Design and Analysis of Experiments (MAL2040)

Project Report

Subject:- Effects of Various Factors on Reading Comprehension: A 2^5 Factorial Design Analysis

Contributors:
Divyamoorti Panthi(B22CY009)
Hritin Raj(B22ES015)
Pranav Bhansali(B22ES003)
Raghavi Epuri(B22ES012)
Samriddhi Jain(B22ES008)
Soumen Kumar(B22ES006)

Date: May 17, 2024

1 Introduction

Reading comprehension is a crucial skill that plays a vital role in academic and professional success. However, numerous factors can potentially influence an individual's ability to comprehend written material effectively. This study aimed to investigate the impact of five distinct factors on reading comprehension: listening to music(A), difficulty of the paragraph(B), writing style(C), font size(D), and environmental conditions(E).

1.1 Data Collection

One participant was exposed to various combinations of factor levels, and their reading comprehension scores were recorded in a table.

1.2 Calculating the Response Variable (RCS)

To calculate the 'Reading Comprehension Score', we deployed a simple formula that considers the time the participant took to read the comprehension.

We kept the base score of 0 and set penalties for extra time, starting after 1 minute for easy passages and after 3 minutes for hard passages. Here's how:

- Base Score: 0 points
- Points per Correct Answer: 10 points
- Penalty for Extra Time (Easy Passage): 1 point deducted per minute beyond 1 minute
- Penalty for Extra Time (Hard Passage): 1 point deducted per minute beyond 3 minutes

The empirical formula to calculate the reading comprehension score is:

$$\text{Score} = (\text{Correct Answers} \times 10) - (\text{Penalty for Extra Time})$$

These examples demonstrate how to apply the formula to calculate the Reading Comprehension Score for different runs based on the updated data.

1. Example 1: For Run Number 1 with factors (-, -, -, -, -):
   - Correct Answers: 6
   - Time Taken: 4 minutes
   - Penalty for Extra Time: 3 * 1 = 3 points (since it's beyond 1 minute for an easy passage)
   - Reading Comprehension Score =62
2. Example 2: For Run Number 10 with factors (+, -, -, +, -):
   - Correct Answers: 4
   - Time Taken: 6 minutes
   - Penalty for Extra Time: 5 * 1 = 5 points (since it's beyond 1 minute for an easy passage)
   - Reading Comprehension Score =18
3. Example 3: For Run Number 24 with factors (+, +, +, -, +):
   - Correct Answers: 2
   - Time Taken: 6 minutes
   - Penalty for Extra Time: 5 * 1 = 5 points (since it's beyond 1 minute for an easy passage)
   - Reading Comprehension Score =13
4. Example 4: For Run Number 32 with factors (+, +, +, +, +):
   - Correct Answers: 2
   - Time Taken: 6 minutes
   - Penalty for Extra Time: 5 * 1 = 5 points (since it's beyond 1 minute for an easy passage)
   - Reading Comprehension Score =13

The data was then uploaded to the statistical software R for further analysis.

1.3 Snippets from Data Collection

Here are also some of the snippets taken by us while collecting the data:

1.4 Collected Data

The following table presents the collected data from the experiment:

Table 1: Collected Data

Run	A	B	C	D	E	Time Taken (min)	Correct Answers	Reading Comprehension Score
1	0	0	0	0	0	4	6	62
2	1	0	0	0	0	5	4	38
3	0	1	0	0	0	3	8	82
4	1	1	0	0	0	6	3	23
5	0	0	1	0	0	4	7	71
6	1	0	1	0	0	5	5	47
7	0	1	1	0	0	3	9	93
8	1	1	1	0	0	6	2	13
9	0	0	0	1	0	4	6	58
10	1	0	0	1	0	6	4	18
11	0	1	0	1	0	3	8	77
12	1	1	0	1	0	5	3	27
13	0	0	1	1	0	6	7	52
14	1	0	1	1	0	5	5	42
15	0	1	1	1	0	4	9	88
16	1	1	1	1	0	6	2	13
17	0	0	0	0	1	5	6	47
18	1	0	0	0	1	6	4	18
19	0	1	0	0	1	4	8	63
20	1	1	0	0	1	5	3	27
21	0	0	1	0	1	6	7	52
22	1	0	1	0	1	5	5	42
23	0	1	1	0	1	4	9	88
24	1	1	1	0	1	6	2	13
25	0	0	0	1	1	6	6	48
26	1	0	0	1	1	5	4	23
27	0	1	0	1	1	6	8	67
28	1	1	0	1	1	4	3	33
29	0	0	1	1	1	6	7	52
30	1	0	1	1	1	5	5	42
31	0	1	1	1	1	4	9	88
32	1	1	1	1	1	6	2	13

2 Problem Statement

Reading comprehension is a crucial skill that can be influenced by various factors. A factorial experiment was conducted to investigate the effects of five factors on an individual's reading ability: listening to music (**A**), difficulty of the paragraph (**B**), writing style (**C**), font size (**D**), and environmental conditions (**E**). Each factor was present at two levels (e.g., with music and without music for factor A).

2.1 Methodology

A 2^5 factorial design with a single replication was employed, resulting in a total of 32 experimental runs. The factors under investigation were:

1. Listening to Music (**A**): Two levels - With music and without music.
2. Difficulty of Paragraph (**B**): Two levels - Easy and difficult.
3. Style of Writing (**C**): Two levels - Formal and informal.
4. Font Size (**D**): Two levels - Small and large.
5. Environment (**E**): Two levels - Quiet and noisy.

The primary response variable measured was the reading comprehension score, which quantified an individual's understanding of the given text.

