Topic 1: Introduction and Preparation

1. Please find the paper "Reshaping Adolescents' Gender Attitudes: Evidence from a School-Based Experiment in India" by Diva Dhar, Tarun Jain, and Seema Jayachandran and read it as much as you can. It will be the paper on which our further course and exercise are based. Please complete the reading for the paper.
2. Get your hands dirty with data!
    a. Please download the "Replication Package" for the paper you read this morning, "Reshaping Adolescents' Gender Attitudes: Evidence from a School-Based Experiment in India". You can access the "Replication Package" on the American Economic Association website through our ND account. If you haven't registered, you can do so for free and quickly.
    b. The "Replication Package" should be downloaded as a folder. Open the folder and locate the "readme.txt" file.
    c. The "readme.txt" is generally referred to as a README file. Please read the README file carefully and use your own words to explain its purpose.
    d. Use the organization structure provided in the README file to build our own replication folder.
    e. Follow the instructions to insert all coding files into corresponding subfolders. And insert all data files into the "Analysis data" subfolders. How many files are in the "_ado" subfolders?
    f. (Bonus) Update the directory path in "0_master_run.do" as directed in the README file. Please note that the line index may be different.
3. According to the paper that you read, Dhar et al. (2022), please answer the following questions:
    a. What is the "intervention" in this social experiment?
    b. What are the outcome variables of interest?
    c. How many schools are in the treatment group, and how many are in the control group? (Bonus) On page 905, the authors say, "The unit of randomization was the school", what does that mean?

d. Do the authors take a balance check between treatment and control groups? Which table shows the balance check?
e. Which table shows the main short-run results of this intervention?
f. What are the concerns in interpreting the changes in self-reported outcomes? How do the authors deal with this issue?
g. (Bonus) On page 906, the authors say that "In cases where a village had more than one government secondary school, we chose at most one of them for the sample to minimize the possibility of spillovers". Could you try to explain the reason behind it based on the logic of RCTs?
h. (Bonus) On page 919, paragraph 1, the authors say that "the 11 percent fade-out in the treatment effect is not because the treatment group held less progressive attitudes at the second end line than at the first end line; rather, their attitudes improved less between the two waves than the control group's attitudes did." Can you combine the numbers in Table 2 and Table 8 to explain this sentence in your own words? (Hint: think about what the estimated results 0.18 and 0.16 mean)

Topic 2: Intro to STATA

Question 2.1: Basic commands

Import data set "*baseline_census_cleaned.dta*" from the "Analysis Data" subfolder to Stata.

1. Use a command to get the number of observations and the number of variables in this dataset.
2. It is a large dataset, and we do not need to use that many variables for today's practice. Just keep the following the six variables: Ctot_p, Cp_lit, Cp_ill, treatment, tru, subdistt
3. What is each of the six variables' label? According to their labels, can you understand what these six variables are?

4. How many subdistricts are in this dataset?
5. How many observations are in the treatment group? Which variable gives you this information?
6. How many people are in the **most** populated village?
7. Generate a variable "ill_rate" that shows the illiteracy rate in each village, and label it "the illiteracy rate in each village".
8. In "subdistt" 00379, how many villages in this sample are located in urban areas?
9. What is the data type of variable "subdistt"? Convert it to the other data type.
10. (Bonus) Can you check this data by adding variable "Cp_lit" and variable "Cp_ill", and then see whether the sum is equal to variable "Ctot_p"? Are they should be equal?

Question 2.2: Do- and log-files

Please re-write Question 2.1, but this time, do it by using do-files and log-files (including the data importing command).

- Create a subfolder, "Exercise", in the replication main folder "Main Analysis and Paper" and save your dofile document in that "Exercise" folder.
- Try to use commands to import data (cd, import) using both absolute and relative paths. This is a good way to get familiar with the directory and path!
- Write at least three pieces of comments in your dofile.
- The final submission should include one dofile and one log-file

Question 2.3: More tricks

Continue with the "*baseline_census_cleaned.dta*" data you played with in previous questions:

1. Import the data to Stata and find the two variables that relate to the information of "district".
2. Keep the six Ctot_p, Cp_lit, Cp_ill, treatment, tru, subdistt variables, AND the two new district variables in this dataset.

3. In this dataset, how many sub-districts are in each district?
4. What are the average village population numbers in treatment and control groups?
5. Please generate a series of dummy variables (locate) according to the value of "subdistt". How many "locate" variables do you generate?
6. Use ONE line of command to summarize those "locate" variables separately.
7. This dataset provides us with the name of each district. Please attach the district name to the district index through label define/label value (there is a bit of inconsistency that one district code corresponds to more than one district name. Use the main one for exercise purposes. But if you encounter the same problem during actual research, you need to find out why does this situation happen and then decide what to do)

Question 2.4: Programming

Part A

- Import the data "lfs_examples_class08.dta" to Stata
- keep the variables: UqNr age earnings_week black female hours Indus_Sep2003

Questions:

1. Find relevant variables to calculate the hourly wage variable wage_hr
2. Generate a histogram of wage_hr
3. Generate a new variable that equals the natural log value of wage_hr: ln_wage_hr
4. Generate a histogram of ln_wage_hr
5. Compare the number of variables in these two histograms. Are they the same?
6. Can you guess why we usually take log to variables like wages?
7. Can you find the average hourly wage level for black female workers in this sample?
8. Can you find the average hourly wage level for male workers above-average age? (Hint: use scalar to store the mean value for the judgment in the next step)

Part B

- Let's go back to our replication folder of Dhar et al. (2022)
    1. Please open the main do-file in the replication folder: 0_master_run.do
    2. What does the macro "main_loc" represent? Can you explain this line of code: "global deid "$main_loc/Main Analysis and Paper/Analysis data"?

