Plots to publish

library(ggplot2)  
library(ggpubr)

## Loading required package: magrittr

library(CDM)

## Loading required package: mvtnorm

## \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  
## \*\* CDM 7.3-17 (2019-03-18 18:33:40)   
## \*\* Cognitive Diagnostic Models \*\*  
## \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

library(boot)  
library(tidyverse)

## -- Attaching packages ----------------------------------------------------------------------------------------------------------------------------------------------------- tidyverse 1.2.1 --

## v tibble 2.1.1 v purrr 0.3.2   
## v tidyr 0.8.3 v dplyr 0.8.0.1  
## v readr 1.3.1 v stringr 1.4.0   
## v tibble 2.1.1 v forcats 0.4.0

## -- Conflicts -------------------------------------------------------------------------------------------------------------------------------------------------------- tidyverse\_conflicts() --  
## x tidyr::extract() masks magrittr::extract()  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()  
## x purrr::set\_names() masks magrittr::set\_names()

library(stringi)  
library(stringr)  
library(ggridges)

##   
## Attaching package: 'ggridges'

## The following object is masked from 'package:ggplot2':  
##   
## scale\_discrete\_manual

library(ggrepel)

rm(list = ls())  
options(max.print=100)

df.error.ec <- read\_rds("..\\data\\FirstYearProject\\report\\df\_error\_ecpe.rds")  
  
df.error.ec.agg <- read\_rds("..\\data\\FirstYearProject\\report\\df\_error\_agg\_ecpe.rds")  
  
df.error.ec.fit <- read\_rds("..\\data\\FirstYearProject\\report\\df\_error\_fit\_ecpe.rds") %>% mutate(dataset = "ECPE")

## Warning: Detecting old grouped\_df format, replacing `vars` attribute by  
## `groups`

df.error.ec.plot <- df.error.ec %>% inner\_join(  
 df.error.ec.agg  
) %>% mutate(dataset = "ECPE")

## Joining, by = c("group", "parameter", "sample\_size", "student\_question\_ratio")

df.error.ec.plot

## # A tibble: 1,120 x 13  
## group parameter questions total\_count sampling\_mean sampling\_error  
## <chr> <chr> <chr> <int> <dbl> <dbl>  
## 1 Part~ Guess E23 500 0.467 0.433   
## 2 Part~ Slip E21 500 0.0561 0.0277  
## 3 Part~ Guess E16 500 0.753 0.0430  
## 4 Part~ Slip E11 500 0.131 0.0768  
## 5 Part~ Slip E22 500 0.326 0.0553  
## 6 Part~ Guess E2 500 0.477 0.0610  
## 7 Part~ Slip E15 500 0.238 0.223   
## 8 Part~ Guess E19 500 0.491 0.156   
## 9 Part~ Guess E12 500 1 0   
## 10 Part~ Slip E17 500 0.347 0.0207  
## # ... with 1,110 more rows, and 7 more variables: sampling\_variance <dbl>,  
## # attempts <dbl>, sample\_size <dbl>, student\_question\_ratio <dbl>,  
## # sampling\_error\_mean <dbl>, avg\_attempts <dbl>, dataset <chr>

df.error.ti <- read\_rds("..\\data\\FirstYearProject\\report\\df\_error\_timss.rds")  
  
df.error.ti.agg <- read\_rds("..\\data\\FirstYearProject\\report\\df\_error\_agg\_timss.rds")  
  
df.error.ti.fit <- read\_rds("..\\data\\FirstYearProject\\report\\df\_error\_fit\_timss.rds") %>% mutate(dataset = "TIMSS")

## Warning: Detecting old grouped\_df format, replacing `vars` attribute by  
## `groups`

df.error.ti.plot <- df.error.ti %>% inner\_join(  
 df.error.ti.agg  
) %>% mutate(dataset = "TIMSS")

## Joining, by = c("group", "parameter", "sample\_size", "student\_question\_ratio")

df.error.ti.plot

## # A tibble: 1,692 x 13  
## group parameter questions total\_count sampling\_mean sampling\_error  
## <chr> <chr> <chr> <int> <dbl> <dbl>  
## 1 Part~ Slip M041158 500 0.470 0.209   
## 2 Part~ Slip M051091 500 0.0721 0.0721  
## 3 Part~ Guess M031346C 500 0.384 0.182   
## 4 Part~ Slip M051055 500 0.189 0.196   
## 5 Part~ Slip M041010 500 0.0443 0.0513  
## 6 Part~ Guess M051123 500 0.0215 0.0150  
## 7 Part~ Slip M041158 500 0.547 0.0859  
## 8 Part~ Guess M031346A 500 0.157 0.0453  
## 9 Part~ Guess M051006 500 0.230 0.141   
## 10 Part~ Guess M031185 500 0.0190 0.0227  
## # ... with 1,682 more rows, and 7 more variables: sampling\_variance <dbl>,  
## # attempts <dbl>, sample\_size <dbl>, student\_question\_ratio <dbl>,  
## # sampling\_error\_mean <dbl>, avg\_attempts <dbl>, dataset <chr>

df.error.ss <- read\_rds("..\\data\\FirstYearProject\\report\\df\_error\_upsample\_socialscience.rds")  
  
df.error.ss.agg <- read\_rds("..\\data\\FirstYearProject\\report\\df\_error\_agg\_upsample\_socialscience.rds")  
  
df.error.ss.fit <- read\_rds("..\\data\\FirstYearProject\\report\\df\_error\_fit\_upsample\_socialscience.rds") %>% mutate(dataset = "Social Psychology (Upsample)")

## Warning: Detecting old grouped\_df format, replacing `vars` attribute by  
## `groups`

df.error.ss.plot <- df.error.ss %>% inner\_join(  
 df.error.ss.agg  
) %>% mutate(dataset = "Social Psychology (UpSampled)")

