Appendix: Data Quality Checks

2026-01-30

Contents

Setup

Load Libraries

```r
library(dplyr)
library(readxl)
```

Set File Paths

```r
root <- "/Users/pei-chin/Research/Behavioral Science and Marketing_data_task"

data_raw       <- file.path(root, "data_raw")
data_processed <- file.path(root, "data_processed")
data_clean     <- file.path(root, "data_clean")
scripts        <- file.path(root, "scripts")
output         <- file.path(root, "output")
figures        <- file.path(output, "figures")
```

Load Data

Load the Excel file. Sheet 1 contains variable definitions; Sheet 2 contains the actual data.

```r
df_variable <- read_excel(
  file.path(data_raw, "Data - 2026.xlsx"),
  sheet = 1
)

raw_data <- read_excel(
  file.path(data_raw, "Data - 2026.xlsx"),
  sheet = 2
)
```

Section A: Sanity-Check Ranges and Missingness

Before getting into 01_Clean Data

```r
dim(raw_data)
```

```
## [1] 61 70
```

Check Feeling Variables

Feeling variables should range from -30 to +30 based on the survey design. No observations are removed in this step.

```r
feeling_vars <- c(
  "feelings_youalone",
  "feelings_bothyoufirst",
  "feelings_themalone",
  "feelings_boththemfirst",
  "feelings_neither",

  "feelings_youaloneforgiven"
)

for (var in feeling_vars) {
  min_val   <- min(raw_data[[var]], na.rm = TRUE)
```

```
  max_val   <- max(raw_data[[var]], na.rm = TRUE)
  n_missing <- sum(is.na(raw_data[[var]]))
  in_range  <- min_val >= -30 & max_val <= 30

  print(
    data.frame(
      variable = var,
      min      = min_val,
      max      = max_val,
      missing  = n_missing,
      in_range = in_range
    )
  )
}
```

```
##            variable min max missing in_range
## 1 feelings_youalone -30  10      12     TRUE
##                variable min max missing in_range
## 1 feelings_bothyoufirst -30  30      12     TRUE
##             variable min max missing in_range
## 1 feelings_themalone -30  30      12     TRUE
##                 variable min max missing in_range
## 1 feelings_boththemfirst -20  30      12     TRUE
##           variable min max missing in_range
## 1 feelings_neither -30  20      12     TRUE
##                    variable min max missing in_range
## 1 feelings_youaloneforgiven -30  22      12     TRUE
```

Recompute Emotional Positivity Rankings

The platform-generated `feeling_DO` variables do not reflect the true ordering of emotional positivity. Here we recompute rankings based on the raw feeling scores and overwrite the original ranking variables.

```
# Raw feeling score variables (used to compute rankings)
feeling_vars <- c(
  "feelings_youalone",
  "feelings_bothyoufirst",
  "feelings_themalone",
  "feelings_boththemfirst",
  "feelings_neither",
  "feelings_youaloneforgiven"
)

# Original (incorrect) ranking variables
do_vars <- c(
  "feelings_DO_1",
  "feelings_DO_2",
  "feelings_DO_3",
  "feelings_DO_4",
  "feelings_DO_5",
  "feelings_DO_6"
)

n <- nrow(raw_data)
```

```r
for (i in 1:n) {
  # Extract raw feeling scores for participant i
  x <- as.numeric(raw_data[i, feeling_vars])

  # Rank feelings in descending order

  # Higher emotional positivity receives a lower rank value
  # Ties are assigned the minimum rank if the emotional positivity is the same
  raw_data[i, do_vars] <- as.list(rank(-x, ties.method = "min"))
}
```

Handle Outcome Ordering for Binary Choice Questions

Randomization order doesn't match, so we create binary indicators for each choice option.

