Validation & Replication Report

Rachel Ganly

10/4/2021

Description This is a replication report for the paired replication exercise, completed for the CaND3 Fellowship 2021-22 Replication Module. Replication Data code is stored at https://github.com/rganly/CAnD3-Data-Activity

Author of original data exercise: Tyler Bruefach Completed: September 2021

Report template outline:

1. Data sources
2. Analysis data files
3. Code description
4. Stated requirements
5. Missing requirements
6. Computing Environment
7. Replication steps
8. Findings
9. Data Preparation Code
10. Tables
11. Classification
12. Reason for incomplete reproducibility

1. Data Sources Dataset is not provided since the data are not publically accessible. A link to the data and codebook is provided. Access conditions are described. The data are cited in the README. All original data and codes are available at https://github.com/tbruefach/CAnD3-Data-Activity

2. Analysis Data Files A codebook was provided for GSS Canada data via an online download link

Link to codebook: http://odesi1.scholarsportal.info/documentation/GSS31/c31pumf_families_codebook_E.pdf

3. Code description There are 4 provided R code .r files, one master .Rmd file and one read me .md file. 'README.md' is a README file 'Shell File.Rmd' is a master file which runs all other codes. 'Cleaning Data.R' and 'Handling Missing Data.R' are data preparation and analysis code. 'Table 1.R' and 'Table 2.R' create tables.

Shortcomings:

- Shell file does not successfully run all other codes.

4. Stated Requirements The following software programs are required to reproduce these analyses:

R and Rstudio (version 1.4.1717) and the following packages as of 9/22/21 + tidyverse
+ haven
+ skimr

+ naniar
+ Hmisc
+ sjlabelled
+ gt
+ gtsummary

These analyses were conducted using Mac OS Catalina (version 10.15.7):
2.9 GHz Dual-Core Intel Core i7 Processor 8 GB 1600 MHz DDR3 Memory

5. Missing Requirements • Computational Requirements were not specified • Time Requirements were not specififed

6. Computing Environment of the Replicator

```
Mac Laptop:MacBookAir7,2
Version: MacOS Big Sur Version 11.2.3
```

Processor Name: Dual-Core Intel Core i7 Processor Speed: 2.2 GHz Number of Processors: 1 Total Number of Cores: 2 L2 Cache (per Core): 256 KB L3 Cache: 4 MB Hyper-Threading Technology: Enabled Memory: 8 GB

7. Replication steps In the file named 'Rachel_CleanData&HandleMissing.Rmd' I did the following

1. Downloaded code from URL provided.
2. Load required packages for analyses.
3. Convert data from the GSS into a tibble.
4. Clean and code variables to be used in the analyses, but I had to read the R code to find a list of variables to be used.
5. Dropping all variables but the ones used in analyses, but I could only find the list of variables to be used by looking at the code. I had already dropped most values earlier so I initially skipped this step but went back to add it later. 6.Label variables, but I looked at the code as there were no notes on labelling of variables.
6. Recode missing values of self-rated health and self-rated mental health (other measures were assigned missing values during the cleaning phase) but I could not find out how to do this using the instructions or comments so I had to read and copy the code.
7. Create an index called "sampmiss" that is a count of how many variables that each respondent has missing values.
8. Create a data set called "sample" that only contains cases with no missing values.

In the file named 'Rachel_CreateDescData&RegressionTables.Rmd' I did the following

1. Load required packages for analyses.
2. Create summary statistics table, I ran Tyler's code 'Table 1.R' to do this as there were no instructions about which variables to use.
3. Create regression results table, I ran Tyler's code 'Table 2.R' to do this as there were no instructions about which variables to use in the regression or which standard. errors to use.

8. Findings Almost matched the numbers in the tables one and two, with some small differences.

Find that there are some slight differences in the distribution of mental health and education across people identifying as male and female. Male and female respondents had nearly identical distributions of self-rated health. Although a greater proportion of female respondents had more than a secondary degree than male respondents, they also had lower household income.

9. Data Preparation Code • 'Shell File.Rmd' ran with many errors and failed to produce output • 'Cleaning Data.R' ran without error, output expected data • 'Handling Missing Data.R' ran without error, output expected data • 'Table 1.R' ran without error, output expected data • 'Table 2.R' ran without error, output expected data

10. Tables • Table 1: Looks almost the same but total individuals by sex is slightly different; perhaps because I coded these differently • Table 2: These are almost the same.

11. Classification Full reproduction with minor issues

12. Reason for incomplete reproducibility The written programme did not make it clear how to re-code all variables and exactly which variables to keep or drop.