## Joining, by = c("group", "parameter", "sample\_size", "student\_question\_ratio")

df.error.ss.plot

## # A tibble: 29,684 x 13  
## group parameter questions total\_count sampling\_mean sampling\_error  
## <chr> <chr> <chr> <int> <dbl> <dbl>  
## 1 Part~ Slip Q920 100 1.10e- 1 3.14e- 1  
## 2 Part~ Guess Q749 100 3.77e- 1 3.22e- 1  
## 3 Part~ Guess Q545 100 1.00e- 2 1.00e- 1  
## 4 Part~ Slip Q959 100 8.08e- 1 2.09e- 1  
## 5 Part~ Slip Q462 100 2.02e- 1 1.88e- 1  
## 6 Part~ Slip Q710 100 2.46e- 1 1.03e- 1  
## 7 Part~ Slip Q256 100 -1.43e-18 3.22e-17  
## 8 Part~ Slip Q6 100 1.44e- 1 2.55e- 1  
## 9 Part~ Slip Q692 100 1.00e- 1 3.02e- 1  
## 10 Part~ Guess Q589 100 6.26e- 2 2.12e- 1  
## # ... with 29,674 more rows, and 7 more variables:  
## # sampling\_variance <dbl>, attempts <dbl>, sample\_size <dbl>,  
## # student\_question\_ratio <dbl>, sampling\_error\_mean <dbl>,  
## # avg\_attempts <dbl>, dataset <chr>

# Merge all Data

df.error.plot <- df.error.ec.plot %>% bind\_rows(df.error.ti.plot) %>% bind\_rows(df.error.ss.plot) %>%  
 mutate(ratio\_n = student\_question\_ratio) %>%  
 mutate(student\_question\_ratio = factor(student\_question\_ratio, levels = sort(unique(as.numeric(student\_question\_ratio))), ordered = TRUE)) %>%  
 mutate(dataset = factor(dataset, levels = c("ECPE", "TIMSS", "Social Psychology (UpSampled)"))) %>%  
 mutate(questions = factor(questions)) %>%   
 group\_by(dataset, group, parameter, sample\_size) %>%  
 mutate(question\_n = n(), qsr = round(n()/sample\_size)) %>%   
   
 #mutate(sampling\_error = sampling\_error / sqrt(1000)) %>% # To get Smulation Error  
   
 # Average avg\_attempts between partitions  
 group\_by(dataset, parameter, sample\_size) %>%  
 mutate(avg\_attempts = round(mean(avg\_attempts))) %>%  
   
 ungroup() %>%  
 arrange(dataset, group, parameter, sample\_size)  
  
df.error.plot <- df.error.plot %>% mutate(qsr = factor(qsr, ordered = TRUE),   
 sample\_size = factor(sample\_size, ordered = TRUE),   
 avg\_attempts = factor(avg\_attempts, ordered = TRUE)  
 )  
df.error.plot %>% group\_by(group, parameter, attempts) %>% top\_n(2) %>% select(questions, sampling\_mean, sampling\_error)

## Selecting by qsr

## Adding missing grouping variables: `group`, `parameter`, `attempts`

## # A tibble: 15,806 x 6  
## # Groups: group, parameter, attempts [908]  
## group parameter attempts questions sampling\_mean sampling\_error  
## <chr> <chr> <dbl> <fct> <dbl> <dbl>  
## 1 Partition 1 Guess 6 E12 1 0   
## 2 Partition 1 Guess 6 E1 1 0   
## 3 Partition 1 Guess 6 E4 0.084 0.278  
## 4 Partition 1 Guess 6 E26 0.398 0.416  
## 5 Partition 1 Guess 6 E16 0.616 0.412  
## 6 Partition 1 Guess 6 E10 1 0   
## 7 Partition 1 Guess 6 E24 1 0   
## 8 Partition 1 Guess 6 E19 0.602 0.416  
## 9 Partition 1 Guess 6 E18 0.532 0.427  
## 10 Partition 1 Guess 6 E6 1 0   
## # ... with 15,796 more rows

df.fit.plot <- df.error.ec.fit %>% bind\_rows(df.error.ss.fit) %>% bind\_rows(df.error.ti.fit)  
df.fit.plot

## # A tibble: 1,228 x 9  
## # Groups: group, parameter [2]  
## group parameter fit sampling\_mean sampling\_error sample\_size  
## <chr> <chr> <chr> <dbl> <dbl> <dbl>  
## 1 Part~ SRMSR fit\_~ 0.0864 0.00429 280  
## 2 Part~ SRMSR fit\_~ 0.374 0.0519 8  
## 3 Part~ SRMSR fit\_~ 0.236 0.0192 28  
## 4 Part~ SRMSR fit\_~ 0.178 0.0126 56  
## 5 Part~ SRMSR fit\_~ 0.0712 0.00330 420  
## 6 Part~ SRMSR fit\_~ 0.0478 0.00155 1461  
## 7 Part~ SRMSR fit\_~ 0.244 0.0186 28  
## 8 Part~ SRMSR fit\_~ 0.119 0.00644 140  
## 9 Part~ SRMSR fit\_~ 0.113 0.00605 140  
## 10 Part~ SRMSR fit\_~ 0.182 0.0122 56  
## # ... with 1,218 more rows, and 3 more variables:  
## # student\_question\_ratio <dbl>, attempts <dbl>, dataset <chr>