2.2 Labeled Table for Analysis

Table 2: Response Data

Run Number	A	B	C	D	E	Run Label	Reading Comprehension Score
1	-	-	-	-	-	(1)	62
2	+	-	-	-	-	a	38
3	-	+	-	-	-	b	82
4	+	+	-	-	-	ab	23
5	-	-	+	-	-	c	71
6	+	-	+	-	-	ac	47
7	-	+	+	-	-	bc	93
8	+	+	+	-	-	abc	13
9	-	-	-	+	-	d	58
10	+	-	-	+	-	ad	18
11	-	+	-	+	-	bd	77
12	+	+	-	+	-	abd	27
13	-	-	+	+	-	cd	52
14	+	-	+	+	-	acd	42
15	-	+	+	+	-	bcd	88
16	+	+	+	+	-	abcd	13
17	-	-	-	-	+	e	47
18	+	-	-	-	+	ae	18
19	-	+	-	-	+	be	63
20	+	+	-	-	+	abe	27
21	-	-	+	-	+	ce	52
22	+	-	+	-	+	ace	42
23	-	+	+	-	+	bce	88
24	+	+	+	-	+	abce	13
25	-	-	-	+	+	de	48
26	+	-	-	+	+	ade	23
27	-	+	-	+	+	bde	67
28	+	+	-	+	+	abde	33
29	-	-	+	+	+	cde	52
30	+	-	+	+	+	acde	42
31	-	+	+	+	+	bcde	88
32	+	+	+	+	+	abcde	13

3 Statistical Analysis using R

3.1 Stage 1: Data Collection and Storage

Comprehension Score (RCS) Data

The reading comprehension scores are stored in a matrix named RCS. The matrix is defined as follows:

$$\text{RCS} = \left(62, 38, 82, 23, 71, 47, 93, 13, 58, 18, 77, 27, 52, 42, 88, 13, 47, 18, 63, 27, 52, 42, 88, 13, 48, 23, 67, 33, 52, 42, 88, 13\right)$$

Design Matrix

The design matrix represents the experimental design and includes columns for factors, interaction terms, and the reading comprehension scores. It is defined as follows:

> Design.matrix

	I	A	B	C	D	E	AB	AC	AD	AE	BC	BD	BE	CD	CE	DE	ABC	ABD	ABE	ACD	ACE	ADE	BCD	BCE	BDE	CDE	ABCD	ABCE	ABDE	ACDE	BCDE	ABCDE	rcs
(1)	1	-1	-1	-1	-1	-1	1	1	1	1	1	1	1	1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1	1	-1	62
a	1	1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1	1	1	1	1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	38
b	1	-1	1	-1	-1	-1	-1	1	1	1	-1	-1	-1	1	1	1	1	1	1	-1	-1	-1	1	1	1	-1	-1	-1	-1	1	-1	1	82
c	1	1	1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	1	1	1	-1	-1	-1	1	1	1	1	1	1	-1	1	1	1	-1	-1	-1	23
d	1	-1	-1	1	-1	-1	1	-1	1	1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	1	1	-1	1	-1	-1	1	-1	-1	1	71
e	1	1	-1	1	-1	-1	-1	1	-1	-1	-1	1	1	-1	-1	1	-1	1	1	-1	-1	1	1	1	-1	1	1	1	-1	1	-1	-1	47
ab	1	-1	1	1	-1	-1	-1	-1	1	1	1	-1	-1	-1	-1	1	-1	1	1	1	1	-1	-1	-1	1	1	1	1	-1	-1	1	-1	93
ac	1	1	1	1	-1	-1	1	1	-1	-1	1	-1	-1	-1	-1	1	1	-1	-1	-1	-1	1	-1	-1	1	1	-1	-1	1	1	1	1	13
ad	1	-1	-1	-1	1	-1	1	1	-1	1	1	-1	1	-1	1	-1	-1	1	-1	1	-1	1	1	-1	1	1	-1	1	-1	-1	-1	1	58
ae	1	1	-1	-1	1	-1	-1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	1	1	-1	1	1	-1	-1	18
bc	1	-1	1	-1	1	-1	-1	1	-1	1	-1	1	-1	-1	1	-1	1	-1	1	1	-1	1	-1	1	-1	1	1	-1	1	-1	1	-1	77
bd	1	1	1	-1	1	-1	1	-1	1	-1	-1	1	-1	-1	1	-1	-1	1	-1	-1	1	-1	-1	1	-1	1	-1	1	-1	1	1	1	27
be	1	-1	-1	1	1	-1	1	-1	-1	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	1	1	-1	1	-1	-1	1	1	-1	52
cd	1	1	-1	1	1	-1	-1	1	1	-1	-1	-1	1	1	-1	-1	-1	-1	1	1	-1	-1	-1	1	1	-1	-1	1	1	-1	1	1	42
ce	1	-1	1	1	1	-1	-1	-1	-1	1	1	1	-1	1	-1	-1	-1	-1	1	-1	1	1	1	-1	-1	-1	-1	1	1	1	-1	1	88
de	1	1	1	1	1	-1	1	1	1	-1	1	1	-1	1	-1	-1	1	1	-1	1	-1	-1	1	-1	-1	-1	1	-1	-1	-1	-1	-1	13
abc	1	-1	-1	-1	-1	1	1	1	1	-1	1	1	-1	1	-1	-1	-1	-1	1	-1	1	1	-1	1	1	1	1	-1	-1	-1	-1	1	47
abd	1	1	-1	-1	-1	1	-1	-1	-1	1	1	1	-1	1	-1	-1	1	1	-1	1	-1	-1	-1	1	1	1	-1	1	1	1	-1	-1	18
abe	1	-1	1	-1	-1	1	-1	1	1	-1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	1	-1	-1	1	-1	1	1	-1	1	-1	63
acd	1	1	1	-1	-1	1	1	-1	-1	1	-1	-1	1	1	-1	-1	-1	-1	1	1	-1	-1	1	-1	-1	1	1	-1	-1	1	1	1	27
ace	1	-1	-1	1	-1	1	1	-1	1	-1	-1	1	-1	-1	1	-1	1	-1	1	1	-1	1	1	-1	1	-1	-1	1	-1	1	1	-1	52
ade	1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	-1	1	-1	-1	1	-1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	1	42
bcd	1	-1	1	1	-1	1	-1	-1	1	-1	1	-1	1	-1	1	-1	-1	1	-1	1	-1	1	-1	1	-1	-1	1	-1	1	1	-1	1	88
bce	1	1	1	1	-1	1	1	1	-1	1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	-1	1	-1	-1	-1	1	-1	-1	-1	-1	13
bde	1	-1	-1	-1	1	1	1	1	-1	-1	1	-1	-1	-1	-1	1	-1	1	1	1	1	-1	1	1	-1	-1	-1	-1	1	1	1	-1	48
cde	1	1	-1	-1	1	1	-1	-1	1	1	1	-1	-1	-1	-1	1	1	-1	-1	-1	-1	1	1	1	-1	-1	1	1	-1	-1	1	1	23
abcd	1	-1	1	-1	1	1	-1	1	-1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1	-1	1	-1	1	1	-1	1	-1	1	67
abce	1	1	1	-1	1	1	1	-1	1	1	-1	1	1	-1	-1	1	-1	1	1	-1	-1	1	-1	-1	1	-1	-1	-1	1	-1	-1	-1	33
abde	1	-1	-1	1	1	1	1	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1	-1	-1	-1	-1	-1	-1	1	1	1	1	-1	-1	1	52
acde	1	1	-1	1	1	1	-1	1	1	1	-1	-1	-1	1	1	1	-1	-1	-1	1	1	1	-1	-1	-1	1	-1	-1	-1	1	-1	-1	42
bcde	1	-1	1	1	1	1	-1	-1	-1	-1	1	1	1	1	1	1	-1	-1	-1	-1	-1	-1	1	1	1	1	-1	-1	-1	-1	1	-1	88
abcde	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	13

