Brief Exploration of ELS:2002 Questionnaries as Rasch Items to Compare Security Measures in U.S. High Schools

Adapted From: Misbehavior, Suspensions and Security Measures in High School: Racial/Ethnic and Gender Differences (Finn & Servoss, 2014)

Joshua Sinamo

Appendix Overview

I. Preface

This is a mini assessment comparing two IRT (or Rasch) models involving two different sets of questions. These models are purposed for calculation, comparison and indexing of security measures across U.S. high schools. The only difference between the two models are the set of questions that are used to measure security level. Finn & Servoss (2014) in their paper: Misbehavior, Suspensions and Security Measures in High School: Racial/Ethnic and Gender Differences, proposed a set of 7 questions to measure security level. In my perspective as an independent analyst with no domain knowledge in school security measurements, their set of questions seems to be qualitatively redundant and therefore I experimented with an alternative set by selecting questions that seemingly unrelated with each other. Finn & Servoss used the Education Longitudinal Study (NCES, 2002) survey data, which is available via NCES website. In ELS:2002 survey, school administrators were asked about their schools' demographics, academic performance, and among other things security measures in their institution.

II. List of ELS:2002 Admin Questionnaire (security questions)

38. During this school year (2001-2002), is it a practice of your school to do the following? (If your school changed its practices in the middle of the school year, please answer regarding your most recent practice.) (MARK ONE RESPONSE ON EACH LINE) (Yes; No)

A. Control access to buildings during school hours
B. Control access to grounds during school hours
C. Require students pass through metal detector
D. Random metal detector checks on students
E. Close campus for students during lunch
F. Random dog sniffs to check for drugs
G. Random sweeps for contraband
H. Require drug testing for any students
I. Require students to wear uniforms
J. Enforce strict dress code
K. Require clear book bags/ban book bags
L. Require students to wear badges/picture ID
M. Require faculty/staff to wear badges/picture ID
N. Use security cameras to monitor school
O. Telephones in most classrooms
P. Emergency call button in classrooms

39. Which of the following does your school do to involve or help parents deal with school discipline issues? (MARK ONE RESPONSE ON EACH LINE) (Yes; No)

A. Process to get parent input on discipline policies
B. Training parents to deal with problem behavior
C. Program involves parents in school discipline

40. During the 2001-2002 school year, did your school regularly use paid law enforcement or security services at school at the following times? (MARK ONE RESPONSE ON EACH LINE) (Yes; No)

A. Use paid security at any time during school hours
B. Use paid security as students arrive or leave
C. Use paid security at school activities
D. Use paid security outside of school hours/activities
E. Use paid security at other time

III. Methods

a. Sources

- On Security Questions and Literature

  Finn, Jeremy D. & Servoss, Timothy J. (2014). Misbehavior, Suspensions, and Security Measures in High Schools: Racial/Ethnic and Gender Differences. Journal of Applied Research on Children Volume 5 Issue 2 https://digitalcommons.library.tmc.edu/cgi/viewcontent.cgi?article=1211&context=childrenatrisk

- On Item Response Theory using LMER package in R

  de Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. Journal of Statistical Software, 39(12), 1-28.

- On Bias-Corrected & Acceleraated Confidence Interval - Source

  Efron, B., Tibshirani, R. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. Statist. Sci. 1 (1986), no. 1, 54-75. doi:10.1214/ss/1177013815. https://projecteuclid.org/euclid.ss/1177013815

- On Bias-Corrected & Accelerated Confidence Interval - R Implementation

  Kropko, J., & Harden, J. (2020). Beyond the Hazard Ratio: Generating Expected Durations from the Cox Proportional Hazards Model. British Journal of Political Science, 50(1), 303-320. doi:10.1017/S000712341700045X. https://cran.r-project.org/web/packages/coxed/coxed.pdf

- On lme4 Package

  Bates, D., Maechler, M., Bolker, B. Walker, S., Christensen, R., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., Green, P., Fox, J. (2020). Linear Mixed-Effects Models using 'Eigen' and S4. Comprehensive R Archive Network. https://cran.r-project.org/web/packages/lme4/lme4.pdf

- On ML vs REML

  Faraway, Julian J. "Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models, Second Edition." CRC Press, 2005, www.routledge.com/Extending-the-Linear-Model-with-R-Generalized-Linear-Mixed-Effects-and/Faraway/p/book/9781498720960.

