe that the median is smaller than the mean, meaning that the data is skewed to the right. The summary statistics from this stage show that across all of our presently cleaned data, there is on average approximately 9 white spaces left on each board when the time is up. We see a maximum of 34 white spaces spaces left (out of a total of 36 white spaces at the beginning of the game). The median number of white spaces when the respondent stops the time around five minutes is 5.

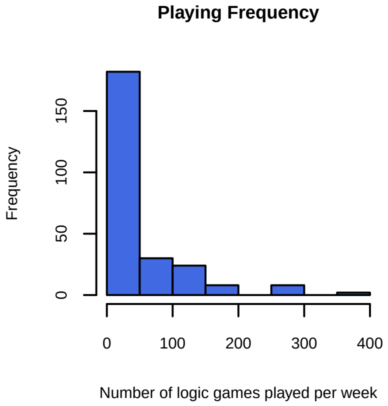
In the first game, we see 9.35 on average, while in the second game, the number of white spaces improves a bit to 8.55 on average. We will do more casual investigation, but overall we do see a slight trend in the direction that we would expect in that participants improve during their second game.

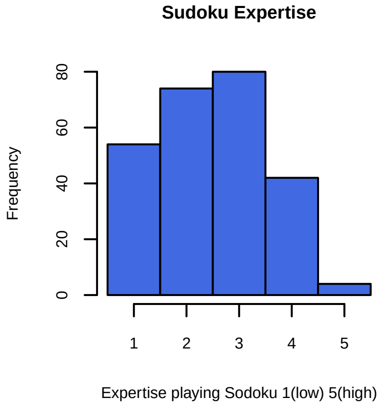
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.000   0.000   0.000   0.437   1.000   1.000
```

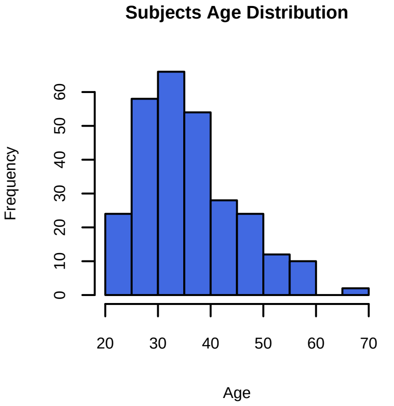
With respect to the completion feature, we observe that in approximately 44% of the games in this cleaned data set, the boards were entirely completed. The completion rate of the first game (that is, no white spaces left) in our cleaned data set is approximately 44.09%. Showing a similar trend that we are seeing with regards to slight improvement in the second game, we see that in our cleaned data set, the completion of the second game is approximately 43.31%.

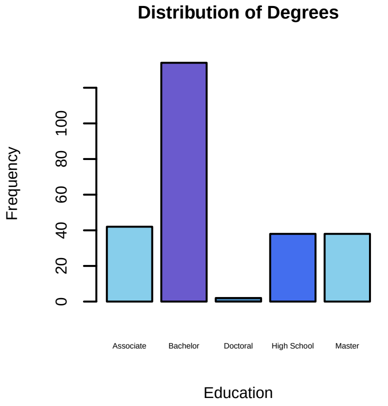
Self-reported Pre-test Behavioral Data











Including all of the subjects, with respect to expertise in logic games, we see a gaussian-like distribution with a signficant majority of subjects reporting average expertise in logic games. However, we observe a heavily right-skewed distribution of individuals who self-reported the time they spend in an average week playing logic games.

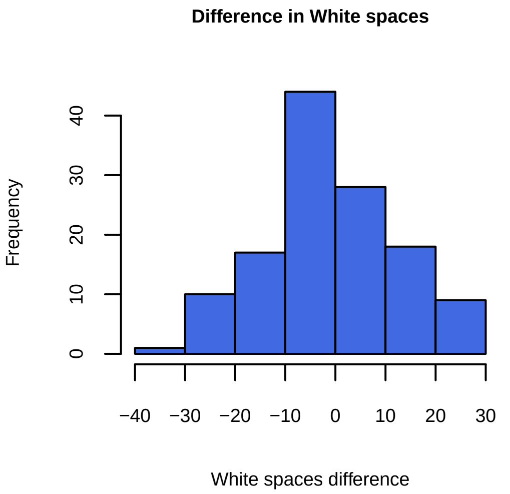
One individual reported spending 360 minutes, or 6 hours per week playing logic games. While this is conceptually reasonable, it did represent the upper-bounds of our respondents. From our distribution, we observe that 75% of our respondents spent an hour or less per week playing logic games.

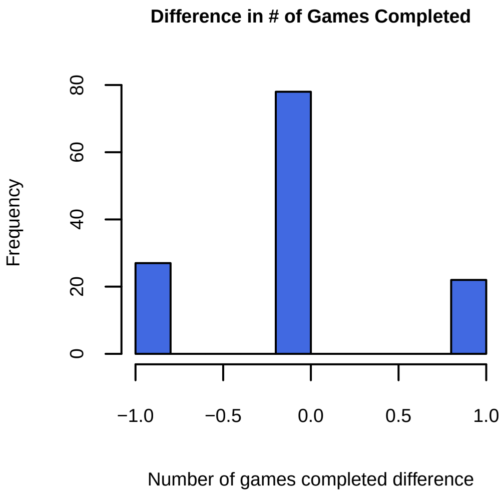
With respect to the self-reported metric on expertise playing Sudoku, the majority of individuals reported a level 1, 2, or 3 representing novice to medium farmiliarity of the game. Only a few individuals self-identified as Sudoku masters.

The self-reported age metric is centered generally around 35 years of age with fewer and fewer respondants citing older ages. This makes intuitive sense as most individuals who are computer savvy and on Mechnical Turk in the first place would likely skew younger.

Within-Subject Outcome Review (Tretment vs. Control)

We see the difference within subjects tend to have a distribution closer to the normal distribution. Negative difference in the graph to the left means that the control game had more white spaces left on the board than the treatment. The graph of difference of white spaces is skewed to the right. Negative difference in the graph to the right means that the control game had a completion and the treatment game did not. The graph of difference of completion is slightly skewed to the left.





Result Review

Covariate Balance Checks

First, we need to change the degree variable from a categorial variable to an interval variable so that we can perform proper covariance balance checks and use this variable within our regressions later on. Having all of the degree variables as separate dummy variables was helpful for us to see the data earlier on.

