HSLS Analysis

```
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(gridExtra)
```

```
Attaching package: 'gridExtra'

The following object is masked from 'package:dplyr':

    combine
```

```
library(psych)
```

```
Warning: package 'psych' was built under R version 4.3.3
```

```
Attaching package: 'psych'

The following objects are masked from 'package:ggplot2':

    %+%, alpha
```

```
library(tidyr)
```

```
Warning: package 'tidyr' was built under R version 4.3.2
```

```
dathsls <- haven::read_sav("HSLS6.11.21.sav")

hsls <- dathsls

hsls <- hsls %>%
  rename(stu_id = STU_ID, # change column names
         sch_id = SCH_ID,
         # excellentTests = S1MTESTS,
         # understandTexts = S1MTEXTBOOK,
         # masterSkills = S1MSKILLS,
         # excellentAssign = S1MASSEXCL,
         sex = X1SEX,
         race = X1RACE,
         hispanic = X1HISPANIC,
         white = X1WHITE,
         black = X1BLACK,
         asian = X1ASIAN,
         pacificIsland = X1PACISLE,
         SES = X1SES,
         hsls_w_cohort_g9 = W1STUDENT,
         hsls_w_cohort_g12 = W2STUDENT)
```

Give only the names that start with S1 or S2

S1 = 9th Grade 2009 S2 = 11th Grade 2012

```
filtered_S1 <- names(hsls)[grep("^S1", names(hsls))]
filtered_S1
```

```
 [1] "S1MPERSON1"   "S1MPERSON2"   "S1MENJOYS"    "S1MENJOYING"  "S1MWASTE"
 [6] "S1MBORING"    "S1MUSELIFE"   "S1MUSECLG"    "S1MUSEJOB"    "S1MTESTS"
[11] "S1MTEXTBOOK"  "S1MSKILLS"    "S1MASSEXCL"   "S1SPERSON1"   "S1SPERSON2"
[16] "S1SENJOYS"    "S1SENJOYING"  "S1SWASTE"     "S1SBORING"    "S1SUSELIFE"
[21] "S1SUSECLG"    "S1SUSEJOB"    "S1STESTS"     "S1STEXTBOOK"  "S1SSKILLS"
[26] "S1SASSEXCL"   "S1SAFE"       "S1PROUD"      "S1TALKPROB"   "S1SCHWASTE"
[31] "S1GOODGRADES" "S1NOHWDN"     "S1NOPAPER"    "S1NOBOOKS"    "S1LATE"
[36] "S1FAVSUBJ"    "S1LEASTSUBJ"
```

```
filtered_s2 <- names(hsls)[grep("^S2", names(hsls))]
filtered_s2
```

```
 [1] "S2FAVSUBJ"     "S2MENJOYS"     "S2MATTENTION"  "S2MONTIME"
 [5] "S2MSTOPTRYING" "S2MGETBY"      "S2MENJOYING"   "S2MTEXTBOOK"
 [9] "S2MWASTE"      "S2MSKILLS"     "S2MTESTS"      "S2MBORING"
[13] "S2MASSEXCL"    "S2SENJOYS"     "S2SATTENTION"  "S2SONTIME"
[17] "S2SSTOPTRYING" "S2SGETBY"      "S2SENJOYING"   "S2STEXTBOOK"
[21] "S2SWASTE"      "S2SSKILLS"     "S2STESTS"      "S2SBORING"
[25] "S2SASSEXCL"    "S2MPERSON1"    "S2MPERSON2"    "S2MUSELIFE"
[29] "S2MUSECLG"     "S2MUSEJOB"     "S2SPERSON1"    "S2SPERSON2"
[33] "S2SUSELIFE"    "S2SUSECLG"     "S2SUSEJOB"     "S2LATESCH"
[37] "S2ABSENT"      "S2WOHWDN"      "S2WOPAPER"     "S2WOBOOKS"
[41] "S2SKIPCLASS"   "S2INSCHSUSP"
```

Create subset of dataset with only math efficacy items

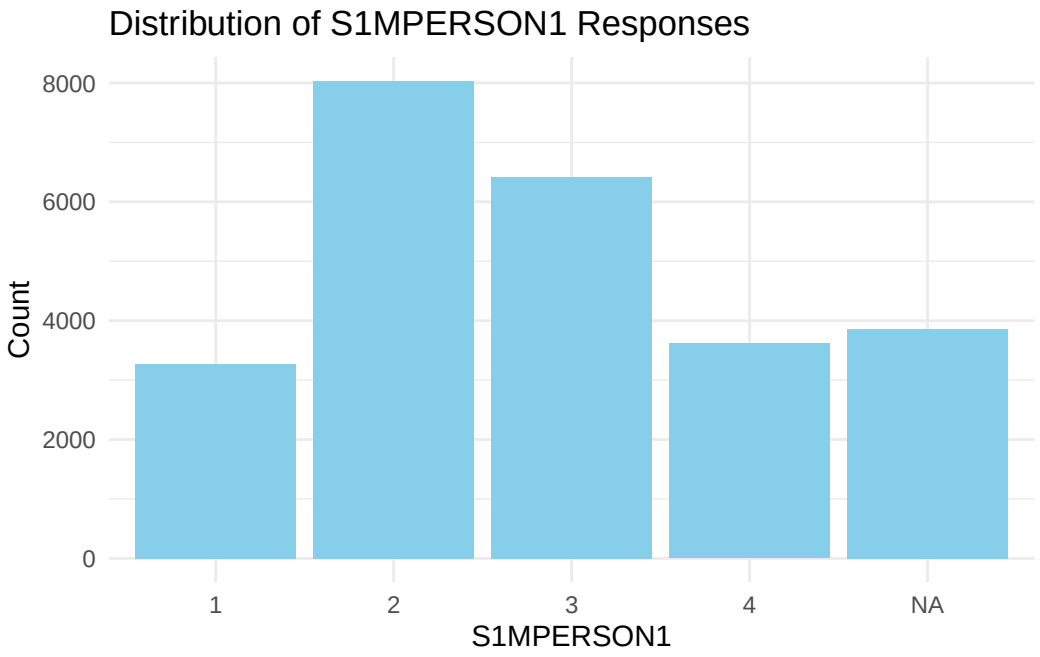
```
names(hsls)[grep("^S1M|^S2M", names(hsls))]
```

```
 [1] "S1MPERSON1"    "S1MPERSON2"    "S1MENJOYS"     "S1MENJOYING"
 [5] "S1MWASTE"      "S1MBORING"     "S1MUSELIFE"    "S1MUSECLG"
 [9] "S1MUSEJOB"     "S1MTESTS"      "S1MTEXTBOOK"   "S1MSKILLS"
[13] "S1MASSEXCL"    "S2MENJOYS"     "S2MATTENTION"  "S2MONTIME"
[17] "S2MSTOPTRYING" "S2MGETBY"      "S2MENJOYING"   "S2MTEXTBOOK"
[21] "S2MWASTE"      "S2MSKILLS"     "S2MTESTS"      "S2MBORING"
[25] "S2MASSEXCL"    "S2MPERSON1"    "S2MPERSON2"    "S2MUSELIFE"
[29] "S2MUSECLG"     "S2MUSEJOB"
```