b. Description

The model utilized to mimic Item Response Theory (or rasch model) for this paper is mixed-effect logistic regression via glmer in R. The process began by finding the proper nAGQ (number of adaptive Gauss-Hermite quadrature points as hyperparameter for the two models. More nAGQ will produce greater accuracy in the evaluation of log-likelihood at the expense of speed (Bates et.al., 2020). In making such comparison, I estimated the AIC, BIC, root of mean-squared error (RMSE), mean of random effects, and fixed effect coefficients for nAGQ = 1 to 20. In general, lower AIC, BIC, RMSE are better, but the main objective is to find the nAGQ value where these estimates stabilizes. The validity of using AIC and BIC (likelihood based measurements) stem from the fact that glmer (which uses maximum likelihood) was in use instead of lmer in R which uses restricted maximum likelihood estimation. Faraway (2016) stated that the likelihood of the two models which have differing fixed effects produced by REML estimation are not directly comparable. Upon finding the proper nAGQ where these coefficients stabilizes, I then commenced the analysis by comparing the indexing methods of the two models using Kendall's Rank Correlation test. Following that, I utilized resampling bootstrap to estimate the bias-corrected confidence intervals of the rasch indices (security measures coefficients produced by the two models) as well its distribution via density plots. This would give us the bigger picture of how many statistically significant comparison of rasch indices (school security measures) can be made using each model.

Results (computational details)

a. External Requirements

- What : Loading required libraries & dataset

```
# to clean up global environment, run this
rm(list = ls())
```

```
library(gridExtra)    ; library(knitr)
library(lme4)         ; library(tidyr)
library(dplyr)        ; library(ggplot2)
library(reshape2)     ; library(Kendall)
library(parallel)     ; library(gplots)
library(graphics)     ; library(dendextend)
library(grid)         ; library(directlabels)
library(RColorBrewer) ; library(coxed)
```

```
## Warning: package 'survival' was built under R version 3.6.2
```

```
dat <- read.csv("./Downloads/project/els_02_12_byf3pststu_v1_0.csv")
```

b. General Data Cleaning for All Analysis

- What : Data cleaning in accordance to NCES ELS 2002 codebook

```
dataSecurity <- dat %>% dplyr::select("F1SCH_ID", "BYA38A", "BYA38B", "BYA38C", "BYA38D", "BYA38E",
                                      "BYA38F", "BYA38G", "BYA38H", "BYA38I", "BYA38J", "BYA38K",
                                      "BYA38L", "BYA38M", "BYA38N", "BYA38O", "BYA38P", "BYA39A",
                                      "BYA39B", "BYA39C", "BYA40A", "BYA40B", "BYA40C", "BYA40D",
                                      "BYA40E")
dataSecurity <- unique(dataSecurity[dataSecurity$F1SCH_ID > 1000 &
                                    dataSecurity$BYA38A != -8 &
                                    dataSecurity$BYA38A != -7 &
                                    dataSecurity$BYA38A != -4, ])
names(dataSecurity)[2:25] <- substring(colnames(dataSecurity[,c(2:25)]), first = 4)
names(dataSecurity)[1] <- "school"
rownames(dataSecurity) <- NULL
dataSecurity[dataSecurity == -9] <- NA
kable(head(dataSecurity), caption="This is the snippet of data that we'll be working with")
```

This is the snippet of data that we’ll be working with

school	38A	38B	38C	38D	38E	38F	38G	38H	38I	38J	38K	38L	38M	38N	38O	38P	39A	39B	39C	40A	40B	40C	40D	40E
1011	1	1	0	0	0	0	0	1	0	0	0	0	0	0	1	1	1	0	0	1	1	0	0	0
1021	1	1	1	1	1	0	1	0	1	1	0	1	1	1	1	0	1	1	1	1	1	1	0	1
1022	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	1	0	0
1031	1	1	0	0	1	0	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
1033	1	1	0	0	1	0	0	0	0	0	0	0	0	1	1	0	1	0	0	1	1	1	1	0
1041	1	0	0	0	0	0	1	0	0	1	0	0	1	1	1	0	1	1	0	1	1	1	0	0

c. Finn & Servoss and Proposed Alternative Question Sets - Initialization

- What : Imitating Finn & Servoss' set of questions & transform the data into item-response set
- Why : Assess & compare the two models later
- How : Literature analysis from Finn & Servoss (2014)

Blue = Finn & Servoss' | Red = Alternative | Magenta = Both

38. During this school year (2001-2002), is it a practice of your school to do the following? (If your school changed its practices in the middle of the school year, please answer regarding your most recent practice.) (MARK ONE RESPONSE ON EACH LINE) (Yes; No)

A. Control access to buildings during school hours
B. Control access to grounds during school hours
C. Require students pass through metal detector
D. Random metal detector checks on students
E. Close campus for students during lunch
F. Random dog sniffs to check for drugs
G. Random sweeps for contraband
H. Require drug testing for any students
I. Require students to wear uniforms
J. Enforce strict dress code
K. Require clear book bags/ban book bags
L. Require students to wear badges/picture ID
M. Require faculty/staff to wear badges/picture ID
N. Use security cameras to monitor school
O. Telephones in most classrooms
P. Emergency call button in classrooms