```
d2$Degree_Interval = ifelse(d2$Degree == "High School", 0.25,
    ifelse(d2$Degree == "Associate", 0.5, ifelse(d2$Degree ==
        "Bachelor", 0.75, 1)))
```

```
##                              Y10        Y00       Y01        Y11
## Treat                   1.000000  0.0000000  0.000000  1.0000000
## Puzzle_Num              0.000000  0.0000000  1.000000  1.0000000
## mean_Age               35.500000 36.5714286 35.500000 36.5714286
## mean_Sudoku_skills      2.562500  2.3968254  2.562500  2.3968254
## mean_time_playing      62.281250 33.0952381 62.281250 33.0952381
## mean_logic_game_skills  3.078125  2.8571429  3.078125  2.8571429
## mean_morning            0.593750  0.5079365  0.593750  0.5079365
## mean_Degree             0.687500  0.6587302  0.687500  0.6587302
## N                      64.000000 63.0000000 64.000000 63.0000000
```

The table above shows very similar means among those who started treatment first and those who started control first. The first two columns are the same as the second two columns because all participants had to complete two games within our experiment. We would expect the covariates to be fairly similar as these results were finalized pre-treatment and should be completely irrespective of the eventual treatment order.

Just to be extra sure that we had covariate balance, we have looked at all of the p-values of t-tests between the covariates in each treatment case below.

