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)
- STUDENT 5 (NetID 5)

Due Date

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

1 Introduction

- Clear statement of the objective or research question.
- Brief context or motivation.

2 Experimental Design

- Description of factors and treatment structure.
- Clearly state what you are measuring and the units. Examples: Number of words recalled (count), reaction time (seconds), taste rating (1–5 scale).
- Identify which factors are fixed vs. random.
- Description of design type (CRD, RCBD, factorial, etc.).
- Explain how randomization, replication, and (if used) blocking were implemented.
- Sample size: Provide number of observations per condition. Guideline: 5–10 per treatment for CRD, 3–5 blocks for RCBD, total feasible within 1 hour.

(You are encouraged to explore more resources for determining the sample size)

Randomized Complete Block Design

Each person is a block, where they do 10 reaction tests for each treatment level (iPad, iPhone, Macbook Trackpad). Each experimental unit is an individual single reaction speed test.

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

- **Exploratory Data**: Start with basic statistics (mean, SD) and visualizations (e.g., boxplots) to understand the data.

4.1 Hypothesis Testing

4.1.1 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.

4.1.2 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)

4.1.3 Preliminary Visualization of Treatment Effects

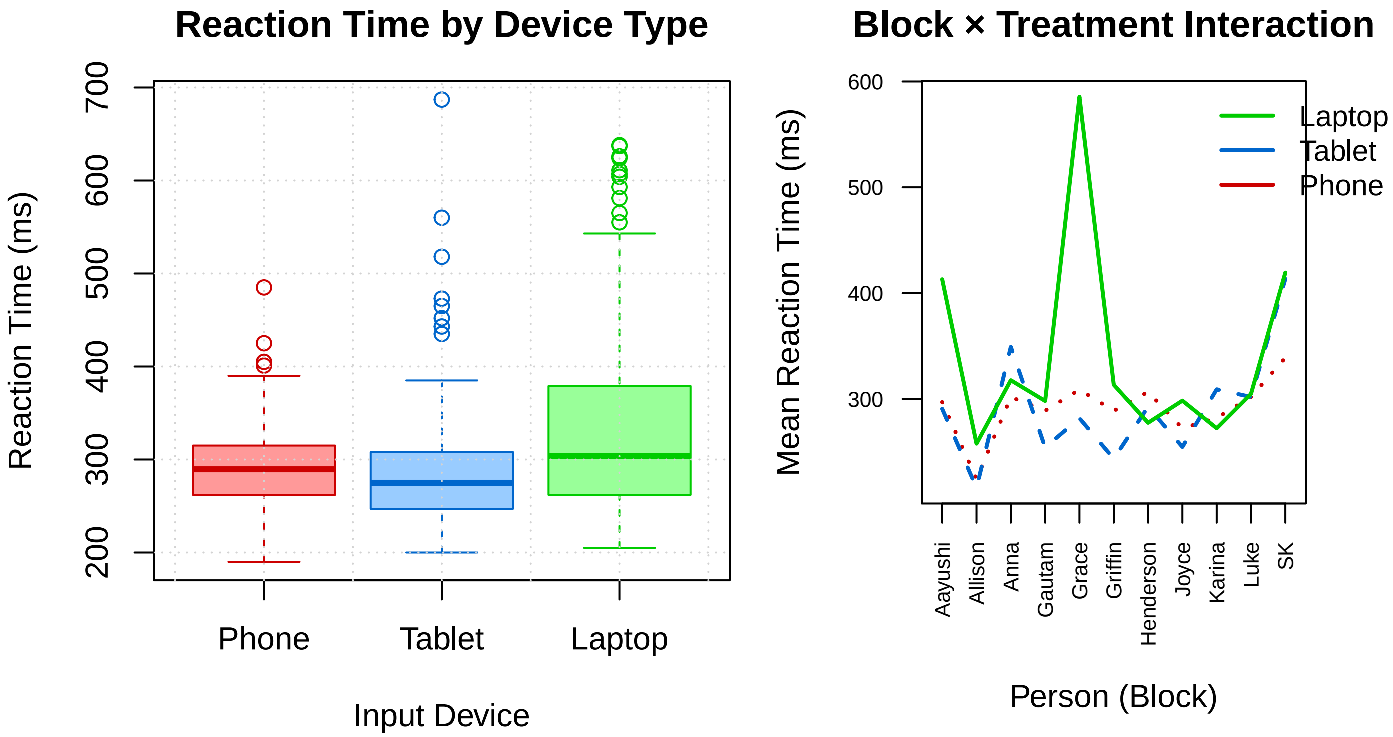

Figure 1 displays two key visualizations. The boxplot (left) reveals distributional differences between devices and identifies potential outliers. The interaction plot (right) demonstrates how participants (blocks) responded differently to each device, with non-parallel lines indicating individual variability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

