Latest Dataset Cleaning

Julia

2024-07-01

```
# Load the tidyverse (including ggplot2) and janitor
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.4.4     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.0
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts
to become errors
```

```
library(janitor)
```

```
##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
# Upload data
new_post_dat <- read_tsv("new_posttest.txt") |>  janitor::clean_names()
```

```
## Rows: 21355 Columns: 34
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: "\t"
## chr  (14): Sample, Anon Student Id, Problem Hierarchy, Problem Name, Step Du...
## dbl  (15): Row, Problem View, Step Name, Incorrects, Hints, Corrects, Opport...
## lgl   (1): Predicted Error Rate (Unique-step)
## dttm  (4): Step Start Time, First Transaction Time, Correct Transaction Time...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
new_practice_dat <- read_tsv("new_practice.txt") |>  janitor::clean_names()
```

```
## Rows: 44810 Columns: 34
## ── Column specification ────────────────────────────────────────────────────
## Delimiter: "\t"
## chr  (14): Sample, Anon Student Id, Problem Hierarchy, Problem Name, Step Du...
## dbl  (15): Row, Problem View, Step Name, Incorrects, Hints, Corrects, Opport...
## lgl   (1): Predicted Error Rate (Unique-step)
## dttm  (4): Step Start Time, First Transaction Time, Correct Transaction Time...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
new_pre_dat <- read_tsv("new_pretest.txt") |>  janitor::clean_names()
```

```
## Rows: 22945 Columns: 34
## ── Column specification ────────────────────────────────────────────────────
## Delimiter: "\t"
## chr  (14): Sample, Anon Student Id, Problem Hierarchy, Problem Name, Step Du...
## dbl  (15): Row, Problem View, Step Name, Incorrects, Hints, Corrects, Opport...
## lgl   (1): Predicted Error Rate (Unique-step)
## dttm  (4): Step Start Time, First Transaction Time, Correct Transaction Time...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# merge pre-post-practice datasets
combined_prepost <- bind_rows(new_post_dat, new_practice_dat, new_pre_dat)
```

```
# separate the "Conditions" column into three different columns and then change the vari
ables to easier-to-understand terms
new_df <- combined_prepost |>
  separate(condition, into = c("Learning objective spacing", "Question variability", "To
pic spacing"), sep = "~~") |>
  mutate(
    `Learning objective spacing` = ifelse(`Learning objective spacing` == "High Learning
objective spacing (Learning objective spacing)", "High LO", 'Low LO'),
    `Question variability` = ifelse(`Question variability` == "High Question variability
(Question variability)", "High Q Variability", "Low Q Variability"),
    `Topic spacing` = ifelse(`Topic spacing` == "High Topic spacing (Topic spacing)", "H
igh Topic", "Low Topic")
  )
```

Check Spacing

need to calculate spacing separately for topic and LO because the same problem can be opportunity 3 for LO but opportunity 1 for topic

```
new_df_topic <- new_df |>
  filter(sample == "Practice") |>
  filter(`opportunity_topic` != 1) |>   # Exclude rows where Opportunity (Topic) equals 1
  arrange(anon_student_id, step_start_time) |>
  group_by(anon_student_id) |>
  mutate(
    PreviousStepEndTime_topic = lag(step_end_time), # show the previous step end time from the opportunity before on the most recent line - for display purposes (easy to read)
    SpacingTime_topic = difftime(step_start_time, PreviousStepEndTime_topic, units = "mins")) |>
  ungroup()

new_df_lo <- new_df |>
  filter(sample == "Practice") |>
  filter(`opportunity_lo` != 1) |>
  arrange(anon_student_id, step_start_time) |>
  group_by(anon_student_id) |>
  mutate(
    PreviousStepEndTime_lo = lag(step_end_time),
    SpacingTime_lo = difftime(step_start_time, PreviousStepEndTime_lo, units = "mins")
  ) |>
  ungroup()

write_csv(new_df_lo, "new_spacing_time.csv")
```

