High School Longitudinal Study (HSLS) Comparison

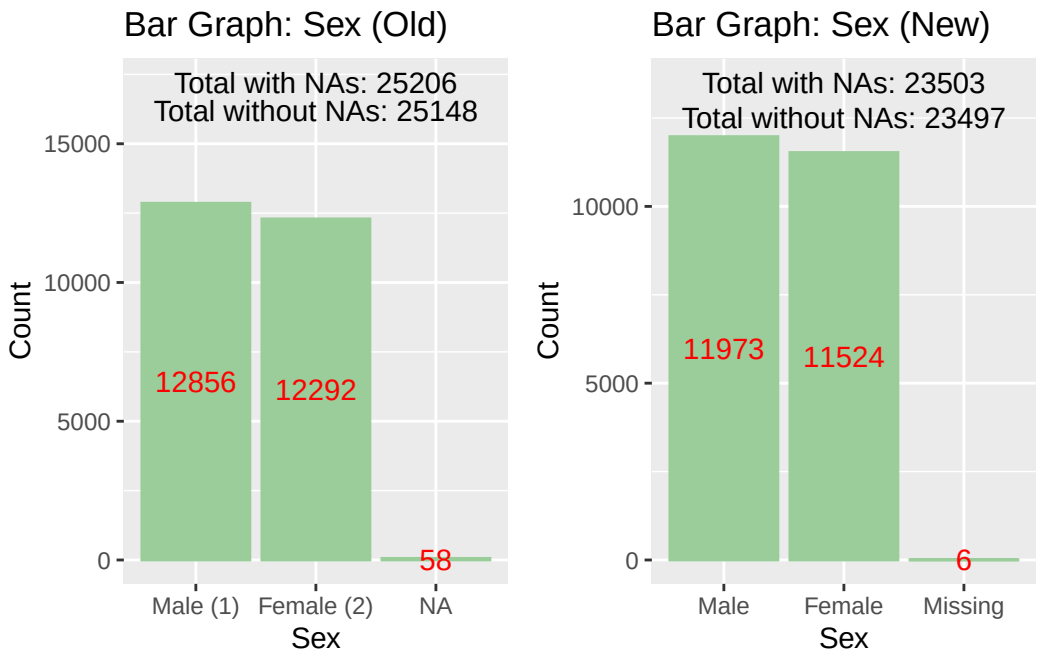
The new dataset (hsls, retrieved from https://nces.ed.gov/datalab/onlinecodebook/session/codebook/fa65d339-7237-4a01-95c6-aa4a10fb02f6) has 23503 observations with 9614 variables. The old dataset (From what we had before) has 25206 observations with 185 variables. The first and last student ID of both datasets are 10000 and 35206.

Difference of 1703. The new dataset is skipping rows, going from (e.g.) 10017 to 10019, where the old dataset includes 10018. For the rows that were skipped in the new dataset, all values (except for demographics) are either 0 or NA.

Demographic variables (sex, race) usually have a value, but seems random as to when it does or doesn't.