Binary Choice 1: I apologize first vs. Neither apologizes

```r
raw_data$outcome_binary1_DO_1 <- ifelse(
  grepl("^I apologize first", raw_data$outcome_binary1),
  1, 0
)

raw_data$outcome_binary1_DO_2 <- ifelse(
  grepl("Neither I nor ", raw_data$outcome_binary1),
  1, 0
)
```

Binary Choice 2: Other doesn't apologize after vs. Neither apologizes

```r
raw_data$outcome_binary2_DO_1 <- ifelse(
  grepl("does not apologize after that", raw_data$outcome_binary2),
  1, 0
)

raw_data$outcome_binary2_DO_2 <- ifelse(
  grepl("Neither I nor", raw_data$outcome_binary2),
  1, 0
)
```

Rename Variables

```r
raw_data <- raw_data %>%
  rename(
    prefer_I_apologize_first  = outcome_binary1_DO_1,
    prefer_Neither_I_nor      = outcome_binary1_DO_2,
    prefer_feelings_youalone  = outcome_binary2_DO_1,
    prefer__Neither_I_nor_2   = outcome_binary2_DO_2
  )
```

Check Blame Variable

The blame variable should range from 0 to 100 based on the survey slider.

```r
max_blame <- max(raw_data$blame_1, na.rm = TRUE)
max_blame
```

```
## [1] 75
```

```r
min_blame <- min(raw_data$blame_1, na.rm = TRUE)
min_blame
```

```
## [1] 0
```

```r
n_missing <- sum(is.na(raw_data$blame_1))
n_missing
```

```
## [1] 13
```

Check Demographic Variables

```r
print(table(raw_data$target_sex, useNA = "ifany"))
```

```
##
## Female   Male  Other   <NA>
##     27     20      1     13
```

```r
print(table(raw_data$sex, useNA = "ifany"))
```

```
##
## Female   Male   <NA>
##     28     20     13
```

```r
print(min(raw_data$age, na.rm = TRUE))
```

```
## [1] 18
```

```r
print(max(raw_data$age, na.rm = TRUE))
```

```
## [1] 149
```

```r
summary(raw_data$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   18.00   26.75   29.00   33.44   32.25  149.00      13
```

```r
print(sum(is.na(raw_data$age)))
```

```
## [1] 13
```

Drop Invalid Observations

```r
processed_data <- raw_data[complete.cases(raw_data$feelings_youalone), ]
processed_data <- processed_data %>%
  filter(age < 100)  # age over 100 is suspicious
```

Overall Missingness Summary

```r
count_na <- function(x) {
  sum(is.na(x))
}
```

```r
sapply(processed_data, count_na)
```

```
##                                     StartDate
##                                             0
##                                       EndDate
##                                             0
##                                        Status
##                                             0
##                                     IPAddress
##                                             0
##                                      Progress
##                                             0
##                         Duration (in seconds)
##                                             0
##                                      Finished
##                                             0
##                                  RecordedDate
##                                             0
##                                    ResponseId
##                                             0
##                               RecipientLastName
##                                            47
##                              RecipientFirstName
##                                            47
##                                 RecipientEmail
##                                            47
##                              ExternalReference
##                                            47
##                               LocationLatitude
##                                             1
##                              LocationLongitude
##                                             1
##                            DistributionChannel
##                                             0
##                                  UserLanguage
##                                             0
##                                       consent
##                                             0
##                                   Q26_Browser
##                                             0
##                                   Q26_Version
##                                             0
##                               Q26_Operating System
##                                             0
##                                 Q26_Resolution
##                                             0
##                              QID54_First Click
##                                             0
##                               QID54_Last Click
##                                             0
##                              QID54_Page Submit
##                                             0
##                              QID54_Click Count
```

```
##                                     0
##                          real_imaginary
##                                     0
##                            initials_box
##                                     0
##                                describe
##                                     0
##                           feelings_youalone
##                                     0
##                         feelings_bothyoufirst
##                                     0
##                            feelings_themalone
##                                     0
##                        feelings_boththemfirst
##                                     0
##                              feelings_neither
##                                     0
##                     feelings_youaloneforgiven
##                                     0
##                                feelings_DO_1
##                                     0
##                                feelings_DO_2
##                                     0
##                                feelings_DO_3
##                                     0
##                                feelings_DO_4
##                                     0
##                                feelings_DO_5
##                                     0
##                                feelings_DO_6
##                                     0
##                                 feelings_exp
##                                       37
##                               outcome_binary1
##                                     0
##                      prefer_I_apologize_first
##                                     0
##                          prefer_Neither_I_nor
##                                     0
##                               outcome_binary2
##                                     0
##                      prefer_feelings_youalone
##                                     0
##                       prefer__Neither_I_nor_2
##                                     0
##                                    blame_1
##                                     0
##                                attention_1
##                                       44
##                              attention_1_TEXT
##                                       46
##                                attention_2
##                                       44
##                              attention_2_TEXT
```

```
##                                           46
##                                  attention_3
##                                           44
##                                attention_3_TEXT
##                                            0
##                                   target_sex
##                                            0
##                                target_sex_3_TEXT
##                                           47
##                                          sex
##                                            0
##                                   sex_3_TEXT
##                                           47
##                                          age
##                                            0
##                                     comments
##                                           45
##                                  mainControl
##                                            0
##                                    passedattn
##                                            0
##                                lottery_draw
##                                            0
##                                      winner
##                                            0
##                                    initials
##                                            0
##                                initiator_type
##                                            0
## InitiatorType-binarychoice_DO_outcome_binary2
##                                            0
## InitiatorType-binarychoice_DO_outcome_binary1
##                                            0
##             InitiatorType-binarychoice_DO_Q30
##                                            0
```

