

Data Management

Table of contents

0.1	Introduction	1
0.2	Filtering Data	3
0.2.1	Check data prior to filtering	3
0.2.2	Filtering the data	5
0.2.3	Check data after filtering	6
0.2.4	Drop variables	8
0.3	Addressing Missing Values	8
0.4	Creating Total Score for Measurements	9
0.4.1	Statistics Anxiety, STARS	9
0.4.2	Mathematics Anxiety, R-MARS	9
0.4.3	Trait Anxiety, STICSA	10
0.4.4	Test Anxiety, Revised Test Anxiety Scale (R TAS)	10
0.4.5	Fear of Negative Evaluation, Brief Fear of Negative Evaluation Scale - Straightforward (BNFE-S)	10
0.4.6	Social Interaction Anxiety and Performance Anxiety, Liebowitz Social Anxiety Scale - Self Report (LSAS-SR)	10
0.4.7	Intolerance of Uncertainty Scale - Short Form (IUS-SF)	11
0.4.8	Creativity Anxiety Scale (CAS)	15
0.4.9	Analytic Thinking, Cognitive Reflection Task (CRT)	18
0.4.10	Self efficacy, New General Self Efficacy Scale (NGSE)	18
0.4.11	Persistence, Attitude Towards Mathematics Survey (ATMS)	18
0.5	Split Data	18

0.1 Introduction

Goal:

- Create an organized dataset that focuses on my variables of interests.

- Create additional variables that I will need to analyze.

Dataset: International Multi-Centre Study of Statistics and Mathematics Anxieties and Related Variables in University Students (the SMARVUS Dataset)

Source: “This large, international dataset contains survey responses from $N = 12,570$ students from 100 universities in 35 countries, collected in 21 languages. We measured anxieties (statistics, mathematics, test, trait, social interaction, performance, creativity, intolerance of uncertainty, and fear of negative evaluation), self-efficacy, persistence, and the cognitive reflection test, and collected demographics, previous mathematics grades, self-reported and official statistics grades, and statistics module details. Data reuse potential is broad, including testing links between anxieties and statistics/mathematics education factors, and examining instruments’ psychometric properties across different languages and contexts.” (<https://osf.io/mhg94/>)

Variables of Interest:

- ‘Var1’: [Variable description]
- ‘Var2’: [Variable description]

Goals:

- Filter dataset to English speakers (my language) and Psychology Majors (majority of participants).
- Create total scores for measures.
- Create a new file with my filtered data.

```
-- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
v dplyr      1.1.4      v readr      2.1.5
v forcats    1.0.0      v stringr    1.5.1
v ggplot2    3.5.1      v tibble     3.2.1
v lubridate  1.9.4      v tidyr      1.3.1
v purrr      1.0.4
-- 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
```

```
source("loadData.R")
```

```
[1] "Data successfully loaded."
```

```
head(df)
```

```
# A tibble: 6 x 788
  unique_id survey_id country language incentive participation_context progress
  <chr>      <chr>      <chr>    <chr>    <chr>      <chr>          <dbl>
1 01057178  airlangga Indones~ Bahasa ~ Prize Dr~ Remotely, in own time      100
2 0300b5f2  airlangga Indones~ Bahasa ~ Prize Dr~ Remotely, in own time      100
3 03f6503b  airlangga Indones~ Bahasa ~ Prize Dr~ Remotely, in own time      100
4 0601d699  airlangga Indones~ Bahasa ~ Prize Dr~ Remotely, in own time       78
5 08b6cb02  airlangga Indones~ Bahasa ~ Prize Dr~ Remotely, in own time      100
6 0903d70d  airlangga Indones~ Bahasa ~ Prize Dr~ Remotely, in own time      100
# i 781 more variables: duration <dbl>, start_date <chr>, end_date <chr>,
#   university <chr>, eligibility_ug <chr>, eligibility_stats <chr>,
#   degree_major <chr>, degree_minor <chr>, degree_level_maths <chr>,
#   degree_level_maths_specify <chr>, degree_year <chr>, age <chr>,
#   gender <chr>, spld <chr>, attention_amnesty <chr>, Q7.1_1 <dbl>,
#   Q7.1_2 <dbl>, Q7.1_3 <dbl>, Q7.1_4 <dbl>, Q7.1_5 <dbl>, Q7.1_6 <dbl>,
#   Q7.1_7 <dbl>, Q7.1_8 <dbl>, Q7.1_9 <dbl>, Q7.1_10 <dbl>, Q7.1_11 <dbl>, ...
```

```
# Alternative example

# Define the file path
# Load the CSV file
#df <- file.path("../", "data", "data.csv") %>%
#   read_csv()

# View the first few rows
#head(df)
```