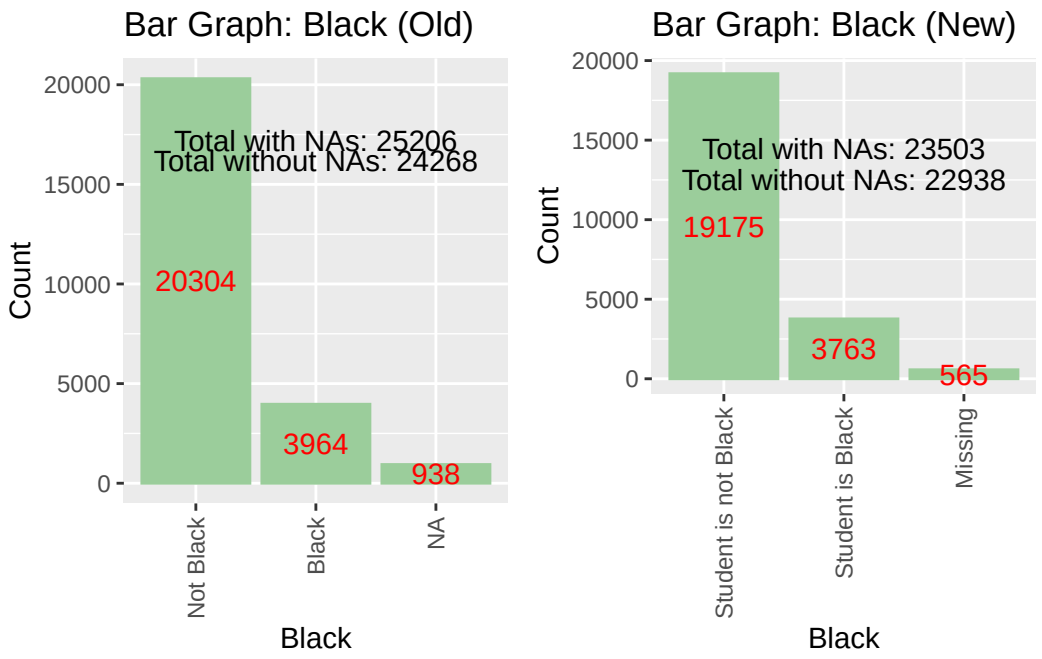
However, even rows that both datasets have (such as student id 10031) have different values. Student 10031 for the new dataset has an SES value of -8. Looking at the X1SES_IM variable, all SES values that were imputed were suppressed with -8 in the public dataset. This does not apply to the efficacy scores. These are all completely equal, excluding NA values.

Bar Graphs of Sex



The 6 "Missing" values in the new dataset are all marked as NA in the old. The proportions are very close, but off due to most of the skipped rows including sex.

Bar Graphs of the unrestricted Races (Black or White)

