1 Instructions for Replication

These files are the code to replicate "Air Filters, Pollution, and Student Achievement" by Michael Gilraine. Only some (limited) school-level characteristic and testing data are provided as the student-level data used in the project are not publicly-available. This document describes the code (in Stata) that is available to replicate the paper. All files are available from the Harvard Dataverse at https://doi.org/10.7910/DVN/SQ0TBY.

1.1 Obtaining the Data

Unfortunately, all of the data used for the paper are from the Los Angeles Unified School District (LAUSD). The URL of the website (as of January 2023) for data applications is: https://achieve.lausd.net/Page/15939. This website lists the current requirements to obtain these data, which includes the submission of a research proposal that must be approved by the district.

The raw data files provided by the LAUSD come in either excel (.xlsx) or comma separated (.csv) format. These raw files were directly read-in by STATA for analysis. All analysis was done on STATA 15.1.

2 Analysis Dataset and .Do Files Explanation

All .do files have a number in front of them. If you run the do files in the sequence of the numbers all the results from the paper will be created (.do files with the same number but differentiated with a letter can be run in any sequence). I have placed the .do files into two folders: the "Code to Generate Analysis Dataset" are .do files that create data, while the ".Do Files" folder contains .do files that run the analysis. We now go through what each .do file does and which results it generates.

2.1 Dataset Generating .Do Files

The dataset used for the main analysis is "main_analysis_data.dta." The dataset is constructed using the LAUSD data and so cannot be shared. Two .do files construct this dataset:

1. "1 - Get_Test_Data.do": constructs the data set "VA_test_scores.dta." The .do file takes the raw test score data provided by the LAUSD and outputs a dataset with math and English test scores at the student-year level for grades 3-8 from 2002-03 through 2012-13 and 2013-14 to 2015-16.

2. "2 - Get_Demo_and_Create_Analysis_Data.do": constructs the data set "main_analysis_data.dta." The .do file combines the student demographic and enrollment data provided by the LAUSD from 2002-03 through 2015-16 at the student-year level and then merges on two datasets. First, it merges on the test score data "VA_test_scores.dta." Second, it merges on school characteristics (crucially distance to gas leak) which were gathered by the author. These data are publicly-available and so are provided in the same folder as these .do files under the file titled "lausd_distance_chars_bound_data.dta." Once all merged together, the .do file saves the resulting data as the main analysis data: "main_analysis_data.dta." This .do file also creates Table 1 (summary stats).

3. "3a - Create_School_Stability_VA.do": relatively minor .do file that constructs two data sets that contain school-level stability rates and value-added measures ("school_stability_rate.dta" and "school_VA.dta") which are used for summary statistics and as controls.

4. "3b - Create_ClassSize_TeacherExp.do": relatively minor .do file that constructs a data set that contains school-level teacher experience and class size measures ("class_size_teacher_data.dta") which are used for summary statistics and as controls.

As stated, the .do file "2 - Get_Demo_and_Create_Analysis_Data.do" calls an additional data file, which is generated from publicly-available data and so is provided in the folder "Dataset Generating .Do Files." I describe this additional data file here:

1. "lausd_distance_chars_bound_data.dta": Dataset created by the author using publicly-available data sources. These data contain: school latitudes and longitudes, distances of the school to the gas leak calculated using STATA from the point (34.315467, -118.563731) which is considered the gas leak location, and school characteristics (e.g., charter and magnet status).

2.2 Analysis .Do Files

The following .do files generate all results in the paper. We state each analysis .do file and what Table/Figure it creates here:

1. "4 - Main Analysis.do": Creates most of the main results in the paper. Includes results in Figures 3, B.3, B.4, B.5, B.6, B.8(b) and Tables 3, 4, 5, 6, 7, B.3, B.4. The inputs required are the main analysis data set "main_analysis_data.dta" along with "school_stability_rate.dta," "school_VA.dta," and "class_size_teacher_data.dta."

2. "5 - Event_Study.do": Creates the event study and difference-in-differences results. Includes results in Figures 4, B.7 and Table B.2. The inputs required are the three

data sets that generate the main analysis data set ("VA_test_scores.dta" from "1 - Get_Test_Data.do," "VA_demographics.dta" from "2 - Get_Demo_and_Create_Analysis_Data.do" and "lausd_distance_chars_bound_data.dta") along with the data sets "main_analysis_data.dta," "school_stability_rate.dta" and "class_size_teacher_data.dta."

2.3 Figure 2 and Testing Data

The code and accompanying data can be fully released for the school-level testing data (as well as the map shown in Figure 2). These files and data are provided in the folder "Air Testing Data and Code."

1. "Figure 2 Code.txt": Creates Figure 2. Does not require any restricted data and can be run by anybody. The file creates Figure 2 by using the google maps web developer interface. The code is written in javascript and the map can be generated by copy and pasting the contents of the .txt file into the appropriate JS fiddle 'columns' (see instructions in .txt file) here: https://jsfiddle.net/.

2. "air_testing.do": Creates Table 2 using the provided data file "Testing_Data.xlsx."

3. "Attendance_zones.do": Creates Figure B.8(a). To do so, it uses the shapefile data from the subfolder "School Zones" which is the shapefile from the 2015-16 school attendance boundary survey (primary schools shapefile) available here: https://nces.ed.gov/programs/edge/sabs.

As stated, the .do file "air_testing.do" calls the data file "Testing_Data.xlsx" which is generated from publicly-available data and so is provided in the same folder. I briefly describe this additional data file here:

1. "Testing_Data.xlsx": Dataset created by the author using publicly-available air quality testing data. The original data can be found in .pdf format at https://achieve.lausd.net//site/Default.aspx?PageID=10329. Table B.1 (basic testing summary statistics) is directly derived from this data.