AI vs Human Selection Process : Data Cleaning

Elham

2025-10-30

```r
#install libraries
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.5.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(ggplot2)
library(dplyr)
library(likert)
```

```
## Loading required package: xtable
##
## Attaching package: 'likert'
##
## The following object is masked from 'package:dplyr':
##
##     recode
```

```r
library(stats)
library(lavaan)
```

```
## This is lavaan 0.6-16
## lavaan is FREE software! Please report any bugs.
```

```r
library(psych)
```

```
##
## Attaching package: 'psych'
##
## The following object is masked from 'package:lavaan':
##
```

```
##     cor2cov
##
## The following objects are masked from ’package:ggplot2’:
##
##     %+%, alpha
```

```r
library(Hmisc)
```

```
##
## Attaching package: ’Hmisc’
##
## The following object is masked from ’package:psych’:
##
##     describe
##
## The following objects are masked from ’package:xtable’:
##
##     label, label<-
##
## The following objects are masked from ’package:dplyr’:
##
##     src, summarize
##
## The following objects are masked from ’package:base’:
##
##     format.pval, units
```

```r
library(broom)
```

```r
#upload the data

JAR_Social_Invitees_raw <- read_csv("JAR_Social_Invitees_synthetic.csv")
```

```
## Rows: 10 Columns: 42
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (39): Gender, Race, Education, Atten_AI, Atten_HR, Org_Attraction_1, Org...
## dbl  (3): Condition, Age, Attention Loop
##
## 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.
```

```r
#Clean Social_Invitees data set
JAR_Social_Invitees_Clean <- JAR_Social_Invitees_raw %>%
  #Remove rows with NA
  filter(!if_any(Intent_Engag_1:Intent_Engag_4,is.na))
```

```r
#clean the data
###########################################################
#convert character values to factor

Agree_Disagree_Levels = c("Strongly disagree",
                                      "Somewhat disagree",
                                      "Neither agree nor disagree",
                                      "Somewhat agree",
                                      "Strongly agree")

Likely_Unlikely_Levels = c("Very unlikely",
                           "Somewhat unlikely",
                           "Neutral",
                           "Somewhat likely",
                           "Very likely")

Famailiar_Unfamiliar_Levels = c("Not familiar at all",
                                "Not so familiar",
                                "Somewhat familiar",
                                "Very familiar",
                                "Extremely Familiar")

#for reverse coding
Agree_Disagree_Levels_Trust_4 = c("Strongly agree",
                                  "Somewhat agree",
                                  "Neither agree nor disagree",
                                  "Somewhat disagree",
                                  "Strongly disagree")


# change data type to factor, create new columns as numeric

JAR_Social_Invitees <- JAR_Social_Invitees_Clean %>%
  mutate(across(Org_Attraction_1:Trust_3,
                ~ factor(., levels = Agree_Disagree_Levels))) %>%
  mutate(across(Trust_4,
                ~ factor(., levels = Agree_Disagree_Levels_Trust_4))) %>%
  mutate(across(Communication_1:Consistency_HR_3,
                ~ factor(., levels = Agree_Disagree_Levels))) %>%
  mutate(across(Intent_Engag_1:Intent_Engag_4,
                ~ factor(., levels = Likely_Unlikely_Levels))) %>%
  mutate(across(AI_Knowledge_Experei_1:AI_Knowledge_Experei_5,
                ~ factor(., levels = Famailiar_Unfamiliar_Levels))) %>%
  mutate(across(c(Org_Attraction_1:Intent_Engag_4,
                  AI_Knowledge_Experei_1:AI_Knowledge_Experei_5),
                ~ as.numeric(.),
                .names = "{.col}_num"))

###########################################################
#Create Age as numeric
#age
JAR_Social_Invitees <- JAR_Social_Invitees %>%
  mutate(Age = as.numeric(Age))
###########################################################
```

```
#Demographic Information (binary and factor version)

Education_Levels = c("Less than high school",
                               "High school degree or equivalent",