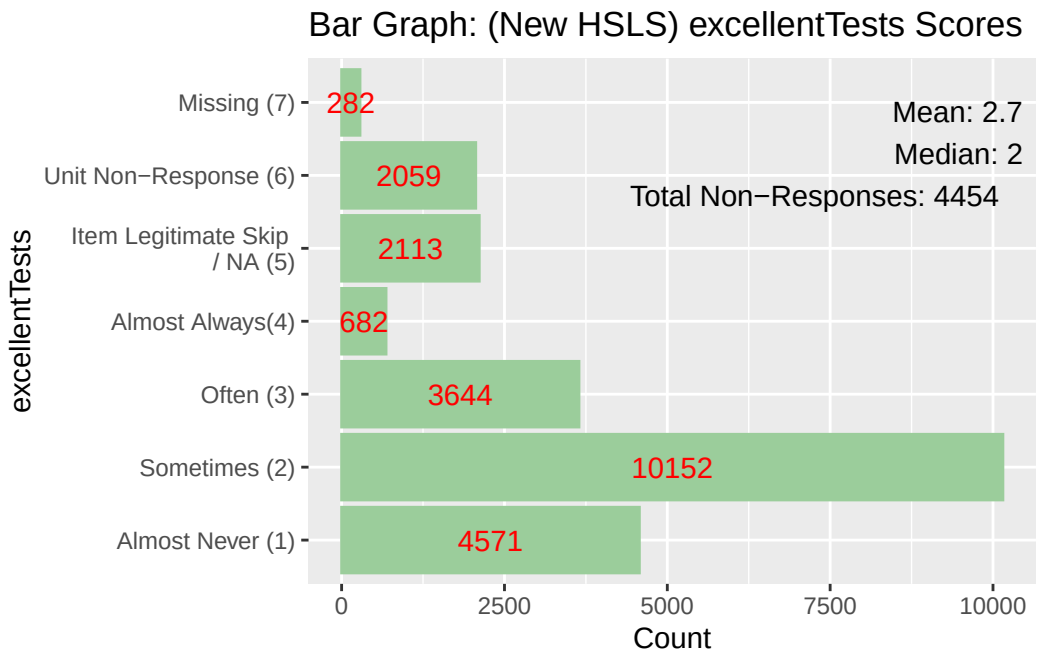


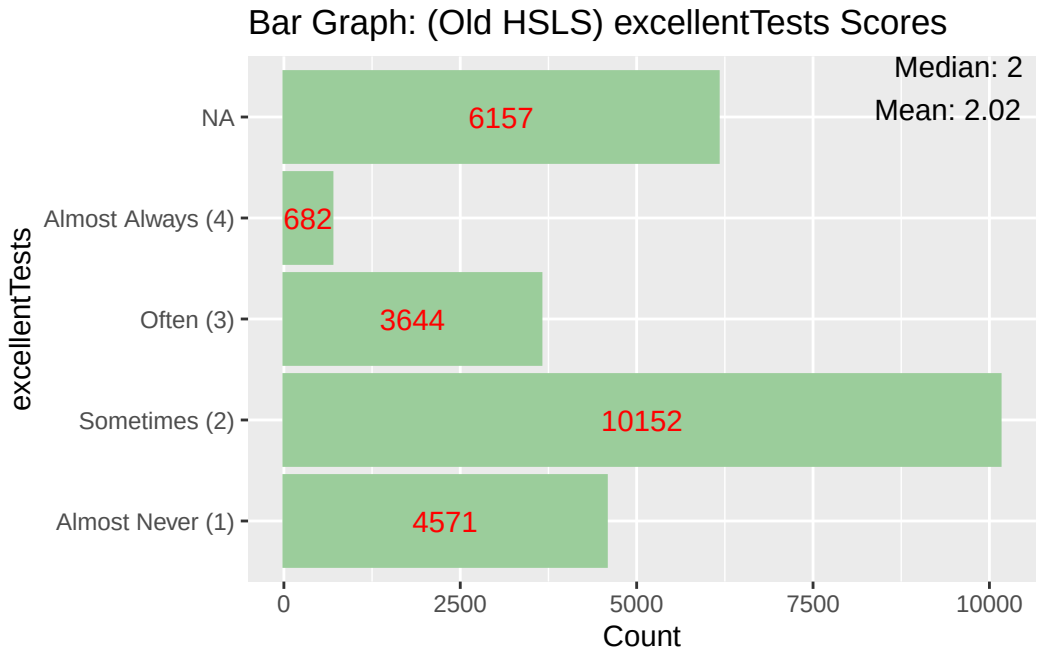
Function to calculate summaries and generate graphs

```
calculate_and_plot_hsls <- function(variable_name, data=hsls) {
  # Calculate mean and median
  mean_value <- mean(data[[variable_name]], na.rm = TRUE)
  median_value <- median(data[[variable_name]], na.rm = TRUE)
  total_non_response <- sum(table(data[[variable_name]], useNA = "always")[5:7])


  # Create the plot
  ggplot(data = data, aes(x = as.factor(data[[variable_name]]))) +
     geom_bar(fill = "darkseagreen3", color = "darkseagreen3") +
     scale_x_discrete(labels = c(`1` = "Almost Never (1)",
                                 `2` = "Sometimes (2)",
                                 `3` = "Often (3)",
                                 `4` = "Almost Always(4)",
                                 `5` = "Item Legitimate Skip \n / NA (5)",
                                 `6` = "Unit Non-Response (6)",
                                 `7` = "Missing (7)")) +
```

```
    labs(title = paste0("Bar Graph: (New HSLS) ", variable_name, " Scores"),
         x = variable_name,
         y = "Count") +
    geom_text(aes(label = after_stat(count)), stat = 'count',
              position = position_stack(vjust = 0.5), color = "red") +
    annotate("text", x = Inf, y = Inf,
             label = paste("Mean:", round(mean_value, 2)),
             hjust = 1.1, vjust = 3, size = 4, color = "black") +
    annotate("text", x = Inf, y = Inf,