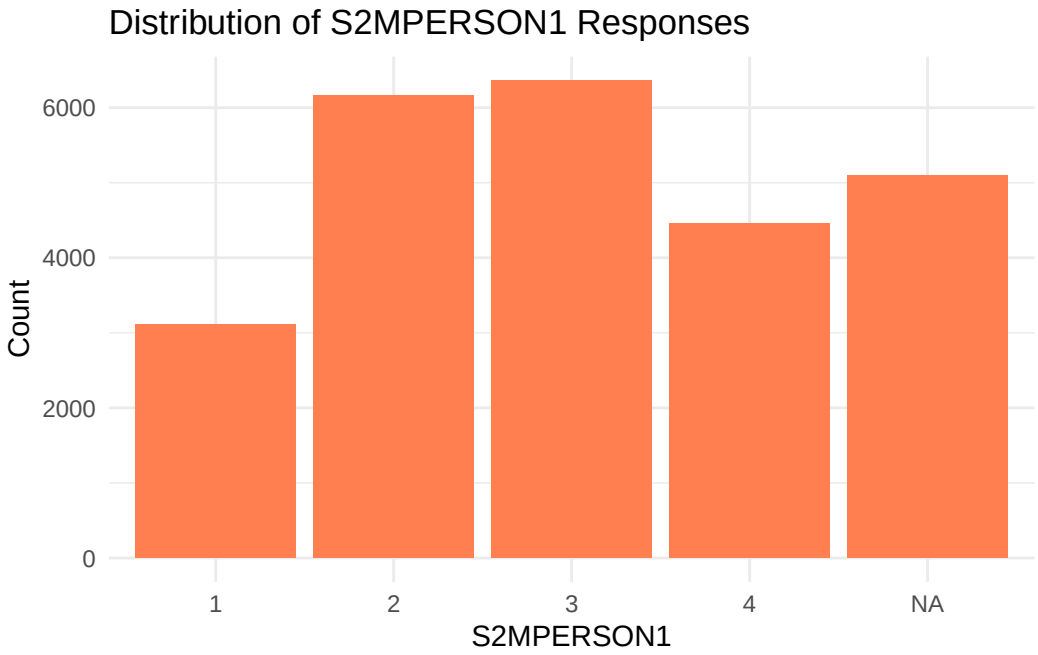
```r
math_eff <- hsls[, grep("^S1M|^S2M", names(hsls))]

# Ensures that the dataset was created successfully, difference should be 0
setdiff(names(hsls)[grep("^S1M | $S2M", names(hsls))], names(math_eff))
```

```
character(0)
```

```r
# 1 Strongly agree
# 2 Agree
# 3 Disagree
# 4 Strongly Disagree

ggplot(math_eff, aes(x = factor(S1MPERSON1))) +
  geom_bar(fill = "skyblue") +
  labs(title = "Distribution of S1MPERSON1 Responses",
       x = "S1MPERSON1",
       y = "Count") +
  theme_minimal()
```



```
ggplot(hsls, aes(x = factor(S2MPERSON1))) +
  geom_bar(fill = "coral") +
  labs(title = "Distribution of S2MPERSON1 Responses",
       x = "S2MPERSON1",
       y = "Count") +
  theme_minimal()
```

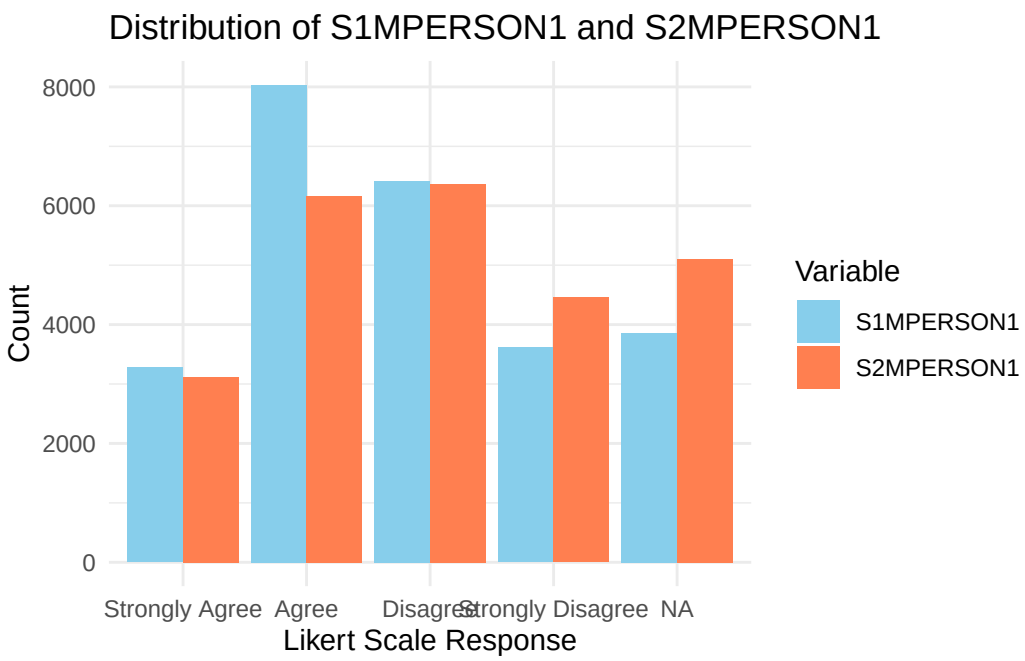



```
long_SMPERSON1 <- gather(math_eff, key = "variable", value = "value", S1MPERSON1, S2MPERSO
```

```
Warning: attributes are not identical across measure variables; they will be
dropped
```

```
# Create a grouped bar plot
ggplot(long_SMPERSON1, aes(x = factor(value), fill = variable)) +
  geom_bar(position = "dodge") +
  labs(title = "Distribution of S1MPERSON1 and S2MPERSON1",
       x = "Likert Scale Response",
```

```
       y = "Count",
       fill = "Variable") +
  theme_minimal() +
  scale_fill_manual(values = c("skyblue", "coral")) +
  scale_x_discrete(labels = c("Strongly Agree", "Agree", "Disagree", "Strongly Disagree",
```