0.2 Filtering Data

0.2.1 Check data prior to filtering

Review the majors and minors of participants. I will use this information to be aware of the levels of the variable when filtering.

```
freq(as.ordered(df$degree_major), plot = FALSE)
```

```
as.ordered(df$degree_major)
      Frequency  Percent Valid Percent Cum Percent
```

Arts & Humanities	232	1.8457	2.1007	2.101
Business & Finance	768	6.1098	6.9540	9.055
Education	397	3.1583	3.5947	12.649
Maths	153	1.2172	1.3854	14.035
Psychology	8759	69.6818	79.3100	93.345
Sciences	456	3.6277	4.1289	97.474
Social Sciences	247	1.9650	2.2365	99.710
Uncategorised	32	0.2546	0.2898	100.000
NA's	1526	12.1400		
Total	12570	100.0000	100.0000	

```
freq(as.ordered(df$degree_minor), plot = FALSE)
```

as.ordered(df\$degree_minor)				
	Frequency	Percent	Valid Percent	Cum Percent
Arts & Humanities	15	0.11933	6.4935	6.494
Business & Finance	30	0.23866	12.9870	19.481
Education	61	0.48528	26.4069	45.887
Maths	4	0.03182	1.7316	47.619
Psychology	107	0.85123	46.3203	93.939
Sciences	2	0.01591	0.8658	94.805
Social Sciences	12	0.09547	5.1948	100.000
NA's	12339	98.16229		
Total	12570	100.00000	100.0000	

```
freq(as.ordered(df$language), plot = 0)
```

as.ordered(df\$language)			
	Frequency	Percent	Cum Percent
Arabic	1490	11.8536	11.85
Bahasa Indonesia	697	5.5449	17.40
Chinese	323	2.5696	19.97
Dutch	394	3.1344	23.10
English	5427	43.1742	66.28
Estonian	98	0.7796	67.06
French	432	3.4368	70.49
German	736	5.8552	76.35
Greek	99	0.7876	77.14
Hebrew	285	2.2673	79.40
Hungarian	206	1.6388	81.04
Italian	248	1.9730	83.02

Polish	69	0.5489	83.56
Portuguese	68	0.5410	84.11
Romanian	317	2.5219	86.63
Serbian	117	0.9308	87.56
Slovakian	88	0.7001	88.26
Slovenian	94	0.7478	89.01
Spanish	523	4.1607	93.17
Turkish	834	6.6348	99.80
Ukrainian	25	0.1989	100.00
Total	12570	100.0000	

```
freq(as.ordered(df$attention_amnesty), plot = 0)
```

```
as.ordered(df$attention_amnesty)
      Frequency Percent Valid Percent Cum Percent
No           172   1.368           1.645       1.645
Yes        10281  81.790          98.355      100.000
NA's         2117  16.842
Total        12570 100.000          100.000
```

0.2.2 Filtering the data

Eliminate any participants that did not pass the attention check items. Participants were directed to provide a particular response to each of the attention check items.

```
filtered_df <- df %>%
  filter(
    attention_amnesty == "Yes") %>%
  filter(
    Q7.1_24 == 1,
    Q8.1_21 == 5,
    Q9.1_22 == 1,
    Q11.1_9 == 3,
    Q13.1_17 == 2,
    Q15.1_9 == 4) %>%
  filter(
    degree_major == "Psychology" |
    degree_minor == "Psychology") %>%
  filter(
    progress == 100) %>%
  filter(
    language == "English")
```

```
#View(filtered_df)
```

