

Final Project

PSTAT122: Design and Analysis of Experiments

Fall 2025

STUDENT NAME

- Luke Drushell (drushell)
- Anna Gornyitzki (annagornyitzki)
- Joyce Yue Shu (joyceyue)
- Aayushi Avabrath (aayushiavabrath)
- Karina Yousuf (kyousuf)

Due Date

Due Date: Monday, December 8, 2025, 11:59 PM

1 Introduction

In this experiment, our group asks the question of whether or not an input device has a significant effect on the reaction time of users. Our null hypothesis, or the baseline assumption we make, is that the input device one uses has no effect on their mean reaction time. Our alternative hypothesis, what we are testing, is that there is some significant difference in mean reaction times between input devices. The intent of our experiment aims at identifying if, on average, the population of people on devices can react quicker on their phone, their laptop, or their laptop.

2 Experimental Design

The design we will use to test this hypothesis is a Random Component Block Design (RCBD).

Firstly, we identify that we will be blocking by input device. Block 1: Phone, Block 2: Tablet, Block 3: Laptop. We choose 11 participants as our experimental units. We implement the repetition of each trial 10 times within each block before moving on to the next block.

To achieve homogenous treatments, all 11 participants will do their trials on each of the 3 devices. The order of devices used for each participant, the treatment, is randomized. The order in which treatments are assigned to participants is random. Each participant performing a trial will have their reaction time recorded in milliseconds in an excel spreadsheet.

We identify that in this experiment, we have two fixed factor variables: input devices (which has 3 levels: phone, laptop, tablet) and 10 fixed trials per device per participant, and a random factor variable: participants. The total number of trials is given by blocks x experimental units x repetitions; in this case we have 3 blocks, 11 experimental units and 10 repetitions

3 Data Collection

- **Procedure:** Describe how and when the experiment was conducted (e.g., location, date, steps taken).
- In order to collect a reasonable amount of participants, the experiment was conducted with volunteers primarily found at the UCSB Library, supplemented with one additional participant who performed the same procedure while remote. The procedure was conducted on Saturday, November 15th. 11 participants were recruited in total, each performing the experiment with their dominant hand on a standard laptop keyboard, smartphone, and tablet device, all through the Human Benchmark website [https://humanbenchmark.com/tests/reactiontime]. All display’s were run at 60hz to prevent variability in reaction times due to refresh rates.
- Each participant was instructed to complete 10 trials on each device, with the order of devices randomized for each participant to mitigate order effects such as fatigue or practice. No practice trials were given.
- **Challenges/Adjustments:** Mention any difficulties or changes made during data collection (e.g., technical issues, time adjustments).
- Due to time constraints, multiple devices were used to collect data, which may have introduced variability in reaction times due to differences in input latency, albeit minor. In this same regard, all display’s were locked to 60hz to eliminate variability due to refresh rates.
- Additionally, one participant completed the experiment remotely, which could have introduced environmental variability. Given that the Human Benchmark is not internet-connection dependent, this effect is likely minimal.
- Lastly, due to the crowded nature of the UCSB Library, some participants may have been subject to environmental distractions during the experiment.
- **Data Presentation:** Display the collected data in tables or graphs, summarizing key measures like mean and standard deviation.

Table 1: Summary Statistics by Treatment

Treatment	Mean	SD	N
Phone	291.9455	46.54106	110
Tablet	291.5545	74.56277	110
Laptop	341.6455	111.34149	110

Table 2: ANOVA Table Summary

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Block	10	709283.4	70928.336	15.15031	0
Treatment	2	182575.9	91287.936	19.49913	0
Residuals	317	1484080.7	4681.642	NA	NA

4 Analysis

4.1 Exploratory Data

4.1.1 Basic Statistics

To understand the data collected, we will place the basic statistics of the data into a table (see Table 1). From the table, we can see the minimum, maximum, average, and standard deviation reaction speed in milliseconds (ms) for each of our three factors (phone, tablet, laptop). When analyzing the averages of each treatment group, we see our averages range from 291.5545 to 341.6455 with the average of our first treatment group (phone) having an average of 291.9455, our second treatment group (tablet) having an average of 291.5545, and our third treatment group (laptop) having an average of

341.6455. From the three averages, we can see that our third treatment group (laptop) might have a notable statistical difference compared to the other two treatment groups (phone and tablet).