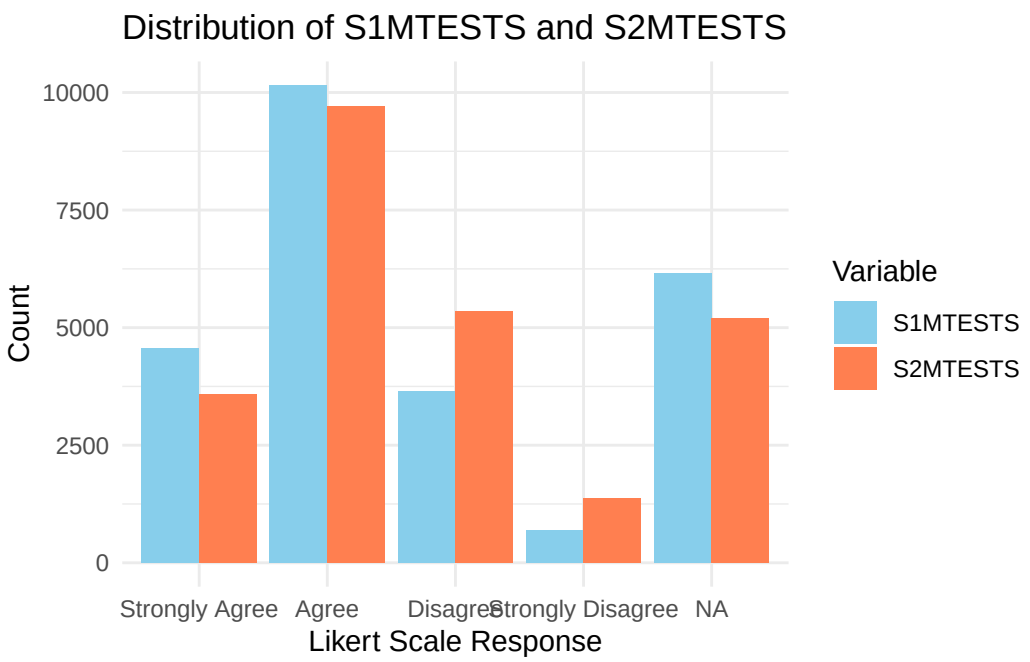



```
long_SMTESTS <- gather(math_eff, key = "variable", value = "value", S1MTESTS, S2MTESTS)
```

```
Warning: attributes are not identical across measure variables; they will be
dropped
```

```
# Create a grouped bar plot
ggplot(long_SMTESTS, aes(x = factor(value), fill = variable)) +
  geom_bar(position = "dodge") +
  labs(title = "Distribution of S1MTESTS and S2MTESTS",
       x = "Likert Scale Response",
       y = "Count",
       fill = "Variable") +
  theme_minimal() +
  scale_fill_manual(values = c("skyblue", "coral")) +
```

```
  scale_x_discrete(labels = c("Strongly Agree", "Agree", "Disagree", "Strongly Disagree",
```



Creating a function to automatically create graph of desired variables

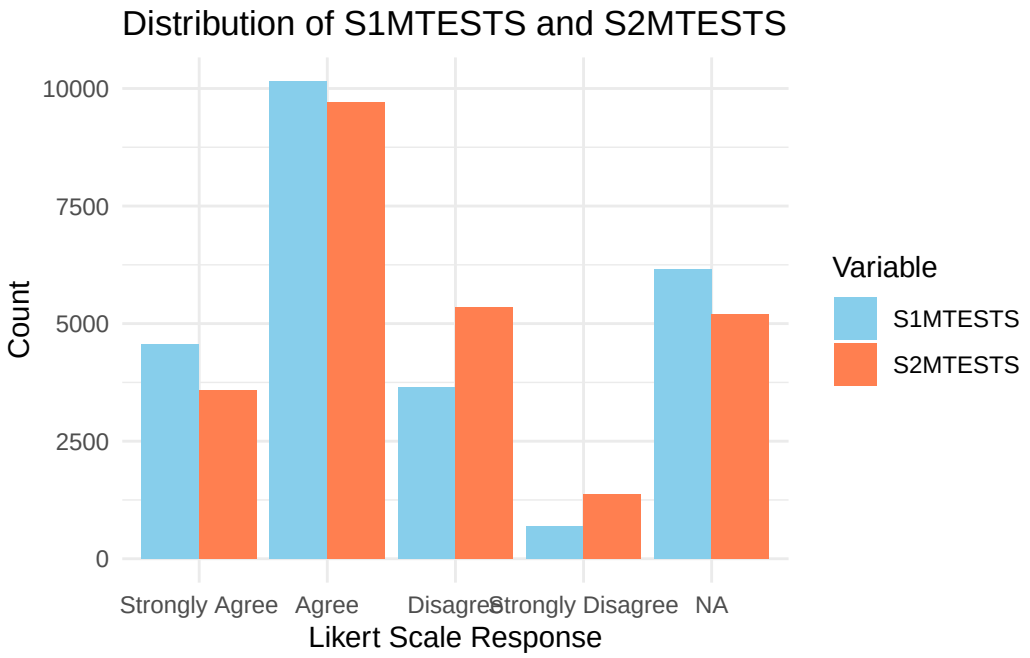
```
create_grouped_bar_plot <- function(data, missing_word) {
  # Dynamically gather the columns based on the missing word
  long_data <- gather(data, key = "variable", value = "value",
                      !!paste0("S1M", missing_word), !!paste0("S2M", missing_word))

  # Create the grouped bar plot
  ggplot(long_data, aes(x = factor(value), fill = variable)) +
    geom_bar(position = "dodge") +
    labs(title = paste("Distribution of S1M", missing_word, " and S2M", missing_word, sep
         x = "Likert Scale Response",
         y = "Count",
         fill = "Variable") +
    theme_minimal() +
    scale_fill_manual(values = c("skyblue", "coral")) +
    scale_x_discrete(labels = c("Strongly Agree", "Agree", "Disagree", "Strongly Disagree"
}
```