0.2.3 Check data after filtering

Use frequency tables to check filtered data.

```
attention_check_items <- c(
  "Q7.1_24",
  "Q8.1_21",
  "Q9.1_22",
  "Q11.1_9",
  "Q13.1_17",
  "Q15.1_9"
)

# Generate frequency tables for each variable
invisible(lapply(attention_check_items, function(var) {
  cat("\n-----\n")
  cat("Frequency Table for", var, "\n")
  print(table(filtered_df[[var]]))
})))
```

```
-----
Frequency Table for Q7.1_24
```

```
  1
3077
```

```
-----
Frequency Table for Q8.1_21
```

```
  5
3077
```

```
-----
Frequency Table for Q9.1_22
```

```
  1
```

3077

Frequency Table for Q11.1_9

3
3077

Frequency Table for Q13.1_17

2
3077

Frequency Table for Q15.1_9

4
3077

```
freq(as.ordered(filtered_df$attention_amnesty), plot = 0)
```

```
as.ordered(filtered_df$attention_amnesty)
      Frequency Percent Cum Percent
Yes           3077      100         100
Total          3077      100
```

```
freq(as.ordered(filtered_df$degree_major), plot = 0)
```

```
as.ordered(filtered_df$degree_major)
      Frequency Percent Cum Percent
Psychology     3077      100         100
Total           3077      100
```

```
freq(as.ordered(filtered_df$degree_minor), plot = 0)
```

```
as.ordered(filtered_df$degree_minor)
      Frequency Percent Valid Percent Cum Percent
NA's           3077      100
Total           3077      100              0
```

```
freq(as.ordered(filtered_df$progress), plot = 0)
```

```
as.ordered(filtered_df$progress)
      Frequency Percent Cum Percent
100          3077      100         100
Total          3077      100
```

```
freq(as.ordered(filtered_df$language), plot = 0)
```

```
as.ordered(filtered_df$language)
      Frequency Percent Cum Percent
English          3077      100         100
Total          3077      100
```

0.2.4 Drop variables

Remove variables that have been used in the filter and are no longer meaningful to include.

```
filtered_df <- filtered_df %>%
  select(
    -Q7.1_24,
    -Q8.1_21,
    -Q9.1_22,
    -Q11.1_9,
    -Q13.1_17,
    -Q15.1_9,
    -attention_amnesty,
    -degree_major,
    -degree_minor,
    -language,
    -progress)
```

```
#View(filtered_df)
```

0.3 Addressing Missing Values

Missing values do not appear to be an issue for the scale variables.

0.4 Creating Total Score for Measurements

My dataset contains the individual items on several measures. I will need to create a total score for total scores and sub-scales. Variables appear to be doubles, which will work for creating the total scores and subscales. Output will be hidden below to save space.

```
#Reverse coded items - NEW variables
filtered_df <- filtered_df %>%
  mutate(Q16.1_2_rev = 6 - Q16.1_2)

filtered_df <- filtered_df %>%
  mutate(across(c(Q16.1_3, Q16.1_4, Q16.1_5, Q16.1_7), ~ 6 - .,
    .names = "{.col}_rev"))

#Check that new reverse coded columns exist
print(colnames(filtered_df)[endsWith(colnames(filtered_df), "_rev")])
```

```
[1] "Q16.1_2_rev" "Q16.1_3_rev" "Q16.1_4_rev" "Q16.1_5_rev" "Q16.1_7_rev"
```

0.4.1 Statistics Anxiety, STARS

(Cruise et al., 1985; Hanna et al., 2008; Papousek et al., 2012)

Scale consists of 23 items: tests and class anxiety (8 items), interpretation anxiety (11 items), and fear of asking for help (4 items).

Uses Likert scale ranging from 1 = “no anxiety” to 5 = “a great deal of anxiety.”

0.4.1.1 STARS modified for math

0.4.2 Mathematics Anxiety, R-MARS

(Baloğlu & Zelhart, 2007)

Subscales: Mathematics test anxiety (15 items), numerical task anxiety (5 original plus 4 modified items), and mathematics course anxiety (5 items).

Uses Likert-type scale ranging from 1 = “no anxiety” to 5 = “a great deal of anxiety”.

0.4.2.1 R-MARS modified for stats

0.4.3 Trait Anxiety, STICSA

(Ree et al., 2008)

Subscales: cognitive (10 items) and somatic symptoms (11 items).

Uses Likert scale ranging from 1 = “not at all” to 4 = “very much so”.

0.4.4 Test Anxiety, Revised Test Anxiety Scale (R TAS)

(Benson & El-Zahhar, 1994)

Subscales: worry (7 items), tension (6 items), test-irrelevant thinking (5 items), bodily symptoms (7 items). Secondary items may be removed. See citation above.