Save Processed Data

Ready to get into 01_Clean Data

The purpose here is to drop invalid observations

The purpose of the next step, 01_Clean Data, is to drop the variables we need

```
save(
  processed_data,
  file = file.path(data_processed, "processed_data.RData")
)
```

Section B: Within- vs Between-Subject Structure

(some code adapted from 02_descriptive_analysis)

Here, we use response patterns (i.e., non-missing values) to confirm that the observed data structure aligns with the intended survey design.

Feeling Ratings: Within-Subject

To verify this, we count the number of non-missing responses for each feeling variable. The presence of valid responses across multiple conditions for the same participants confirms a within-subject design. We also inspect basic descriptive statistics for these variables.

```
feeling_vars <- c(
  "feelings_youalone",
  "feelings_bothyoufirst",
  "feelings_themalone",
  "feelings_boththemfirst",
  "feelings_neither",
  "feelings_youaloneforgiven"
)

load(file.path(data_clean, "clean_data.RData"))

summary(clean_data[, feeling_vars])
```

```
##  feelings_youalone feelings_bothyoufirst feelings_themalone
##  Min.   :-30.00    Min.   :-30.000       Min.   :-30.000
##  1st Qu.:-30.00    1st Qu.:  0.000       1st Qu.:-20.000
##  Median :-20.00    Median : 10.000       Median :-10.000
##  Mean   :-18.36    Mean   :  7.844       Mean   : -5.533
##  3rd Qu.:-10.00    3rd Qu.: 20.000       3rd Qu.: 10.000
##  Max.   : 10.00    Max.   : 30.000       Max.   : 30.000
##  feelings_boththemfirst feelings_neither feelings_youaloneforgiven
##  Min.   :-20.00         Min.   :-30.00   Min.   :-30.00
##  1st Qu.: 10.00         1st Qu.:-23.00   1st Qu.:-30.00
##  Median : 20.00         Median :-15.00   Median :-19.00
##  Mean   : 17.42         Mean   :-14.38   Mean   :-13.91
##  3rd Qu.: 26.00         3rd Qu.: -8.00   3rd Qu.: -5.00
##  Max.   : 30.00         Max.   : 20.00   Max.   : 22.00
```

```
# Check how many participants have complete data for each feeling variable
feeling_complete_n <- sapply(
  feeling_vars,
  function(var) sum(!is.na(clean_data[[var]]))
)

print(feeling_complete_n)
```

```
##         feelings_youalone     feelings_bothyoufirst        feelings_themalone
##                        45                        45                        45
##    feelings_boththemfirst          feelings_neither feelings_youaloneforgiven
##                        45                        45                        45
```

Self-Blame Grouping: Between-Subject

Self-blame is treated as a stable, individual-level characteristic and is used to classify participants into high vs. low self-blame groups. Each participant should belong to exactly one group, indicating a between-subject structure.