```
# Only code used for the initial t-test is shown
paste("Age covariate p-value:", round(t.test(d2[Treat == 1 &
    Puzzle_Num == 0, Age], d2[Treat == 0 & Puzzle_Num == 0, Age])$p.value,
    digits = 2))
```

```
## [1] "Age covariate p-value: 0.49"
```

```
## [1] "Sudoku Skills covariate p-value: 0.38"
## [1] "Time Playing Logic Games covariate p-value: 0.02"
```

```
## [1] "Logic Game Skills covariate p-value: 0.09"
```

```
## [1] "Degree covariate p-value: 0.48"
```

From the p-values above, we do not see any cause for concern that these covariates are correlated with our treatment. The covariate of time playing logic games appears to be significant. However, we discussed above how the maximum value here is a large outlier. Therefore, we will not worry about this one significant covariate. We will make sure not to use this covariate as a control in our regression equations though as it does not pass the covariate balance test.

Within-Subject Outcome Results

When discussing each experiment, we will state our null hypothesis, H_o, and our alternative hypothesis, H_a. The hypotheses and model equations for all outcome measures are below.

For our within-subject outcome results section, we will not be showing our `coeftest` model output summaries due to space concerns. We are clustering by each individual participant so the outputs are incredibly long. However, we have drawn all of our conclusions through in line code and further review can be observed in our Stargazer outputs.

Does treatment affect the amount of white spaces left at five minutes of game play within the same subjects?

H_o: There is no significant difference in outcome between the treatment and control groups. The number of white spaces at the end of game play between treatment and control are very similar.

H_a: The number of white spaces between the treatment and control groups is statistically significantly different, and this difference is due to the treatment.

$Model$ 1: Clustered by the subject ID (not shown in the equation)

$$White\_spaces = \beta_0 + \beta_1 * Treat + \beta_2 * PuzzleNum + \epsilon$$

Coefficient	Interpretation (within subjects)
β_0	Baseline
β_1	Difference in number of white spaces between treatment and control group
β_2	Difference in number of white spaces between first and second puzzles

```r
# Model for within-subject white space and completion
# analysis
mod_white_space <- lm(White_spaces ~ Treat + Puzzle_Num + as.factor(ID),
    data = d3)

# Calculating robust standard errors with clustering
cvcov_white_space <- vcovCL(mod_white_space, cluster = d3[, ID])

# Heteroskedastic errors to account for robustness
se_white_space <- sqrt(diag(cvcov_white_space))

# Estimating the p-values
```

```
p_value_white_space <- signif(coeftest(mod_white_space, vcovCL(mod_white_space,
    cluster = d3[, ID]))[2, 4], 3)
```

The treatment (our Barney music) increases the number of white spaces by 0.868 so just under one additional white space, which is very low and not practically significant. In addition, we observe a 36.3% chance that the difference in white spaces between the subject's two games happened randomly, which is not statistically significant. Therefore, we cannot reject our null hypothesis that the treatment impacts the number of white spaces left on the board after five minutes of game play.

Does treatment affect the likelihood of completing the puzzle in five minutes within the same subjects?

H_o: There is no significant difference in outcome between the treatment and control groups. The completion rate with treatment or control music is the same.

H_a: The completion rate between the treatment and control groups is statistically significantly different, and this difference is due to the treatment.

$Model$ 2: Clustered by the subject ID (not shown in the equation)

$$Completion = \beta_0 + \beta_1 * Treat + \beta_2 * PuzzleNum + \epsilon$$

Coefficient	Interpretation (within subjects)
β_0	Baseline
β_1	Difference in completion between treatment and control group
β_2	Difference in completion between first and second puzzles