Tests

Teen (9th / 11th grader) confident can do excellent job on (fall 2009 / spring 2012) math tests

```
create_grouped_bar_plot(math_eff, "TESTS")
```

```
Warning: attributes are not identical across measure variables; they will be
dropped
```

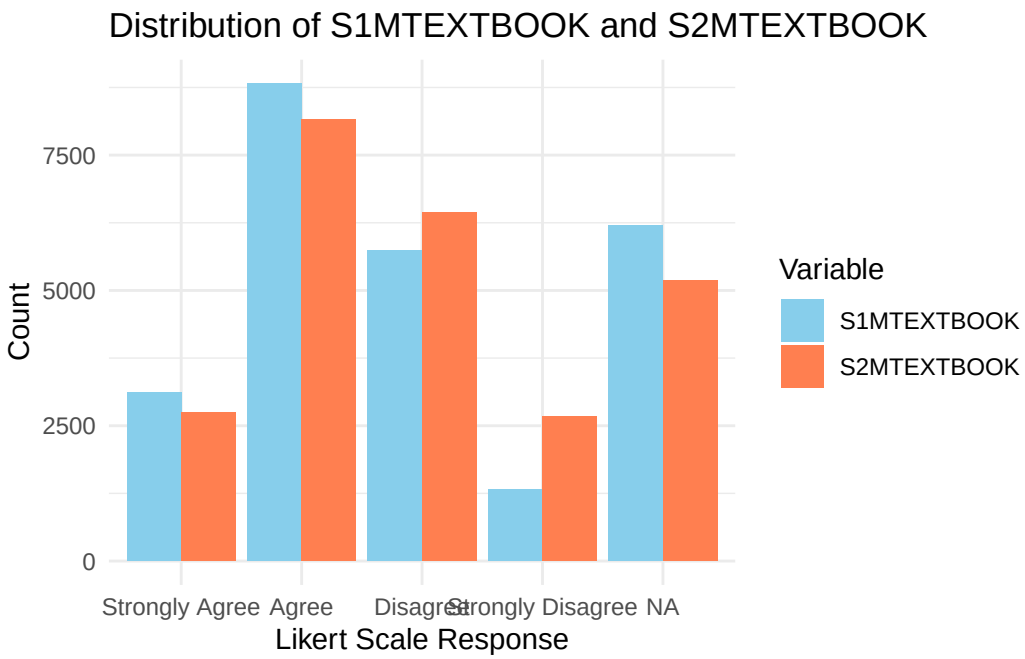


Textbook

Teen (9th / 11th grader) certain can understand (fall 2009 / spring 2012) math textbook

```
create_grouped_bar_plot(math_eff, "TEXTBOOK")
```

```
Warning: attributes are not identical across measure variables; they will be
dropped
```

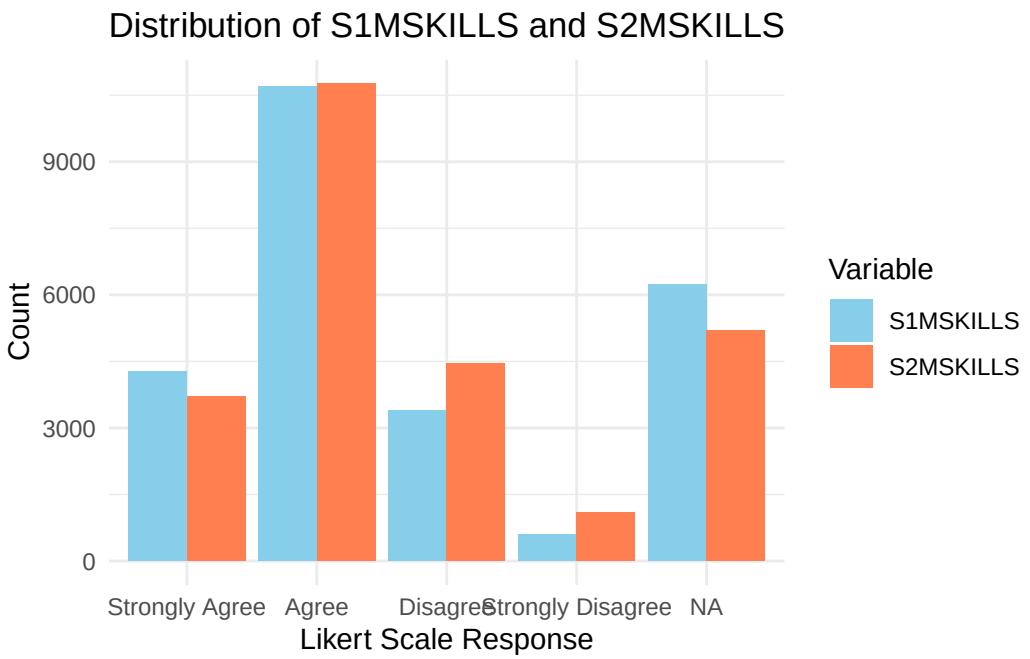


Skills

Teen certain can master skills in math course

```
create_grouped_bar_plot(math_eff, "SKILLS")
```

```
Warning: attributes are not identical across measure variables; they will be
dropped
```