Stage 2: Pre-Analysis

In this stage, we conduct pre-analysis to identify important factors from the chosen set of factors (A, B, C, D, E) using the Half Normal Plot method.

Half Normal Plot

We first calculate the effects using the Yates method and create a Half Normal Plot to visualize the magnitude of the effects. The code snippet for generating the Half Normal Plot is as follows:

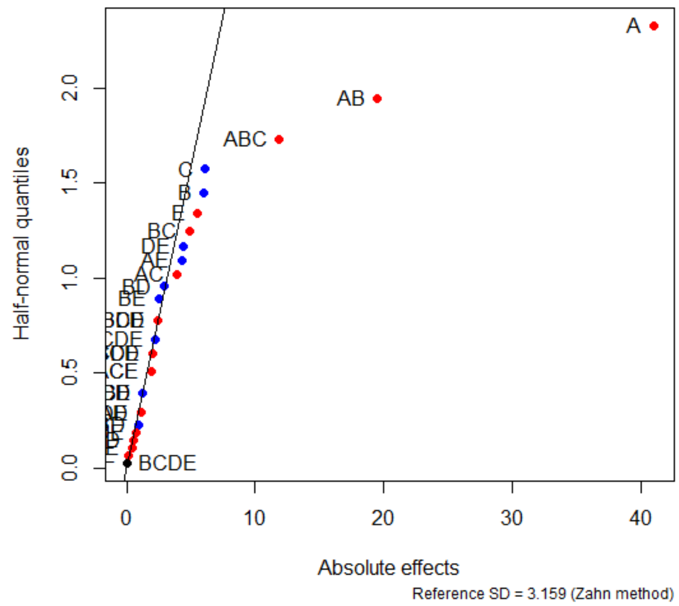
```r
# Importing the required library
library(unrepx)

# Calculating effects using Yates method
G <- Design.matrix[, 33]
pilotEff <- yates(G, labels = c("A", "B", "C", "D", "E"))

# Creating the Half Normal Plot
hnplot(pilotEff, ID = 0)
```

The Half Normal Plot helps us identify the significant factors based on the magnitude of their effects.

```
> pilotEff
      A       B      AB       C      AC      BC     ABC       D      AD      BD     ABD      CD     ACD     BCD    ABCD       E      AE      BE     ABE
-41.000   6.000 -19.500   6.125  -3.875  -4.875 -11.875  -2.375   1.125   2.875   0.875  -1.250   1.250  -0.500  -2.000  -5.500   4.250   2.500   1.250
     CE     ACE     BCE    ABCE      DE     ADE     BDE    ABDE     CDE    ACDE    BCDE   ABCDE
  1.875  -1.875  -0.125  -2.375   4.375  -0.375  -2.375  -1.125  -0.750  -2.000   0.000   2.250
attr(,"mean")
```



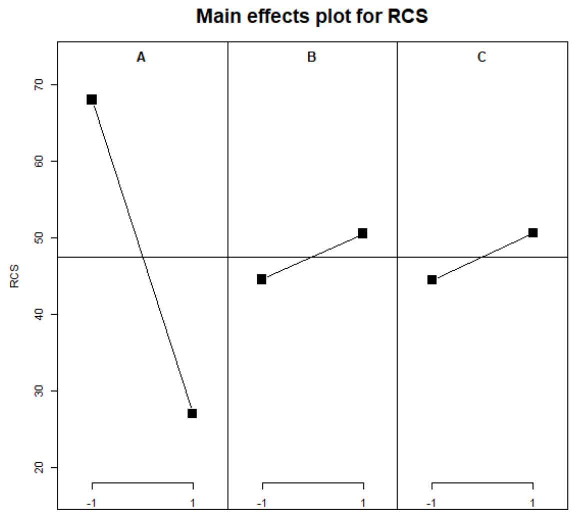
Conclusion from this plot:
From the value of estimate effects and visually from the half-normal plot, we observe that A = -41.000, B = 6.000, C = 6.125. Hence, this analysis revealed that factors A (Listening to Music), B (Difficulty of Paragraph), and C (Style of Writing) had the most substantial effects on reading comprehension scores.