39. Which of the following does your school do to involve or help parents deal with school discipline issues? (MARK ONE RESPONSE ON EACH LINE) (Yes; No)

A. Process to get parent input on discipline policies
B. Training parents to deal with problem behavior
C. Program involves parents in school discipline

40. During the 2001-2002 school year, did your school regularly use paid law enforcement or security services at school at the following times? (MARK ONE RESPONSE ON EACH LINE) (Yes; No)

A. Use paid security at any time during school hours
B. Use paid security as students arrive or leave
C. Use paid security at school activities
D. Use paid security outside of school hours/activities
E. Use paid security at other time

Previously, I determined that this set is qualitatively redundant because, for example, 38C and 38D seemingly covers the same things, also 38F and 3H in the same manner.

```
# FINN AND SERVOS SET OF QUESTIONS
FandS_set <- dataSecurity %>% select("school", "38C", "38D", "38H", "38G", "38N", "40A", "38F")
FandS_rasch <- FandS_set %>% tidyr::gather("item", "response", -school)
```

```
# MY ALTERNATIVE SET OF QUESTIONS
alt_set <- dataSecurity %>% select("school", "38H", "38K", "38C", "38E", "40E", "39C", "38J")
alt_rasch <- alt_set %>% tidyr::gather("item", "response", -school)
```

d. Hyperparameter Tuning - Computation

- What : Selecting the minimal nAGQ where the metrics start to stabilize; although the difference is minimal, I'm looking for precision. Also, catching failure to converge
- Why : Need appropriate nAGQ to have a more precise & unbiased model comparison (more nAGQ = longer computation)
- How : Comparison of RMSE, fixed effect coefficients & means of random effect coefficients w.r.t nAGQ

```
n = 20
RMSE_ori       <- rep(NA, n) ; RMSE_alt       <- rep(NA, n)
RANEF_mean_ori <- rep(NA, n) ; RANEF_mean_alt <- rep(NA, n)
FIXEF_ori      <- rep(NA, n) ; FIXEF_alt      <- rep(NA, n)
AIC_ori        <- rep(NA, n) ; AIC_alt        <- rep(NA, n)
BIC_ori        <- rep(NA, n) ; BIC_alt        <- rep(NA, n)

for (i in 1:n){

  tryCatch({rasch_lmer_ori_NAGQ <- glmer(response ~ 0 + item + (1|school),
                                         family = binomial, data = FandS_rasch, nAGQ = i-1)},
           warning = function(warn2) {cat("Finn & Servoss at nAGQ = ", i-1, "\n")
                                      warning(warn2)})

  tryCatch({rasch_lmer_alt_NAGQ <- glmer(response ~ 0 + item + (1|school),
                                         family = binomial, data = alt_rasch, nAGQ = i-1)},
           warning = function(warn2) {cat("Alternative at nAGQ    = ", i-1, "\n")
                                      warning(warn2)})

  RMSE_ori[i]  <- sqrt(mean(resid(rasch_lmer_ori_NAGQ)^2))
  RMSE_alt[i]  <- sqrt(mean(resid(rasch_lmer_alt_NAGQ)^2))
  FIXEF_ori[i] <- list(fixef(rasch_lmer_ori_NAGQ))
  FIXEF_alt[i] <- list(fixef(rasch_lmer_alt_NAGQ))
  RANEF_mean_ori[i] <- mean(ranef(rasch_lmer_ori_NAGQ)$school[["(Intercept)"]])
  RANEF_mean_alt[i] <- mean(ranef(rasch_lmer_alt_NAGQ)$school[["(Intercept)"]])
  AIC_ori[i]        <- AIC(rasch_lmer_ori_NAGQ) ; AIC_alt[i]        <- AIC(rasch_lmer_alt_NAGQ)
  BIC_ori[i]        <- BIC(rasch_lmer_ori_NAGQ) ; BIC_alt[i]        <- BIC(rasch_lmer_alt_NAGQ)

}
```

```
## Finn & Servoss at nAGQ =  6
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.00199317 (tol = 0.001, component 1)
```

```
## Finn & Servoss at nAGQ =  9
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.00108106 (tol = 0.001, component 1)
```

- Result : nAGQ = 6 and 9 failed to converge. We will exclude these two candidates for analysis

e. Hyperparameter Tuning - Visualization (AIC, BIC, RMSE, Mean of RanEf Intercept Coeffs.)

