This read me document describes the structure of some of the data files used in the **Teacher Performance Pay: Experimental Evidence from India** paper published in the Journal of Political Economy. Only those files that may need explanation are described here.

To run these analysis files all you need to is to unzip the folder “Teacher Performance Pay - Replication Files”. The required folder and sub folders will automatically be created and the only extra step is to open STATA on your computer and change the working directory to the location of the downloaded files.

Note that the bootstrap routines (for the confidence intervals for Figures 3-5) can take a long time to run and so you may want to change the number of repetitions from 1000 to 10 before running the files (especially for Figure 5 with the teacher fixed effects).

1. **Incentives\_JPE\_analysis\_starter\_file.dta**

The main data file for all the tables in the paper (except sample balance in table 1 , and teacher behavior in table 8) is Incentives\_JPE\_analysis\_starter\_file.dta. A few notes on the file:

1. The file is a student level file with test score data, teacher data and school data merged in.
2. Each row in the file is unique by student, subject and level where level refers to one of the two rounds of endline testing that was administered during the year. Level 1 refers to the first round of testing which tests material from the previous class and is known as lower endline (LEL) while level 2 is the second round of testing which tests material from the current class and is known as a higher endline (HEL).[[1]](#footnote-1)
3. The key identifying codes are: apfschoolcode is the unique school code; hplstudentkey is the unique student code
4. The year of the test is indicated in the variable name, i.e. y2\_perc is the year 2 percentage score. From the subject and level column one can understand whether the particular score is a math or telugu test and whether it was a lower endline or higher endline test.
5. The file contains normalized test scores. Test scores in year 0 were normalized based on the distribution of all schools, while test scores in years 1 and 2 are normalized based on the distribution of the control schools in the project. The do file for the normalization can be provided upon request.

Explanation of some key variables:

1. **y1\_nts\_level\_mean** and **y2\_nts\_level\_mean:** These two variables are the mean normalized test scores over the lower and higher endline of that particular subject. These are the key test score variables used in the paper.
2. **repeat\_perc\_with\_HEL\_y1, non\_repeat\_perc\_with\_HEL\_y1, repeat\_perc\_with\_HEL\_y2, non\_repeat\_perc\_with\_HEL\_y2:** The label of this variable in the data set is self explanatory. A few things to note :
   1. Each student has ONE repeat score per year per subject test that they took. That is, we calculate the student’s percentage correct on repeat questions in BOTH LEL and HEL rounds.
   2. On the “\_with\_HEL” part of the variable name : This was to indicate that this captures a student’s score on repeats that occur within the year (i.e. between the HEL and LEL rounds) AND repeats that occur across years, i.e. questions that showed up in a previous year.
3. **teacher\_class\_y1.dta and teacher\_class\_y2.dta**

These files map the teachers in the schools to which classes they have taught. Most of the time a teacher teaches more than one class and occasionally (about 5% of the time) more than one teacher teaches the same class. These two files are used solely to calculate teacher turnover and teacher attrition.

1. **y0\_y2\_assessmentfile.dta**

This file is also a test score file structured slightly differently with one observation per student and is thus used to calculate student turnover and attrition. It is only used once in table 1: sample balance.

1. **enrollment\_cleanup\_y1\_y2.dta**

List of students who should not count in the calculation of student turnover and attrition. After data cleaning they were deemed as not part of the enrollment list. They are merged into the test score file in (3) above and removed from the calculation of student turnover and attrition.

1. **apf\_treatmentcode.dta**

This file lists the school code of the schools in the project and the treatment that was assigned to them. This information (treatment code variable) is already in the main analysis file but is there for reference and is used once in the do file Incentives\_JPE\_table1B.do to create the teacher turnover and attrition variables.

1. Class 1 has only level 1 of testing (although it should probably be called level 2 since it is based on the current class material, but in terms of timing it was administered during the lower endline and is hence labeled as such. [↑](#footnote-ref-1)