                               "Some college (if currently an undergraduate student, select this option)
                               "Associate (2 year) degree",
                               "Bachelor's (4 year) degree",
                               "Some graduate school",
                               "Master's degree",
                               "Professional degree (e.g., JD, MD)",
                               "Doctorate (PhD)")


#factor in right order, then numeric, then if else greater than or equal
JAR_Social_Invitees <- JAR_Social_Invitees %>%
  # Gender
  mutate(Gender_fct = factor(Gender),
         Gender_female = if_else(Gender == "Female", 1, 0),
         Gender_male = if_else(Gender == "Male", 1, 0),
  # Race
         Race_fct = factor(Race),
        Race_white = if_else(Race == "White or European American", 1, 0),
  # Education (bachelor degree or higher)
        Education_fct = factor(Education,
                               levels = Education_Levels,
                               labels = Education_Levels),
        Education_num = as.numeric(Education_fct),
        Education_college = if_else(Education_num >= "3", 1, 0))

# report
JAR_Social_Invitees %>%
  dplyr::select(Gender_fct,
         Race_fct,
         Education_fct,
         Education_num,
         Education_college) %>%
map(table)
```

```
## $Gender_fct
##
##                                   Female                                    Male
##                                        2                                       3
## Non-binary / Genderqueer / Gender fluid                       Prefer not to say
##                                        3                                       2
##
## $Race_fct
##
##    Asian or Asian American  Black or African American
##                          2                         1
##     Latino/a/x or Hispanic             Middle Eastern
##                          1                         1
##          Prefer not to say White or European American
##                          1                         4
##
## $Education_fct
##
```

```
##                                                     Less than high school
##                                                                         1
##                                         High school degree or equivalent
##                                                                         0
## Some college (if currently an undergraduate student, select this option)
##                                                                         1
##                                                Associate (2 year) degree
##                                                                         1
##                                               Bachelor’s (4 year) degree
##                                                                         2
##                                                     Some graduate school
##                                                                         0
##                                                          Master’s degree
##                                                                         3
##                                       Professional degree (e.g., JD, MD)
##                                                                         0
##                                                          Doctorate (PhD)
##                                                                         2
##
## $Education_num
##
## 1 3 4 5 7 9
## 1 1 1 2 3 2
##
## $Education_college
##
## 0 1
## 1 9
```

```r
#number of Female and Male based on condition
Female_Male_Count <- JAR_Social_Invitees %>%
  group_by(Condition) %>%
  summarise(
    Female_Count = sum(Gender_fct == "Female", na.rm = TRUE),
    Male_Count = sum(Gender_fct == "Male", na.rm = TRUE))

#number of "White or European American" race based on condition
White_European_Count <- JAR_Social_Invitees %>%
  group_by(Condition) %>%
  summarise(
    White_European_Count = sum(Race_fct == "White or European American", na.rm = TRUE))

#number of "bachelor degree or higher" based on condition
Education_Count <- JAR_Social_Invitees %>%
  group_by(Condition) %>%
  summarise(
    Education_Count = sum(Education_num >= "3", na.rm = TRUE))