Uses Likert scale ranging from 1 = “almost never” to 4 = “almost always”.

0.4.5 Fear of Negative Evaluation, Brief Fear of Negative Evaluation Scale - Straightforward (BNFE-S)

(Leary, 1983; Rodebaugh et al., 2004)

8 items.

Uses Likert scale ranging from 1 = “not at all characteristic of me” to 5 = “extremely characteristic of me”.

0.4.6 Social Interaction Anxiety and Performance Anxiety, Liebowitz Social Anxiety Scale - Self Report (LSAS-SR)

(Baker et al., 2002; Liebowitz, 1987)

Subscales: interaction anxiety (12 items) and performance anxiety (12 items).

Uses Likert scale ranging from 0 = “not at all” to 3 = “very much so”.

0.4.7 Intolerance of Uncertainty Scale - Short Form (IUS-SF)

(Carleton et al. 2007)

Subscales: prospective anxiety (fear of the future and uncertainty-related anticipation; 6 items) and inhibitory anxiety (avoidance behavior due to uncertainty; 6 items).

Uses Likert scale ranging from 1 = “not at all characteristic of me” to 5 = “entirely characteristic of me”.

```
IUS_prospective_items <- c(
  "Q14.1_1",
  "Q14.1_2",
  "Q14.1_4",
  "Q14.1_5",
  "Q14.1_8",
  "Q14.1_9"
)

IUS_inhibitory_items <- c(
  "Q14.1_3",
  "Q14.1_6",
  "Q14.1_7",
  "Q14.1_10",
  "Q14.1_11",
  "Q14.1_12"
)

filtered_df <- filtered_df %>%
  mutate(
    IUS_Prospective = rowSums(select(., all_of(IUS_prospective_items)), na.rm = TRUE),
    IUS_Inhibitory = rowSums(select(., all_of(IUS_inhibitory_items)), na.rm = TRUE)
  )

filtered_df <- filtered_df %>%
  mutate(IUS_Total = IUS_Prospective + IUS_Inhibitory)

print("IUS Prospective Anxiety")
```

```
[1] "IUS Prospective Anxiety"
```

```
summary(filtered_df$IUS_Prospective)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	13.00	18.00	17.73	22.00	30.00

```
print("IUS Inhibitory Anxiety")
```

```
[1] "IUS Inhibitory Anxiety"
```

```
summary(filtered_df$IUS_Inhibitory)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0	12.0	17.0	16.7	21.0	30.0

```
print("Total IUS-SF Score")
```

