**DOCUMENTATION FOR REPLICATION OF RESULTS IN *SCIENCE ADVANCES* ARTICLE**

**Part I: Data Needed**

The majority of the data used in this paper comes from student-level education records maintained by the Michigan Department of Education (MDE). This data is not publicly available, but can be requested by researchers who agree to adhere to privacy and security protocols. This section explains what data is used in the analysis and how to obtain it.

(*Note* *that the data files maintained by MDE are “living” in the sense that the department continually cleans and improves them as they get additional information about students. The vast, vast majority of records will be constant over time. However, a tiny number may undergo changes. For example, in the version of the data available to researchers in 2021, student A might be listed in school 73 in year 2013. However, in the 2024 version of the data, student A might be listed in school 76 in year 2013. Again, such changes are* ***extraordinarily*** *rare. We note this simply to alert other researchers that there may be very minor differences between some of the numbers shown in the* Science Advances *article and the numbers that they obtain during replication. The original data creation for this paper took place primarily in 2019. As part of pulling together this replication package, we recreated the necessary data files using the data available to researchers in spring 2024.* *As an example of these kinds of small data changes, 149 (or 0.72%) of the Michigan public school students reported to reside within the geographic borders of Flint in 2013-2014 or 2014-2015 whose addresses were able to be matched to the “pipes data” in 2019 (see below) were no longer recorded in the data in 2024 as Flint residents. However, even when using updated MEDC data with these kinds of minor differences, we were able to replicate patterns nearly identical to those shown in Figures 2 (mean educational outcomes over time in Flint) and 4 (synthetic control estimates of the effect of the FWC on student outcomes).*

1. Student Longitudinal Data from the Michigan Education Data Center (MEDC). You can find details on making such a request [here](https://medc.miedresearch.org/research-application). You will need to request the following data files. Note that these files are versions of the Michigan education data that have undergone additional processing by researchers at the University of Michigan Education Policy Initiative (EPI) and are informally known as “EPI” files.  
   1. **K-12 Student Demographic and Enrollment Data:** this student by school by year-level dataset contains information about K-12 graduation, enrollment, demographics, and program participation for the population of public school students in Michigan from 2004 to the present. It is currently called **k12\_student.dta**; more details on the data used to create this file is available [here](https://medc.miedresearch.org/dataset/k-12-student-demographic-and-enrollment-data).
   2. **K-12 Student Assessments**: these two student by year by assessment-level datasets contain information about the assessments which are (typically) required to be taken by Michigan students. The first, currently called **assessment\_standard\_student\_year**, includes those student assessments intended to capture the performance of the K-12 educational system and spans 2007 to the present for public school students in Michigan in certain grades; the second, currently titled **assessment\_readiness\_student\_year**, includes those assessments pertaining to college and career readiness for public high school students in Michigan from 2008 to the present. More details on the data used to create these files are presently available [here](https://medc.miedresearch.org/dataset/k-12-student-assessments) and [here](https://medc.miedresearch.org/dataset/k-12-student-assessments-career-and-college-readiness), respectively.
   3. **K-12 Educational Institutions:** this entity by year by collection-level dataset contains information on the various educational entities in the state, including location data (longitude, latitude, address, county), entity type, and education programs provided (virtual education, adult education, grade levels taught, early middle college, etc.) for all K-12 education buildings in the state of Michigan from 2006 to the present. It is currently named **school\_level.dta**; more details on the data used to create this file are presently available [here](https://medc.miedresearch.org/dataset/k-12-educational-institutions).
2. Pipe data from inspections conducted by FAST Start and researchers at the University of Michigan and the Georgia Institute of Technology (provided in the “data” folder).
   1. The “raw” data is contained in a file named **parcel\_inspections.dta**. This dataset is unique by parcel (or “lot” of land – essentially a property). It contains information on the address of the building along with a variety of its characteristics such as size, location, year built, value. Most importantly, it contains information on the material used in the plumbing lines for the building (e.g., copper, lead, galvanized steel); the preferred operationalization of this information is contained in the ‘dangerous’ indicator variable.

**Part II: Matching of “pipes data” (i.e., parcel\_inspections.dta) to Student Records**

Several analyses in the *Science Advances* article report subgroup effects for students who were (and were not) living in homes with lead pipes prior to the Flint Water Crisis and thus may have been at higher risk for lead exposure. In order to conduct these analyses, researchers must request that MEDC match Flint students to the “pipes data” via their home address. The directions below walk researchers through the necessary steps to achieve this task; merging the ‘pipes data’ to student addresses in the longitudinal MEDC student data involves a multi-step process, which we describe in detail. Relevant code is provided in the “syntax/pipes” folder. You need to run the .do files in the order shown below.