             label = paste("Median:", median_value),
             hjust = 1.1, vjust = 5, size = 4, color = "black") +
    annotate("text", x = Inf, y = Inf,
             label = paste("Total Non-Responses:", total_non_response),
             hjust = 1.1, vjust = 7, size = 4, color = "black") +
    coord_flip()
}
calculate_and_plot_hsls_old <- function(variable_name, data=hsls_old) {
  # Calculate mean and median
  mean_value <- mean(data[[variable_name]], na.rm = TRUE)
  median_value <- median(data[[variable_name]], na.rm = TRUE)

  # Create the plot
  ggplot(data, aes(x = as.factor(data[[variable_name]]))) +
    geom_bar(fill = "darkseagreen3", color = "darkseagreen3") +
    scale_x_discrete(labels = c(`1` = "Almost Never (1)",
                                `2` = "Sometimes (2)",
                                `3` = "Often (3)",
                                `4` = "Almost Always (4)")) +
    labs(title = paste0("Bar Graph: (Old HSLS) ", variable_name, " Scores"),
         x = variable_name,
         y = "Count") +
    geom_text(aes(label = after_stat(count)), stat = 'count',
              position = position_stack(vjust = 0.5), color = "red") +
    annotate("text", x = Inf, y = Inf,
             label = paste("Mean:", round(mean_value, 2)),
             hjust = 1.1, vjust = 3, size = 4, color = "black") +
    annotate("text", x = Inf, y = Inf,
             label = paste("Median:", median_value),
             hjust = 1.1, vjust = 1, size = 4, color = "black") +
    coord_flip()
}
```