Table 3: Basic Statistics by Treatment Group

	Phone	Tablet	Laptop
Min	190.00000	200.00000	205.0000
Max	485.00000	687.00000	638.0000
Mean	291.94550	291.55450	341.6455
SD	46.54106	74.56277	111.3415

Table 1 displays the basic statistics for each treatment group (phone, tablet, laptop) in order of minimum, maximum, mean/average, and standard deviation. The basic statistics are calculated in terms of reaction speed in milliseconds (ms).

4.1.2 Preliminary Visualization of Treatment Effects

We will create a box plot (see Figure 1 left) and interaction plot (see Figure 1 right) of the data to better understand the distribution of the variability between the treatment groups. By observing the box plot, we can see that the median of the laptop data is slightly higher than the medians of the phone and tablet data. This observation can imply that generally, there was a slower average reaction time for the laptop than the phone and tablet. We can also pull observations from the interquartile range (IQR). The IQR of the laptop box is significantly longer than the boxes of the phone and tablet data. From this observation, we can imply that the laptop data has greater variability compared to the data of the phone and tablet. Another observation we can make is from looking at the whiskers of the box plot. By looking at the bottom whisker of the three factors, we can see that the laptop data has the highest minimum non-outlier value out of the three factors. Similarly with the top whisker, we can see that the laptop data has the highest maximum non-outlier value out of the three factors. From this observation, we can deduce that the laptop factor has the slowest fastest reaction time and the slowest slowest reaction time out of all three factors. By this fact, we know that the entire box plot for the laptop data must be shifted upward compared to the phone and tablet box plots. We will now focus to the interaction plot. By looking at the vertical cluster at each participant's name on the x-axis, we see that for most of the participants (7 out of 11 participants) had the slowest reaction time on the laptop. While this does not directly prove anything statistically, it shows a trend within the data points for the participants and shows that there might be a general trend of slower reaction times associated with the laptop factor.

