Don't Mind the Music... Or Should We?

I. Introduction

In a world full of distractions, many seek an elusive mental edge—a way to enhance performance in studying, working, or simply remembering where they left their keys. Working memory plays a crucial role in reasoning, retainment, and learning. While it allows our brains to hold and manipulate information, it is particularly sensitive to attentional load and environmental interference. Studies have postulated that different types of sound have different influences on cognitive performance through mechanisms such as cortical arousal regulation, cognitive load theory, and attentional resource allocation. For instance, white noise may improve focus and reduce distractions through stochastic resonance—enhanced signal detection in the brain. Additionally, calming instrumental music could optimize mental performance by lowering stress without acting as a distraction. High-intensity lyrical music, however, might provide an increased level of cognitive interference, demanding more mental resources and potentially impairing working memory performance.

This study seeks to explore the question: *Does the type of background sound—white noise, calming piano music, or high-intensity, lyrical music—significantly affect performance on a working memory task as measured by a Dual N-Back memory test?* Our parameter of interest is the mean score on the Google Dual N-Back memory benchmark for each of the three randomly induced auditory conditions. With approximately 50-60% of students in the U.S. reporting that they listen to sound while studying, understanding how different auditory inputs interact with cognitive functioning could have significant implications for learning strategies and productivity.

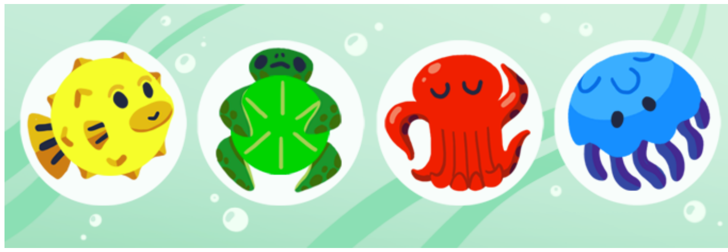

Figure 1: The Google Dual N-Back Memory Benchmark

II. Statistical Question

Is there a difference between the means of scores on the Google Dual N-Back memory benchmark game across groups of students at our high school like the ones in the study who would listen to *Power (Clean)* by Kanye West, *River Flows in You* by Yiruma, or white noise? Although we are aware that one song cannot be truly representative of the population of a genre of music, we chose these audios (*Power* by Kanye West, *River Flows in You* by Yiruma, and white noise) because we think they are fairly representative of their respective genres: hip-hop/pop songs, classical background music, and neutral sounds.

<u>Null Hypothesis:</u> There is no difference between the means of scores on the Google Dual N-Back Memory Benchmark game across groups of students at our high school like the ones in the study who would listen to *Power*, *River Flows In You*, or white noise.

<u>Alternative Hypothesis:</u> At least one pair of means of scores on the Google Dual N-Back Memory Benchmark game is different across groups of students at our high school like the ones in the study who would listen to *Power*, *River Flows In You*, or white noise.

III. Data Collection

We used an experiment for our data collection to see if we could conclude a cause and effect relationship, not aiming to generalize to a population. Therefore, we took a convenience sample of 90 students at our high school. Our three treatment groups were listening to *Power* by Kanye West, *River Flows in You* by Yiruma, and white noise. We chose these audios because

Power is a hip-hop song with lyrics, rhythm, and high emotions which could be more distracting; *River Flows in You* provides calming piano without lyrics, which is engaging but has low cognitive interference; and white noise which has a consistent background sound with little to no interference, serving as a control. These sounds represented typical audios that students listen to while studying. Each subject was approached using a standardized script that explained the experiment's purpose, provided instructions, and outlined the conditions of participation. The script emphasized voluntary consent, anonymity, and consistency.

To randomly assign treatments to subjects, we wrote down a column of 30 cells for each of the three treatment groups (90 cells total). We then randomized the column. For each trial, we did the following:

1. Approach a student at our high school and ask if they would be willing to participate in an experiment (outlined in the script).
2. Follow the script in explaining how to play the Google Dual N-Back Memory Benchmark and participate in the trial.
3. Identify the treatment by corresponding the row number to the subject number (determined chronologically) and finding the appropriate cell.
4. On our personal computer, open the Google Dual N-Back Memory Benchmark (muted), and the appropriate audio.
5. Give noise-cancelling headphones for the subject to use, control for potential confounding variables (listed below), and start the audio.
6. Have the subject play until they lose, at which point they will indicate to the researcher they have completed the trial.
7. Record their score at the appropriate trial number row (1-90).