Interaction and Main Effects

Next, we compute the Interaction and Main Effects based on the Comprehension Scores (RCS). We calculate these effects using linear regression models. Here's a summary of the Interaction and Main Effects:

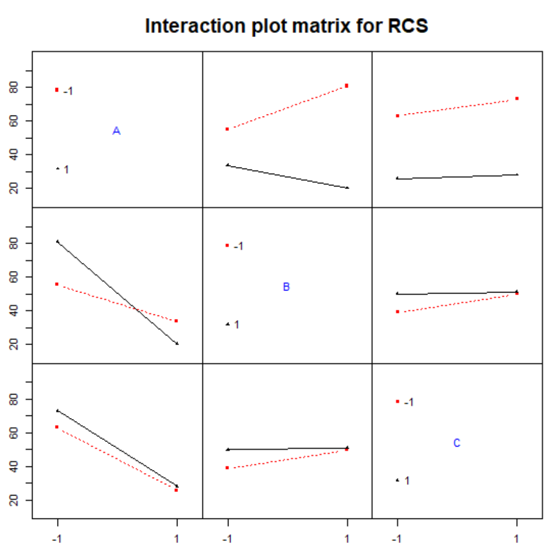
```
# Computing Interaction and Main Effects
n <- 1 # Replication
Feff <- t(RCS) %*% cbind(A, B, AB, C, AC, BC, ABC, D, AD, BD, ABD, CD, ACD, BCD, ABCD,
                         E, AE, BE, ABE, CE, ACE, BCE, ABCE, DE, ADE, BDE, ABDE, CDE,
                         ACDE, BCDE, ABCDE) / (16 * n)
Ieff <- t(RCS) %*% cbind(I) / (32 * n)
eff <- cbind(Ieff, Feff)
Summary <- rbind(cbind(I, A, B, AB, C, AC, BC, ABC, D, AD, BD, ABD, CD, ACD, BCD, ABCD,
                         E, AE, BE, ABE, CE, ACE, BCE, ABCE, DE, ADE, BDE, ABDE, CDE,
                         ACDE, BCDE, ABCDE), eff)
dimnames(Summary)[[1]] <- c(dimnames(RCS)[[1]], "Effect")
```

This summary provides insight into the effects of different factors on the Comprehension Scores.



The main effects plot visually represents how the different levels of each factor (A, B, and C) influence the response variable (Reading Comprehension Scores, RCS).

- **Factor A (Listening to Music):** The plot shows a decrease in RCS when moving from the low level (-1) to the high level (+1) of factor A. This suggests that listening to music while reading may have a negative impact on reading comprehension scores.
- **Factor B (Difficulty of Paragraph):** The plot clearly demonstrates a significant increase in RCS when moving from the low level (-1, difficult paragraph) to the high level (+1, easy paragraph) of factor B. This makes sense in real life because easier paragraphs lead to better reading comprehension than more difficult ones.
- **Factor C (Style of Writing):** The plot suggests a slight increase in RCS when moving from the low level (-1, informal style) to the high level (+1, formal style) of factor C. This means that a formal writing style may be more conducive to better reading comprehension than an informal one.



The interaction plot matrix visually represents the interactions between the different factors (A, B, and C) and their effects on the response variable (Reading Comprehension Scores, RCS).

- **A:B Interaction:** The plot for the interaction between factors A (Listening to Music) and B (Difficulty of Paragraph) shows slightly diverging lines. This suggests a potential interaction effect. The practical implication is that the impact of listening to music on reading comprehension may depend to some extent on the difficulty level of the paragraph, but this effect is relatively weak.
- **A:C Interaction:** The plot for the interaction between factors A (Listening to Music) and C (Style of Writing) shows nearly parallel lines, indicating a lack of a substantial interaction effect.

- **B:C Interaction:** The plot for the interaction between factors B (Difficulty of Paragraph) and C (Style of Writing) displays a clear non-parallel pattern, with the lines intersecting. The interaction plot indicates that the effect of writing style (factor C) on reading comprehension scores is more pronounced for difficult paragraphs (low level of factor B) compared to easy paragraphs. In other words, the choice of writing style becomes more critical when the paragraph is difficult, as an appropriate writing style can potentially improve reading comprehension for challenging texts.

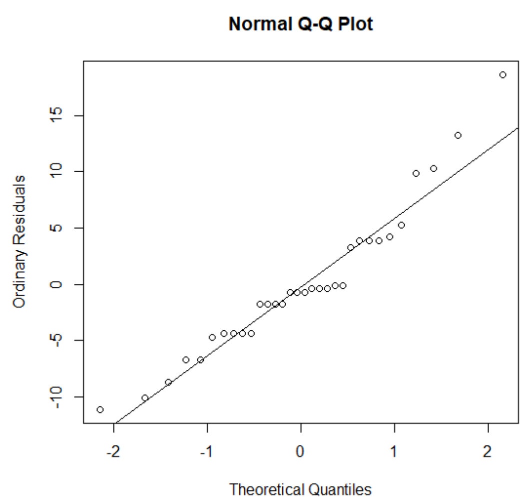
  Therefore, the best reading comprehension scores would be obtained when A is at a low level and B and C are at a high level.

Regression Analysis

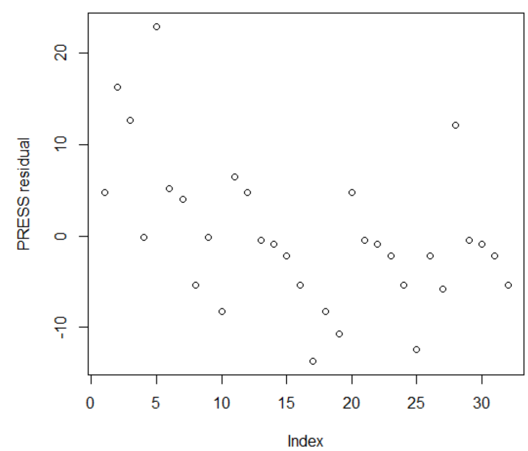
Finally, we perform a regression analysis to model the relationship between the factors (A, B, C) and the Comprehension Scores (RCS). Here's a summary of the regression analysis:

```r
# Regression Analysis
mod <- lm(RCS ~ A + B + C + A:B + A:B:C, data = data.rcs)
```

We fit a linear regression model using the factors A, B, C, and their interactions. The summary of this model provides information about the significance of each factor and their interactions.