To confirm this, we inspect the distribution of the self-blame grouping variable and verify that each observation is assigned to a single category.

```
table(clean_data$high_blame)
```

```
##
##  0  1
## 35 10
```

Binary Choice Questions: Within-Subject

Breaking down feeling ratings by self-blame group:

```
by(
  clean_data[, feeling_vars],
  clean_data$high_blame,
  summary
)
```

```
## clean_data$high_blame: 0
##  feelings_youalone feelings_bothyoufirst feelings_themalone
##  Min.   :-30.00    Min.   :-30.000       Min.   :-30.0
##  1st Qu.:-30.00    1st Qu.:  0.000       1st Qu.:-14.0
##  Median :-22.00    Median : 10.000       Median : -8.0
##  Mean   :-20.89    Mean   :  6.629       Mean   : -4.2
##  3rd Qu.:-12.00    3rd Qu.: 19.000       3rd Qu.:  9.0
##  Max.   :  8.00    Max.   : 27.000       Max.   : 30.0
##  feelings_boththemfirst feelings_neither feelings_youaloneforgiven
##  Min.   :-20.00         Min.   :-30.0    Min.   :-30.00
##  1st Qu.: 11.00         1st Qu.:-23.5    1st Qu.:-30.00
##  Median : 20.00         Median :-19.0    Median :-20.00
##  Mean   : 17.54         Mean   :-15.4    Mean   :-15.66
##  3rd Qu.: 25.50         3rd Qu.: -8.5    3rd Qu.: -5.00
##  Max.   : 30.00         Max.   :  4.0    Max.   : 22.00
## ------------------------------------------------------------
## clean_data$high_blame: 1
##  feelings_youalone feelings_bothyoufirst feelings_themalone
##  Min.   :-30.00    Min.   :-10.00        Min.   :-30.00
##  1st Qu.:-20.00    1st Qu.:  6.25        1st Qu.:-22.50
##  Median :-11.50    Median : 15.50        Median :-20.00
##  Mean   : -9.50    Mean   : 12.10        Mean   :-10.20
##  3rd Qu.:  4.75    3rd Qu.: 20.00        3rd Qu.:  6.25
##  Max.   : 10.00    Max.   : 30.00        Max.   : 20.00
##  feelings_boththemfirst feelings_neither feelings_youaloneforgiven
##  Min.   :-10.00         Min.   :-30.00   Min.   :-30.00
##  1st Qu.:  8.50         1st Qu.:-18.25   1st Qu.:-20.00
##  Median : 20.00         Median :-10.00   Median : -7.00
##  Mean   : 17.00         Mean   :-10.80   Mean   : -7.80
##  3rd Qu.: 27.75         3rd Qu.: -5.75   3rd Qu.:  2.25
##  Max.   : 30.00         Max.   : 20.00   Max.   : 20.00
```

Counting preferences for the first set of binary choice questions:

```
sum(clean_data$prefer_I_apologize_first)    # participants who prefer apologizing first
```

```
## [1] 35
```

```
sum(clean_data$prefer_Neither_I_nor)        # participants who prefer neither option
```

```
## [1] 10
```

Counting preferences for the second set of binary choice questions:

```
sum(clean_data$prefer_feelings_youalone)    # participants who prefer apologizing alone
```

```
## [1] 21
```

```
sum(clean_data$prefer__Neither_I_nor_2)     # participants who prefer neither option
```

```
## [1] 24
```