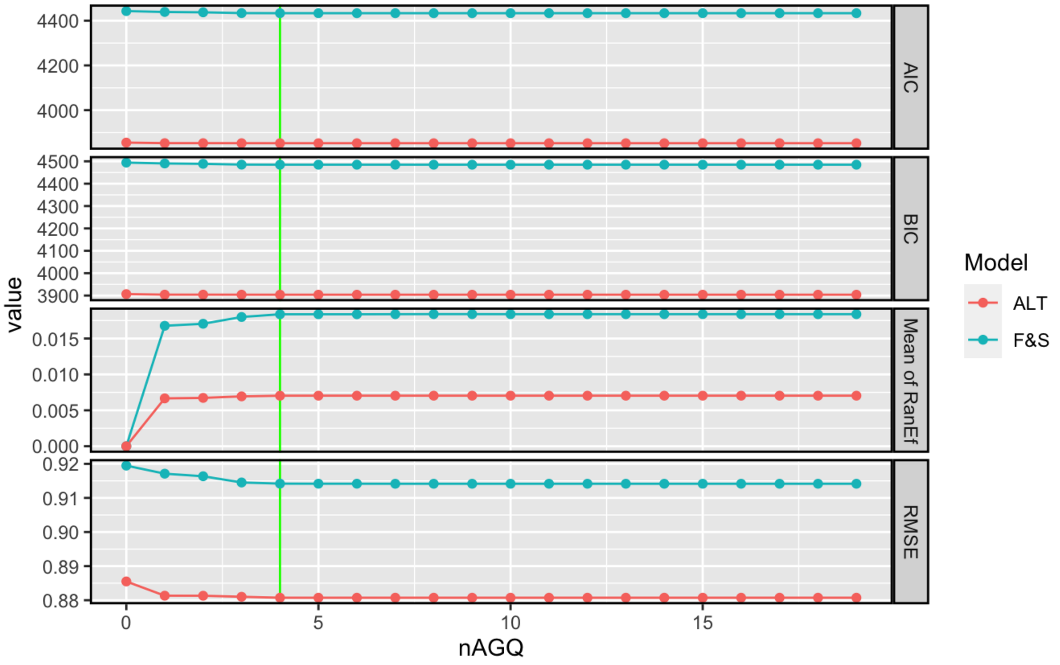
- What : Visualizing the results from nAGQ optimization above
- Why : To see what is the minimum value of nAGQ that can be considered as stable (1)
- How : Find the minimal value where there's no considerable fluctuation of the aforementioned measurements. GLMER ignores REML and uses ML, therefore it is legitimate to use AIC and BIC to compare models with different fixed effects

```
assessment_data <- data.frame("measurement" = rep(c("AIC", "BIC", "RMSE", "Mean of RanEf"),
                                                  each = 40),
                              "model" = rep(c("F&S", "ALT"), each = 20, times = 4),
                              "value" = c(c(AIC_ori), c(AIC_alt), c(BIC_ori), c(BIC_alt),
                                          c(RMSE_ori), c(RMSE_alt),
                                          c(RANEF_mean_ori), c(RANEF_mean_alt)),
                              "nAGQ" = rep(seq(from = 0, to = 19), times = 8))

ggplot(data = assessment_data, aes(x = nAGQ, y = value, color = model)) +
  geom_vline(xintercept = 4, color = "green") +
  facet_grid(measurement~., scales = "free") + geom_line() + geom_point() +
  labs(color    = "Model", caption = "Coefficients stabilizes at nAGQ >= 4",
       title    = "Selected Measurement vs # of Adaptive Gauss-Hermite Quadrature Points",
       subtitle =  "AIC, BIC, Mean of Ranef, and RMSE") +
  theme(panel.spacing     = unit(0.20, "lines"),
        panel.border     = element_rect(color = "black", fill = NA, size = 1),
        strip.background = element_rect(color = "black", size = 1),
        plot.caption     = element_text(hjust = 0.5, size = 9),
        plot.title       = element_text(size = 11),
        plot.subtitle    = element_text(size = 9))
```

Selected Measurement vs # of Adaptive Gauss-Hermite Quadrature Points

AIC, BIC, Mean of Ranef, and RMSE



Coefficients stabilizes at nAGQ >= 4

- Result :
  - Number of Minimal Gauss-Hermit Quadrature needed >= 4
  - My model has better AIC, BIC, RMSE, and mean of random intercepts are closer to 0 (although negligible)

f. Hyperparameter Tuning - Visualization (Fixed Effect Coefficients)