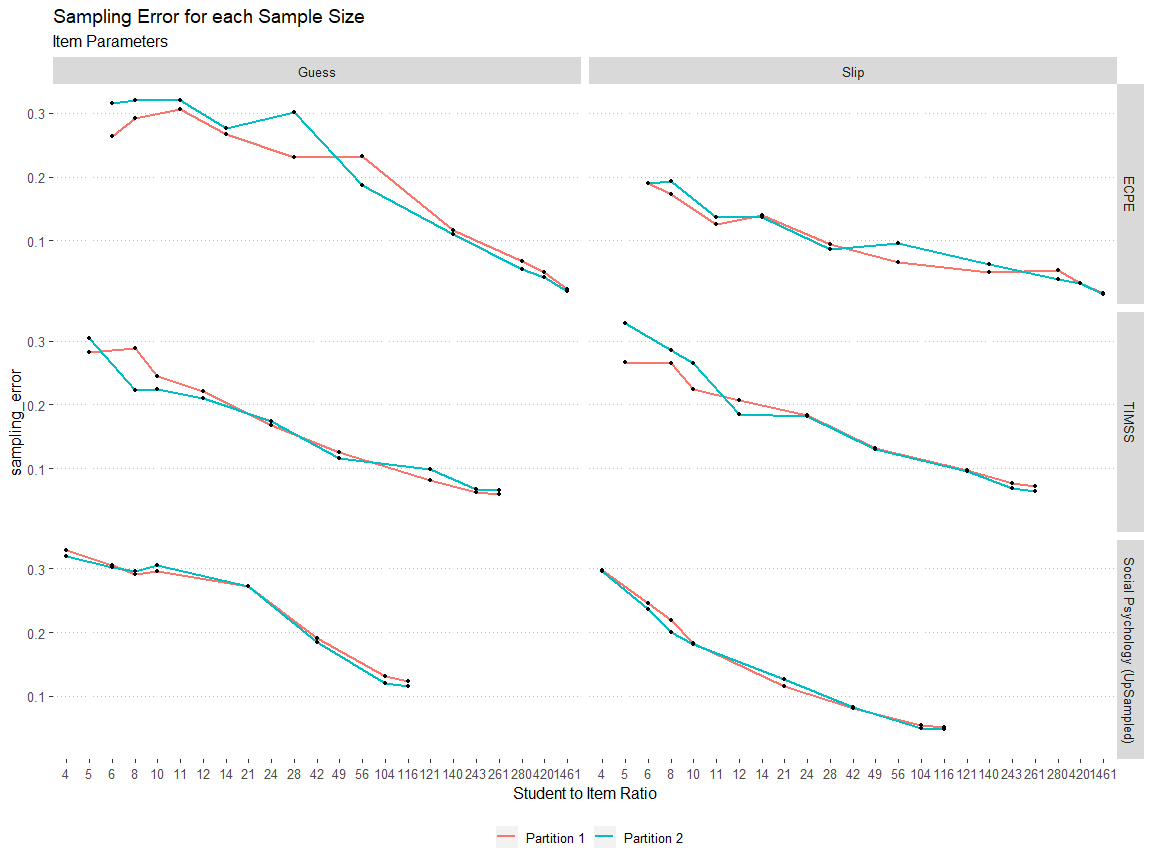
Need to take average because for different sample sizes there can be same number of attempts

df.fit.plot.attempts <- df.fit.plot %>% group\_by(group, parameter, fit, dataset, sample\_size) %>% summarise\_all(.funs = mean)  
df.fit.plot.attempts

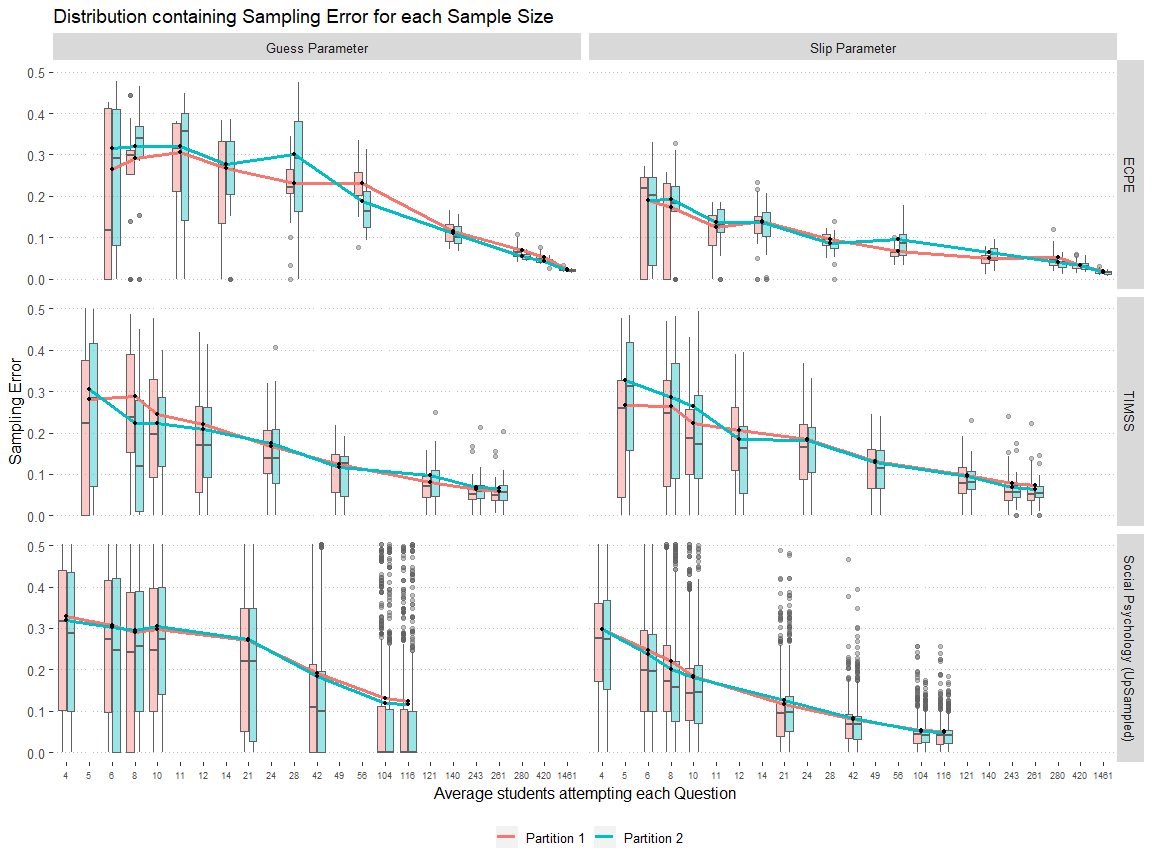
## # A tibble: 54 x 9  
## # Groups: group, parameter, fit, dataset [6]  
## group parameter fit dataset sample\_size sampling\_mean sampling\_error  
## <chr> <chr> <chr> <chr> <dbl> <dbl> <dbl>  
## 1 Part~ SRMSR fit\_~ ECPE 6 0.369 0.0760   
## 2 Part~ SRMSR fit\_~ ECPE 8 0.374 0.0519   
## 3 Part~ SRMSR fit\_~ ECPE 11 0.333 0.0405   
## 4 Part~ SRMSR fit\_~ ECPE 14 0.312 0.0353   
## 5 Part~ SRMSR fit\_~ ECPE 28 0.236 0.0192   
## 6 Part~ SRMSR fit\_~ ECPE 56 0.178 0.0126   
## 7 Part~ SRMSR fit\_~ ECPE 140 0.119 0.00644  
## 8 Part~ SRMSR fit\_~ ECPE 280 0.0864 0.00429  
## 9 Part~ SRMSR fit\_~ ECPE 420 0.0726 0.00307  
## 10 Part~ SRMSR fit\_~ ECPE 1461 0.0478 0.00155  
## # ... with 44 more rows, and 2 more variables:  
## # student\_question\_ratio <dbl>, attempts <dbl>