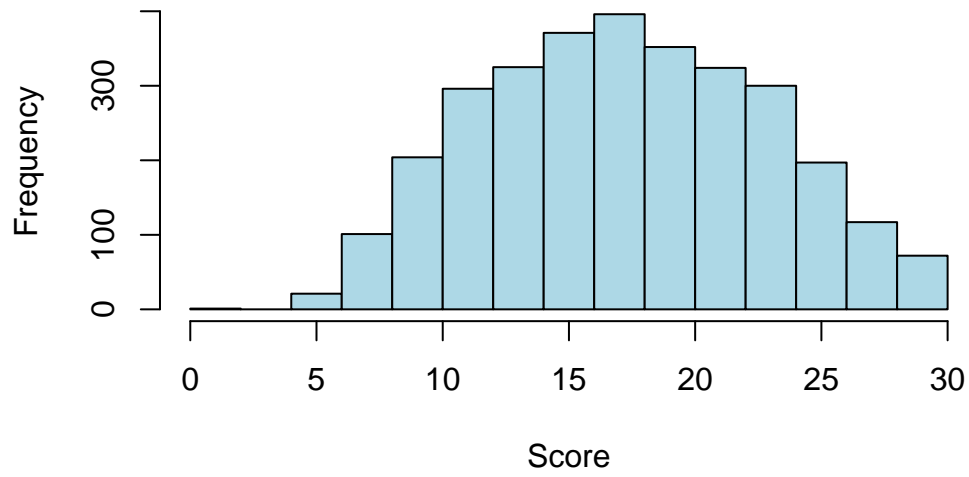
```
[1] "Total IUS-SF Score"
```

```
summary(filtered_df$IUS_Total)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	26.00	34.00	34.43	42.00	60.00

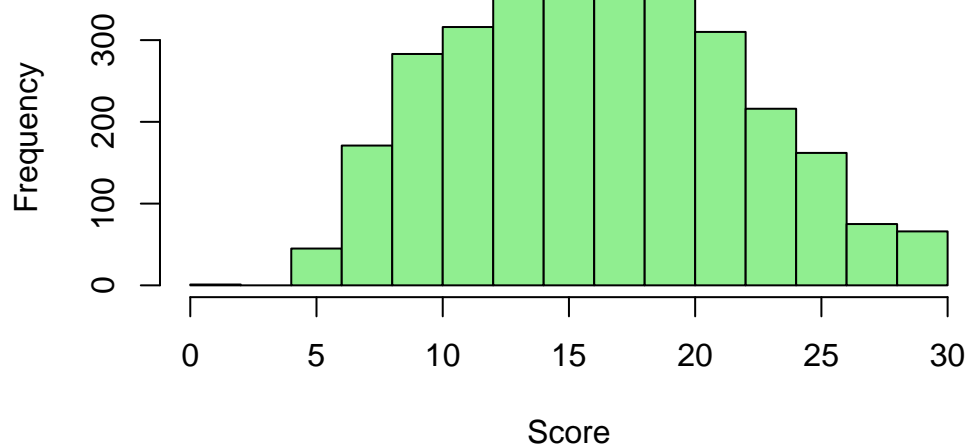
```
hist(filtered_df$IUS_Prospective,  
      main="IUS Prospective Anxiety",  
      xlab="Score",  
      col="lightblue"  
    )
```

IUS Prospective Anxiety

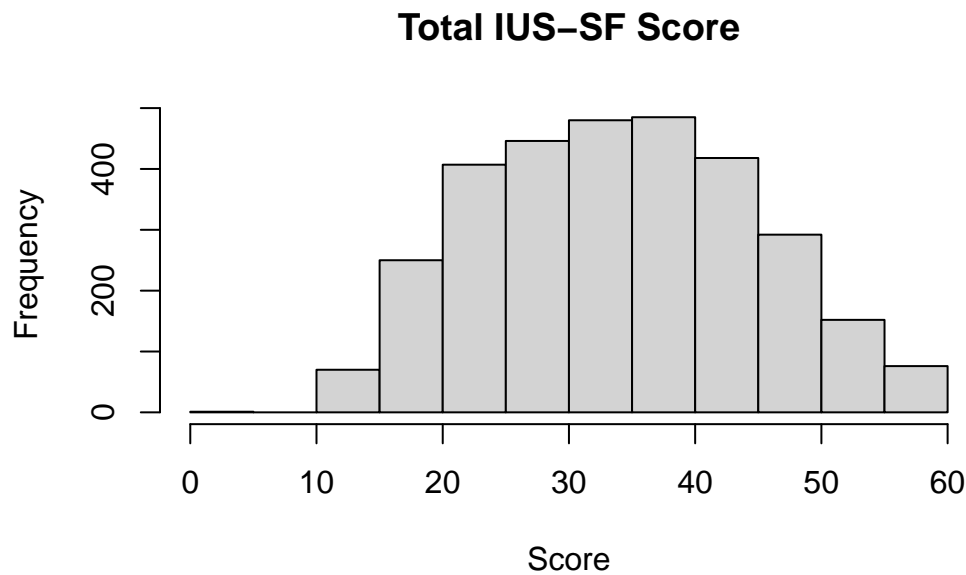


```
hist(filtered_df$IUS_Inhibitory,  
      main="IUS Inhibitory Anxiety",  
      xlab="Score",  
      col="lightgreen"  
    )
```

IUS Inhibitory Anxiety



```
hist(filtered_df$IUS_Total,  
      main="Total IUS-SF Score",  
      xlab="Score",  
      col="lightgray"  
)
```



0.4.8 Creativity Anxiety Scale (CAS)

(Daker et al., 2020)

Subscales: creativity (8 items) and non-creativity (8 items).

Uses Likert scale ranging from 0 = “not at all” to 4 = “very much”.

Sum.