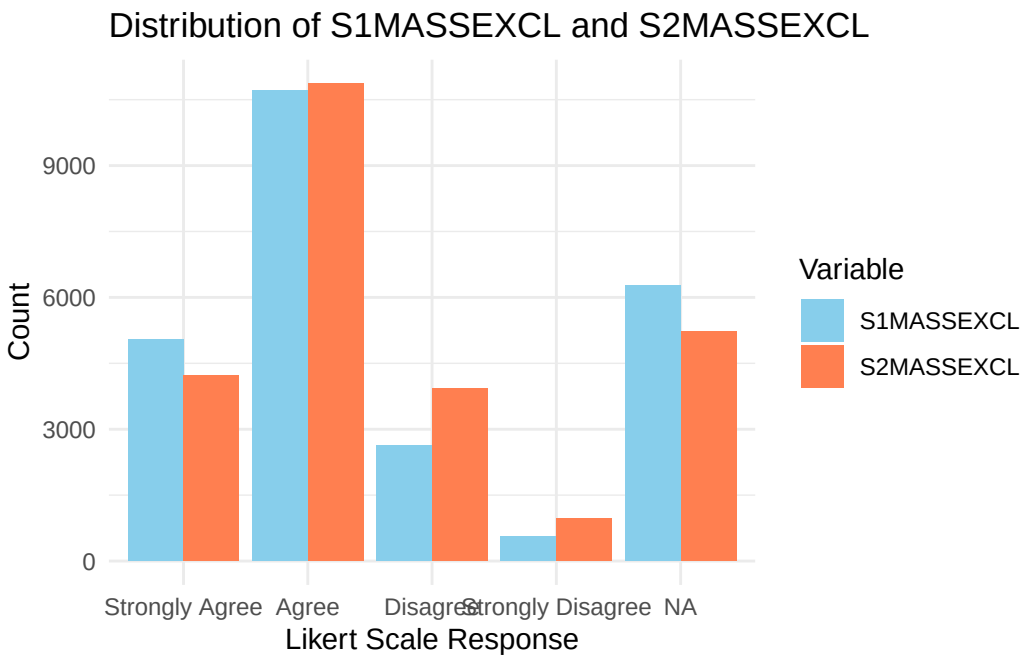


Assignments

Teen confident can do an excellent job on math assignments

```
create_grouped_bar_plot(math_eff, "ASSEXCL")
```

```
Warning: attributes are not identical across measure variables; they will be
dropped
```



Discrepancy!

From 9th to 11th grade, students' math self-efficacy declines. Why?

1. Difficulty of coursework goes up, self-efficacy follows as students do poorer
2. 11th graders had more time to compare with their peers than 9th graders did, more comparison = lower self-efficacy?
3. 11th graders have a better gauge of their math ability, are less over-confident.
4. Higher stakes. 11th graders are thinking about college, where math scores are much more important.
5. Lack of encouragement. 9th graders were highly motivated, just starting high school. 11th graders slack on their assignments / grades, self-efficacy is reflected in this.

Does self-efficacy correlate highly with actual math scores? If yes, could the worst self-efficacy scorers have dropped out? Conveniently, there is a "mathematics ability variable." Let's find out!

Actual and Efficacy

Treating the likert scale as continuous for the correlation

```r
# Renaming math ability for easier calling

hsls <- dathsls %>%
  rename(math_theta1 = X2TXMTH,
         math_theta2 = X1TXMTH)


summary(hsls$math_theta1)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
-2.602   0.007   0.686   0.717   1.433   4.505    4612
```

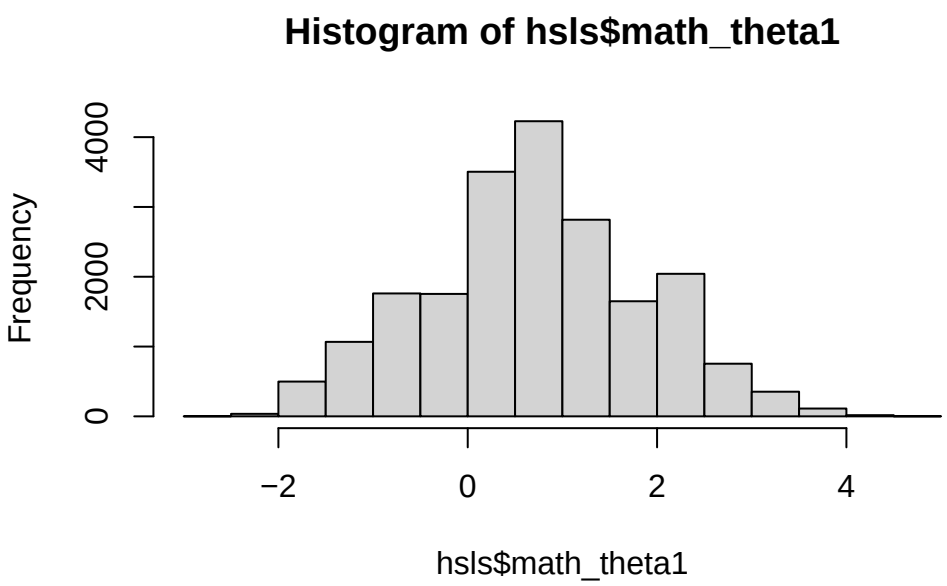
```r
summary(hsls$math_theta2)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
-2.575  -0.557   0.021   0.035   0.714   3.028    3762
```