Part C

- We played with census data and got to know some information about the villages in the sample. Now let's take a look at the students' situation.
- Import the data "baseline_student_raw.dta" into Stata.
    1. Which variable tells you the student's age? How do you find it?
    2. Find what ".s" ".d" means in the data. What will you do if you want to exclude **all** missing values?
    3. There are eleven sib_age* variables asking their siblings' ages. Create a loop using "forvalues" to summarize these variables. There are eleven variables for each sibling-related question. Can you tell me why and provide some data evidence?
    4. Can you count how many siblings each student has? How many students have eight siblings?

Part D

- [These questions may be the tough ones, and answering them requires you to spend some time understanding the data and the coding logic.]
- Let's continue the exploration of "baseline_student_raw.dta" data:

  Questions:

    1. Can you generate a variable that represents the age of each student's **youngest married** siblings (marital status ==1)? Hint: there are many ways to achieve this goal, and you might need to Google some new commands to realize your ideas.
    2. Get to know the label values for the variables sib_high_edu_level*
    3. For families with two children, what's the average highest education level for these two children(for now, just to take an average of the sib_high_edu_level*)?

How about families with five kids? or six kids? (Let's assume each family has only one child represented by child_id entering this dataset, in other words, each child_id can uniquely represent a family. Otherwise, the sibling to child No.1101002 might be the child No.1101006) Hint: these may require data type conversion. Try "des" if you need it.

4. What are the problems if I use the result in the last question arguing children in large families tend to have more/less education?
5. Pick any other two variables related to their attitudes to gender equality or their living standard that you are interested in. Please tell me what the data tells you.

Topic 3: Making Effective Graphs

Dataset using: *baseline_student_raw.dta*

Expected results: do-file and two graphs

1. Draw a picture showing the distribution of the oldest sibling's highest education level. Change the title of the picture and change the picture's theme to whatever you like.
2. There is a variable that shows whether the student has a TV in her house. And for those who have a TV, there are six variables showing what they watch on TV "what_tv_*". Make a bar picture where the horizontal axis shows the type of TV program, and the vertical axis shows the percentage of students who watch it (among the students who have a TV at home). Several other alterations are expected:
    - change the legends to their proper meanings
    - change the unit of the y-axis to the percentage (this should be in "Options" menu if you use GUI)
    - label the bar's height and keep the first two digits after the decimal point
    - Add a note to specify the data source ("Data source: Survey data collected by Dhar et al.(2022)")
    - Add a title to this graph

Topic 4: Linear Regression

Dataset using: *replication folder*

Submission: log-file

1. Run the 0_master_run before "PART 4: FINAL INDEX GENERATION AND ANALYSIS ". (The next line should be: do "$do/04a_merge_indices.do")
2. Find the corresponding "04a_merge_indices.do" do file. Run through all the codes **before** "ENDLINE 2 PRIMARY OUTCOMES: INDICES". This should give you some of the final index and related variables that you need. Save the dataset.(Set the path first, then "save xxxx.dta, replace")
3. Implement the following three OLS and extract the results together by outreg2. The results you would get will not be the same as in the paper because there are still some data processing details we did not execute and because we do not include other control variables.
    - reg E_Sgender_index2_ni_nd B_treat B_Sgender_index2
    - reg E_Saspiration_index2_ni_nd B_treat B_Saspiration_index2_ni
    - reg E_Sbehavior_index2_ni_nd B_treat B_Sbehavior_index2_ni

Topic 5: Relational Datasets

Question A:

Dataset using: *baseline_parent_raw.dta*

1. keep child_id, child_age, and all variables that store information about the child's siblings (sib_*). After this step, you should have 112 variables in your dataset.

2. This survey asks ten questions for each sibling the child possibly has. (You can order them first) Please reshape the database into the following form:

Child_id	Child_age	Sib_count	sib_age	sib_gender	...
1	13	1			
1	13	2			
...	13	...			
1	13	11			
2	14	1			
2	14	2			
....	...	...			
2	14	11			
...					

Question B:

Dataset using: *lfs_examples_class08.dta*

1. Please make a graph that shows the relationship between (age-average) wages and age. (Collapse)
2. Divide the sample into several 10-year age groups, like 25-35, 35-45, etc. You need to choose the age dividing start point as the minimum age level in the database. And then calculate each person's relative hourly wage to the average hourly wage in their 10-year age group. (Suppose one person's hourly wage is 15, and the average wage in her age group is 20, then the relative wage is 15/20=0.75.)
3. After you get the (age-average) wages, save this after-collapse dataset, and merge it with your original dataset. Calculate the relative wage with respect to the average wage of this person's age group. What is the other way that you can directly calculate the relative wage from the original database without “collapse” and “merge”?

Question C:

Dataset using: *replication folder*

1. import the "baseline_parent_raw.dta" into Stata. How many variables and how many observations are in this dataset? Hint: try "des,short"
2. import the "baseline_student_raw.dta" into Stata. How many variables and how many observations are in this dataset?
3. merge the "baseline_parent_raw.dta" and "baseline_student_raw.dta" by the variable "child_id". How many of them are merged? How many of the un-merged are from the parent dataset? How many variables are left? Is there something wrong?
4. Go back to the replication folder and find the do-file "0_master_run.do". Before the merge, the authors have commands like:
5. *rename * P*
6. *rename Pchild_id child_id*
7. Can you explain why they include these two lines of commands for the parent data?
8. Merge the two datasets again, and for this time, please make sure the final variable number = variables from student + variables from parents - 1 +1
9. Continue to merge this dataset with "baseline_school_cleaned.dta"