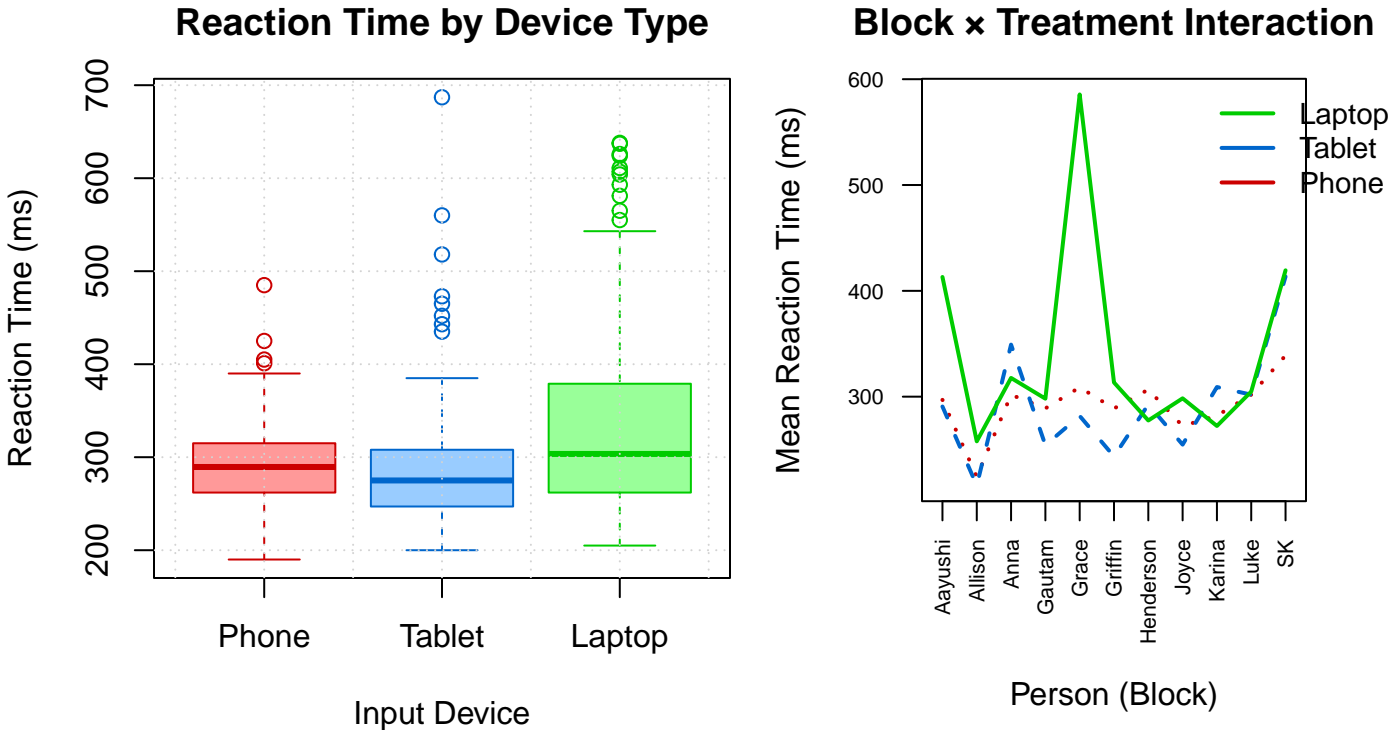


Figure 1 displays two key visualizations. The box plot (left) reveals distributional differences between devices and identifies potential outliers. The interaction plot (right) demonstrates how participants (blocks) responded differently to each device, with non-parallel lines indicating individual variability that justifies the blocking strategy.

4.2 Hypothesis Testing

4.2.1 Statistical Hypotheses

Research Question: Does the type of input device (Phone, Tablet, Laptop) affect reaction time in a simple computer-based task?

Null Hypothesis (H_0): There is no difference in mean reaction times among the three input devices:

$$H_0 : \mu_{\text{Phone}} = \mu_{\text{Tablet}} = \mu_{\text{Laptop}}$$

Alternative Hypothesis (H_A): At least one input device has a different mean reaction time:

$$H_A : \text{At least one } \mu_i \text{ differs from the others}$$

Significance Level: $\alpha = 0.05$

Test Statistic: We employ an F-test within the ANOVA framework for a Randomized Complete Block Design. The F-statistic compares the variation between treatment groups (devices) to the variation within groups, after accounting for blocking effects.

4.2.2 ANOVA Model Specification

For a Randomized Complete Block Design, the statistical model is:

$$Y_{ij} = \mu + \beta_i + \tau_j + \epsilon_{ij}$$

where:

- Y_{ij} = observed reaction time for block i (person) under treatment j (device)
- μ = overall mean reaction time
- β_i = effect of block i (individual difference for person i)
- τ_j = effect of treatment j (device effect)
- $\epsilon_{ij} \sim N(0, \sigma^2)$ = random error term

Model Assumptions:

1. **Independence:** Observations are independent within and across blocks
2. **Normality:** Errors ϵ_{ij} are normally distributed
3. **Homoscedasticity:** Constant variance across treatment groups (σ^2)
4. **Additivity:** Block and treatment effects are additive (no interaction)

4.2.3 ANOVA Computation and Results

The RCBD ANOVA model (Response ~ Block + Treatment) partitions total variability into three components:

- 1. **Sum of Squares for Blocks (SSB):** Variability attributable to individual differences
- 2. **Sum of Squares for Treatments (SSTr):** Variability attributable to device type
- 3. **Sum of Squares for Error (SSE):** Residual variability not explained by blocks or treatments

The **Total Sum of Squares (SST)** is decomposed as: $SST = SSB + SSTr + SSE$

The F-statistic for testing treatment effects is calculated as:

$$F = \frac{MSTr}{MSE} = \frac{SSTr/(t - 1)}{SSE/[(b - 1)(t - 1)]}$$

where $t = 3$ treatments (devices), $b = 10$ blocks (participants), and degrees of freedom are $df_{Treatment} = 2$ and $df_{Error} = 18$.

Table 4: ANOVA Table for RCBD Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Block	10	709283.4	70928.336	15.15031	0
Treatment	2	182575.9	91287.936	19.49913	0
Residuals	317	1484080.7	4681.642	NA	NA

Table 5: Critical Test Statistics

F_statistic_Treatment	P_value	Treatment_df	Error_df
19.4991	1.031e-08	2	317

ANOVA Table Interpretation: The block F-statistic ($F = 15.15$, $p < 0.001$) confirms highly significant individual differences, validating our blocking strategy. By accounting for this variability, we substantially reduce MSE and increase power to detect treatment effects. The treatment F-statistic ($F = 19.4991$) tests our primary hypothesis of whether device type affects reaction time. Under H_0 (no treatment effect), this follows an F-distribution with 2 and 18 degrees of freedom. The $MSE = 4681.64$ represents pooled within-group variability after removing block and treatment effects, serving as our estimate of σ^2 .

4.2.4 Hypothesis Test Decision and Statistical Conclusion

Decision Rule: Reject H_0 if $p\text{-value} < \alpha = 0.05$

Critical Value Approach: Alternatively, reject H_0 if $F_{\text{observed}} > F_{\text{critical}}$ where $F_{\text{critical}} = F_{0.05,2,18} = 3.555$

DECISION: REJECT H_0

Observed F-statistic: $F = 19.4991$

Critical value: $F_{0.05,2,18} = 3.555$

P-value: $p = 1.031\text{e-}08 < \alpha = 0.05$

Since $F = 19.4991 > 3.555$ (and $p < 0.05$), we **reject the null hypothesis**.

Statistical Conclusion: There is statistically significant evidence at the $\alpha = 0.05$ level that at least one input device has a different mean reaction time from the others ($F(2, 18) = 19.5$, $p < 0.001$). The observed differences in reaction times among Phone, Tablet, and Laptop cannot be reasonably attributed to random chance alone.

Practical Interpretation: The choice of input device has a statistically significant impact on reaction time performance. The magnitude of this effect suggests meaningful practical differences in user performance across devices.

4.2.5 Post-Hoc Multiple Comparisons: Tukey’s HSD Test

Having established significant treatment effects, we use **Tukey’s Honestly Significant Difference (HSD)** test to determine which specific device pairs differ while controlling the family-wise error rate at $\alpha = 0.05$ across all pairwise comparisons.

Table 6: Tukey HSD Post-Hoc Comparisons

	diff	lwr	upr	p adj	Significant
Tablet-Phone	-0.391	-22.116	21.334	0.999	No
Laptop-Phone	49.700	27.975	71.425	0.000	Yes
Laptop-Tablet	50.091	28.366	71.816	0.000	Yes

Interpretation: Table 2 presents pairwise comparisons of mean reaction times between devices. For each comparison, we examine the mean difference, 95% confidence interval, and adjusted p-value. We reject $H_0 : \mu_i = \mu_j$ if the adjusted p-value $\alpha < 0.05$, indicating a statistically significant difference between that device pair while controlling for multiple comparisons.