To account for confounding factors, we did the following:

1. We followed the exact same script between treatment groups, ensuring there was no confounding in the way we approached subjects (differing approaches may influence the way subjects responded or performed—creating bias).
2. Immediately after explaining the procedure according to the script, we stepped away, leaving the subject alone and ensuring no confounding from varying degrees of pressure to perform that the subjects may feel (social inhibition caused by researcher surveillance may result in lowered scores).
3. We further assured that the subjects were in a setting they were comfortable with, asking if they would like to move to a location where nobody could see their screen or judge their performance (to ensure consistency across treatments we aimed to place subjects in a comfortable location as they are performing a complex mental task).
4. We normalized the audios in iMovie, ensuring there was no confounding by audio level (louder audio may be more distracting).
5. We looped the audios, ensuring there was no confounding by audio length as some subjects' trials may take longer depending on performance (a change or abrupt stop in the audio may be more distracting).
6. We ensured that subjects were placed in a quiet area (such as a hallway or library, measured at 50 decibels or less using An Online Decibel (dB) Meter) to control for external noise.
7. We used the same personal computer and noise-cancelling headphones for each trial, controlling for differences in experience.

Overall, our methodology ensured minimization of potential confounding variables and variability in our dataset, and differences between subjects would be accounted for by random assignment. This ensured that we could conclude cause and effect properly based solely on the treatment, which audio the subjects would listen to.

IV. Data

Score on benchmark for each group of background audios.

Power			*River Flows in You*			White Noise		
4	6	8	9	10	14	12	9	7
7	10	10	23	11	14	5	9	9
5	6	8	6	14	16	15	15	8
6	11	10	11	13	18	16	12	19
12	7	9	22	14	15	15	15	13
8	9	7	18	13	15	14	13	18
11	10	6	16	14	9	15	19	15
13	8	14	14	15	10	23	13	17
8	7	9	18	13	15	13	12	15
4	8	5	13	17	12	16	12	11

Figure 2: Table of collected data

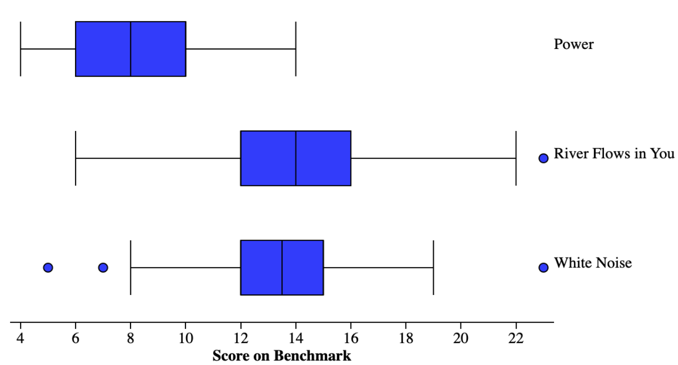

Figure 3: Boxplot of collected data

Group Name	n	mean	SD	min	Q_1	med	Q_3	max
1: Power	30	8.2	2.511	4	6	8	10	14
2: River Flows in You	30	14.067	3.648	6	12	14	16	23
3: White Noise	30	13.5	3.866	5	12	13.5	15	23

Figure 4: Summary statistics of collected data

V. Statistical Test

One-Way Omnibus ANOVA: In the summary statistics above, we can see the means for the sample distribution that used *Power* was significantly less than the means of the other two sample distributions. Thus, we have some evidence of a difference in means between treatment groups, so statistical inference is applicable. Since we are testing for means across more than two groups, we could use a one-way omnibus ANOVA test, a Welch's ANOVA test, a Kruskal-Wallis ANOVA test, or another type of ANOVA test. In this case, the one-way omnibus ANOVA test is the most suitable for this scenario due to the homogeneity of variances, approximately normal distributions, and equal sample sizes as seen below.

Therefore, we will be conducting a one-way omnibus ANOVA test with the aforementioned hypotheses (Section II) at the 0.05 alpha significance level, if the following conditions are met:

1. Random: Each subject was randomly assigned to a treatment by the "randomize range" Google Sheet method. The random condition is satisfied.
2. Independent: Since we are not randomly sampling, the 10% condition does not apply in this case. It is safe to assume that the result of one individual's performance on the google memory benchmark does not affect other individuals' performance on the google memory benchmark within each treatment group. It is also safe to assume that the treatment groups (*Power*, *River Flows in You*, and White Noise) are all

independent from each other since individuals were randomly assigned to each group, and each individual was only present in one of the treatment groups. The independence condition is satisfied.

3. Normal/Large Sample Size: Although there are outliers in the sample data, since the sample sizes of each treatment group is greater than or equal to 30, the Central Limit Theorem applies. Therefore, the sampling distribution of sample mean is approximately normal for each treatment group. The normal condition is satisfied.
4. Equal Variance: As seen in *Figure 3*, the standard deviations are roughly equal across each group. In addition, the ratio of the largest standard deviation of a treatment group to the smallest standard deviation of a treatment group is less than 2: $\hat{\sigma}_{max} / \hat{\sigma}_{min} = 3.866/2.511 = 1.540 < 2$ (as seen in *Figure 4*). Thus, the homoskedastic condition is satisfied. Moreover, since the sample sizes for each of our three samples is the same, ANOVA is robust enough to ignore violations of homogeneity of variances. Thus, even if we weren't able to conclude equal variances between treatment groups, we would still be able to conduct the one-way omnibus ANOVA test allowing us to disregard the potential usage of Levene's test for equal variances.