Normal Q-Q Plot



Residuals vs. Observation Order

```
> lr=lm(RCS ~ A*B*C, data=data.rcs)
> summary(lr)

Call:
lm.default(formula = RCS ~ A * B * C, data = data.rcs)

Residuals:
   Min     1Q Median     3Q    Max 
 -9.25  -4.75  -1.25   3.75  14.25 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   47.500      1.154  41.146  < 2e-16 ***
A            -20.500      1.154 -17.758 2.61e-15 ***
B              3.000      1.154   2.599   0.0158 *  
C              3.063      1.154   2.653   0.0139 *  
A:B           -9.750      1.154  -8.446 1.19e-08 ***
A:C           -1.938      1.154  -1.678   0.1063    
B:C           -2.438      1.154  -2.111   0.0453 *  
A:B:C         -5.938      1.154  -5.143 2.89e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.53 on 24 degrees of freedom
Multiple R-squared:  0.9476,	Adjusted R-squared:  0.9323 
F-statistic: 62.03 on 7 and 24 DF,  p-value: 7.927e-14
```

LR using main factors

```
Call:
lm.default(formula = RCS ~ A + B + C + A:B + A:B:C, data = data.rcs)

Residuals:
    Min      1Q  Median      3Q     Max 
-11.125  -4.375  -0.750   3.875  18.625 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   47.500      1.266  37.516  < 2e-16 ***
A            -20.500      1.266 -16.191 4.25e-15 ***
B              3.000      1.266   2.369   0.0255 *  
C              3.063      1.266   2.419   0.0229 *  
A:B           -9.750      1.266  -7.701 3.59e-08 ***
A:B:C         -5.937      1.266  -4.690 7.63e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.162 on 26 degrees of freedom
Multiple R-squared:  0.9317,	Adjusted R-squared:  0.9186 
F-statistic: 70.98 on 5 and 26 DF,  p-value: 2.56e-14
```

LR using main and interaction factors both

Figure 1: Outputs from Linear Regression Analysis

1. **Image 1**: This is a normal Q-Q plot, which is used to assess the normality of the residuals in a linear regression model. The points in the plot appear to follow a roughly straight line, suggesting that the residuals are normally distributed, which is a desirable assumption for linear regression models.

2. **Image 2**: This plot shows the residuals (differences between the observed values and the fitted values) against the index or observation order. The residuals appear to be randomly scattered around zero, which is an indication that the assumptions of linearity and homoscedasticity (constant variance of residuals) are met.

3. **Image 3**: This is the output from a linear regression analysis, which includes the coefficients, their standard errors, t-values, and p-values. The coefficients labeled A, B, A:B, and A:B:C are statistically significant (indicated by asterisks), suggesting that these terms are important in explaining the response variable (RCS). The adjusted R-squared value of 0.9333 indicates that the model explains a good portion of the variance in the response variable.

4. **Image 4**: This output is from a more complex linear regression model that includes additional interaction terms (A:B and A:B:C). The adjusted R-squared value of 0.9186 is slightly lower than the previous model, but the F-statistic and its associated p-value suggest that the overall model is still statistically significant. The coefficients for some of the terms (A, B, A:B, and A:B:C) are also statistically significant.

   Overall, the images suggest that the linear regression models being analyzed are reasonably well-fitted, with the residuals meeting the assumptions of normality, linearity, and homoscedasticity. The coefficients and their significance levels indicate which terms are important in explaining the response variable (RCS). The Q-Q plot and residual plots are useful diagnostic tools for assessing the validity of the linear regression assumptions.

```
> residuals
   (1)       a       b      c      d      e     ab     ac     ad     ae     bc     bd     be     cd     ce     de    abc    abd    abe
 3.875  13.250  10.250 -0.125 18.625  4.250  3.250 -4.375 -0.125 -6.750  5.250  3.875 -0.375 -0.750 -1.750 -4.375 -11.125 -6.750 -8.750
   acd     ace     ade    bcd    bce    bde    cde   abcd   abce   abde   acde   bcde  abcde
 3.875  -0.375  -0.750 -1.750 -4.375 -10.125 -1.750 -4.750  9.875 -0.375 -0.750 -1.750 -4.375
```

```
> ##  PRESS ##
> x=model.matrix(mod)
> PRESS_res=summary(mod)$res/(1-hat(x))
> print(PRESS_res)
       (1)           a           b           c           d           e          ab          ac          ad          ae          bc          bd          be
 4.7692308  16.3076923  12.6153846  -0.1538462  22.9230769   5.2307692   4.0000000  -5.3846154  -0.1538462  -8.3076923   6.4615385   4.7692308  -0.4615385
        cd          ce          de         abc         abd         abe         acd         ace         ade         bcd         bce         bde         cde
-0.9230769  -2.1538462  -5.3846154 -13.6923077  -8.3076923 -10.7692308   4.7692308  -0.4615385  -0.9230769  -2.1538462  -5.3846154 -12.4615385  -2.1538462
      abcd        abce        abde        acde        bcde       abcde
-5.8461538  12.1538462  -0.4615385  -0.9230769  -2.1538462  -5.3846154
```

```
> ################   R^2 Prediction #######################
> PRESS
[1] 2020.355
> SS_T= sum(anova(mod)$"Sum Sq")
> pred.r.squared = 1 - PRESS/(SS_T)
> pred.r.squared
[1] 0.8966041
```

The value "2020.355" in the third image appears to be the value of the PRESS statistic, which stands for "Predicted Residual Sum of Squares." This statistic is used to evaluate the predictive performance of a regression model.
A high predicted R-squared value close to 1 generally suggests a good fit of the regression model to the data. An R-squared of 0.8966 is considered a relatively high value, indicating that the model explains about 90 percent of the variability in predicting new observations as compared to approximately 93 percent of the variability in the original data explained by the least-squares fit. (R-squared adjusted = 0.9323)
However, it's important to note that a high R-squared alone does not guarantee a reliable or useful model. Other factors, such as the significance of predictor variables, the assumptions of the regression model, and the presence of multicollinearity or other issues, should also be considered when evaluating the model's performance.
In summary, based on the provided information, the regression model appears to have a reasonably good fit to the data, with an R-squared value of 0.8966.This good fit of model considering the factors A , B and C and their interactions shows that these factors are important as they are able to explain the data properly.