```
mod_completion <- lm(Completion ~ Treat + Puzzle_Num + as.factor(ID),
    data = d3)

cvcov_completion <- vcovCL(mod_completion, cluster = d3[, ID])

se_completion <- sqrt(diag(cvcov_completion))

p_value_completion <- signif(coeftest(mod_completion, vcovCL(mod_completion,
    cluster = d3[, ID]))[2, 4], 3)
```

The treatment (our Barney music) decreases the number of completed games by 0.0394 which is a very small coefficient and not practically significant. In addition, we observe a 39.3% chance that the difference in completion between the subject's two games happened randomly, which is not statistically significant either. Therefore, we cannot reject our null hypothesis that treatment had an impact on completion rate.

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Apr 24, 2019 - 02:40:56

Table 10: Within-Subject Outcome Results

	Dependent variable:	
	White_spaces Model of White Spaces	Completion Model of Game Completions
	(1)	(2)
Treat	0.868	−0.039
	(0.951)	(0.046)
Puzzle_Num	−0.710	−0.008
	(0.951)	(0.046)
Constant	−0.079	1.024***
	(0.657)	(0.030)
ID Fixed Effects	Yes	Yes
P-value	0.363	0.393
Observations	254	254
R^2	0.846	0.866
Adjusted R^2	0.688	0.728
Residual Std. Error (df = 125)	5.351	0.259
Note:		*p<0.05; **p<0.01; ***p<0.001

Between Subject First Game Completion Results

Does treatment affect the likelihood of someone to complete their first game in five minutes?

H_o: There is no significant difference in outcome between the treatment and control groups. The initial game completion rates are the same whether someone listened to the control or treatment music first.

H_a: The initial game completion rate between the treatment and control groups is statistically significantly different, and this difference is due to the treatment.

$Model$ 3: Completion of Sodoku game in the first game only

$$Completion = \beta_0 + \beta_1 * Treat + \epsilon$$

Coefficient	Interpretation (between subjects)
β_0	Baseline
β_1	Difference in the number of games completed between treatment and control group in the first game

```r
mod_first_game_completion <- lm(Completion ~ Treat, data = d3[Puzzle_Num ==
    0, .(Completion, Treat)])

se_mod_first_game_completion <- sqrt(diag(vcovHC(mod_first_game_completion)))

p_value_mod_first_game_completion <- signif(coeftest(mod_first_game_completion,
    vcov. = vcovHC(mod_first_game_completion))[2, 4], 3)
```

```
# Showing the table summary
coeftest(mod_first_game_completion, vcov. = vcovHC(mod_first_game_completion))
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  0.460317   0.063808  7.2141 4.598e-11 ***
## Treat       -0.038442   0.089467 -0.4297    0.6682
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From our between subjects outcome in during the first play, we see that the treatment decreases the completion by 0.0384 but this result is not statistically or practically significant, similar to the previous outcomes described above. More specifically, there is a 66.8% chance that this outcome happened due to random chance. The coefficient of the constant or β_0 0.46 represents the mean completion rate when someone begins with the control (white noise) music.

In this model, where we are just looking at the initial game, it may make sense to use all of the possible controls that we screened for in our survey. Therefore, the model below shows a full model, encompassing every control variable besides time spent playing logic games as this variable did not pass the covariate balance check above. We are assuming that our treatment intercept will be very similar in this instance as no covariates should be correlated with treatment. We are assuming that our standard error will be slightly smaller as well because more of the noise in the model may be explained with additional control variables. Our hypothesis is the same as the one stated above.

First we will scale all variables so that all are between 0 and 1 for better model handling and prediction.

```
d3$Age_2 <- d3$Age/100
d3$Expertise_logic_sodoku_2 <- d3$Expertise_logic_sodoku/5
d3$Expertise_logic_games_2 <- d3$Expertise_logic_games/5
```