The Welch's ANOVA test is not appropriate in this case since we are able to conclude homogeneity of variances. Additionally, the non-parametric Kruskal-Wallis test is not appropriate either since we are able to conclude that each of our treatment groups followed an approximately normal distribution and that each of our three sample sizes are the same.

Since all conditions were met, we conducted a one-way omnibus ANOVA test:

Source of Variation	Degrees of Freedom (df)	Sum of Squares (SS)	Mean Squares (MS)	F	P-value
Between	2	628.289	314.144	27.272	< 0.00001
Within	87	1002.167	11.519		
Total	89	1630.456			

Figure 5: Values calculated for conducting one-way omnibus ANOVA test

Since the p-value of < 0.00001 is less than the alpha significance level of 0.05, we reject the null hypothesis. We have significant statistical evidence that at least one pair of means of scores on the Google Dual N-Back Memory Benchmark game is different across groups of students at our high school like the ones in the study who would listen to *Power*, *River Flows In You*, or white noise. Additionally, the F-test-statistic of 27.272 means that there is 27.272 more variability between the treatment groups than within each treatment group which is significantly greater than the critical value of 3.111, thus further proving the conclusion stated above.

Tukey's HSD: In order to identify which pairs of groups are statistically different, we must conduct post-hoc analysis, since the one-way omnibus ANOVA test does not identify which pair-wise group means are different. We could have used several post-hoc tests, including Tukey's HSD (Honestly Significant Difference), Bonferroni test, or Fisher's LSD (Least Significant Difference). Tukey's HSD was most appropriate since the sample sizes and variances were approximately equal across the treatment groups.

Therefore, we will be conducting a Tukey's HSD test at the 0.05 alpha significance level, with the following hypotheses:

Null Hypothesis: For each pair of groups of students at our high school like the ones in the study who would listen to *Power*, *River Flows in You*, or white noise, there is no difference in the means of scores on the Google Dual N-Back memory benchmark.

Alternative Hypothesis: For each pair of groups of students at our high school like the ones in the study who would listen to *Power*, *River Flows in You*, or white noise, there is a difference in the means of scores on the Google Dual N-Back memory benchmark.

Since the conditions for a Tukey's HSD test are the same as for a one-way omnibus ANOVA test (in which the conditions were met), and the ANOVA test concluded significant results, we can proceed with the test: $k = 3$, $df = 87$, $Q_{critical} = 3.3722$

Pairwise Comparisons	Tukey's HSD Test Statistic	Q-statistic	P-value
M_P-M_{RFiY}	5.87	9.47	< 0.00001
M_P-M_{WN}	5.30	8.55	< 0.00001
M_{RFiY}-M_{WN}	0.57	0.91	0.79

Figure 6: Values calculated for conducting Tukey's HSD test

Since the p-values for the pairs M_P-M_{RFiY} and M_P-M_{WN} are < 0.00001 which is less than the alpha significance level of 0.05, we can reject those null hypotheses. We have significant statistical evidence that there is a difference in means of scores on the Google Dual N-Back memory benchmark for the groups of students at our high school like the ones in the study who would listen to *Power* vs. *River Flows in You*, and for the groups of students at our high school like the ones in the study who would listen to *Power* vs. white noise.

VI. Final Conclusion

Through our experiment, we found significant statistical evidence that the type of background audio (*Power*, *River Flows in You*, or white noise) causes students at our high school like the ones in the study to have different average scores on the Google Dual N-Back memory benchmark, which evaluates working memory retention. In addition, through our post-hoc analysis using Tukey's HSD test, we can conclude that the mean score of students at our high school like the ones in the study who would listen to *Power* is lower than the mean score of students at our high school like ones in the study who would listen to *River Flows in You* or white noise. Our findings support the idea that listening to *Power* (which is intense, lyrical, and emotional) disrupts working memory retention, indicative of cognitive interference, when compared to *River Flows in You* (which is calming) and white noise (which is neutral) for students at our high school like the ones in the study.

Our study was designed carefully to minimize confounding variables, by following a specific script, using the same computer and headphones, normalizing audio levels, and limiting external noise. Random assignment allowed us to draw cause-and-effect conclusions about the effect of sound type on working memory, but we can only generalize to students at our high school like the ones in the study since we used a convenience sample.

If we could repeat this experiment, we would consider implementing a stratified random sample across grade levels (9-12) within our high school student population, so that we could limit variability and ensure a sample with more representative age-group proportions relative to the population. We could have also done a randomized block design, blocking our sample into groups based on potential confounding variables—such as performance on standardized tests (a measure of intelligence)—would allow us to have a more precise causation relationship between the treatments and memory game scores, leading to a more statistically significant conclusion. It would also be interesting to survey certain people, like people who play an instrument, or people who have attention disorders, to see how the cause-and-effect relationship potentially changes.

Overall, our study has very practical implications, suggesting that students should listen to calming or neutral audio rather than typical hip-hop music when working on homework or studying, as working memory is essential in academics.