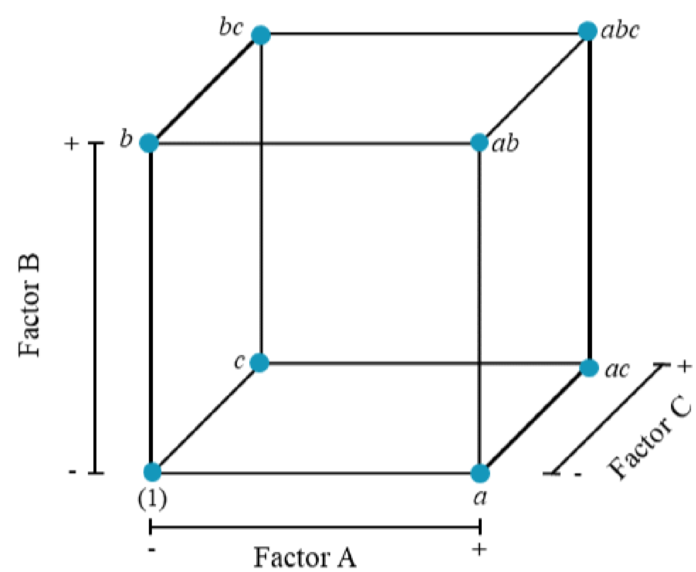
3.2 Stage 3:Analysis based on Design Projection

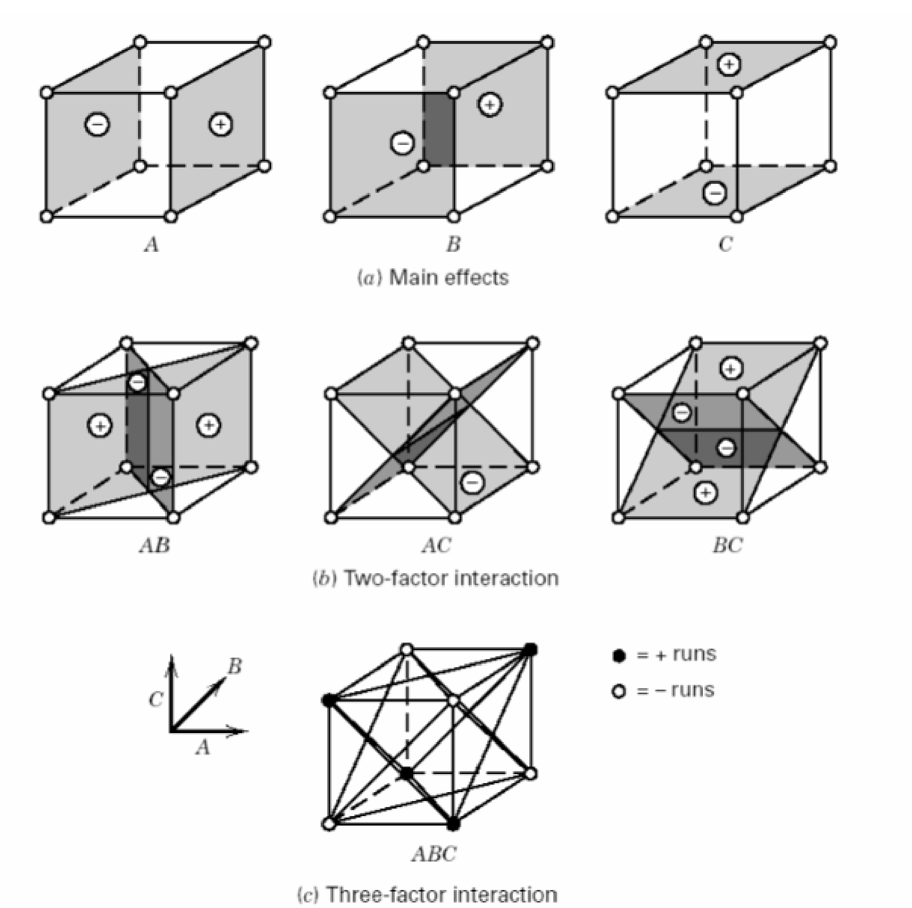
Design Projection

The concept of projecting an unreplicated factorial into a replicated factorial with fewer factors is very useful. In general, if we have a single replicate of a 2^k design, and if h $(h < k)$ factors are negligible and can be dropped, then the original data correspond to a full two-level factorial in the remaining $k - h$ factors with 2^h replicates.

Based on the initial screening, a 2^3 factorial design with four replications was constructed, focusing solely on factors A, B, and C. This design projection allowed for a more in-depth investigation of the significant factors while reducing the complexity of the analysis.

Geometric view of 2^3 design with Four Replications:





Theory of 2^3 Factorial Design: A 2^3 factorial design is a type of experimental design used in factorial experiments where three factors are each set at two levels. This design allows for the investigation of main effects as well as two-way interactions between factors.