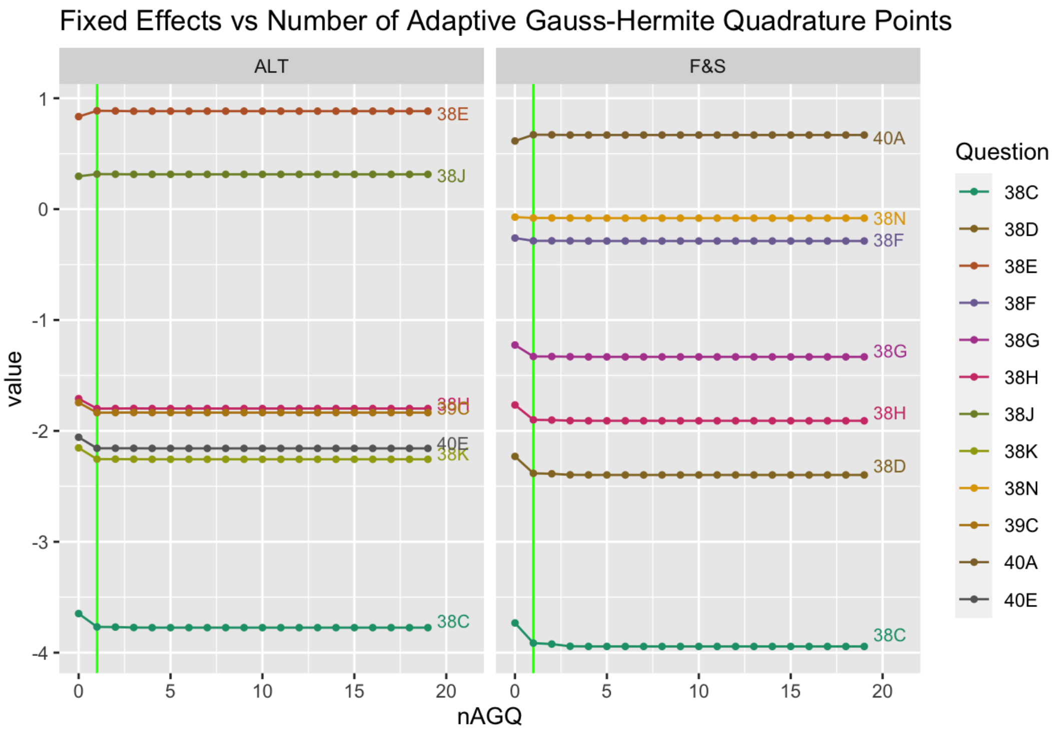
- What : Visualizing the fixed effect coefficients from nAGQ optimization above
- Why : To check what is the minimum value of nAGQ that can be considered as stable (2)
- How : Finding the minimal value where there’s no considerable fluctuation of the fixed effects coefficients

```
FIXEF_comb_ori<- FIXEF_ori[[1]]
FIXEF_comb_alt<- FIXEF_alt[[1]]
for(i in 2:20){
  FIXEF_comb_ori <- rbind(FIXEF_comb_ori, FIXEF_ori[[i]])
  FIXEF_comb_alt <- rbind(FIXEF_comb_alt, FIXEF_alt[[i]])
}

fixef_coef_ori0 <- data.frame(FIXEF_comb_ori);
fixef_coef_ori0$nAGQ <- seq(0,19); fixef_coef_ori0$model <- "F&S"
fixef_coef_ori <- melt(fixef_coef_ori0, id.vars = c("nAGQ", "model"))
fixef_coef_alt0 <- data.frame(FIXEF_comb_alt);
fixef_coef_alt0$nAGQ <- seq(0,19); fixef_coef_alt0$model <- "ALT"
fixef_coef_alt <- melt(fixef_coef_alt0, id.vars = c("nAGQ", "model"))

fixef_graph_data <- rbind(fixef_coef_ori, fixef_coef_alt)
fixef_graph_data$variable <- substring(fixef_graph_data$variable, first = 5)

ggplot(data = fixef_graph_data, aes(x = nAGQ, y = value, colour = variable)) +
  xlim(0,21) + geom_vline(xintercept = 1, color = "green") +
  facet_grid(~model, scales = "free") + geom_line() + geom_point(size = 1) +
  labs(color   = "Question",
       title   = "Fixed Effects vs Number of Adaptive Gauss-Hermite Quadrature Points ",
       caption = "Fixed effect coefficients stabilizes at nAGQ = 1 (Laplace Approximation) or above") +
  theme(plot.caption = element_text(hjust = 0.5)) +
  geom_dl(aes(label = variable, x = 19.5), method = list(dl.combine("last.points"), cex = 0.7)) +
  scale_color_manual(values = colorRampPalette(brewer.pal(name = "Dark2", n = 8))(12))
```



Fixed effect coefficients stabilizes at nAGQ = 1 (Laplace Approximation) or above