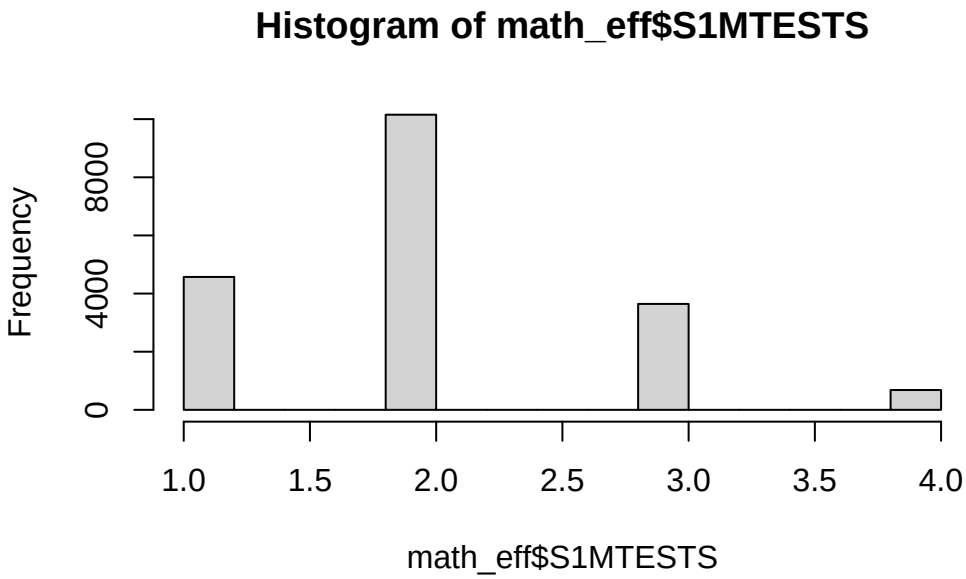
```r
head(math_eff$S1MTESTS)
```

```
<labelled<double>[6]>: S1 C08A 9th grader confident can do excellent job on fall 2009 math te
[1] 1 2 1 2 2 1

Labels:
 value                    label
    -9                  Missing
    -8       Unit non-response
    -7 Item legitimate skip/NA
     1           Strongly agree
     2                    Agree
     3                 Disagree
     4       Strongly disagree
```

```r
hist(hsls$math_theta1)
```




```
hist(math_eff$S1MTESTS)
```



```r
cor.test(hsls$math_theta1, hsls$S1MTESTS)
```

```
    Pearson's product-moment correlation

data:  hsls$math_theta1 and hsls$S1MTESTS
t = -35.519, df = 16653, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.2794331 -0.2511970
sample estimates:
       cor
-0.2653719
```

```r
tests_sagree <- hsls[hsls$S1MTESTS %in% c(1), ]
tests_agree <- hsls[hsls$S1MTESTS %in% c(2), ]
tests_disagree <- hsls[hsls$S1MTESTS %in% c(3), ]
tests_sdisagree <-  hsls[hsls$S1MTESTS %in% c(4), ]

head(tests_sdisagree$S1MTESTS, 20)
```

```
<labelled<double>[20]>: S1 C08A 9th grader confident can do excellent job on fall 2009 math t
 [1] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4

Labels:
 value                   label
    -9                 Missing
    -8       Unit non-response
    -7 Item legitimate skip/NA
     1          Strongly agree
     2                   Agree
     3                Disagree
     4       Strongly disagree
```

```r
print("Strongly disagree")
```

```
[1] "Strongly disagree"
```

```
  summary(tests_sdisagree$math_theta1)
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
-2.09310 -0.77978  0.17130  0.01517  0.66455  2.84380      134
```

```
  print("disagree")
```

```
[1] "disagree"
```

```
  summary(tests_disagree$math_theta1)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
-2.6019 -0.3304  0.4501  0.4151  1.0377  4.1434     549
```

```
  print("agree")
```

```
[1] "agree"
```

```
  summary(tests_agree$math_theta1)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
-2.2935  0.1487  0.7493  0.7792  1.4641  4.1908    1221
```

```
  print("Strongly agree")
```

```
[1] "Strongly agree"
```

```
  summary(tests_sagree$math_theta1)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
-2.1000  0.4741  1.1627  1.2229  2.1238  4.5046     490
```

```
print("Total")
```

```
[1] "Total"
```

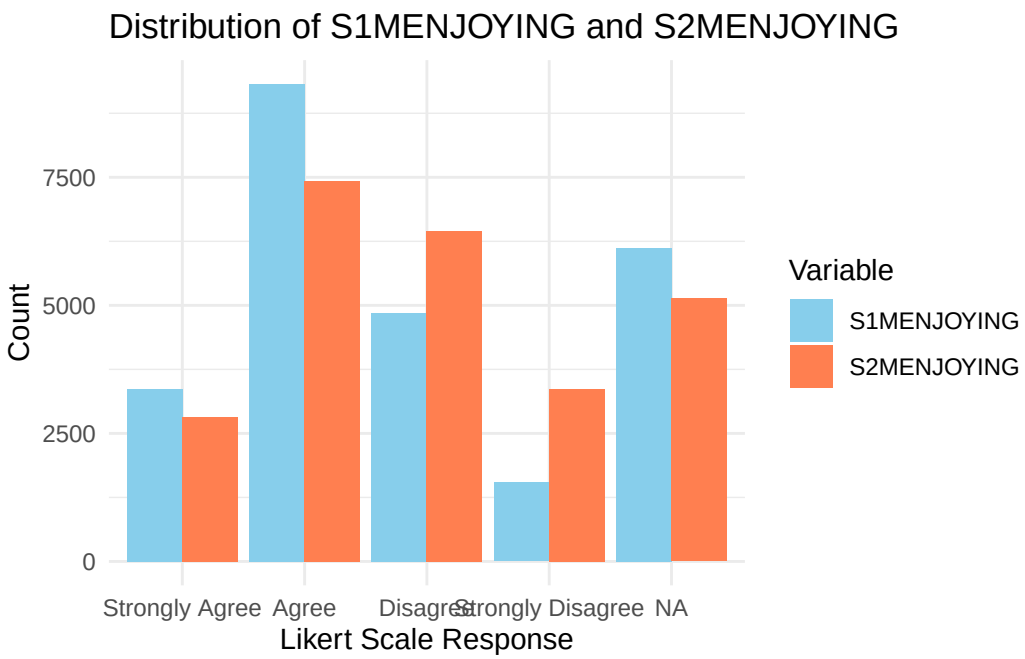
```
summary(hsls$math_theta1)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
-2.602   0.007   0.686   0.717   1.433   4.505    4612
```