Remove all studentids with missing/more trials to only include those participants with all trials completed

```
# List of student IDs to remove
students_to_remove <- c(23925, 23937, 24559, 24652, 24653, 28859, 28903, 28906, 28915, 28925, 28938, 28956, 28968, 28970, 28982, 28983, 28986, 28994, 28999, 29047, 29048, 29047, 29048, 29049, 29050, 29053, 29054, 29057, 29058, 29059, 29060, 29069, 29071, 29074, 29075, 29078, 29081, 29083, 29084, 29088, 29089, 29090, 29091, 29097, 29099, 29128, 'Stu_138ca89cacc7b313af2fd0ec77dafbe6', 'Stu_86faf4453738b08f6b9bdfb293bec8b3', 'Stu_8e810aa779db12acf660d6a4a88f54cf', 'Stu_94193d5580a6b0dd9acd121dd058f7a0')

# Remove students with missing data
filtered_new_df <- new_df |>
  filter(!anon_student_id %in% students_to_remove)
```

Plots

Question 1:

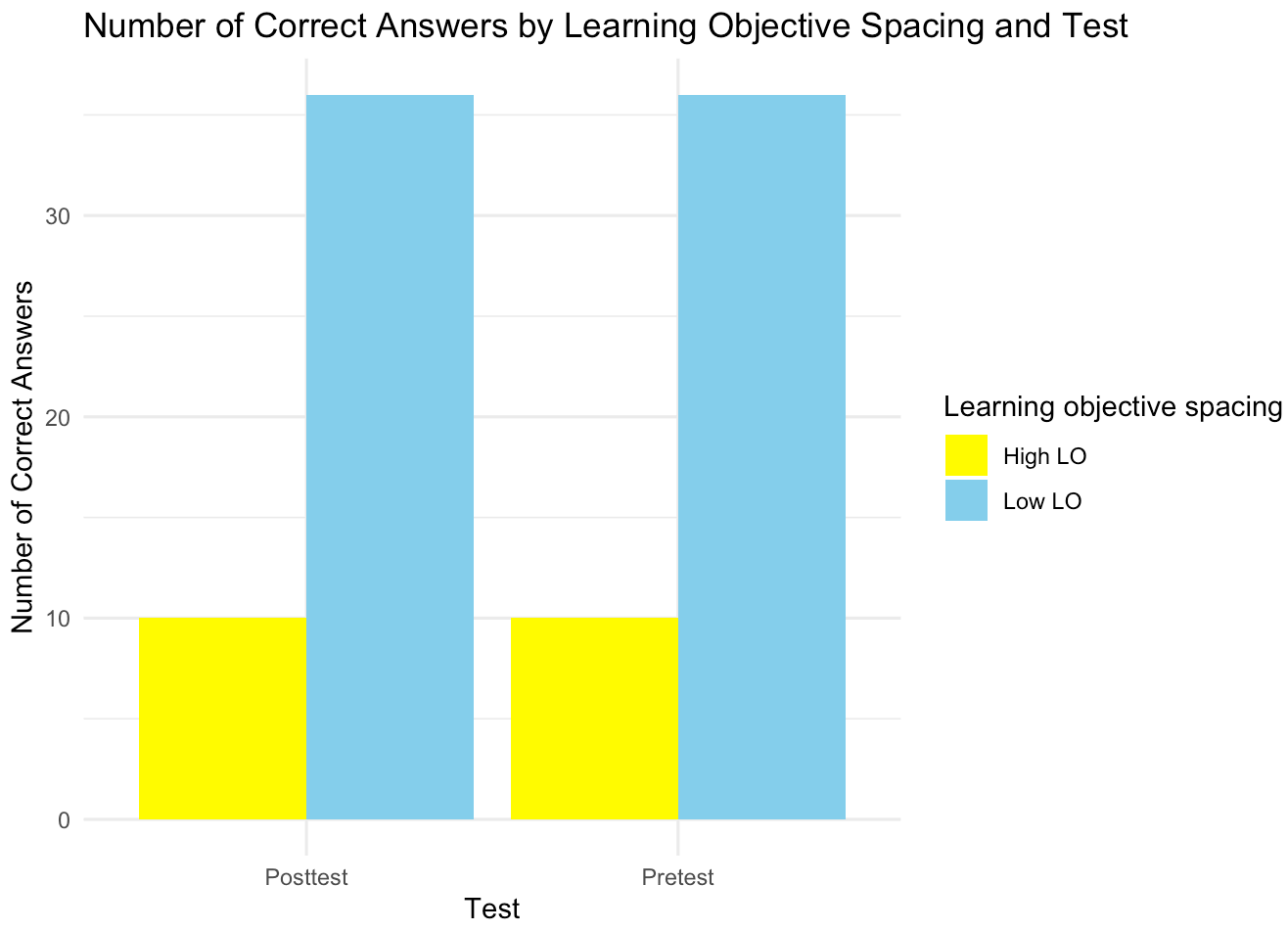
Does learning object spacing (High vs. Low) affect the improvement in the number of correct answers from pretest to posttest?

```
new_df_summary_lo <- filtered_new_df |>
  filter(sample %in% c("Pretest", "Posttest")) |>
  group_by(anon_student_id, `Learning objective spacing`, sample) |>
  summarize(correct_answers = sum(corrects)) |>
  ungroup()
```

```
## `summarise()` has grouped output by 'anon_student_id', 'Learning objective
## spacing'. You can override using the `.groups` argument.
```

```
# ggplot(new_df_summary, aes(x = sample, y = correct_answers, group = anon_student_id, c
olor = `Learning objective spacing`)) +
#   # geom_line() +
#   geom_point() +
#   labs(title = "Improvement from Pretest to Posttest by Condition", x = "Test", y = "N
umber of Correct Answers")
```

```
ggplot(new_df_summary_lo, aes(x = sample, y = correct_answers, fill = `Learning objectiv
e spacing`)) +
  geom_bar(stat = "identity", position = position_dodge()) +
  labs(title = "Number of Correct Answers by Learning Objective Spacing and Test", x =
"Test", y = "Number of Correct Answers") +
  scale_fill_manual(values = c("yellow", "skyblue")) +
  theme_minimal()
```