# Sampling Error

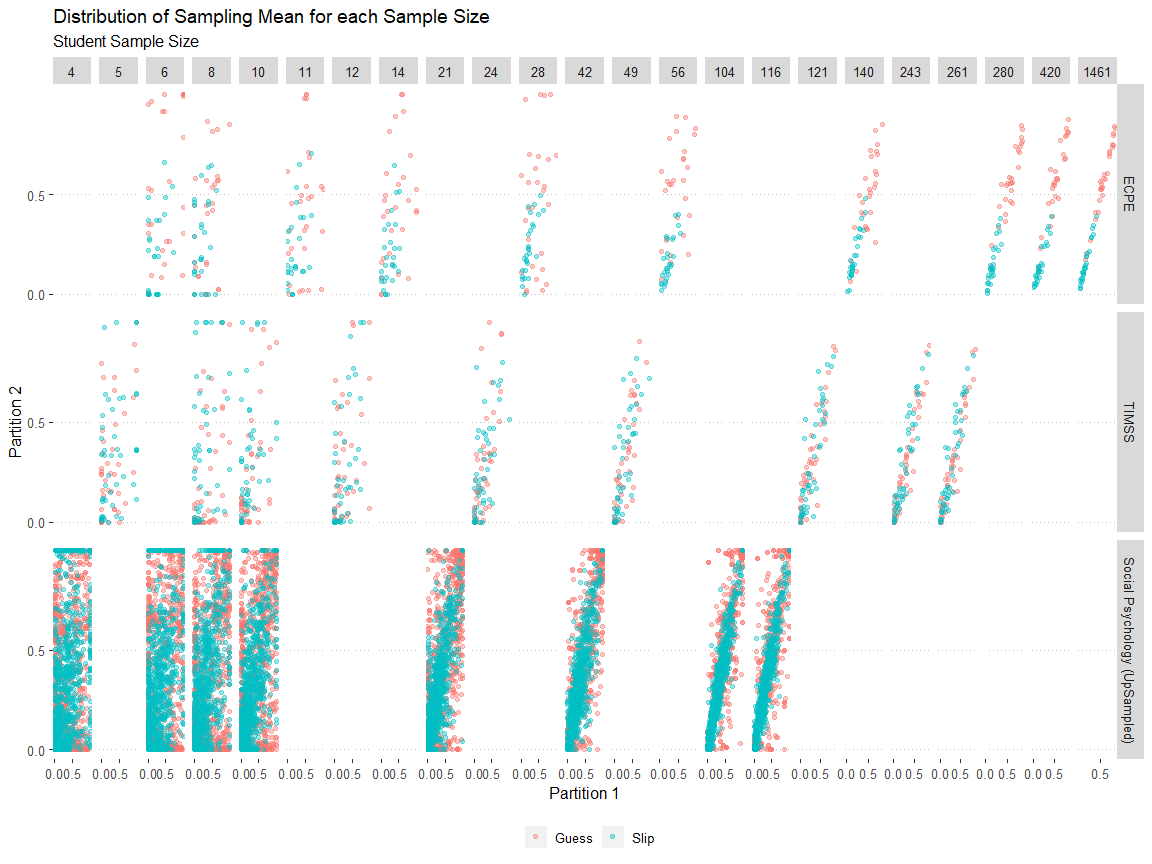
df.error.plot %>% #filter(parameter == "Guess") %>%  
 ggplot() +   
 aes(x=avg\_attempts, y=sampling\_error, fill = group) +  
 #geom\_boxplot(alpha = 0.4, show.legend = FALSE) +  
 geom\_line(aes(y=sampling\_error\_mean, color = group, group = group), linetype = 1, size = 1) +   
 geom\_point(aes(y=sampling\_error\_mean), size = 1, show.legend = FALSE) +   
   
 labs(x="Student to Item Ratio", color="") +  
 facet\_grid(dataset~parameter) +   
 ggtitle("Sampling Error for each Sample Size") + ggpubr::theme\_pubclean() +   
 theme(legend.position = "bottom") +   
 labs(subtitle = "Item Parameters")



df.error.plot %>% #filter(parameter == "Guess") %>%  
 ggplot() +   
 aes(x=avg\_attempts, y=sampling\_error, fill = group) +  
 geom\_boxplot(alpha = 0.4, show.legend = FALSE, color = "grey38") +  
 geom\_line(aes(y=sampling\_error\_mean, color = group, group = group), linetype = 1, size = 1.1) +   
 geom\_point(aes(y=sampling\_error\_mean), size = 1, show.legend = FALSE) +   
   
 labs(x="Average students attempting each Question", color="", y = "Sampling Error") +  
 facet\_grid(dataset~paste(parameter, "Parameter")) +   
 ggtitle("Distribution containing Sampling Error for each Sample Size") + ggpubr::theme\_pubclean() +   
 theme(legend.position = "bottom", axis.text.x = element\_text(size = 7))



df.error.plot %>% select(parameter, group, questions, sampling\_mean, avg\_attempts, student\_question\_ratio, dataset) %>%   
 spread(key="group", value = "sampling\_mean") %>% arrange(avg\_attempts) %>% #filter(parameter == "Guess") %>%  
   