- Result :
  - Number of Minimal Gauss-Hermit Quadrature needed >= 1 from the prior results, I decided to proceed with nAGQ = 4
  - It seems that fixed effect coefficients of Finn & Servoss’ model more well spread, forming a more reasonable hierarchy of security measures than mine

g. Assessment of Indexing Methods between The Two Models

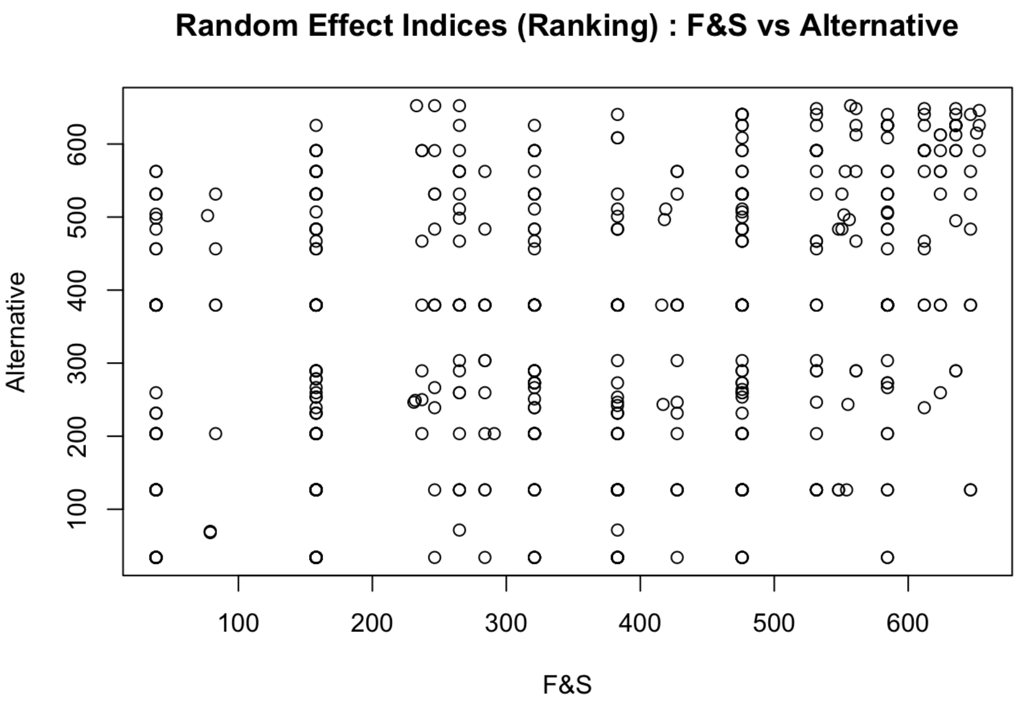
- What : Initializing the two rasch (or IRT) models and transforming the data into item-response format for assessment
- Why : To compare the indexing methods of the two models (my alternative model vs Finn & Servoss') and see whether the two models index schools security "level" differently
- How : Kendall rank test and visualization of the two set of indices

```
rasch_lmer_FandS <- glmer(response ~ 0 + item + (1|school), family = binomial,
                          data = FandS_rasch, nAGQ = 4)
rasch_lmer_alt <- glmer(response ~ 0 + item + (1|school), family = binomial,
                        data = alt_rasch, nAGQ = 4)

cor.test(ranef(rasch_lmer_FandS)$school[["(Intercept)"]],
         ranef(rasch_lmer_alt)$school[["(Intercept)"]], method = "kendal")
```

```
##
##  Kendall's rank correlation tau
##
## data:  ranef(rasch_lmer_FandS)$school[["(Intercept)"]] and ranef(rasch_lmer_alt)$school[["(Intercept)"]]
## z = 7.8767, p-value = 3.361e-15
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.2362383
```

```
plot(rank(ranef(rasch_lmer_FandS)$school[["(Intercept)"]]),
     rank(ranef(rasch_lmer_alt)$school[["(Intercept)"]]),
     xlab = "F&S", ylab = "Alternative",
     main = "Random Effect Indices (Ranking) : F&S vs Alternative")
```