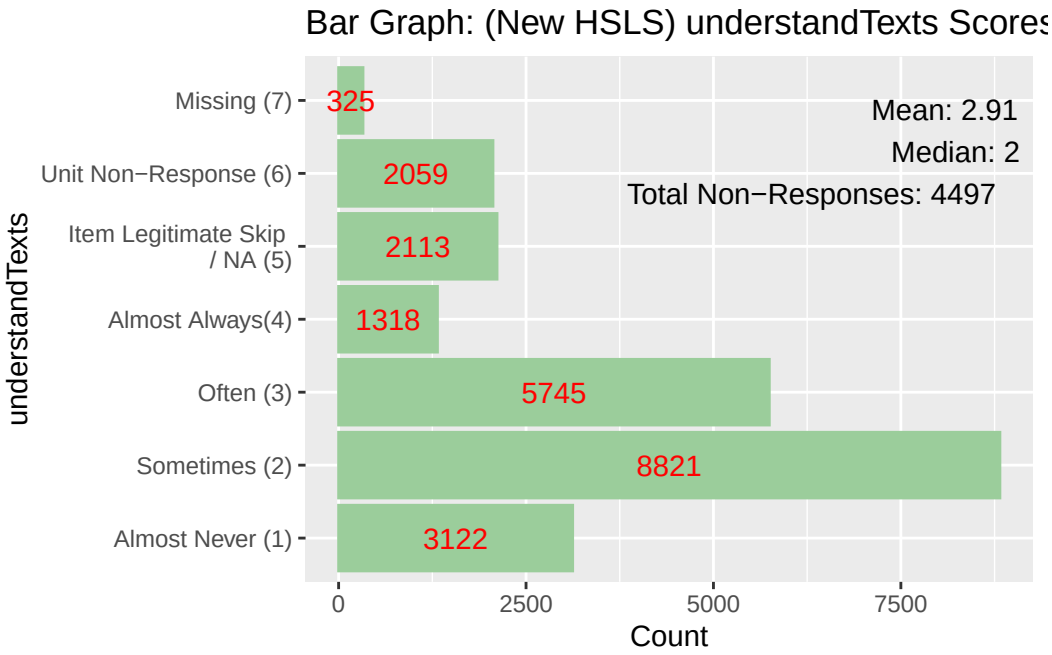
Bar Graphs of Excellent Tests

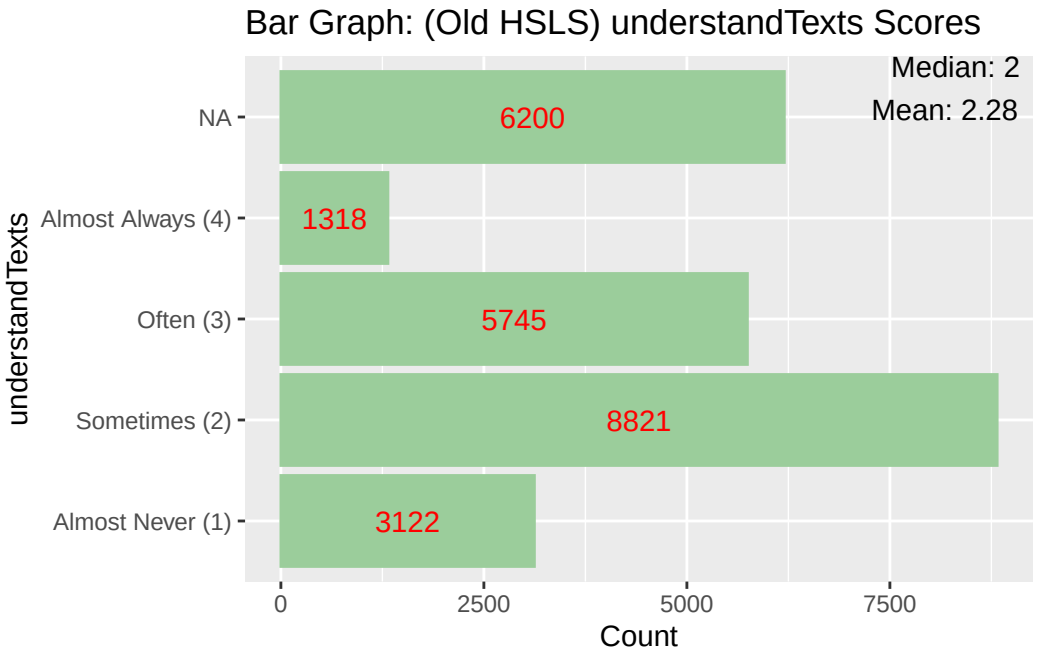




6157 - 4454 is 1703, the difference in rows of the two datasets.