   
   
 ggplot() +   
 aes(x=`Partition 1`, y=`Partition 2`, color = parameter) +  
 geom\_point(show.legend = TRUE, alpha = 0.4) +  
 facet\_grid(dataset~avg\_attempts, scales = "free") +   
 ggtitle("Distribution of Sampling Mean for each Sample Size") +   
 scale\_x\_continuous(breaks = c(0,0.5)) +   
 scale\_y\_continuous(breaks = c(0,0.5)) +   
 ggpubr::theme\_pubclean() +   
 labs(color = "", subtitle = "Student Sample Size") +   
 theme(legend.position = "bottom")



df.error.plot

## # A tibble: 32,496 x 16  
## group parameter questions total\_count sampling\_mean sampling\_error  
## <chr> <chr> <fct> <int> <dbl> <dbl>  
## 1 Part~ Guess E12 500 1 0   
## 2 Part~ Guess E1 500 1 0   
## 3 Part~ Guess E4 500 0.084 0.278  
## 4 Part~ Guess E26 500 0.398 0.416  
## 5 Part~ Guess E16 500 0.616 0.412  
## 6 Part~ Guess E10 500 1 0   
## 7 Part~ Guess E24 500 1 0   
## 8 Part~ Guess E19 500 0.602 0.416  
## 9 Part~ Guess E18 500 0.532 0.427  
## 10 Part~ Guess E6 500 1 0   
## # ... with 32,486 more rows, and 10 more variables:  
## # sampling\_variance <dbl>, attempts <dbl>, sample\_size <ord>,  
## # student\_question\_ratio <ord>, sampling\_error\_mean <dbl>,  
## # avg\_attempts <ord>, dataset <fct>, ratio\_n <dbl>, question\_n <int>,  
## # qsr <ord>

df.t <- df.error.plot #%>% #%>% filter(dataset == "TIMSS")  
#mutate(sampling\_mean = round(sampling\_mean, 2),  
# sampling\_error = round(sampling\_error, 2)  
# )  
  
df.mean\_diff <- df.t %>% select(parameter, group, questions, sampling\_mean, avg\_attempts, dataset) %>%   
 spread(key="group", value = "sampling\_mean") %>% rename("Partition1\_mean" = `Partition 1`, "Partition2\_mean" = `Partition 2`) %>%   
 inner\_join(  
   
 df.t %>% select(parameter, group, questions, sampling\_error, avg\_attempts, dataset) %>%   
 spread(key="group", value = "sampling\_error") %>% rename("Partition1\_error" = `Partition 1`, "Partition2\_error" = `Partition 2`)  
   
 ) %>% group\_by(dataset,parameter,avg\_attempts,questions) %>% mutate(  
 diff = Partition1\_mean - Partition2\_mean,  
 #Z\_test = diff /   
 # ( sqrt( ( (Partition1\_error^2) ) + ((Partition2\_error^2)) + 0.001^2 ) ) #   
 Z\_test = diff / Partition1\_mean,  
 # ( sqrt( ( (Partition1\_error^2) ) + ((Partition2\_error^2)) + 0.001^2 ) )  
   
 ) %>%   
   
 #filter(Z\_test > -250 & Z\_test < 250) %>%  
 group\_by(dataset,parameter,avg\_attempts) %>% mutate(Z\_avg = mean(Z\_test, na.rm = TRUE)) %>% ungroup()

## Joining, by = c("parameter", "questions", "avg\_attempts", "dataset")

df.mean\_diff %>% arrange(dataset, Z\_test, questions, avg\_attempts) %>% filter(parameter == "Guess") %>% select(-Z\_avg)

## # A tibble: 8,124 x 10  
## parameter questions avg\_attempts dataset Partition1\_mean Partition2\_mean  
## <chr> <fct> <ord> <fct> <dbl> <dbl>  
## 1 Guess E22 6 ECPE 0. 0.35   
## 2 Guess E6 8 ECPE 1.07e-313 0.160   
## 3 Guess E4 8 ECPE 2.31e-247 0.445   
## 4 Guess E17 6 ECPE 3.93e- 42 0.00267  
## 5 Guess E23 8 ECPE 3.27e- 32 0.289   
## 6 Guess E3 11 ECPE 7.99e- 27 0.617   
## 7 Guess E4 28 ECPE 1.86e- 9 0.111   
## 8 Guess E2 6 ECPE 1.40e- 2 0.950   
## 9 Guess E11 6 ECPE 1.40e- 2 0.530   
## 10 Guess E14 28 ECPE 2.04e- 2 0.594   
## # ... with 8,114 more rows, and 4 more variables: Partition1\_error <dbl>,  
## # Partition2\_error <dbl>, diff <dbl>, Z\_test <dbl>