#count the number of observations for each condition
Condition_Count <- JAR_Social_Invitees %>%
  group_by(Condition) %>%
  count(Condition)
```

```r
#Attention binary
JAR_Social_Invitees <- JAR_Social_Invitees %>%
# Attention AI binary
  mutate(Attention_AI_binary = if_else(Atten_AI == "Personality based on vocal tone, facial expressions

# Attention HR binary
         Attention_HR_binary = if_else(Atten_HR == "Communication, interpersonal skills, and job-relatec

#Attention loop binary
JAR_Social_Invitees <- JAR_Social_Invitees %>%
  mutate(
    Attention_loop_AI_binary = if_else(
      Condition == 1 & `Attention Loop` == 1, 1, 0),
    Attention_loop_HR_binary = if_else(
      Condition == 2 & `Attention Loop` == 1, 1, 0)
  )

# NOTE: Create an Attention Score variable that combines all of the Attention metrics
JAR_Social_Invitees <- JAR_Social_Invitees %>%
  mutate(Attention_AI_score = Attention_AI_binary + Attention_loop_AI_binary,
         Attention_HR_score = Attention_HR_binary + Attention_loop_HR_binary)


JAR_Social_Invitees %>%
  dplyr::select(Attention_AI_binary,
         Attention_HR_binary,
         Attention_loop_AI_binary,
         Attention_loop_HR_binary,
         Attention_AI_score,
         Attention_HR_score) %>%
  map(table)
```

```
## $Attention_AI_binary
##
## 0 1
## 6 4
##
## $Attention_HR_binary
##
## 0 1
## 6 4
##
## $Attention_loop_AI_binary
##
## 0 1
## 7 3
##
## $Attention_loop_HR_binary
##
## 0 1
## 8 2
##
## $Attention_AI_score
```

```
##
## 0 1 2
## 5 3 2
##
## $Attention_HR_score
##
## 0 1 2
## 6 2 2
```

```
JAR_Social_Invitees %>%
  group_by(Condition) %>%
  count(`Attention Loop`)
```

```
## # A tibble: 6 x 3
## # Groups:   Condition [2]
##   Condition ‘Attention Loop‘     n
##       <dbl>            <dbl> <int>
## 1         1                1     3
## 2         1                2     1
## 3         1                3     2
## 4         2                1     2
## 5         2                2     1
## 6         2                3     1
```

```
############################################################
```

```
#factor analysis


#on whole data set##############################
Cfa_model_all_data <- '
  OrgAttraction_cfa =~ Org_Attraction_1_num + Org_Attraction_2_num + Org_Attraction_3_num + Org_Attracti
  Trust_cfa =~ Trust_1_num + Trust_2_num + Trust_3_num + Trust_4_num
  Communication_cfa =~ Communication_1_num + Communication_2_num + Communication_3_num + Communication_4
  ChancePerform_cfa =~ Chance_Perform_1_num + Chance_Perform_2_num + Chance_Perform_3_num + Chance_Perfo
  Consistency_cfa =~ Consistency_AI_1_num + Consistency_AI_2_num + Consistency_AI_3_num + Consistency_HR
  IntentEngage_cfa =~ Intent_Engag_1_num + Intent_Engag_2_num + Intent_Engag_3_num + Intent_Engag_4_num
  AI_Knowldge_cfa =~ AI_Knowledge_Experei_1_num + AI_Knowledge_Experei_2_num + AI_Knowledge_Experei_3_nu

JAR_Social_Invitees %>%
  dplyr::select(Org_Attraction_1_num:
                AI_Knowledge_Experei_5_num) %>%
  cfa(model = Cfa_model_all_data, missing = "fiml") %>%
  summary()
```

```
## Warning in lav_data_full(data = data, group = group, cluster = cluster, : lavaan WARNING: small numbe
##   nobs = 10 nvar = 34
```

```
## Warning in lav_mvnorm_missing_h1_estimate_moments(Y = X[[g]], wt = WT[[g]], : lavaan WARNING:
##     The smallest eigenvalue of the EM estimated variance-covariance
##     matrix (Sigma) is smaller than 1e-05; this may cause numerical
##     instabilities; interpret the results with caution.
## Warning in lavaan::lavaan(model = Cfa_model_all_data, data = ., missing = "fiml", : lavaan WARNING:
##     the optimizer warns that a solution has NOT been found!