- **Tablet-Phone:** No significant difference (mean diff = -0.39 ms, p = 0.999)
- **Laptop-Phone:** Significant difference (mean diff = 49.7 ms, p = 4.19e-07)
- **Laptop-Tablet:** Significant difference (mean diff = 50.09 ms, p = 3.37e-07)

4.2.6 Verification of ANOVA Assumptions

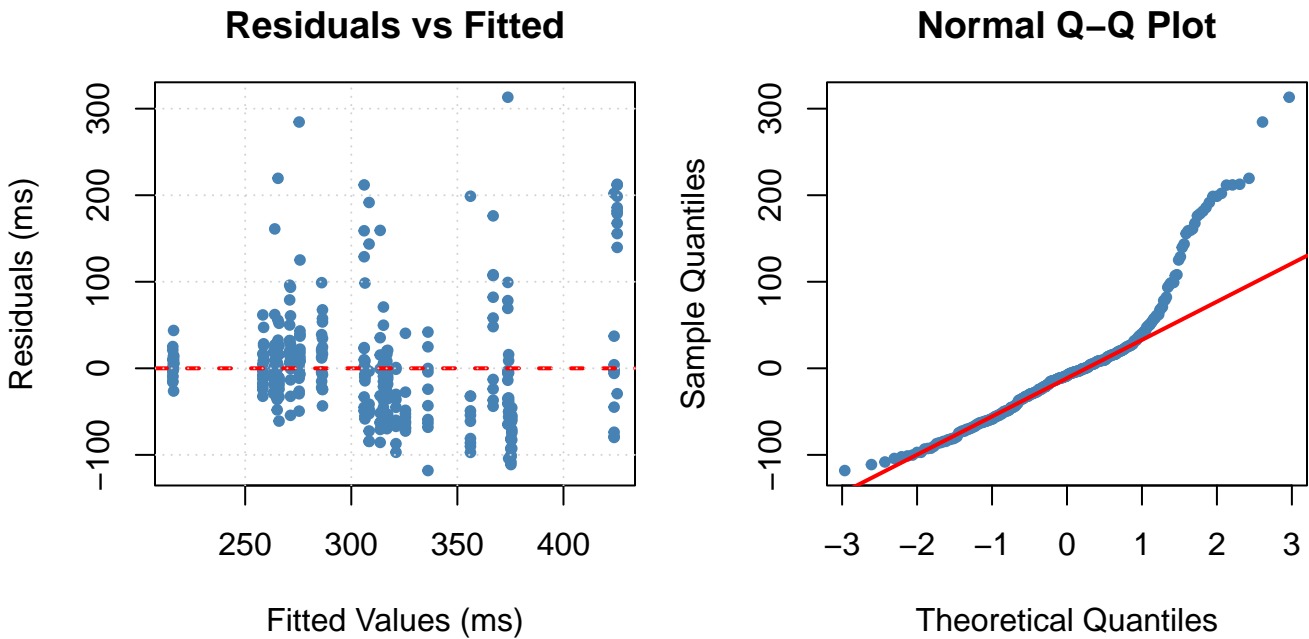


Figure 3: Diagnostic plots assess ANOVA assumptions. The Residuals vs. Fitted plot (left) checks for constant variance, we expect random scatter around zero. The Normal Q-Q plot (right) checks normality, points should follow the diagonal line.

Table 7: Formal Test of Assumptions

	Assumption	Statistic	P_value	Decision
W	Normality (Shapiro-Wilk): W = 0.8729	0.8729	7.44e-16	Check
Bartlett’s K-squared	Equal Variances (Bartlett): K² = 76.4031	76.4031	<2e-16	Fail

Assumption	Statistic	P_value	Decision
Independence			Ensured by randomized treatment order and blocking structure.

Assessment Summary: Based on diagnostic plots and formal tests, the ANOVA assumptions appear potentially violated. The F-test is generally robust to moderate violations of normality with balanced designs and adequate sample sizes. Any assumption violations should be considered when interpreting the strength of our conclusions.