High LO Spacing: This suggests that a significant number of students in the “High Learning objective spacing” condition have lower scores (number of correct answers). This could indicate that many students struggled or did not perform well on the test.

Low LO Spacing: This suggests that the scores of students in the “Low Learning objective spacing” condition are more varied. There is no single common score that many students achieved, indicating greater variability in performance

---

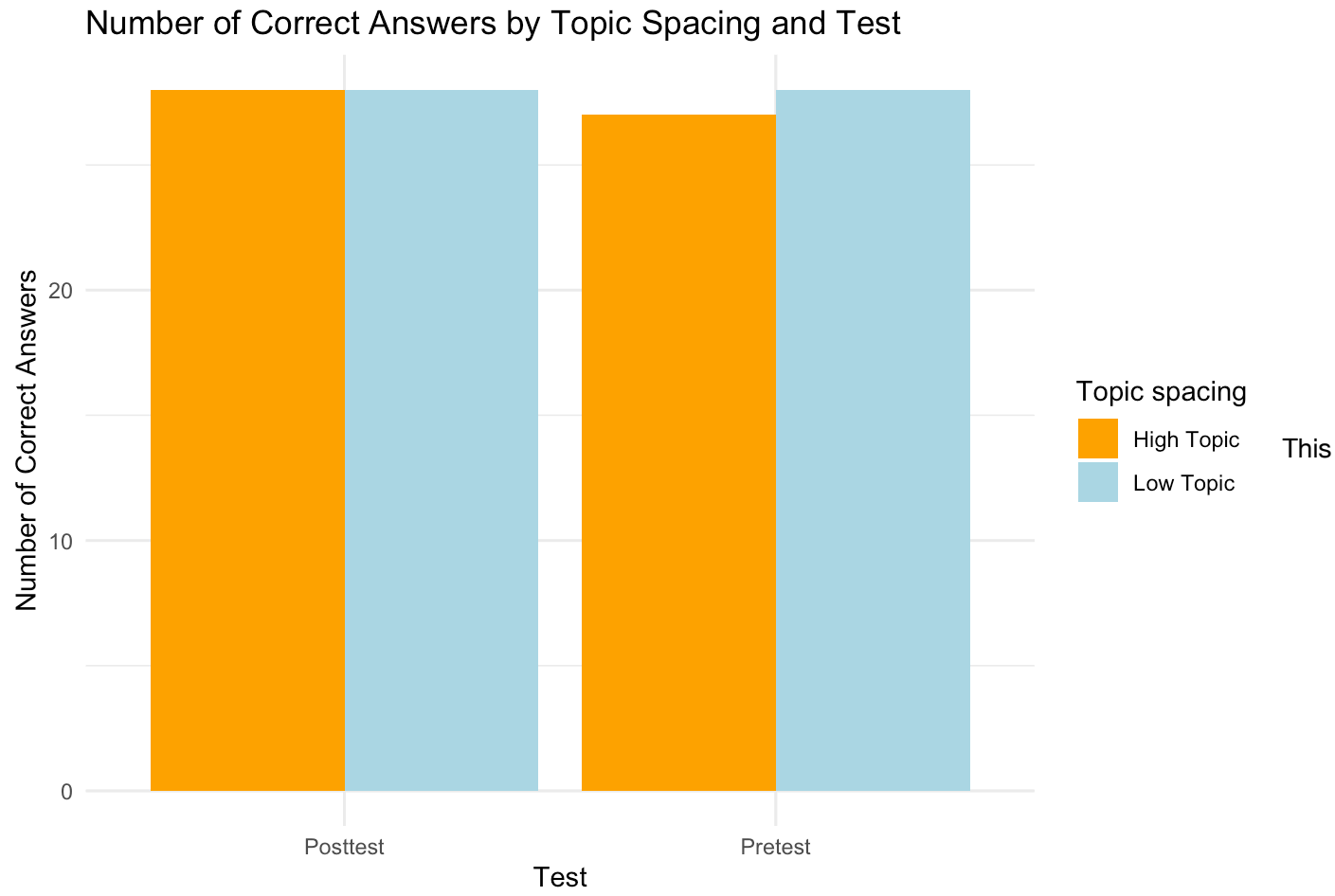
Question 2:

Does topic spacing (High vs. Low) affect the improvement in the number of correct answers from pretest to posttest?

```
new_df_summary_topic <- filtered_new_df |>
  filter(sample %in% c("Pretest", "Posttest")) |>
  group_by(anon_student_id, `Topic spacing`, sample) |>
  summarize(correct_answers = sum(corrects)) |>
  ungroup()
```

```
## `summarise()` has grouped output by 'anon_student_id', 'Topic spacing'. You can
## override using the `.groups` argument.
```

```
ggplot(new_df_summary_topic, aes(x = sample, y = correct_answers, fill = `Topic spacing
`)) +
  geom_bar(stat = "identity", position = position_dodge()) +
  labs(title = "Number of Correct Answers by Topic Spacing and Test", x = "Test", y = "N
umber of Correct Answers") +
  scale_fill_manual(values = c("orange", "lightblue")) +
  theme_minimal()
```



plot suggests that the number of correct answers in the “high topic” and “low topic” spacing conditions barely vary from pre-test to post-test.

Question 3:

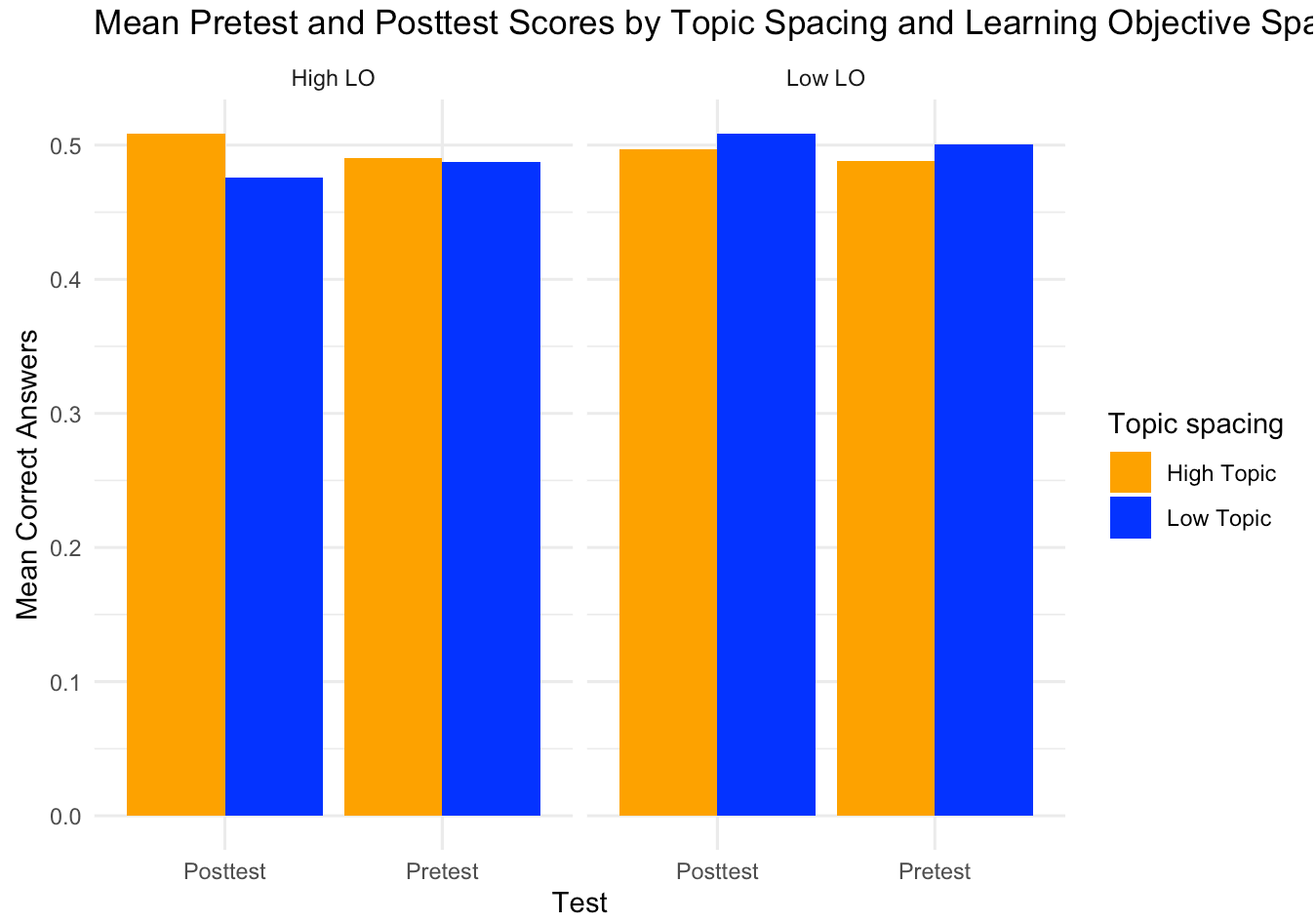
Is there a difference in the mean pretest and posttest scores between different combinations of topic spacing and learning objective spacing conditions?