Related Formulas:

1. Total Number of Experimental Runs:

$$N = 2^3 = 8$$

2. Number of Factors (k): 3
3. Number of Levels per Factor: 2
4. Total Degrees of Freedom (df_T):

$$df_T = N - 1 = 7$$

5. Degrees of Freedom for Each Factor (df_F):

$$df_F = 1$$

2^3Effects from Design Matrix

I:

$$I = \frac{\text{ContrastI}}{8n} = \frac{1}{8n}[(1) + a + b + ab + c + ac + bc + abc]$$

A:

$$A = \frac{\text{ContrastA}}{4n} = \frac{1}{4n}[-(1) + a - b + ab - c + ac - bc + abc]$$

B:

$$B = \frac{\text{ContrastB}}{4n} = \frac{1}{4n}[-(1) - a + b + ab - c - ac + bc + abc]$$

C:

$$C = \frac{\text{ContrastC}}{4n} = \frac{1}{4n}[-(1) - a - b - ab + c + ac + bc + abc]$$

AB:

$$AB = \frac{\text{ContrastAB}}{4n} = \frac{1}{4n}[(1) - a - b + ab + c - ac - bc + abc]$$

AC:

$$AC = \frac{\text{ContrastAC}}{4n} = \frac{1}{4n}[(1) - a + b - ab - c + ac - bc + abc]$$

BC:

$$BC = \frac{\text{ContrastBC}}{4n} = \frac{1}{4n}[(1) + a - b - ab - c - ac + bc + abc]$$

ABC:

$$ABC = \frac{\text{ContrastABC}}{4n} = \frac{1}{4n}[-(1) + a + b - ab + c - ac - bc + abc]$$

ANOVA (Analysis of Variance):

The ANOVA table for a 2^3 factorial design typically includes the following sources of variation:

1. Factor A
2. Factor B
3. Factor C
4. Interaction AB
5. Interaction AC
6. Interaction BC
7. Error (Residual)

Hypotheses:

The hypotheses tested in a 2^3 factorial design are as follows:

Main Effects:

$$H_0 : \text{There is no effect of factor A (or B, C)}$$

$$H_1 : \text{There is an effect of factor A (or B, C)}$$

Interaction Effects:

$$H_0 : \text{There is no interaction between factors A and B (or A and C, B and C)}$$

$$H_1 : \text{There is an interaction between factors A and B (or A and C, B and C)}$$

Mathematical Formulas:
Main Effects:

$$\text{Effect of A} = \frac{Y_{\text{high}} + Y_{\text{low}}}{2}$$

Interaction Effects:

$$\text{Effect of AB} = \frac{(Y_{\text{AB}} - Y_{\text{A}} - Y_{\text{B}})}{2}$$

Hypotheses:
Main Effects:

$$H_0 : \beta_A = 0 \quad \text{(No effect of factor A)}$$

$$H_1 : \beta_A \neq 0 \quad \text{(Effect of factor A)}$$

Interaction Effects:

$$H_0 : \beta_{AB} = 0 \quad \text{(No interaction between factors A and B)}$$

$$H_1 : \beta_{AB} \neq 0 \quad \text{(Interaction between factors A and B)}$$

The linear model is given by:

$$y_{ijkl} = \mu + \tau_i + \tau_i + \beta_j + \gamma_k + (\tau\beta)_{ij} + (\tau\gamma)_{ik} + (\beta\gamma)_{jk} + (\tau\beta\gamma)_{ijk} + \varepsilon_{ijkl}, \quad i, j, k = 1, 2, \quad l = 1, 2, \ldots, n,$$

where:

- y_{ijkl} is the response when the factor A is in the i-th level and the factor B is in the j-th level, factor C is in the k-th level, for the l-th replicate.
- μ is the overall mean effect.
- τ_i is the effect of the i-th level of the factor A.
- β_j is the effect of the j-th level of the factor B.
- γ_k is the effect of the k-th level of the factor C.
- $(\tau\beta)_{ij}$ is the effect of the interaction between i-th level of A and j-th level of B.
- $(\tau\gamma)_{ik}$ is the effect of the interaction between i-th level of A and k-th level of C.
- $(\beta\gamma)_{jk}$ is the effect of the interaction between j-th level of B and k-th level of C.
- $(\tau\beta\gamma)_{ijk}$ is the effect of the interaction between i-th level of A, j-th level of B, and k-th level of C.
- ε_{ijkl} is the random error.

Hypotheses Testing:
Testing Problem 1:

$$H_{10} : \tau_1 = \tau_2 = 0$$

$$H_{11} : \tau_i \neq 0, \text{ for at least one } i$$

Testing Problem 2:

$$H_{20} : \beta_1 = \beta_2 = 0$$

$$H_{21} : \beta_j \neq 0, \text{ for at least one } j$$

Testing Problem 3:

$$H_{30} : \gamma_1 = \gamma_2 = 0$$

$$H_{31} : \gamma_k \neq 0, \text{ for at least one } k$$

Testing Problem 4:

$$H_{40} : (\tau\beta)_{ij} = 0 \text{ for all } i, j$$

$$H_{41} : \text{At least one } (\tau\beta)_{ij} \neq 0$$

Testing Problem 5:

$$H_{50} : (\tau\gamma)_{ik} = 0 \text{ for all } i, k$$

$$H_{51} : \text{At least one } (\tau\gamma)_{ik} \neq 0$$

Testing Problem 6:

$$H_{60} : (\beta\gamma)_{jk} = 0 \text{ for all } j, k$$
$$H_{61} : \text{At least one } (\beta\gamma)_{jk} \neq 0$$

Testing Problem 7:

$$H_{70} : (\tau\beta\gamma)_{ijk} = 0 \text{ for all } i, j, k$$
$$H_{71} : \text{At least one } (\tau\beta\gamma)_{ijk} \neq 0$$

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F Value
A	SS_A	1	MS_A	$F_A=MS_A/MS_E$
B	SS_B	1	MS_B	$F_B=MS_B/MS_E$
C	SS_C	1	MS_C	$F_C=MS_C/MS_E$
AB	SS_{AB}	1	MS_{AB}	$F_{AB}=MS_{AB}/MS_E$
AC	SS_{AC}	1	MS_{AC}	$F_{AC}=MS_{AC}/MS_E$
BC	SS_{BC}	1	MS_{BC}	$F_{BC}=MS_{BC}/MS_E$
ABC	SS_{ABC}	1	MS_{ABC}	$F_{ABC}=MS_{ABC}/MS_E$
Error	SS_E	8(n-1)		
Total	SS_T	8n-1		