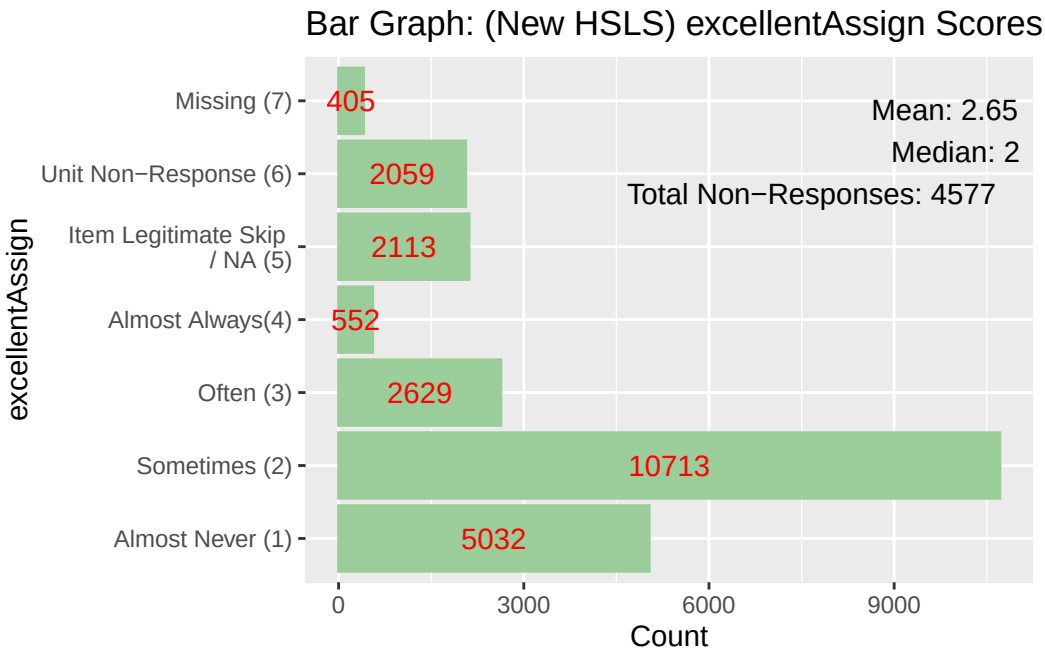
Bar Graphs of Understand Texts

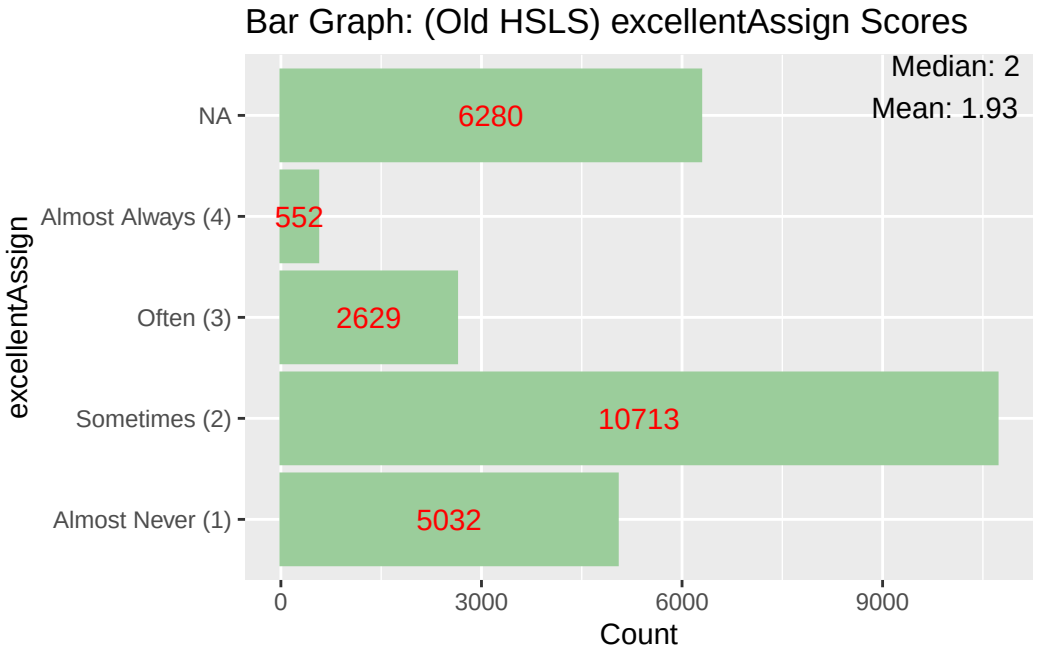




Difference again, 1703.