1. **make\_pipes\_data\_for\_matching.do** - reduces **parcel\_inspections.dta** to a dataset unique by property address, removing parcels without comprehensive or accurate address information and accounting for multi-unit buildings sharing the same inspection data. The intermediate dataset produced by this file is titled **flint\_pipes\_data\_for\_matching.dta**.
2. **make\_medc\_data\_for\_matching.do** - creates a dataset of annual observations of all students residing within the geographic borders of either Flint in the 2013-2014 or 2014-2015 school years. This file produces an intermediate dataset titled **FlintResidentLEAS2014\_and\_2015.dta**
3. The datasets created in steps 1 (**flint\_pipes\_data\_for\_matching.dta**) and 2 (**FlintResidentLEAS2014\_and\_2015.dta**) should be sent to research staff at MEDC who can conduct a probabilistic match using address number, street direction, street name, street name suffix, and zip code. MEDC staff will then provide researchers with a file of matched students, which we will call **matched\_pipes.csv.** This matched file should contain multiple observations per student corresponding to the different time periods (academic years and/or collection periods within academic years) the student’s address was found in the set of parcels in the pipes dat..
4. **clean\_matched\_pipes.do** - lightly cleans the matched data (**matched\_pipes.csv**) and collapses it to the student level, keeping only the few variables that operationalize the presence of lead in-service lines for the heterogeneity analyses described in step 4a of Part II below. The final dataset is saved as **matched\_pipes\_clean.dta**. This data set will be an input for the code listed below.

**Part III: Creation of Analysis Data Files**

The analysis in the *Science Advances* is conducted using a series of district x year level datasets. This section explains how each of these data sets are created. All code is provided in the “syntax/build” folder. All data from Part I should be placed in the “data/raw” folder. All .do files for creating analysis data files should be run in the order presented.

1. **create\_flint\_individualpanel\_nonfixed\_55districts.do**  
   This file combines the raw MEDC datasets and produces an intermediate dataset (**student\_level\_nonfixed\_55districts.dta**) of annual observations of all students residing within the geographic borders of either Flint or any of the 54 other districts in the donor pool.
2. **create\_flint\_geodistpanel\_nonfixed\_55districts.do**  
   This file takes the intermediate dataset of student observations created in the previous step (**student\_level\_nonfixed\_55districts.dta**) and collapses it to the geographic district-year level. This will create **mi\_geodist\_panel\_dy\_nonfixed\_55districts.dta,** a district-year panel using time-varying district assignment containing Flint and our donor pool of 54 potential synthetic control districts (*nonfixed*).
3. **create\_flint\_individualpanel\_nonfixed\_all\_districts\_2014.do**  
   This file combines the raw MEDC datasets and produces an intermediate dataset (**student\_level\_all\_2014.dta**) of annual observations of all students residing within the geographic borders of all districts across Michigan with at least 1,000 students in the 2013-2014 school year.
4. **create\_flint\_geodistpanel\_nonfixed\_alldistricts\_2014.do**  
   This file takes the intermediate dataset of student observations created in the step above (**student\_level\_all\_2014.dta**) and collapses it to the geographic district-year level, creating **mi\_geodist\_panel\_dy\_all\_2014.dta**, a district-year panel using time-varying district assignment containing all Michigan districts that enrolled >1000 students in the 2013-2014 school year (*all*). (*Note that the threshold of >1000 students includes students in ungraded special education classrooms who are ultimately excluded from this panel and is for the 2013-2014 school year only; together, this sample restriction and changes in enrollment over time result in some district-years with <1000 students, even in 2013-2014).*
5. **create\_flint\_individualpanel\_fixed\_55districts.do**  
   This file combines the raw MEDC datasets listed above and produces an intermediate dataset (**student\_level\_fixed\_55districts.dta**) of annual observations of all students residing within the geographic borders of either Flint or any of the 54 other districts in the donor pool in the 2013-2014 school year, holding their residential district constant in all other years.
6. **create\_flint\_geodistpanel\_nonfixed\_55districts.do**  
   This file takes the intermediate dataset of student observations created in the step above (**student\_level\_fixed\_55districts.dta**) and collapses it to the geographic district-year level. This will create **mi\_geodist\_panel\_dy\_fixed\_55districts.dta**, a district-year panel that perpetually assigns students to the district they attended in 2013-2014 school year containing Flint and our donor pool of 54 potential synthetic control districts (*fixed*).
7. **create\_flint\_subgroup\_geodist\_data.do**This file iteratively takes the intermediate dataset of annual observations of all students produced in step #1 above (**student\_level\_nonfixed\_55districts.dta**) and then collapses to a geographic district-year-*subgroup* level, creating five (5) different district-year-subgroup panels with the naming convention **mi\_geodist\_panel\_subgroup\_nonfixed\_55districts.dta** for the following subgroups: gender (*dygender*), grade (*dygradehighlow*), SES (*dyses2impute*), service line material (*dypipes*), and district mobility (*dyevernf*).   
     
   (*Note that, due to a change in the variables in the datasets maintained by MEDC, the SES variable is no longer constructed as described in the article (i.e., the median-dichotomized first principal component of ﻿three variables—economic disadvantage, receipt of Supplemental Nutrition Assistance Program (SNAP) benefits, and receipt of Temporary Assistance for Needy Families (TANF) benefits—averaged across all years in which a student is observed). It is averaged in the same way but now relies only on a single variable—an economically disadvantaged indicator that MEDC now creates by incorporating much of the same information— instead of a principal component.*)

**Part IV: Analysis to Create Tables and Figures that Appear in the *Science Advances* Article**

Once the analysis data sets are placed in the “data” folder, one can recreate all results by running the Stata do file “**\_MAIN\_v10.do**”, which will call a number of other program code files in the “syntax” folder.