# Instructions

You have three hours to complete this test. All files necessary for completing the tasks are included in the zipped folder. Please provide a complete record of your work, which includes a do-file, a log-file, and this Word document with written responses where necessary.

You will be graded on the final responses and code, and on the methodology and reasoning used in the process. This test is meant to judge your general reasoning and problem-solving ablity besides your STATA skills, so if there are any questions that you cannot answer using STATA, please explain what steps you would have taken had you known the appropriate commands for partial credit. Keep in mind that some questions are somewhat open-ended/ambiguous and that there is no "correct" answer.

# Part I. Data Management (40%)

An implementing partner is running a remote phone survey data collection exercise, and is sharing the data collected daily for you to check the quality of incoming data (known as a high-frequency checks).

a) We have obtained data from the partner organization with the age of the survey respondents. However, the file that the organization has sent to you, *respondent\_age.csv*, is not in .dta format. Import and save this dataset as STATA data file with appropriate variable names and data formats. **(5%)**

b) Our main dataset is called *main\_dataset.dta. Use this data as the base file and*  merge in the *respondent\_age* in this data using the file in (a). **(10%)**

c) The most recent data is shared in a separate file *new\_observations.dta.* Add these observations to the main dataset. **(5%)**

d) When working with the survey data, it is important to protect personally identifiable information (PII). Generate a new variable to replace the ‘enumerator’ variable with an ‘enum\_id’ variable that is numeric but does not include the enumerator name. Remove the `enumerator’ variable from the dataset and properly label the new variable. **(5%)**

e) The `surveydate’ variable is in string format. Format the survey date in a recognizable data format and generate a new variable ‘month’ which contains only the survey month. **(5%)**

f) Survey callbacks and response verification are important components of data quality checks. Export into Excel a random sample with 5% survey respondents, who will be called back for response verification. The Excel file should only include the following three variables: userid, surveyid, surveydate. **(10%)**

# Part II. Data Analysis (60%)

One of the objectives of the survey is to measure the labor market impacts of COVID-19 in the country. You have been tasked with preparing a presentation that summarizes the core impacts of COVID-19, and whether it impacted different groups differently. For this task, you can output the data into excel and generate the necessary graphs. Please make sure the output on excel is self-explanatory and sheets well-labelled. Please answer each question below in the same document and use the generated figures as necessary.

A pre-defined indicator variable, called ‘covid\_deterioration’ is included in the dataset to identify whether a respondent faced a labor-related shock due to COVID-19.

1. Identify 6 variables from the dataset that should help your manager understand the impact of COVID-19 labor deterioration on different demographic and socio-economic characteristics. Give a one sentence explanation for each variable chosen. **(15%)**

* v3\_02(Gender): because of cultural expectations. And the burden of responding to a crisis. A male may take on additional burden, and may thus be subject to greater risk. Conversely, a woman may be expected to do the same
* v3\_04(Education level): as a proxy for income, years of schooling serves to tell us whether respondent had the wherewithal to respond to a crisis, either monetarily or rationally
* v3\_09 (City/district/village): due to the contagion and clusters of areas where covid cases spiked and were the highest
* v3\_10(Household size): for the same reasons, having dense households, with less space per person could enable more rapid contagion from carriers.
* Age: this is certainly a contributing factor to covid related illnesses as the experts have made clear
* v6\_02\_1(emergency funds): since covid was an unexpected emergency, this is a more reliable predictor of a risk responses than income range or industry worked in

1. Prepare graphs on Microsoft Excel, showing the relationship between the 6 variables you selected and the covid\_deterioration variable. First, export the required tables from Stata to excel, then use Excel to create the graphs. In a maximum of two paragraphs, explain your choice of graphs to display this relationship. **(15%)**

I chose to use bar graphs to display all of my variables’ relationship to covid deterioration. The primary reason for this is that bar graphs show a side-by-side comparison of binary categories in a visual manner that is best for inferring the effects of each category on the outcome of interest. Here, the outcome of interest, and the binary variable, is covid deterioration. By breaking down covid deterioration by gender, for instance, we see which category was more likely to be affect adversely (gender = M/F, covid\_deterioration = Faced deterioration/No deterioration).

Furthermore, bar graphs can either be ordered. For x variables that are numeric integers, such as household size and age, we can see more of a steady progression of the variable, and breakdown of deterioration at each level. Other variables can be treated as unordered categorical variables in bar graphs.

1. Your team lead is interested in labor force participation after the pandemic. Given the variables at your disposal, calculate the labor force participation rate. Explain your logic in one paragraph below. **(15%)**

For the labor force participation rate, I chose to use survey respondents who had worked in the last 7 days and respondents who had actively been seeking work for the last 7 days to represent the labor force. This served as the numerator of the ratio calculation. The labor force based on this assumption = 3802. Since the population is age 15 and above, I chose not to modify the denominator and to leave it as is. Although typically the base age is 16. This brings us to an overall labor force participation rate of 3802 / 7692 = 49.4279%

1. Using a Linear Probability Model to conduct a regression analysis with covid\_deterioration as the dependent variable, and the 6 variables you chose as independent variables. Export the regression table into Excel. In two paragraphs, explain the relationship between each of the 6 variables and covid\_deterioration. **(15%)**

Of all the variables used as predictors in our analysis, only household size did not bear a significant relation to the overall deterioration experienced from Covid-19. The most significant variable impacting covid deterioration was gender, with being female reducing the likelihood of a deterioration from covid by up to 30 percentage points, at the 99% confidence level. This is perhaps because males constitute a greater percentage of the workforce in developing countries.

Other variables that had a highly significant effect on covid deterioration at the 99% confidence level were access to emergency funds, educational attainment, and age. An access to emergency funds significantly reduced the likelihood of being negatively impacted by covid 19 by 8 percentage points. Each additional level of educational attainment significantly reduced the likelihood of being negatively impacted by covid 19 by 6 percentage points. Age meanwhile, increased the likelihood of being negatively impacted by covid 19 by 6 percentage points for each additional year in age.

1. You have learned that during the survey design, the probability of selecting a female out of the relevant study population was 2%, and the probability of selecting a male was 3%. Create a variable called ‘pweight’ that includes probability weights to adjust for the gender differences in probability of selection. In a brief paragraph, explain your logic below. **(Optional Question, and you may skip it)**