## lavaan 0.6.16 did NOT end normally after 213 iterations
## ** WARNING ** Estimates below are most likely unreliable
##
##   Estimator                                         ML
##   Optimization method                           NLMINB
##   Number of model parameters                       123
##
##   Number of observations                            10
##   Number of missing patterns                         6
##
##
## Parameter Estimates:
##
##   Standard errors                             Standard
##   Information                                 Observed
##   Observed information based on                Hessian
##
## Latent Variables:
##                        Estimate  Std.Err  z-value  P(>|z|)
##   OrgAttraction_cfa =~
##     Org_Attrctn_1_        1.000
##     Org_Attrctn_2_        1.308       NA
##     Org_Attrctn_3_        1.059       NA
##     Org_Attrctn_4_        1.428       NA
##     Org_Attrctn_5_        1.233       NA
##     Org_Attrctn_6_        1.466       NA
##   Trust_cfa =~
##     Trust_1_num           1.000
##     Trust_2_num           1.236       NA
##     Trust_3_num           1.473       NA
##     Trust_4_num           1.126       NA
##   Communication_cfa =~
##     Communctn_1_nm        1.000
##     Communctn_2_nm        1.283       NA
##     Communctn_3_nm        0.747       NA
##     Communctn_4_nm        0.989       NA
##     Communctn_5_nm        1.464       NA
##   ChancePerform_cfa =~
##     Chnc_Prfrm_1_n        1.000
##     Chnc_Prfrm_2_n        0.986       NA
##     Chnc_Prfrm_3_n        1.441       NA
##     Chnc_Prfrm_4_n        1.291       NA
##   Consistency_cfa =~
##     Cnsstncy_AI_1_        1.000
##     Cnsstncy_AI_2_        1.225       NA
##     Cnsstncy_AI_3_        0.878       NA
##     Cnsstncy_HR_1_        1.378       NA
##     Cnsstncy_HR_2_        0.566       NA
##     Cnsstncy_HR_3_        1.596       NA
##   IntentEngage_cfa =~
##     Intnt_Engg_1_n        1.000
##     Intnt_Engg_2_n        1.303       NA
##     Intnt_Engg_3_n        1.201       NA
##     Intnt_Engg_4_n        1.458       NA
##   AI_Knowldge_cfa =~
##     AI_Knwldg_E_1_        1.000
##     AI_Knwldg_E_2_        1.203       NA
##     AI_Knwldg_E_3_        1.073       NA
##     AI_Knwldg_E_4_        1.079       NA
##     AI_Knwldg_E_5_        0.999       NA
##
## Covariances:
##                        Estimate  Std.Err  z-value  P(>|z|)
##   OrgAttraction_cfa ~~
##     Trust_cfa             1.543       NA
##     Communicatn_cf        1.302       NA
##     ChancePrfrm_cf        1.691       NA
##     Consistency_cf        1.341       NA
##     IntentEngag_cf        1.470       NA
##     AI_Knowldge_cf        1.352       NA
##   Trust_cfa ~~
##     Communicatn_cf        1.651       NA
##     ChancePrfrm_cf        1.317       NA
##     Consistency_cf        1.297       NA
##     IntentEngag_cf        1.244       NA
##     AI_Knowldge_cf        1.208       NA
##   Communication_cfa ~~
##     ChancePrfrm_cf        1.348       NA
##     Consistency_cf        1.621       NA
##     IntentEngag_cf        1.346       NA
##     AI_Knowldge_cf        1.466       NA
##   ChancePerform_cfa ~~
##     Consistency_cf        1.485       NA
##     IntentEngag_cf        1.510       NA
##     AI_Knowldge_cf        1.477       NA
##   Consistency_cfa ~~
##     IntentEngag_cf        1.411       NA
##     AI_Knowldge_cf        1.534       NA
##   IntentEngage_cfa ~~
##     AI_Knowldge_cf        1.343       NA
##
## Intercepts:
##                    Estimate  Std.Err  z-value  P(>|z|)
##    .Org_Attrctn_1_    0.938       NA
##    .Org_Attrctn_2_    0.549       NA
##    .Org_Attrctn_3_    0.419       NA
##    .Org_Attrctn_4_    0.381       NA
##    .Org_Attrctn_5_    0.367       NA
##    .Org_Attrctn_6_    0.405       NA
##    .Trust_1_num       1.286       NA
##    .Trust_2_num       0.875       NA
##    .Trust_3_num       0.027       NA
##    .Trust_4_num       0.610       NA
##    .Communctn_1_nm    1.054       NA
##    .Communctn_2_nm   -0.360       NA
##    .Communctn_3_nm    0.955       NA
##    .Communctn_4_nm    0.497       NA
##    .Communctn_5_nm    0.013       NA
##    .Chnc_Prfrm_1_n    1.434       NA
##    .Chnc_Prfrm_2_n    0.208       NA
##    .Chnc_Prfrm_3_n    0.182       NA
##    .Chnc_Prfrm_4_n    0.573       NA
##    .Cnsstncy_AI_1_    0.793       NA
##    .Cnsstncy_AI_2_    0.797       NA
##    .Cnsstncy_AI_3_    0.583       NA
##    .Cnsstncy_HR_1_    0.316       NA
##    .Cnsstncy_HR_2_    0.908       NA
##    .Cnsstncy_HR_3_    0.153       NA
##    .Intnt_Engg_1_n    1.551       NA
##    .Intnt_Engg_2_n    0.270       NA
##    .Intnt_Engg_3_n    0.525       NA
##    .Intnt_Engg_4_n    0.202       NA
##    .AI_Knwldg_E_1_    1.086       NA
##    .AI_Knwldg_E_2_    0.175       NA
##    .AI_Knwldg_E_3_    0.590       NA
##    .AI_Knwldg_E_4_    