Enjoying

```
create_grouped_bar_plot(math_eff, "ENJOYING")
```

```
Warning: attributes are not identical across measure variables; they will be
dropped
```

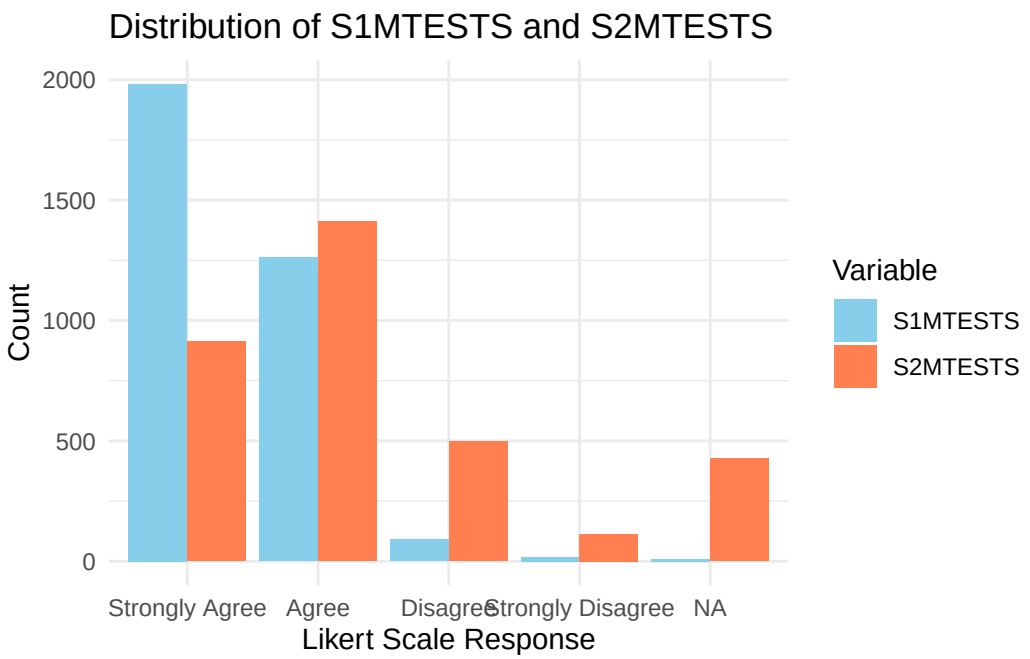



```
#summary(math_eff$S1MENJOYING)
```

Enjoying Math and Tests

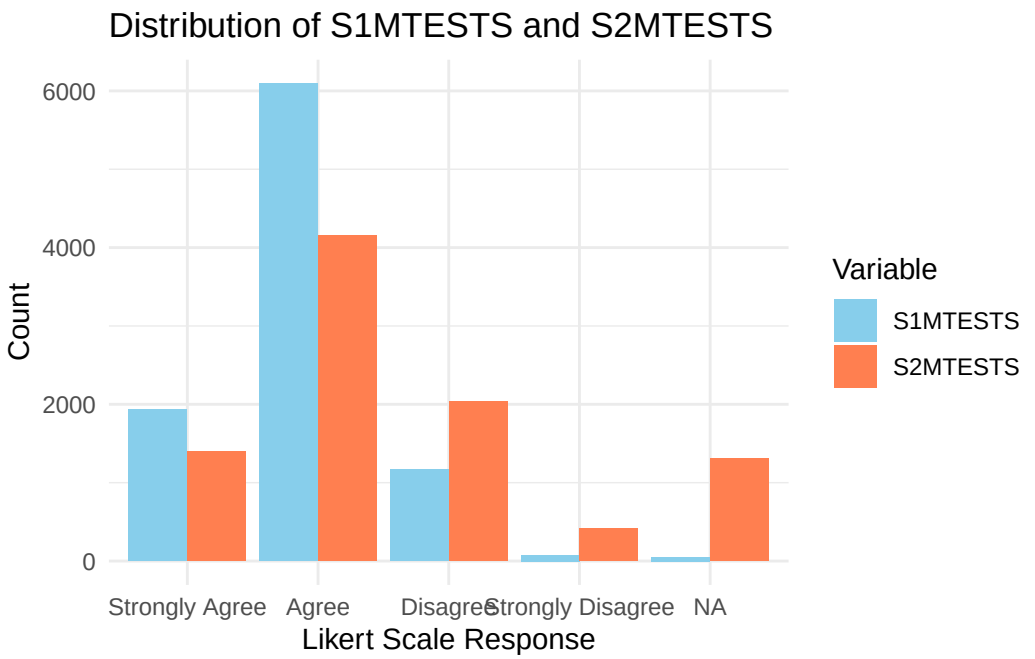
```
seg4 <- hsls[hsls$S1MENJOYING %in% c(4), ]
seg3 <- hsls[hsls$S1MENJOYING %in% c(3), ]
seg2 <- hsls[hsls$S1MENJOYING %in% c(2), ]
seg1 <- hsls[hsls$S1MENJOYING %in% c(1), ]


create_grouped_bar_plot(seg1, "TESTS")
```

```
Warning: attributes are not identical across measure variables; they will be
dropped
```