4.1.4 ANOVA Computation and Results

The RCBD ANOVA model (Response $\sim$ 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: $\mathrm{SST} = \mathrm{SSB} + \mathrm{SSTr} + \mathrm{SSE}$

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

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

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

```
             Df  Sum Sq Mean Sq F value   Pr(>F)    
Block        10  709283   70928   15.15  < 2e-16 ***
Treatment     2  182576   91288   19.50 1.03e-08 ***
Residuals   317 1484081    4682                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


=== Critical Test Statistics ===

F-statistic for Treatment: 19.4991

P-value: 1.031e-08

Degrees of Freedom: Treatment df = 2 , Error df = 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.

4.1.5 Hypothesis Test Decision and Statistical Conclusion

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

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

DECISION: REJECT H_0

Observed F-statistic: F = 19.4991

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

P-value: p = 1.031e-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.

4.1.6 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 3: 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)

4.1.7 Verification of ANOVA Assumptions

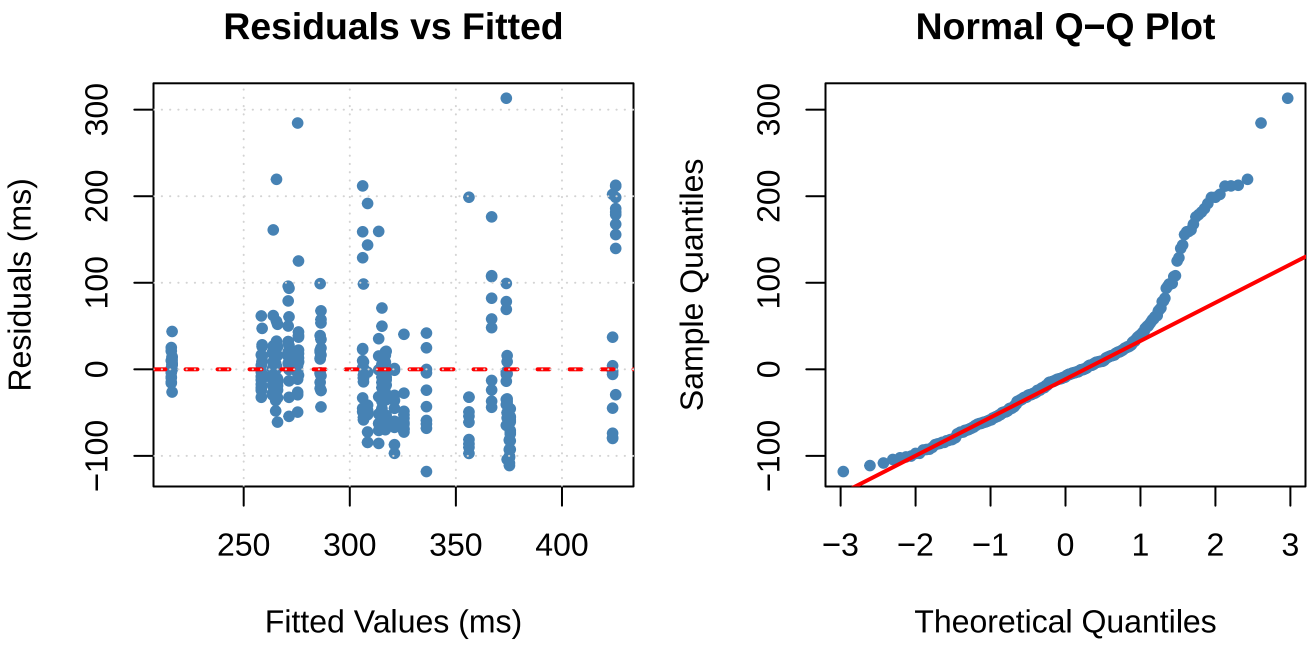

Figure 3: Diagnostic plots assess ANOVA assumptions. The Residuals vs. Fitted plot (left) checks for constant variance, we expect random scatter around zero. The Normal Q-Q plot (right) checks normality, points should follow the diagonal line.