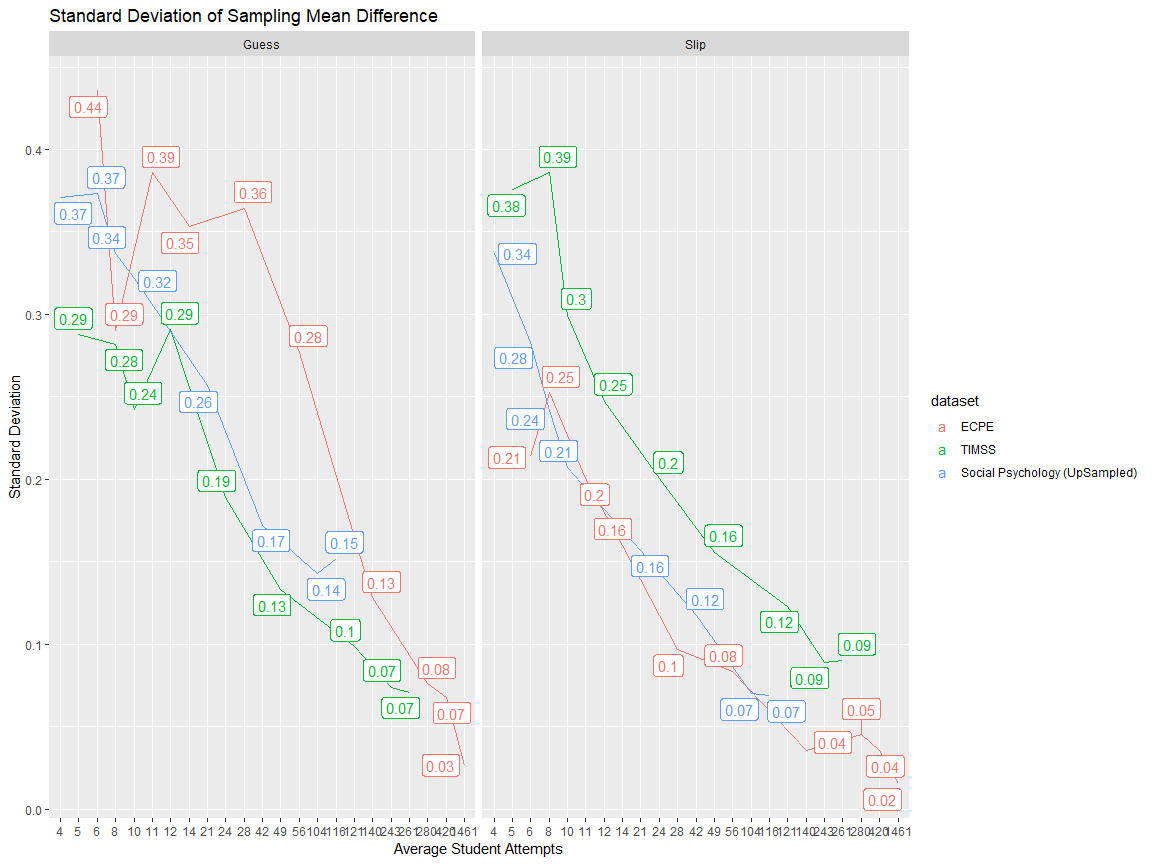
# Plots of mean difference

df.diff <- df.mean\_diff %>% group\_by(dataset, parameter, avg\_attempts) %>%  
 summarize(  
 mn = median(diff),  
 sd = sd(diff),  
 min = min(diff),  
 max = max(diff)  
   
 )  
  
df.diff

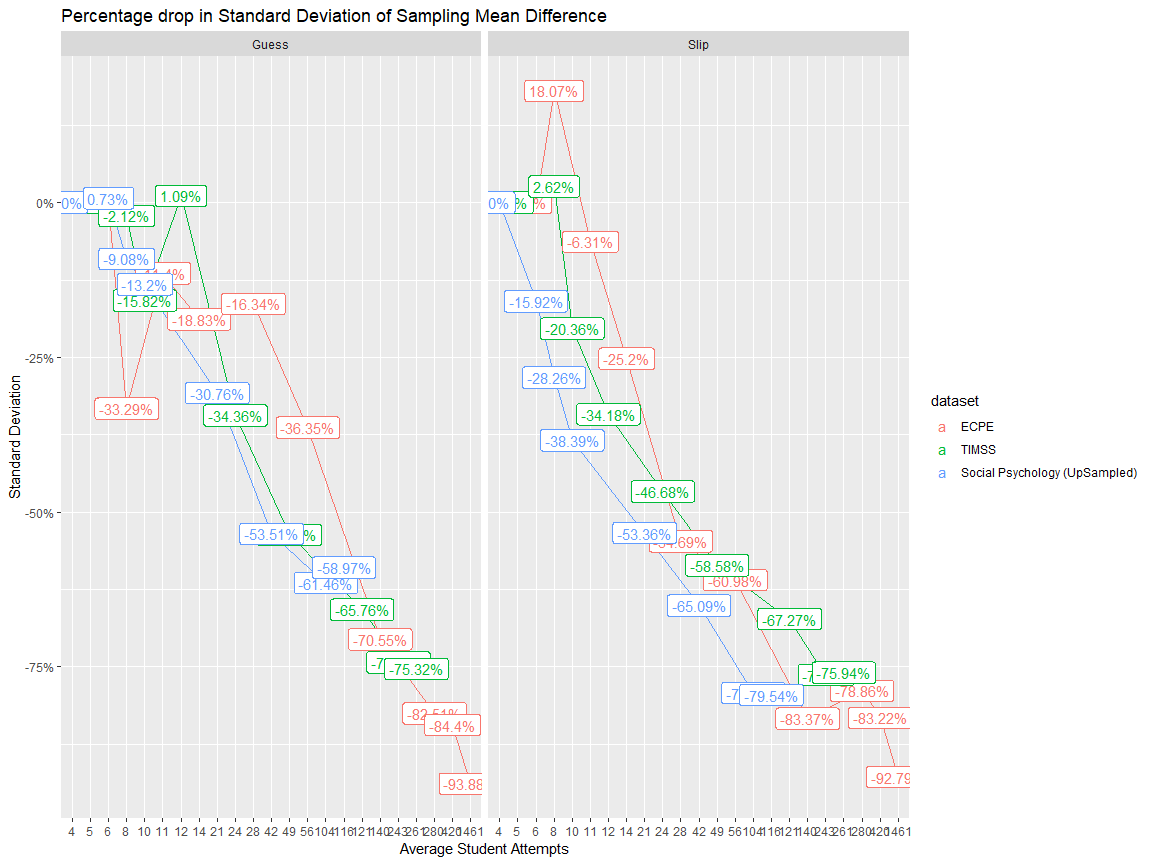
## # A tibble: 54 x 7  
## # Groups: dataset, parameter [6]  
## dataset parameter avg\_attempts mn sd min max  
## <fct> <chr> <ord> <dbl> <dbl> <dbl> <dbl>  
## 1 ECPE Guess 6 0 0.436 -0.936 0.892   
## 2 ECPE Guess 8 0.01 0.291 -0.559 0.665   
## 3 ECPE Guess 11 -0.00364 0.386 -0.617 0.932   
## 4 ECPE Guess 14 -0.0534 0.354 -0.577 0.688   
## 5 ECPE Guess 28 -0.0752 0.364 -0.849 0.688   
## 6 ECPE Guess 56 -0.0491 0.277 -0.559 0.566   
## 7 ECPE Guess 140 -0.00990 0.128 -0.207 0.391   
## 8 ECPE Guess 280 0.0166 0.0762 -0.189 0.147   
## 9 ECPE Guess 420 -0.0348 0.0680 -0.214 0.110   
## 10 ECPE Guess 1461 -0.00333 0.0267 -0.0789 0.0450  
## # ... with 44 more rows