```
#Create list of items for each subscale
```

```
CAS_creativity_items <- c(  
  "Q13.1_1",  
  "Q13.1_2",  
  "Q13.1_3",  
  "Q13.1_4",  
  "Q13.1_5",  
  "Q13.1_6",  
  "Q13.1_7",  
  "Q13.1_8"  
)
```

```
CAS_non_creativity_items <- c(  
  "Q13.2_1",  
  "Q13.2_2",  
  "Q13.2_3",  
  "Q13.2_4",  
  "Q13.2_5",  
  "Q13.2_6",  
  "Q13.2_7",  
  "Q13.2_8"
```

```

"Q13.1_9",
"Q13.1_10",
"Q13.1_11",
"Q13.1_12",
"Q13.1_13",
"Q13.1_14",
"Q13.1_15",
"Q13.1_16"
)

#Generate subscale scores
filtered_df <- filtered_df %>%
  mutate(
    CAS_creativity = rowSums(select(., all_of(CAS_creativity_items)), na.rm = TRUE),
    CAS_noncreativity = rowSums(select(., all_of(CAS_non_creativity_items)), na.rm = TRUE)
  )

freq(as.ordered(filtered_df$CAS_creativity), plot = 0)

```

```

as.ordered(filtered_df$CAS_creativity)

```

	Frequency	Percent	Cum Percent
8	61	1.9825	1.982
9	81	2.6324	4.615
10	72	2.3399	6.955
11	67	2.1774	9.132
12	68	2.2099	11.342
13	100	3.2499	14.592
14	98	3.1849	17.777
15	108	3.5099	21.287
16	134	4.3549	25.642
17	130	4.2249	29.867
18	146	4.7449	34.612
19	148	4.8099	39.422
20	125	4.0624	43.484
21	137	4.4524	47.936
22	142	4.6149	52.551
23	157	5.1024	57.654
24	158	5.1349	62.788
25	162	5.2649	68.053
26	125	4.0624	72.116
27	116	3.7699	75.886
28	124	4.0299	79.916

29	99	3.2174	83.133
30	107	3.4774	86.610
31	74	2.4049	89.015
32	92	2.9899	92.005
33	54	1.7550	93.760
34	43	1.3975	95.158
35	41	1.3325	96.490
36	29	0.9425	97.433
37	21	0.6825	98.115
38	26	0.8450	98.960
39	24	0.7800	99.740
40	8	0.2600	100.000
Total	3077	100.0000	

```
freq(as.ordered(filtered_df$CAS_noncreativity), plot = 0)
```

```
as.ordered(filtered_df$CAS_noncreativity)
```

	Frequency	Percent	Cum Percent
8	176	5.7199	5.72
9	146	4.7449	10.46
10	139	4.5174	14.98
11	127	4.1274	19.11
12	134	4.3549	23.46
13	120	3.8999	27.36
14	166	5.3949	32.76
15	166	5.3949	38.15
16	205	6.6623	44.82
17	169	5.4924	50.31
18	163	5.2974	55.61
19	149	4.8424	60.45
20	160	5.1999	65.65
21	120	3.8999	69.55
22	113	3.6724	73.22
23	103	3.3474	76.57
24	107	3.4774	80.05
25	87	2.8274	82.87
26	73	2.3724	85.25
27	78	2.5349	87.78
28	63	2.0474	89.83
29	59	1.9175	91.75
30	40	1.3000	93.05
31	50	1.6250	94.67

32	37	1.2025	95.87
33	34	1.1050	96.98
34	19	0.6175	97.60
35	22	0.7150	98.31
36	13	0.4225	98.73
37	9	0.2925	99.03
38	13	0.4225	99.45
39	8	0.2600	99.71
40	9	0.2925	100.00
Total	3077	100.0000	

0.4.9 Analytic Thinking, Cognitive Reflection Task (CRT)

(Frederick, 2005; Shenhav et al. 2012)

Responses must be coded. Skipped.

0.4.10 Self efficacy, New General Self Efficacy Scale (NGSE)

(Chen et al., 2001)

8 items.

Uses Likert scale of 1 = “strongly disagree” to 5 = “strongly agree”.

0.4.11 Persistence, Attitude Towards Mathematics Survey (ATMS)

(Miller et al., 1996)

8 items.

Uses Likert scale of 1 = “strongly disagree” to 5 = “strongly agree”.

0.5 Split Data