Now, we will run our model itself below.

```
mod_first_game_completion_2 <- lm(Completion ~ Treat + Age_2 +
    Expertise_logic_sodoku_2 + Expertise_logic_games_2 + Morning +
    Degree_Interval, data = subset(d3, Puzzle_Num == 0))

se_mod_first_game_completion_2 <- sqrt(diag(vcovHC(mod_first_game_completion_2)))

p_value_mod_first_game_completion_2 <- signif(coeftest(mod_first_game_completion_2,
    vcov. = vcovHC(mod_first_game_completion_2))[2, 4], 3)

coeftest(mod_first_game_completion_2, vcov. = vcovHC(mod_first_game_completion_2))
```

```
##
## t test of coefficients:
##
##                           Estimate Std. Error t value  Pr(>|t|)
## (Intercept)               0.434921   0.286046  1.5205    0.1310
## Treat                    -0.037933   0.089647 -0.4231    0.6730
## Age_2                     0.026162   0.519296  0.0504    0.9599
## Expertise_logic_sodoku_2  1.028843   0.218436  4.7100 6.715e-06 ***
## Expertise_logic_games_2  -0.627916   0.407575 -1.5406    0.1260
## Morning                  -0.021845   0.089498 -0.2441    0.8076
## Degree_Interval          -0.163128   0.194796 -0.8374    0.4040
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The model above with all control covariates tells a very similar story to the basic model within this section. We see that the coefficient for treatment is only -1.3020833% less than in the prior regression. We should have very similar coefficients given that none of the covariates were conditioned on treatment, so our confidence in the model increases because this is the case. Our treatment standard error and p-values remained incredibly consistent from this model to the last one as well, showing that these pre-treatment coefficients helped to explain very little of the model.

While the `Expertise_logic_sodoku_2` variable is significant, this does not tell us very much. This variable was not conditioned on or randomized due to treatment, so it is not casual in any way. We know that even by chance we may observe significance, and this is what is happened with regards to this variable. We tried taking this coefficient out of the model (specifics not shown) and observed a very similar story, so we decided to leave it in here for the sake of completeness.

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Apr 24, 2019 - 02:41:05

Table 12: Between Subject Outcome Result

	Dependent variable:	
	Completion	
	First Game Completion: Basic Model	First Game Completion: Full Model
	(1)	(2)
Treat	−0.038	−0.038
	(0.089)	(0.090)
Age_2		0.026
		(0.519)
Expertise_logic_sodoku_2		1.029***
		(0.218)
Expertise_logic_games_2		−0.628
		(0.408)
Morning		−0.022
		(0.089)
Degree_Interval		−0.163
		(0.195)
Constant	0.460***	0.435
	(0.064)	(0.286)
ID Fixed Effects	No	No
P-value	0.668	0.673
Observations	127	127
R^2	0.001	0.154
Adjusted R^2	−0.006	0.111
Residual Std. Error	0.500 (df = 125)	0.470 (df = 120)
Note:		*p<0.05; **p<0.01; ***p<0.001

Conclusion

With the data we obtained from our experiment and subsequent evaluation running longitudinal and trasversal analysis we found no evidence of the impact of music on the performace of Sodoku games. We found no evidence (in statistical or practical significance) that the I Love You song by Barney and Friends had any effect on the percentage of completion rate of the game or partial completion of the game counting the number white spaces left on the game board, when compared to white noise.

One potential reason for our lack of effect is that Barney really may not be annoying in short doses. Participants were only listening to the song for five minutes at most, which may not have been long enough to observe an effect. Conversely, white noise may also have been irritating to many of our participants. While articles have shown that white noise is quite soothing and can help cognitive function, our sample of participants may not have felt that way.

Another reason for our lack of effect could simply be the compliance. Participants only had to listen for 30 - 60 seconds of the audio track to get the compliance-check number. Participants may have turned off the music at that point and simply focused on the Sudoku game only for the remainder of the time.

If given more time, our team would build out a custom Sudoku solution of sorts to automatically check for correctness. The Sudoku game itelf would also play the different types of music, forcing participants to listen to it while they played. We would also conduct the experiment with in-person monitoring to make sure that participants did not turn off their volume while playing, or give up and just let the clock run out during the five minutes of game play.