# Plot for poster presentation

df.diff %>% ggplot() +   
 aes(x=avg\_attempts, y = sd, color = dataset) +   
 geom\_line(aes(group = dataset)) +   
 facet\_grid(.~parameter,scales = "free") + labs(x = "Average Student Attempts", y = "Standard Deviation", title = "Standard Deviation of Sampling Mean Difference") +  
 geom\_label\_repel(aes(label = round(sd,2)))



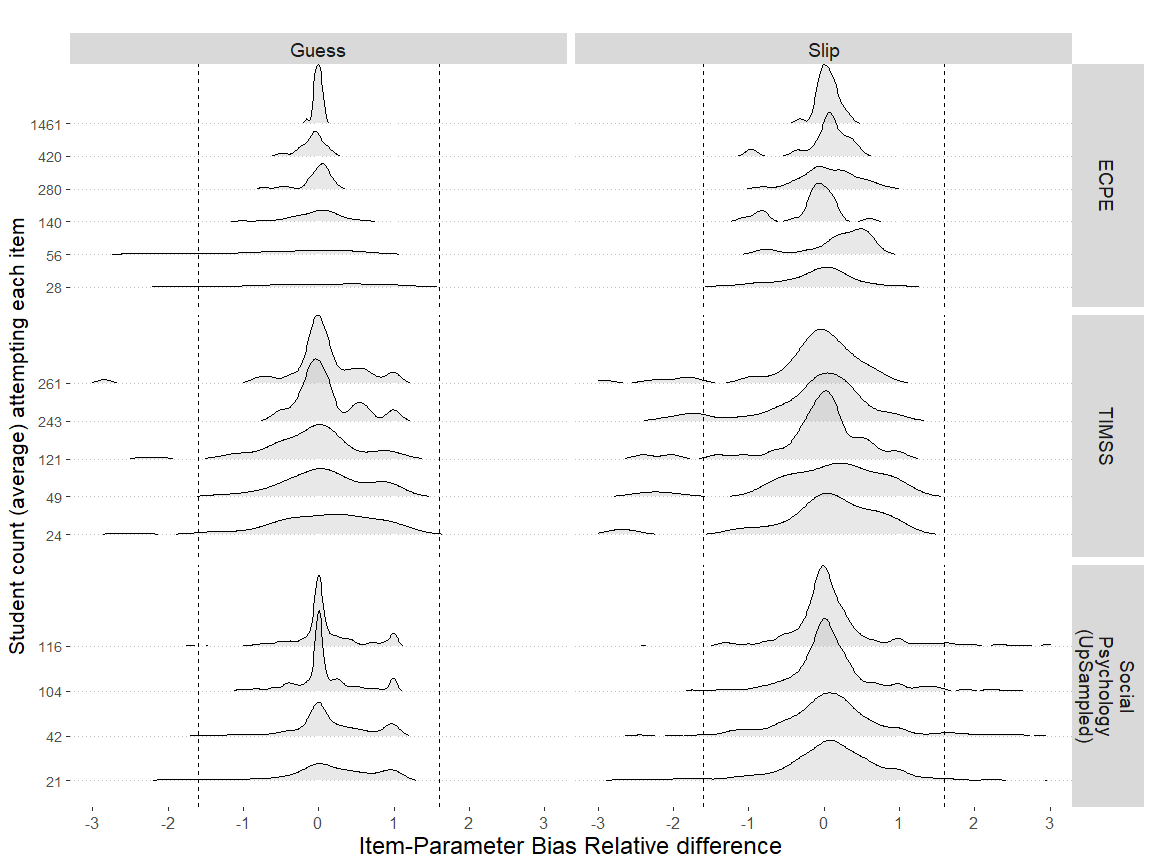
df.diff %>% group\_by(dataset, parameter)%>% arrange(dataset, parameter, avg\_attempts) %>% mutate(first = first(sd)) %>% mutate(diff\_from\_first = ((sd-first)/first) ) %>%   
   
 ggplot() +   
 aes(x=avg\_attempts, y = diff\_from\_first, color = dataset) +   
 geom\_line(aes(group = dataset)) +   
 facet\_grid(.~parameter,scales = "free") + labs(x = "Average Student Attempts", y = "Standard Deviation", title = "Percentage drop in Standard Deviation of Sampling Mean Difference") +  
 geom\_label(aes(label = paste0(round(diff\_from\_first\*100,2),"%"))) +   
 scale\_y\_continuous(labels=scales::percent\_format(accuracy = 1))



# 2 Sample Z Test

df.mean\_diff %>% filter(dataset == "Social Psychology" | avg\_attempts > 14) %>%  
   
   
 ggplot() +   
 aes(y=avg\_attempts, x=Z\_test) +  
   
 geom\_density\_ridges(aes(height = ..density..),   
 stat = "density", rel\_min\_height=0.01,   
 alpha = 0.3  
 ) +  
   
 #labs(y = "Student count (average) attempting each item", x = "Mean Difference between partitions", title = "") +   
 labs(y = "Student count (average) attempting each item", x = "Item-Parameter Bias Relative difference", title = "") +   
 geom\_vline(aes(xintercept = 1.6), linetype = 2) +   
 geom\_vline(aes(xintercept = -1.6), linetype = 2) +   
 theme\_pubclean() +   
 facet\_grid(dataset ~ parameter, scales = "free",   
 labeller = label\_wrap\_gen(width = 10, multi\_line = FALSE)) +   
 theme(axis.text.x = element\_text(size = 12),   
 axis.text.y = element\_text(size = 10.5),  
 strip.text.x = element\_text(size = 14),  
 strip.text.y = element\_text(size = 14),  
 axis.title.x = element\_text(size = 18),  
 axis.title.y = element\_text(size = 16)  
   
 ) + scale\_x\_continuous(limits = c(-3,3), breaks = seq(-3,3,1))