Break preferences down by self-blame group:

```
# First preference set
clean_data %>%
  group_by(high_blame) %>%
  summarise(
    I_apologize_first_1 = sum(prefer_I_apologize_first, na.rm = TRUE),
    Neither_1           = sum(prefer_Neither_I_nor, na.rm = TRUE),
    n                   = n()
  )
```

```
## # A tibble: 2 x 4
##   high_blame I_apologize_first_1 Neither_1     n
##        <dbl>               <dbl>     <dbl> <int>
## 1          0                  26         9    35
## 2          1                   9         1    10
```

```
# Second preference set - used to check whether patterns are consistent
clean_data %>%
  group_by(high_blame) %>%
  summarise(
    I_apologize_alone_2 = sum(prefer_feelings_youalone, na.rm = TRUE),
    Neither_2           = sum(prefer__Neither_I_nor_2, na.rm = TRUE),
    n                   = n()
  )
```

```
## # A tibble: 2 x 4
##   high_blame I_apologize_alone_2 Neither_2     n
##        <dbl>               <dbl>     <dbl> <int>
## 1          0                  14        21    35
## 2          1                   7         3    10
```

Section C: Verify Levels of Variables

Verify that categorical and grouping variables have the expected levels and coding before analysis.

Binary Choice Variables

Original text responses:

```r
table(clean_data$outcome_binary1, useNA = "ifany")
```

```
##
## I apologize first, then ${e://Field/initials} apologizes.
##                                                       35
##           Neither I nor ${e://Field/initials} apologizes.
##                                                       10
```

```r
table(clean_data$outcome_binary2, useNA = "ifany")
```

```
##
## I apologize first, but ${e://Field/initials} does not apologize after that.
##                                                                         21
##                             Neither I nor ${e://Field/initials} apologizes.
##                                                                         24
```

Recoded binary indicators:

```r
table(clean_data$prefer_I_apologize_first, useNA = "ifany")
```

```
##
##  0  1
## 10 35
```

```r
table(clean_data$prefer_Neither_I_nor, useNA = "ifany")
```

```
##
##  0  1
## 35 10
```

```r
table(clean_data$prefer_feelings_youalone, useNA = "ifany")
```

```
##
##  0  1
## 24 21
```

```r
table(clean_data$prefer__Neither_I_nor_2, useNA = "ifany")
```

```
##
##  0  1
## 21 24
```

Self-Blame Grouping Variable

Distribution of original self-blame score:

```r
summary(clean_data$blame_1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   30.00   40.00   38.09   50.00   75.00
```

Gender Variables

```r
table(clean_data$sex, useNA = "ifany")
```

```
##
## Female   Male
##     26     19
```

```r
table(clean_data$target_sex, useNA = "ifany")
```

```
##
## Female   Male
##     25     20
```

Section D: Missing Data Handling

Before running any analyses, the data were organized into three stages: raw, processed, and clean. This structure is used to keep data-cleaning steps transparent and to clearly separate data quality decisions from analysis-related decisions.

Raw → Processed Data

The transition from the raw dataset to the processed dataset is documented in Section A (Sanity-check ranges and missingness). At this stage, the data were screened for obvious validity issues by confirming that scale-based variables fell within their expected ranges and that demographic values were plausible (e.g., excluding ages over 100).

Several variables were found not to align with the intended survey logic due to randomization order (e.g., emotional positivity rankings and binary choice variables). These variables were reconstructed directly from the original response values.

Importantly, no variables were removed at this stage; only observations with clearly invalid data were excluded. The resulting dataset is referred to as the **processed dataset**.

Processed → Clean Data

The clean dataset was constructed by applying the study's inclusion criteria, including survey completion, consent, and attention checks. At this stage, variables that were not required for analysis were removed, and analysis-specific variables were created.

The **clean dataset** represents the final analysis sample. It is used for all subsequent analyses, as well as for verifying the within- versus between-subject structure of the study and confirming that all key variables have the expected levels and coding prior to analysis.