- Result :
  - Despite the plot showing that the relationship looks undefined, we can reject the null hypothesis that the ranking system between the two models is different. Consequently, there's a low positive correlation between them

h. Model Comparison via Bootstrapping - Initialization

- What : Parallel nonparametric bootstrapping
- Why : To calculate and compare the expected length of the 95% CIs, means of random effect coefficients, and time spent for bootstrapping the two models
- How : Resampling bootstrap using parallel processing in R using “Parallel” package. I’m not using bootMer because we need to sample schools not school-answers (e.g. there’s a data format modification step needed to be executed after resampling but before bootMer execute/update the model.

```
# Initialize parallel clusters
cl <- makeCluster(detectCores()-1)
invisible(clusterEvalQ(cl, c(library(lme4), library(dplyr), library(tidyr))))
clusterSetRNGStream(cl)
```

```
ori_rasch_wide <- dataSecurity %>% dplyr::select(school, "38C", "38D", "38H", "38G", "38N", "40A", "38F")
alt_rasch_wide <- dataSecurity %>% dplyr::select(school, "38H", "38K", "38C", "38E", "40E", "39C", "38J")

ori_rasch <- ori_rasch_wide %>% tidyr::gather("item", "response", -school)
rasch_lmer <- glmer(response ~ 0 + item + (1 | school),
                             family = binomial, data = ori_rasch, nAGQ = 4)

resamplerWide = function(i, questionSet){
  whichrows <- sample(1L:nrow(questionSet), nrow(questionSet), replace = TRUE)
  bootspl <- questionSet[whichrows,] %>% tidyr::gather("item", "response", -school)
  refitted.mod <- update(rasch_lmer, data = bootspl)
  ranef(refitted.mod)
}

clusterExport(cl, c("ori_rasch_wide", "alt_rasch_wide", "rasch_lmer"))
```

```
bootreps <- 500; cat("B = ", bootreps, "\n")
```

```
## B =  500
```

```
start.time1 <- Sys.time()
bootstats_ori  <- parSapply(cl, 1:bootreps, FUN = resamplerWide, questionSet = ori_rasch_wide)
end.time1 <- Sys.time(); cat("Finn & Servoss : Processing Time =  ", end.time1 - start.time1, " min\n")
```

```
## Finn & Servoss : Processing Time =   49.38469  min
```

```
start.time2 <- Sys.time()
bootstats_alt  <- parSapply(cl, 1:bootreps, FUN = resamplerWide, questionSet = alt_rasch_wide)
end.time2 <- Sys.time(); cat("Alternative    : Processing Time =  ", end.time2 - start.time2, " min\n")
```

```
## Alternative    : Processing Time =   45.02447  min
```

```
stopCluster(cl)
```

- Note : Finn & Servoss’ model takes more time to converge

i. Model Comparison via Bootstrapping - Data Cleaning & Confidence Interval

- What : Data cleaning of the bootstrapped data and calculations of 95% bias-corrected & accelerated CI length
- Why : The bootsrapped data in replicate is in a form of list of multiple dataframes, need to combine them into one to then compute the confidence intervals of our random intercepts
- How : Simple dplyr group-by and summarise for confidence interval using bias-corrected and accelerated confidence intervals (DiCiccio & Efron, 1996). Implementation in R (bca) from Kropko & Harden (package = “coxed”; 2014)

```
# "pasting" other bootstrap results into the data
dataCleaning <- function(bootstrappedData) {

  lmerBootContainer  <- data.frame(bootstrappedData[1])
  lmerBootContainer$index <- rownames(lmerBootContainer)
  lmerBootContainer$rep  <- 1
  colnames(lmerBootContainer)[1] <- "value"
  for (i in 2:bootreps){
    lmerBootAdd <- data.frame(bootstrappedData[i])
    lmerBootAdd$index <- rownames(lmerBootAdd)
    lmerBootAdd$rep <- i
    colnames(lmerBootAdd)[1] <- "value"
    lmerBootContainer <- rbind(lmerBootContainer, lmerBootAdd)
  }

  # statistics for the bootstrap coefficients
  lmerBoot  <- lmerBootContainer %>% dplyr::group_by(index) %>%
                  dplyr::summarise("mean" = mean(value, na.rm = TRUE),
                                   "LB_BCA_CI" = bca(value, conf.level = 0.95)[1],
                                   "UB_BCA_CI" = bca(value, conf.level = 0.95)[2],
                                   "CI_length" = UB_BCA_CI - LB_BCA_CI)
  return(list(lmerBootContainer, lmerBoot))
}
```

```
ori_boot <- dataCleaning(bootstats_ori)
alt_boot <- dataCleaning(bootstats_alt)
```

```
mean(alt_boot[[2]]$CI_length)/ mean(ori_boot[[2]]$CI_length) - 1
```

```
## [1] 0.04561288
```

- Results : On average, security indices produced using Finn & Servoss’ model is 4.6% more precise than mine in a sense that on average their CIs are approx. 4.6% tighter

j. Density of Rasch Indices

- What : Comparing the density plots of Rasch indices (random intercepts) from the two models
- How : Density plots of Rasch indices two models