0.001       NA
##    .AI_Knwldg_E_5_    0.644       NA
##     OrgAttractn_cf    0.000
##     Trust_cfa         0.000
##     Communicatn_cf    0.000
##     ChancePrfrm_cf    0.000
##     Consistency_cf    0.000
##     IntentEngag_cf    0.000
##     AI_Knowldge_cf    0.000
##
## Variances:
##                    Estimate  Std.Err  z-value  P(>|z|)
##    .Org_Attrctn_1_    2.539       NA
##    .Org_Attrctn_2_    1.849       NA
##    .Org_Attrctn_3_    1.136       NA
##    .Org_Attrctn_4_    1.554       NA
##    .Org_Attrctn_5_    1.916       NA
##    .Org_Attrctn_6_    1.235       NA
##    .Trust_1_num       1.847       NA
##    .Trust_2_num       2.115       NA
##    .Trust_3_num       1.993       NA
##    .Trust_4_num       1.885       NA
##    .Communctn_1_nm    2.286       NA
##    .Communctn_2_nm    1.590       NA
##    .Communctn_3_nm    2.726       NA
##    .Communctn_4_nm    2.501       NA
##    .Communctn_5_nm    1.899       NA
##    .Chnc_Prfrm_1_n    2.818       NA
##    .Chnc_Prfrm_2_n    2.420       NA
##    .Chnc_Prfrm_3_n    1.622       NA
##    .Chnc_Prfrm_4_n    0.632       NA
##    .Cnsstncy_AI_1_    2.079       NA
##    .Cnsstncy_AI_2_    1.970       NA
##    .Cnsstncy_AI_3_    1.986       NA
##    .Cnsstncy_HR_1_    1.659       NA
##    .Cnsstncy_HR_2_    1.434       NA
##    .Cnsstncy_HR_3_    1.358       NA
##    .Intnt_Engg_1_n    1.117       NA
##    .Intnt_Engg_2_n    1.176       NA
##    .Intnt_Engg_3_n    1.574       NA
##    .Intnt_Engg_4_n    0.947       NA
##    .AI_Knwldg_E_1_    0.936       NA
##    .AI_Knwldg_E_2_    1.801       NA
##    .AI_Knwldg_E_3_    1.183       NA
##    .AI_Knwldg_E_4_    0.548       NA
##    .AI_Knwldg_E_5_    1.675       NA
##     OrgAttractn_cf    1.676       NA
##     Trust_cfa         1.691       NA
##     Communicatn_cf    1.507       NA
##     ChancePrfrm_cf    1.501       NA
##     Consistency_cf    1.439       NA
##     IntentEngag_cf    1.712       NA
##     AI_Knowldge_cf    1.318       NA
```

```
#mean of each variable based on condition
#hist for those means
#########################################################
JAR_Social_Invitees <- JAR_Social_Invitees %>%
  rowwise() %>%
  mutate(Org_Attraction =
           mean(c_across(c(Org_Attraction_1_num:
                           Org_Attraction_6_num)),
              na.rm = TRUE),
         Trust =
           mean(c_across(c(Trust_1_num:
                           Trust_4_num)),
              na.rm = TRUE),
         Communication =
           mean(c_across(c(Communication_1_num:
                           Communication_5_num)),
              na.rm = TRUE),
         Chance_Perform =
           mean(c_across(c(Chance_Perform_1_num:
                           Chance_Perform_4_num)),
              na.rm = TRUE),
         Consistency =
           mean(c_across(c(Consistency_AI_1_num:
                           Consistency_HR_3_num)),
              na.rm = TRUE),
         Intent =
           mean(c_across(c(Intent_Engag_1_num:
                           Intent_Engag_4_num)),
              na.rm = TRUE),
         AI_Knowledge =
           mean(c_across(c(AI_Knowledge_Experei_1_num:
                           AI_Knowledge_Experei_5_num)),
              na.rm = TRUE))
```

```r
JAR_Social_Invitees_means <- JAR_Social_Invitees %>%
  group_by(Condition) %>%
  summarise(
    mean_org_attraction = mean(Org_Attraction, na.rm = TRUE),
    mean_trust = mean(Trust, na.rm = TRUE),
    mean_communication = mean(Communication, na.rm = TRUE),
    mean_chance_perform = mean(Chance_Perform, na.rm = TRUE),
    mean_consistency = mean(Consistency, na.rm = TRUE),
    mean_intent = mean(Intent, na.rm = TRUE),
    mean_AI_knowledge = mean(AI_Knowledge, na.rm = TRUE),
    sd_org_attraction = sd(Org_Attraction, na.rm = TRUE),
    sd_trust = sd(Trust, na.rm = TRUE),
    sd_communication = sd(Communication, na.rm = TRUE),
    sd_chance_perform = sd(Chance_Perform, na.rm = TRUE),
    sd_consistency = sd(Consistency, na.rm = TRUE),
    sd_intent = sd(Intent, na.rm = TRUE),
    sd_AI_knowledge = sd(AI_Knowledge, na.rm = TRUE))
```

```r
# export data
save(JAR_Social_Invitees, file = "JAR_Social_Invitees_clean.RData")
write_csv(JAR_Social_Invitees, file = "JAR_Social_Invitees_clean.csv")
```