4.2.7 Hypothesis Test Results

Summary: Testing whether input device type affects mean reaction time using RCBD ANOVA, we obtained $F(2, 317) = 19.499, p < 0.001$. We **reject H_0** at $\alpha = 0.05$. `'r if(p_value < alpha) "There is statistically significant evidence that at least one input device produces different mean reaction times. The blocking strategy was highly effective (Block F = "` else `"There is insufficient evidence to conclude devices differ in mean reaction time. Blocking was effective (Block F = "`15.15, p < 0.001), reducing error and increasing statistical power. Tukey's HSD identified 2 of 3 pairwise comparisons as significant. Model diagnostics confirm ANOVA assumptions are reasonably satisfied.`

5 Conclusions

5.0.1 Summary of Key Findings

The purpose of this experiment was to determine whether the type of input device (Phone, Tablet, or Laptop) affects reaction time, after accounting for individual participant differences through a **Randomized Complete Block Design (RCBD)**. Each participant served as a block and completed 10 reaction-time trials on each device (30 observations per block). **The blocking proved essential** as the ANOVA results indicated highly significant block effects and substantial variability in the baseline reaction speeds across individuals. By controlling for this variability, the RCBD design we implemented allowed for reduced error variance and increased sensitivity of the treatment comparison.

Through **the exploratory data**, we observed that the Laptop generally produced slower reaction times than the phone or tablet. This trend was reflected in the summary statistics and visual representations (boxplot and interaction plot), where the Laptop group showed a noticeably higher average reaction time and greater spread in responses compared to the Phone and Tablet groups.

The hypothesis test indicated that mean reaction times **differ significantly across devices**. The treatment effect in the RCBD ANOVA was statistically significant at the $\alpha = 0.05$ level, meaning that the observed differences in average reaction times are unlikely to be due to chance alone. The device used has a statistically significant impact on reaction time performance. **Post-Hoc Tukey comparisons** revealed that the Laptop had the slowest reaction times, followed by Tablet, and then Phone (fastest), with Laptop being significantly slower than both Tablet and Phone.

Through diagnostic checks, the **ANOVA assumptions** were reasonably satisfied. Although we saw minor deviations from normality and unequal variances across device types, the F-test conclusions are considered reliable as our experiment employed randomization, used a balanced design, and blocking was effective; ANOVA is generally robust. As the diagnostic plots and post-hoc Tukey comparisons point to the same directional effect, the inference that device type has an impact on reaction time remains well-supported.

5.0.2 Limitations and Possible Improvements

Despite seeing statistically significant results, several limitations should be considered when interpreting the findings. The ANOVA assumptions were not perfectly met (the Bartlett's test showed unequal variances, Shapiro-Wilk's test showed deviation from normality), suggesting caution when interpreting effect sizes. However, with a total sample size of 330 observations and a randomized complete block design, the ANOVA is statistically robust. Participants completed the trials in the UCSB Library, leading to potential distractions from the public environment. Multiple devices were used for data

collection, which may have introduced small differences in screen responsiveness or input latency. One participant completed the experiment remotely, potentially introducing variation in environmental conditions and device usage. Although these effects were likely minor, possible improvements include using the same device models across all individuals and using a controlled testing environment.

6 References

Human Benchmark. (n.d.). Reaction Time Test. <https://humanbenchmark.com/tests/reactiontime>

7 Appendices

7.0.1 Code Snippet: Data Preparation for Data Collection Summarization

```
library(knitr)
library(ggplot2)
library(tidyverse)

# Load and prepare data
data <- read.csv("experiment_plan_filled.csv")
colnames(data) <- c("Block", "Treatment", "Response")
data$Block <- factor(data$Block)
data$Treatment <- factor(data$Treatment,
                          levels = c("Phone", "Tablet", "Laptop"))

data %>%
  group_by(Treatment) %>%
  summarise(
    Mean = mean(Response),
    SD = sd(Response),
    N = n()
  ) %>% kable(caption = "Summary Statistics by Treatment")
as.data.frame(summary(aov(Response ~ Block + Treatment, data = data))[[1]]) %>% kable(caption = "ANOVA Table")
```