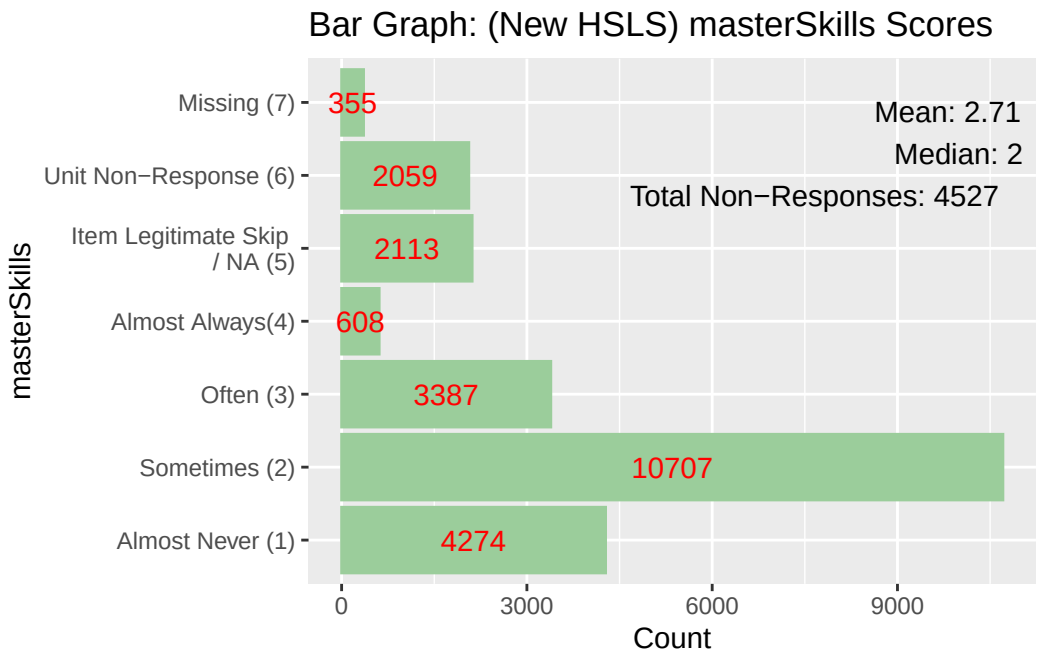
Bar Graphs of Excellent Assign

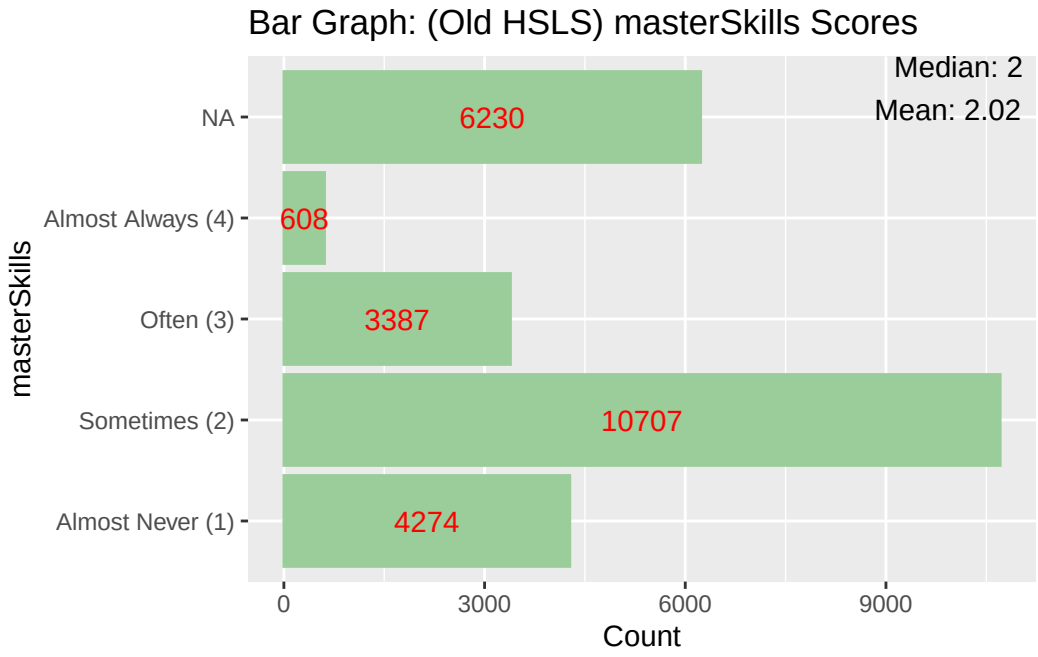




Difference again, 1703.

Bar Graphs of Master Skills





Difference again, 1703.

Example of imputed data: SES

```
summary(hsls$SES[hsls$SES != -8.0])
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-1.93020 -0.50115 -0.01090  0.05409  0.56480  2.88070
```

```
summary(hsls_old$SES)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 -1.930  -0.515  -0.026   0.042   0.550   2.881    3214
```

```
SES_imputed_rows_new <- hsls %>% filter(X1SES_IM == "Imputed entirely" | (X1SES_IM == "Com
nrow(SES_imputed_rows_new)
```

```
[1] 6809
```

```
SES_imputed_rows_suppressed_new <- SES_imputed_rows_new %>% filter(SES == -8)
nrow(SES_imputed_rows_suppressed_new)
```

```
[1] 172
```

The old dataset contains NAs and imputed SES values. The new dataset codes both of these with a value of -8.0. When we remove all instances of -8 in the old dataset, the descriptive statistics are quite similar.

Of the 6809 rows that had either the "Imputed entirely" or "Components imputed" flag, 172 of them have been suppressed with a -8. The slight discrepancy in the descriptive statistics of the SES values of both datasets are due to these 172 values that are coded as -8 in the new (public) dataset, whereas they have non-NA, meaningful values in the old dataset.

Conclusion

As far as I can tell, the difference in the two datasets is due to the new dataset being the public use version, where many non-efficacy variables are suppressed. Some variables have a random number of observations suppressed due to these observations being imputed, and observations that were mostly NA except for a few demographics were skipped in the public version.

The demographic variables all have slightly different proportions. The new dataset "skips" some student ID's, and these skipped observations include demographic variables but most other variables are NA.

The efficacy score variables all have the same counts and proportions when missing values are not counted.