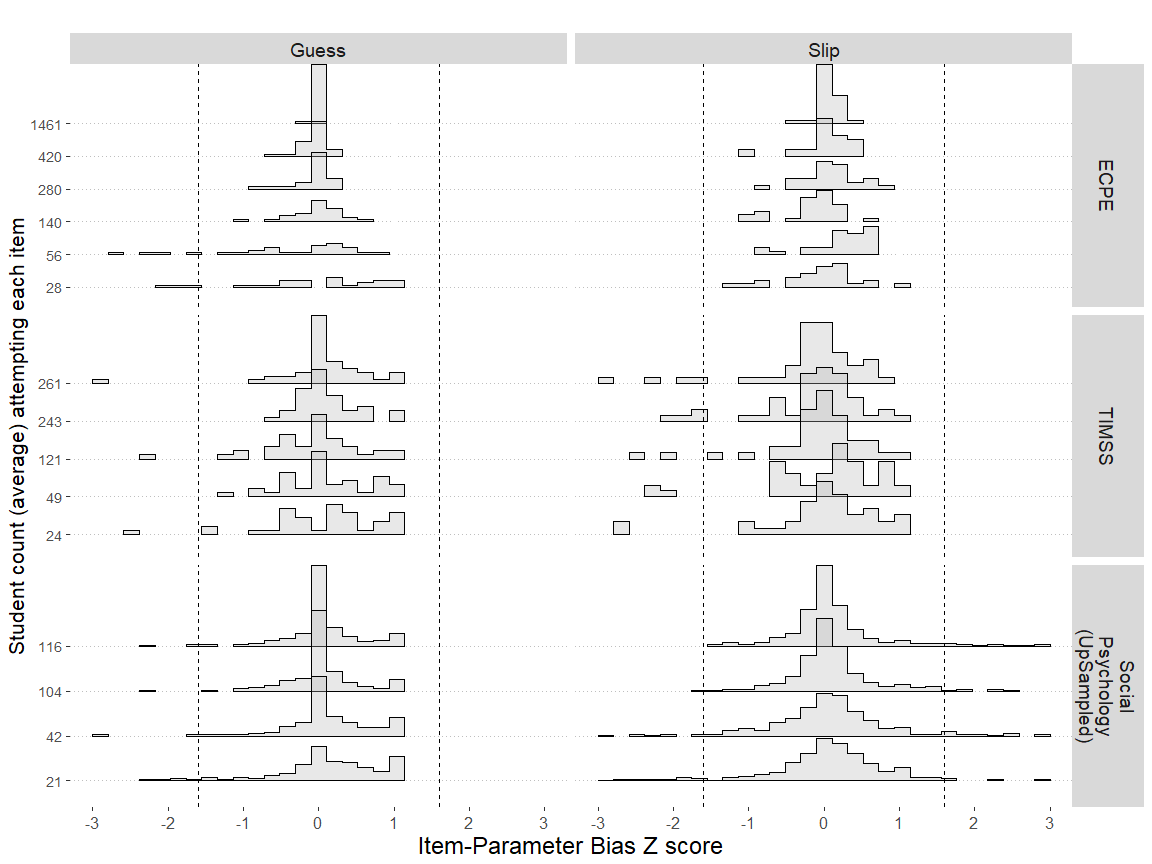
## Warning: Removed 1616 rows containing non-finite values (stat\_density).



df.mean\_diff %>% filter(dataset == "Social Psychology" | avg\_attempts > 14) %>%  
   
   
 ggplot() +   
 aes(y=avg\_attempts, x=Z\_test) +  
   
 geom\_density\_ridges2(aes(height = ..density..),   
 stat = "binline", rel\_min\_height=0.01,   
 alpha = 0.3  
 ) +  
   
 #labs(y = "Student count (average) attempting each item", x = "Mean Difference between partitions", title = "") +   
 labs(y = "Student count (average) attempting each item", x = "Item-Parameter Bias Z score", title = "") +   
 geom\_vline(aes(xintercept = 1.6), linetype = 2) +   
 geom\_vline(aes(xintercept = -1.6), linetype = 2) +   
 theme\_pubclean() +   
 facet\_grid(dataset ~ parameter, scales = "free",   
 labeller = label\_wrap\_gen(width = 10, multi\_line = FALSE)) +   
 theme(axis.text.x = element\_text(size = 12),   
 axis.text.y = element\_text(size = 10.5),  
 strip.text.x = element\_text(size = 14),  
 strip.text.y = element\_text(size = 14),  
 axis.title.x = element\_text(size = 18),  
 axis.title.y = element\_text(size = 16)  
   
 ) + scale\_x\_continuous(limits = c(-3,3), breaks = seq(-3,3,1))

## `stat\_binline()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 1616 rows containing non-finite values (stat\_binline).



# EDA Model Fit

df.fit.plot.attempts %>% arrange(sample\_size) %>% distinct(group, sample\_size, dataset) %>% filter(dataset == "Social Psychology (Upsample)") %>% filter(group == "Partition 2")

## # A tibble: 8 x 5  
## # Groups: group, parameter, fit, dataset [1]  
## group sample\_size dataset parameter fit   
## <chr> <dbl> <chr> <chr> <chr>   
## 1 Partition 2 188 Social Psychology (Upsample) SRMSR fit\_SRMSR  
## 2 Partition 2 282 Social Psychology (Upsample) SRMSR fit\_SRMSR  
## 3 Partition 2 376 Social Psychology (Upsample) SRMSR fit\_SRMSR  
## 4 Partition 2 470 Social Psychology (Upsample) SRMSR fit\_SRMSR  
## 5 Partition 2 939 Social Psychology (Upsample) SRMSR fit\_SRMSR  
## 6 Partition 2 1878 Social Psychology (Upsample) SRMSR fit\_SRMSR  
## 7 Partition 2 4695 Social Psychology (Upsample) SRMSR fit\_SRMSR  
## 8 Partition 2 5250 Social Psychology (Upsample) SRMSR fit\_SRMSR

df.fit.plot.attempts %>% filter(group == "Partition 1") %>%  
   
 ggplot() +  
 aes(x=attempts, y=sampling\_error, color = dataset, group = group) +  
 geom\_line() +  
 geom\_label(aes(label = round(sampling\_error,6)))+  
 geom\_hline(aes(yintercept=0.05),linetype="dotted") +  
 #scale\_x\_continuous(limits = c(0,300)) +   
 facet\_wrap(dataset~.,scales = "free")