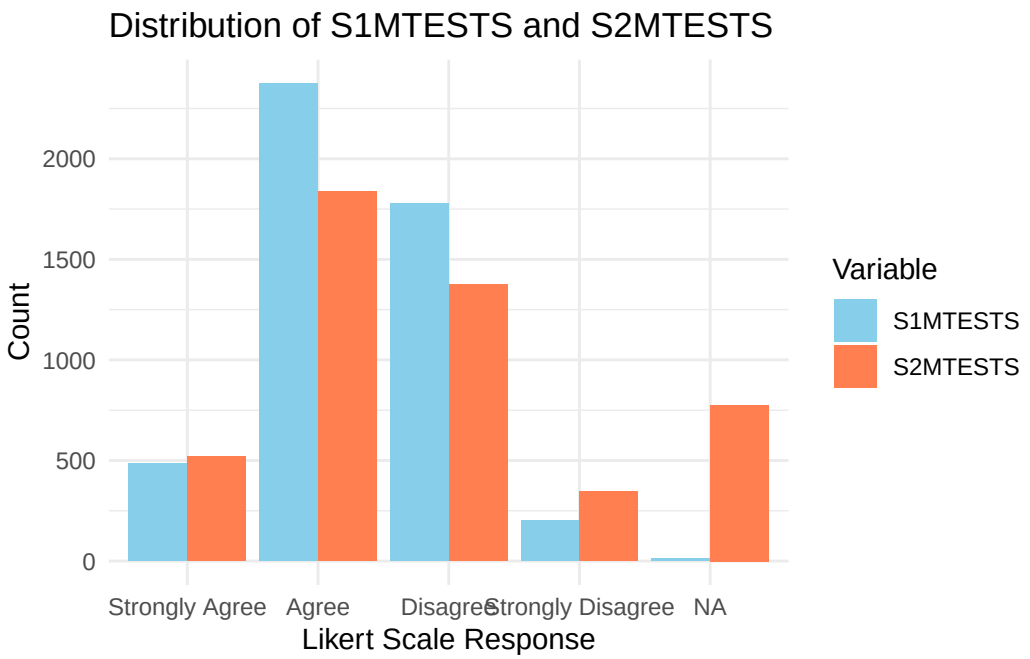

```
create_grouped_bar_plot(seg2, "TESTS")
```

```
Warning: attributes are not identical across measure variables; they will be
dropped
```



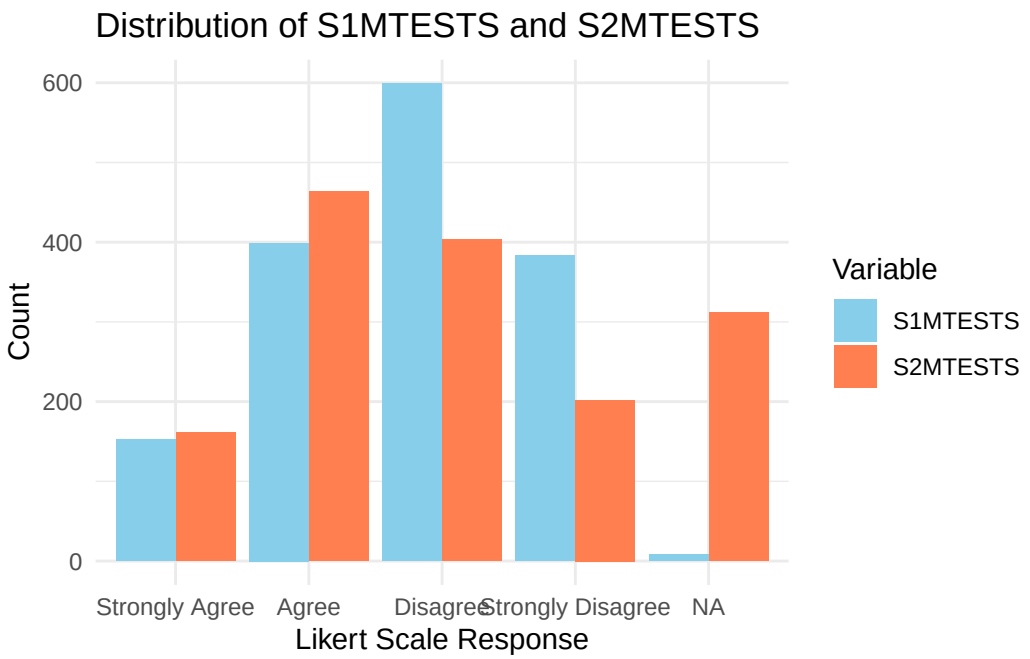

```
create_grouped_bar_plot(seg3, "TESTS")
```

```
Warning: attributes are not identical across measure variables; they will be
dropped
```




```
create_grouped_bar_plot(seg4, "TESTS")
```

```
Warning: attributes are not identical across measure variables; they will be
dropped
```




```
mean(is.na(seg1$S2MTESTS))
```

```
[1] 0.1267103
```

```
mean(is.na(seg2$S2MTESTS))
```

```
[1] 0.1403283
```

```
mean(is.na(seg3$S2MTESTS))
```

```
[1] 0.1598352
```

```
mean(is.na(seg4$S2MTESTS))
```

```
[1] 0.2023346
```

Looking at 9th graders that strongly agree that they are enjoying math, proportion of missing for tests for 11th grade is 12%. For strongly disagree, 20%

Missing at random? Does this mean anything?

```
  print("Assignments")
```

```
[1] "Assignments"
```

```
  mean(is.na(seg1$S2MASSEXCL))
```

```
[1] 0.1296847
```

```
  mean(is.na(seg2$S2MASSEXCL))
```

```
[1] 0.1390409
```

```
  mean(is.na(seg3$S2MASSEXCL))
```

```
[1] 0.161277
```

```
  mean(is.na(seg4$S2MASSEXCL))
```

```
[1] 0.2029831
```

```
  print("Skills")
```

```
[1] "Skills"
```

```
  mean(is.na(seg1$S2MTESTS))
```

```
[1] 0.1267103
```

```
  mean(is.na(seg2$S2MTESTS))
```

```
[1] 0.1403283
```

```
  mean(is.na(seg3$S2MTESTS))
```

```
[1] 0.1598352
```

```
  mean(is.na(seg4$S2MTESTS))
```

```
[1] 0.2023346
```

```
  mean(is.na(seg1$S2MTESTS))
```

```
[1] 0.1267103
```

```
  mean(is.na(seg2$S2MTESTS))
```

```
[1] 0.1403283
```

```
  mean(is.na(seg3$S2MTESTS))
```

```
[1] 0.1598352
```

```
  mean(is.na(seg4$S2MTESTS))
```

```
[1] 0.2023346
```