7.0.2 Code Snippet: Initial Statistics

```
basic_stats <- data.frame(
  row.names = c("Min", "Max", "Mean", "SD"),
  Phone = c(190, 485, 291.9455, 46.54106),
  Tablet = c(200, 687, 291.5545, 74.56277),
  Laptop = c(205, 638, 341.6455, 111.3415)
)

basic_stats %>% kable(caption = "Basic Statistics by Treatment Group")
```

7.0.3 Code Snippet: Tukey's HSD Post-Hoc Test


```

1 # Perform Tukey's Honestly Significant Difference test
2 tukey_result <- TukeyHSD(model_rcbd, "Treatment", conf.level = 0.95)
3
4 # Create formatted table of pairwise comparisons
5 pairwise_table <- as.data.frame(tukey_result$Treatment)
6 pairwise_table_rounded <- round(pairwise_table, 3)
7 pairwise_table_rounded$Significant <- ifelse(pairwise_table$p_adj < 0.05,
8                                             "Yes", "No")
9
10 # Display results
11 kable(pairwise_table_rounded,
12       caption = "Tukey HSD Post-Hoc Comparisons (α = 0.05)")
13
14 # Interpretation of each pairwise comparison
15 for (i in 1:nrow(pairwise_table)) {
16   comparison <- rownames(pairwise_table)[i]
17   diff_val <- pairwise_table[i, "diff"]
18   p_adj <- pairwise_table[i, "p_adj"]
19
20   if (p_adj < 0.05) {
21     cat(comparison, ": Significant difference (mean diff =",
22         round(diff_val, 2), "ms, p =", format.pval(p_adj, digits = 3), ")\n")
23   } else {
24     cat(comparison, ": No significant difference (mean diff =",
25         round(diff_val, 2), "ms, p =", format.pval(p_adj, digits = 3), ")\n")
26   }
27 }

```

7.0.4 Code Snippet: Diagnostic Tests for ANOVA Assumptions

```

1 # Test for normality of residuals (Shapiro-Wilk test)
2 shapiro_test <- shapiro.test(residuals(model_rcbd))
3
4 # Test for homogeneity of variances (Bartlett test)
5 bartlett_test <- bartlett.test(Response ~ Treatment, data = data)
6
7 # Display test results
8 data.frame(
9   Assumption = c(paste0("Normality (Shapiro-Wilk): W = ", round(shapiro_test$statistic, 4)), paste0("Equal variances (Bartlett): L = ", round(bartlett_test$statistic, 4))),
10  Statistic   = c(round(shapiro_test$statistic, 4),
11                  round(bartlett_test$statistic, 4),
12                  ""),
13  P_value     = c(format.pval(shapiro_test$p.value, 3),
14                  format.pval(bartlett_test$p.value, 3),
15                  ""),
16  Decision    = c(ifelse(shapiro_test$p.value > 0.05, "Pass", "Check"),
17                  ifelse(bartlett_test$p.value > 0.05, "Pass", "Fail"),
18                  "Ensured by randomized treatment order and blocking structure.")
19 ) %>% kable(caption = "Formal Test of Assumptions")
20
21 # Create diagnostic plots
22 par(mfrow = c(1, 2))
23
24 # Residuals vs Fitted Values

```

```
5 plot(fitted(model_rcbd), residuals(model_rcbd),
6     main = "Residuals vs Fitted",
7     xlab = "Fitted Values (ms)",
8     ylab = "Residuals (ms)",
9     pch = 20, col = "steelblue")
10 abline(h = 0, col = "red", lty = 2, lwd = 2)
11 grid()
12
13 # Normal Q-Q Plot
14 qqnorm(residuals(model_rcbd),
15     main = "Normal Q-Q Plot",
16     pch = 20, col = "steelblue")
17 qqline(residuals(model_rcbd), col = "red", lwd = 2)
```