Description: This plot will show the mean pretest and posttest scores for each combination of “topic spacing” and “learning objective spacing”

```
# Summarize the data
df_mean_scores <- filtered_new_df |>
  filter(sample %in% c("Pretest", "Posttest")) |>
  group_by(`Topic spacing`, `Learning objective spacing`, sample) |>
  summarize(mean_correct = mean(corrects)) |>
  ungroup()
```

```
## `summarise()` has grouped output by 'Topic spacing', 'Learning objective
## spacing'. You can override using the `.groups` argument.
```

```
# Create the bar plot
ggplot(df_mean_scores, aes(x = sample, y = mean_correct, fill = `Topic spacing`)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_wrap(~ `Learning objective spacing`) +
  labs(title = "Mean Pretest and Posttest Scores by Topic Spacing and Learning Objective
Spacing", x = "Test", y = "Mean Correct Answers") +
  scale_fill_manual(values = c("orange", "blue")) +
  theme_minimal()
```



Question 4:

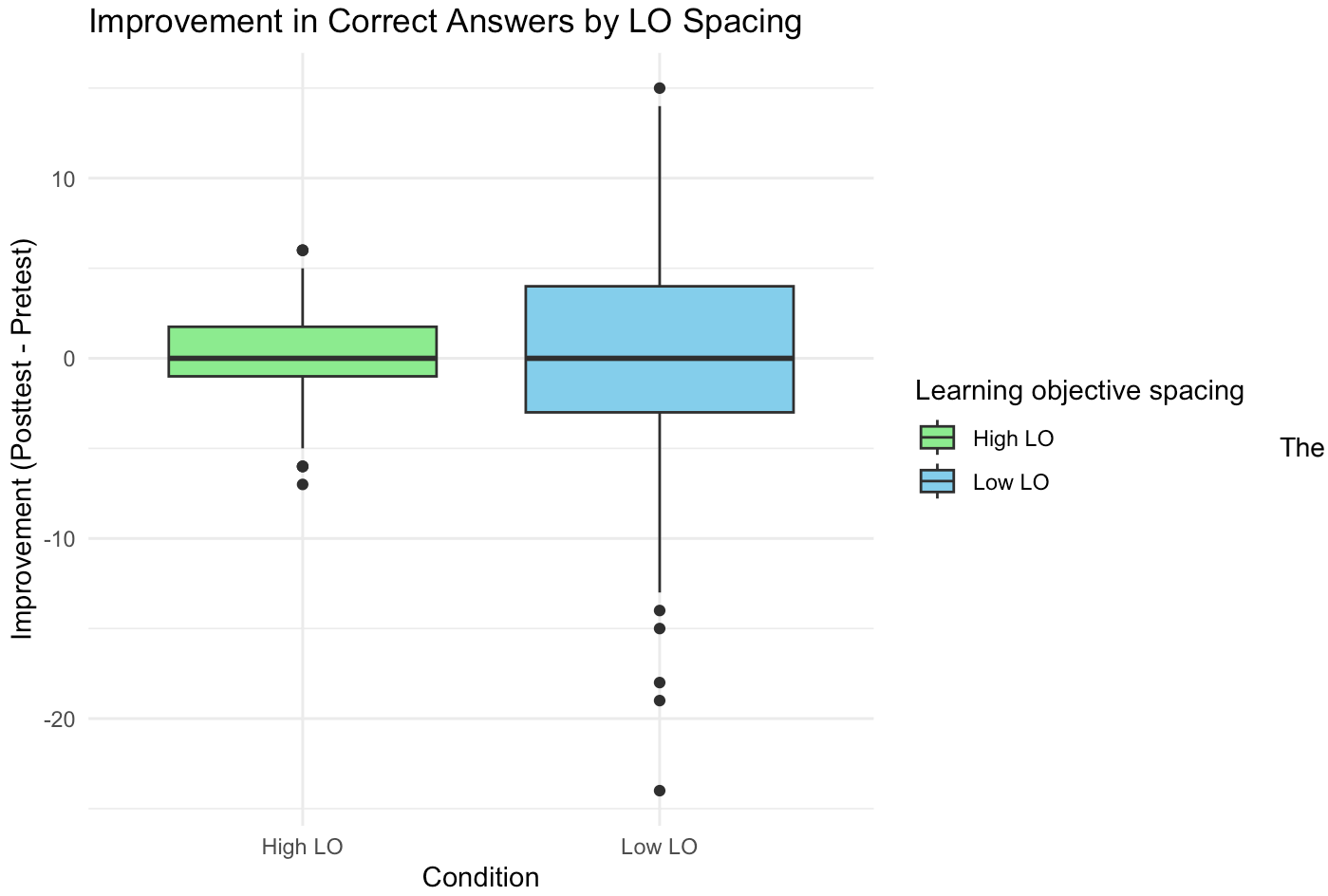
Is there a difference in the variability of improvement in correct answers between conditions (specifically, Learning Obejective Spacing)?

Description of the plot below: Boxplot of the difference in the number of correct answers between pretest and posttest by condition (LO spacing).

```
df_improvement <- new_df_summary_lo |>
  spread(key = sample, value = correct_answers) |> # reshape data from long to wide form
at
  mutate(improvement = Posttest - Pretest) # show the change in the number of correct an
swers from the pretest to the posttest

ggplot(df_improvement, aes(x = `Learning objective spacing`, y = improvement, fill = `Le
arning objective spacing`)) +
  geom_boxplot() +
  labs(title = "Improvement in Correct Answers by LO Spacing", x = "Condition", y = "Imp
rovement (Posttest - Pretest)") +
  scale_fill_manual(values = c("lightgreen", "skyblue")) +
  theme_minimal()
```

```
## Warning: Removed 46 rows containing non-finite values (`stat_boxplot()`).
```



The short and squished boxplot for the “High LO spacing” condition suggests that the majority of the data points are very close to each other, indicating low variability in the improvement scores. The “taller” boxplot for “Low LO spacing” indicates higher variability in the improvement scores; students in this condition had a wider range of improvements generally compared to the “High LO” boxplot.

Question 5:

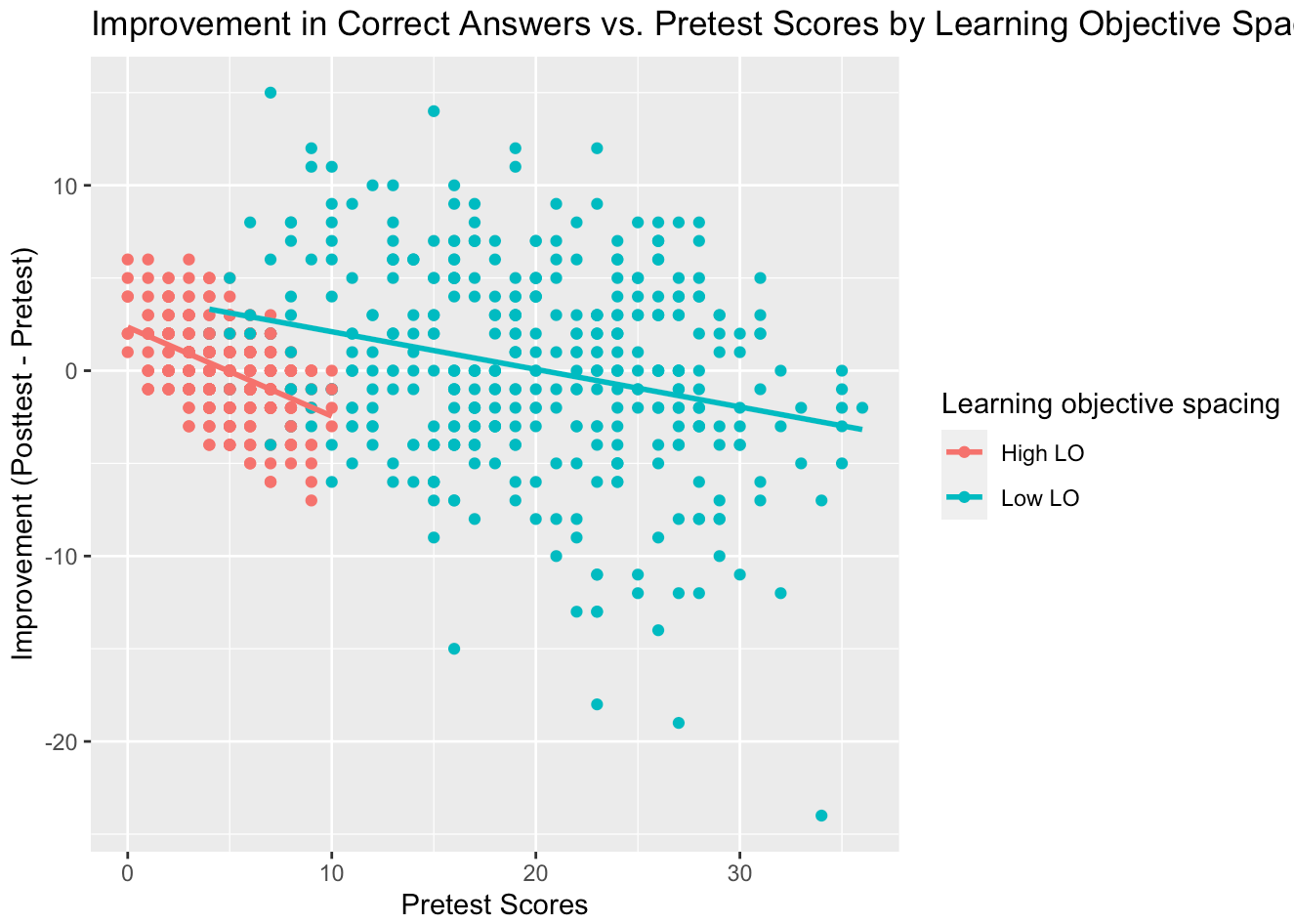
How does the improvement in test scores differ between students with different levels of prior knowledge (e.g., based on pretest scores) across conditions?

```
ggplot(df_improvement, aes(x = Pretest, y = improvement, color = `Learning objective spa
cing`)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Improvement in Correct Answers vs. Pretest Scores by Learning Objective
Spacing", x = "Pretest Scores", y = "Improvement (Posttest - Pretest)")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 46 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 46 rows containing missing values (`geom_point()`).
```



Hypothesis: Students with higher pretest scores will show less improvement compared to students with lower pretest scores, regardless of the learning objective spacing condition.

Question 6:

Is there a difference in the variability of improvement in correct answers between conditions (specifically, Topic Spacing)?

Description of the plot below: Boxplot of the difference in the number of correct answers between pretest and posttest by condition (Topic spacing).

