**Appendix 1**

**R code for generating test\_score\_data**

# Set seed for reproducibility

set.seed(123)

# Number of students

n <- 1000

# Generate age, sex, hours studied, and tutor status

age <- round(runif(n, min = 20, max = 30)) # Uniform distribution for age (20 to 30)

sex <- sample(c("Male", "Female"), n, replace = TRUE, prob = c(0.5, 0.5)) # Randomly sample Male and Female

hours\_studied <- round(runif(n, min = 0, max = 100)) # Random uniform distribution for hours studied (0 to 100)

tutor <- sample(c("Yes", "No"), n, replace = TRUE, prob = c(0.3, 0.7)) # Randomly assign tutor status (30% Yes, 70% No)

# Function to generate test scores based on inputs

generate\_test\_score <- function(age, sex, hours\_studied, tutor) {

base\_score <- 30 # Base score (minimum score)

age\_effect <- (age - 20) \* 2 # Older students get slightly higher score (up to 20 points more)

sex\_effect <- ifelse(sex == "Female", 10, 0) # Females get +10 score advantage

study\_effect <- 0.015 \* hours\_studied^2 - 0.0001 \* hours\_studied^3 # Adjusted quadratic relationship

tutor\_effect <- ifelse(tutor == "Yes", 15, 0) # Students with tutor get +15 score advantage

# Calculate final test score

test\_score <- base\_score + age\_effect + sex\_effect + study\_effect + tutor\_effect

# Ensure test score is at least 30 and at most 100

test\_score <- pmax(test\_score, 30)

test\_score <- pmin(test\_score, 100)

return(test\_score)

}

# Generate test score using the function

test\_score <- generate\_test\_score(age, sex, hours\_studied, tutor)

# Create data frame

test\_score\_data <- data.frame(age = age, sex = sex, hours\_studied = hours\_studied, tutor = tutor, test\_score = test\_score)

# Print first few rows of the data frame

head(test\_score\_data)