```
=== Formal Tests of Assumptions ===

**Normality (Shapiro-Wilk):** W = 0.8729 , p = 7.44e-16

 → Minor departure from normality (ANOVA is robust)
```

```
**Equal Variances (Bartlett):** K² = 76.4031 , p = <2e-16

 → Unequal variances detected

**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.

4.1.8 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** at $\alpha = 0.05$. There is statistically significant evidence that at least one input device produces different mean reaction times. The blocking strategy was highly 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

- Summarize key findings.
- Comment on limitations and possible improvements.

6 References

(If needed.)

7 Appendices

7.0.1 Code Snippet: Data Preparation for Data Collection Summarization

```r
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 Ta
```

7.0.2 Code Snippet: Data Preparation for RCBD Model Fitting

```r
# Load required libraries
library(knitr)
library(ggplot2)

# 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"))

# Fit Randomized Complete Block Design (RCBD) ANOVA model
model_rcbd <- aov(Response ~ Block + Treatment, data = data)

# Extract ANOVA table and key statistics
anova_table <- summary(model_rcbd)
f_stat <- anova_table[[1]]["Treatment", "F value"]
p_value <- anova_table[[1]]["Treatment", "Pr(>F)"]
alpha <- 0.05

# Display ANOVA results
print(anova_table)
cat("\nF-statistic for Treatment:", round(f_stat, 4), "\n")
cat("P-value:", format.pval(p_value, digits = 4), "\n")
```

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

```r
# Perform Tukey's Honestly Significant Difference test
tukey_result <- TukeyHSD(model_rcbd, "Treatment", conf.level = 0.95)

# Create formatted table of pairwise comparisons
pairwise_table <- as.data.frame(tukey_result$Treatment)
pairwise_table_rounded <- round(pairwise_table, 3)
pairwise_table_rounded$Significant <- ifelse(pairwise_table$`p adj` < 0.05,
                                              "Yes", "No")

# Display results
kable(pairwise_table_rounded,
      caption = "Tukey HSD Post-Hoc Comparisons (  = 0.05)")

# Interpretation of each pairwise comparison
for (i in 1:nrow(pairwise_table)) {
  comparison <- rownames(pairwise_table)[i]
  diff_val <- pairwise_table[i, "diff"]
  p_adj <- pairwise_table[i, "p adj"]

  if (p_adj < 0.05) {
    cat(comparison, ": Significant difference (mean diff =",
        round(diff_val, 2), "ms, p =", format.pval(p_adj, digits = 3), ")\n")
  } else {
    cat(comparison, ": No significant difference (mean diff =",
        round(diff_val, 2), "ms, p =", format.pval(p_adj, digits = 3), ")\n")
  }
}
```

7.0.4 Code Snippet: Diagnostic Tests for ANOVA Assumptions

```
# Test for normality of residuals (Shapiro-Wilk test)
shapiro_test <- shapiro.test(residuals(model_rcbd))

# Test for homogeneity of variances (Bartlett test)
bartlett_test <- bartlett.test(Response ~ Treatment, data = data)

# Display test results
cat("=== Formal Tests of Assumptions ===\n\n")

cat("Normality (Shapiro-Wilk Test):\n")
cat("  W =", round(shapiro_test$statistic, 4), "\n")
cat("  p-value =", format.pval(shapiro_test$p.value, digits = 4), "\n")
if (shapiro_test$p.value > 0.05) {
  cat("  Interpretation: Normality assumption satisfied\n\n")
} else {
  cat("  Interpretation: Minor departure from normality detected\n\n")
}

cat("Homogeneity of Variances (Bartlett Test):\n")
cat("  K² =", round(bartlett_test$statistic, 4), "\n")
cat("  p-value =", format.pval(bartlett_test$p.value, digits = 4), "\n")
if (bartlett_test$p.value > 0.05) {
  cat("  Interpretation: Equal variance assumption satisfied\n\n")
} else {
  cat("  Interpretation: Unequal variances detected\n\n")
}

# Create diagnostic plots
par(mfrow = c(1, 2))

# Residuals vs Fitted Values
plot(fitted(model_rcbd), residuals(model_rcbd),
     main = "Residuals vs Fitted",
     xlab = "Fitted Values (ms)",
     ylab = "Residuals (ms)",
     pch = 20, col = "steelblue")
abline(h = 0, col = "red", lty = 2, lwd = 2)
grid()

# Normal Q-Q Plot
qqnorm(residuals(model_rcbd),
       main = "Normal Q-Q Plot",
       pch = 20, col = "steelblue")
qqline(residuals(model_rcbd), col = "red", lwd = 2)
```