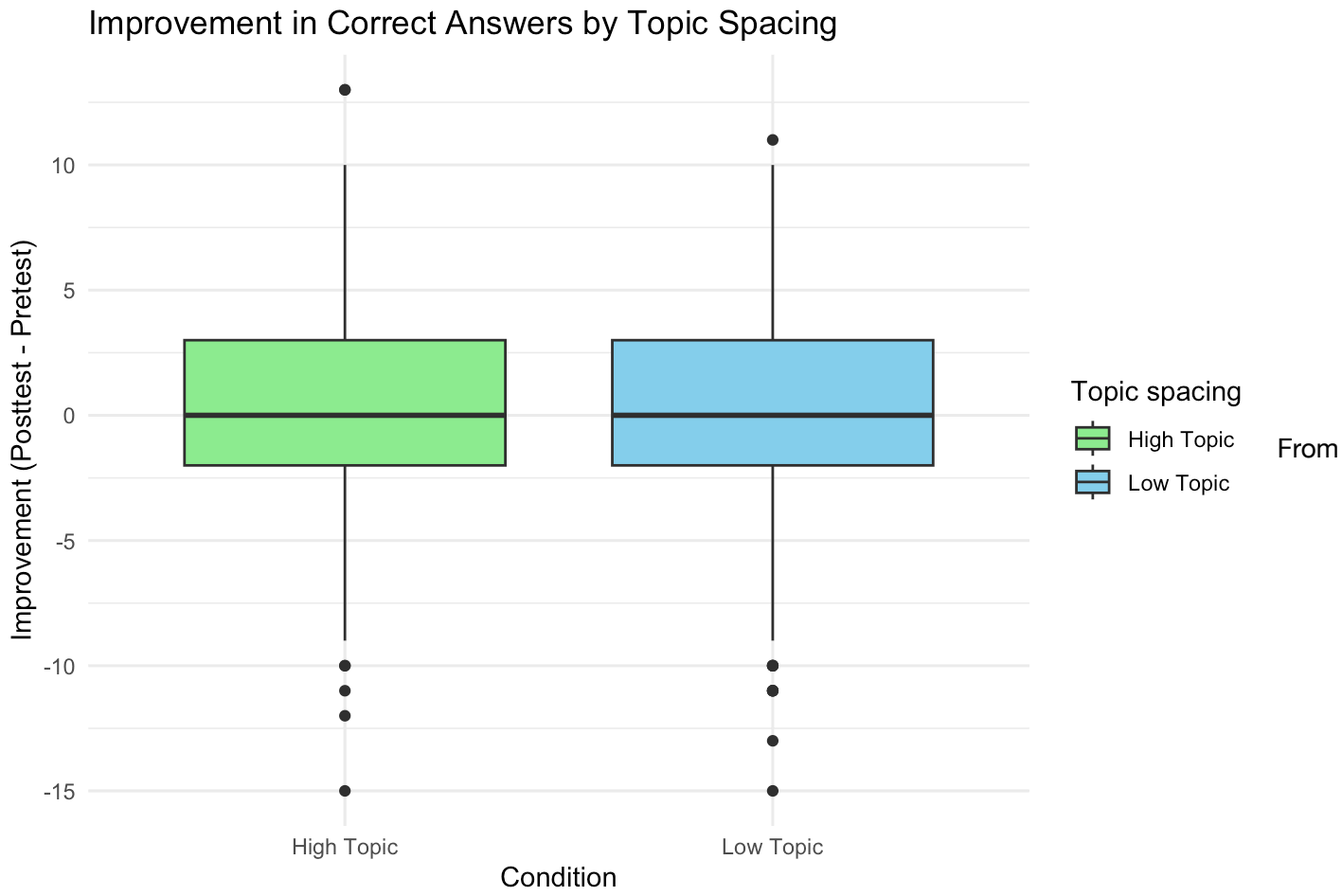
```
new_df_summary_topic <- filtered_new_df |>
  filter(sample %in% c("Pretest", "Posttest")) |>
  group_by(anon_student_id, `Topic spacing`, sample) |>
  summarize(correct_answers = sum(corrects)) |>
  ungroup()
```

```
## `summarise()` has grouped output by 'anon_student_id', 'Topic spacing'. You can
## override using the `.groups` argument.
```

```
df_improvement_topic <- new_df_summary_topic |>
  spread(key = sample, value = correct_answers) |> # reshape data from long to wide form
at
  mutate(improvement = Posttest - Pretest) # show the change in the number of correct an
swers from the pretest to the posttest

ggplot(df_improvement_topic, aes(x = `Topic spacing`, y = improvement, fill = `Topic spa
cing`)) +
  geom_boxplot() +
  labs(title = "Improvement in Correct Answers by Topic Spacing", x = "Condition", y =
"Improvement (Posttest - Pretest)") +
  scale_fill_manual(values = c("lightgreen", "skyblue")) +
  theme_minimal()
```

```
## Warning: Removed 48 rows containing non-finite values (`stat_boxplot()`).
```



From the plot, it looks like there is low variability in the improvement with the topic spacing condition.