R Code for Design Projection and its ANOVA:

```
#Design Projection
Fil= matrix(c(62, 58, 47, 48 , 38, 18, 18, 23, 82, 77, 63, 67, 23, 27, 27, 33,
              71, 52, 52, 52, 47, 42, 42, 42, 93, 88, 88, 88, 13, 13, 13, 13),byrow=T,ncol=4)
dimnames(Fil) = list(c("(1)","a","b","ab","c","ac","bc","abc"),
                     c("Rep1","Rep2","Rep3","Rep4"))
A= rep(c(-1,1),4)
B =rep(c(-1,-1,1,1),2)
C= c(rep(-1,4),rep(1,4))
Total = apply(Fil,1,sum)
data2=data.frame(A,B,C,Fil,Total)
data2
data3=data.frame(A,B,C,Fil)
data3
#design matrix for updated df
I=c(rep(1,8))
AB=A*B
AC=A*C
BC=B*C
ABC=A*B*C
dm2=cbind(I,A,B,AB,C,AC,BC,ABC,Total)
dm2

#anova for updated df
Ft= c(t(Fil))
Ft
Af= rep(as.factor(A),rep(2,8))
Bf= rep(as.factor(B),rep(2,8))
Cf= rep(as.factor(C),rep(2,8))
dm=data.frame(Af,Bf,Cf,Ft)
dm
an=aov(Ft ~ Af*Bf*Cf, data=dm)
summary(an)
```

Outputs:

	A	B	C	Rep1	Rep2	Rep3	Rep4	Total
(1)	-1	-1	-1	62	58	47	48	215
a	1	-1	-1	38	18	18	23	97
b	-1	1	-1	82	77	63	67	289
ab	1	1	-1	23	27	27	33	110
c	-1	-1	1	71	52	52	52	227
ac	1	-1	1	47	42	42	42	173
bc	-1	1	1	93	88	88	88	357
abc	1	1	1	13	13	13	13	52

Figure 2: Updated Dataframe

	I	A	B	AB	C	AC	BC	ABC	Total
(1)	1	-1	-1	1	-1	1	1	-1	215
a	1	1	-1	-1	-1	-1	1	1	97
b	1	-1	1	-1	-1	1	-1	1	289
ab	1	1	1	1	-1	-1	-1	-1	110
c	1	-1	-1	1	1	-1	-1	1	227
ac	1	1	-1	-1	1	1	-1	-1	173
bc	1	-1	1	-1	1	-1	1	-1	357
abc	1	1	1	1	1	1	1	1	52

Figure 3: Updated Design Matrix

```
            Df Sum Sq Mean Sq F value   Pr(>F)    
Af           1    242     242   2.511   0.1261    
Bf           1  13448   13448 139.538 1.73e-11 ***
Cf           1    288     288   2.988   0.0967 .  
Af:Bf        1    144     144   1.499   0.2327    
Af:Cf        1     50      50   0.519   0.4783    
Bf:Cf        1   3042    3042  31.564 8.77e-06 ***
Af:Bf:Cf     1     13      13   0.130   0.7219    
Residuals   24   2313      96                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 4: ANOVA

Figure 5: Design Projection Analysis

Results and Discussion

The ANOVA results revealed the following significant effects: Based on the ANOVA results, we can draw the following conclusions:

Main Effects:

1. Factor A (Listening to Music): The p-value for factor A is 0.1261, which is greater than the significance level of 0.05. This indicates that the main effect of listening to music is not statistically significant.

2. Factor B (Difficulty of Paragraph): The p-value for factor B is highly significant (1.73e-11), which is much lower than the significance level of 0.05. This means that the difficulty of the paragraph has a substantial impact on reading comprehension scores. Specifically, easier paragraphs will likely result in better reading comprehension than more difficult ones.

3. Factor C (Style of Writing): The p-value for factor C is 0.0097, which is lower than the significance level of 0.05. This indicates that the style of writing (formal or informal) has a statistically significant effect on reading comprehension scores. The practical implication is that the choice of writing style can influence an individual's ability to comprehend the material effectively.

Interaction Effects:

1. A:B (Listening to Music x Difficulty of Paragraph): The p-value for this interaction is 0.2327, which is greater than the significance level of 0.05. This suggests that the interaction between listening to music and the difficulty of the paragraph is not statistically significant.

2. A:C (Listening to Music x Style of Writing): The p-value for this interaction is 0.4783, which is greater than the significance level of 0.05. This indicates that the interaction between listening to music and the writing style is not statistically significant.

3. B:C (Difficulty of Paragraph x Style of Writing): The p-value for this interaction is highly significant (8.77e-06), which is much lower than the significance level of 0.05. This implies that the combined effect of paragraph difficulty and writing style on reading comprehension scores is statistically significant. **The practical implication is that the choice of writing style should be tailored to the difficulty level of the paragraph to optimize reading comprehension.**

Overall, the results suggest that the difficulty of the paragraph and the writing style are the most significant factors influencing reading comprehension scores. Additionally, the interaction between these two factors is also crucial, indicating that the **choice of writing style should be adjusted based on the difficulty level of the paragraph to promote effective reading comprehension.**

NOTE THAT: The main effects plot visually confirms the ANOVA results, highlighting the significant impact of paragraph difficulty and writing style on reading comprehension scores. It also suggests a potential negative effect of listening to music, although this effect was not statistically significant in the ANOVA analysis. Therefore, the ANOVA results and hn-plot do not perfectly align for all the hypotheses, but they do for our interested significant factors A,B, and C.

The interaction plot matrix confirms the ANOVA findings, highlighting the significant interaction between paragraph difficulty and writing style. It also visually depicts the lack of substantial interactions between other factor pairs (A: B and A:C). The practical implication is that optimizing the combination of paragraph difficulty and writing style is crucial for promoting effective reading comprehension.

Conclusion:

This study provides valuable insights into the factors that influence reading comprehension. By systematically investigating the effects of listening to music, difficulty of the paragraph, writing style, font size, and environmental conditions, we have identified the most critical factors impacting reading comprehension scores: paragraph difficulty and writing style. These findings can inform educational practices, workplace environments, and personal study habits to optimize effective reading and learning conditions.