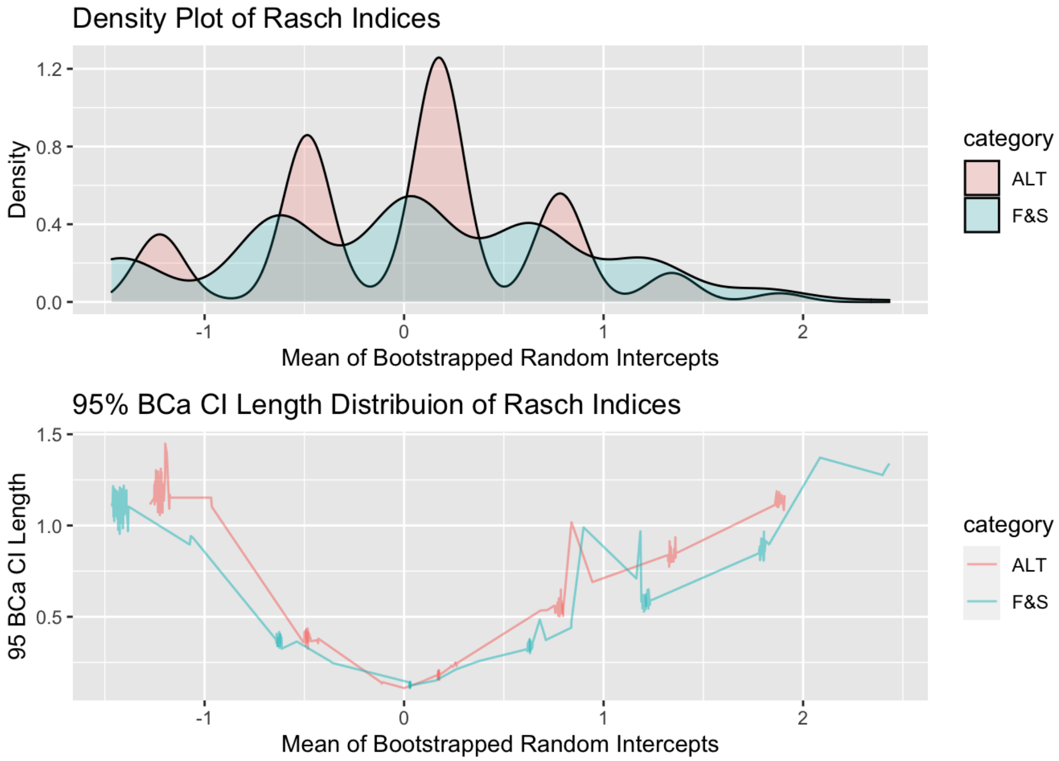
```
ori_boot[[2]]$category <- "F&S"
alt_boot[[2]]$category <- "ALT"

boot_all <- rbind(ori_boot[[2]], alt_boot[[2]])

plot1 <- ggplot(boot_all, aes(x = mean, fill = category)) +
  geom_density(alpha = .2) +
  ggtitle("Density Plot of Rasch Indices") +
  labs(x = "Mean of Bootstrapped Random Intercepts", y = "Density")

plot2 <- ggplot(boot_all, aes(x = mean, y = CI_length, color = category)) +
  geom_line(alpha = .5) +
  ggtitle("95% BCa CI Length Distribuion of Rasch Indices") +
  labs(x = "Mean of Bootstrapped Random Intercepts", y = "95 BCa CI Length")

grid.arrange(plot1, plot2)
```



- Results : It seems that Finn & Servoss' model enables more statistically significant comparisons between schools security indices because my model's random intercepts are more concentrated in few points rather than scattered out while also at the same time having larger expected CI length. Moreover, the 95% BCa CI of my security indices are wider almost at all points compared to Finn & Servoss'. In general, both models are more precise in the midpoints and less as indices approach the edges

V. Conclusion & Future Analysis

My model performs quite well against Finn & Servoss' w.r.t AIC, BIC and RMSE under 4 gaussian quadrature points. However, having a better AIC, BIC and RMSE doesn't guarantee a better model (it is contextual and depending on application) and it is evident by the distribution of Rasch indices between the two models. Moreover, my model on average is 4.6% less precise in its indexing process. The two models rank security levels at schools in a roughly similar way (statistically significant positive but low Kendall tau). Finn & Servoss' model also allows more statistically significant comparison of security measures between schools.

Previously, we chose the questions based on literature analysis (e.g. what are the security questions that seemingly the least qualitatively overlapping with each other). Based on our prior analyses, it seems that constructing questions that forms a stable hierarchy (e.g. pick n questions that ranges from most common among schools to the least common, as evident in the fixed effect coefficients of Finn & Servoss' model). The previous inference seems align with the intuition in constructing an exam; form a set of questions consists from easy to hard challenges. The aforementioned strategy can be achieved by ordering all the security questions ascendingly w.r.t how many schools answer yes and split them into n groups and then select kth question in each group.