NBER WORKING PAPER SERIES

# AIR POLLUTION AND STUDENT PERFORMANCE IN THE U.S.

Michael Gilraine
Angela Zheng

Working Paper 30061
http://www.nber.org/papers/w30061

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2022

We would like to thank Claudia Persico and Robert McMillan along with seminar participants at the 11th European meeting of the Urban Economics Association for helpful comments. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Air Pollution and Student Performance in the U.S.
Michael Gilraine and Angela Zheng
NBER Working Paper No. 30061
May 2022
JEL No. I14,I24,Q53

## ABSTRACT

We combine satellite-based pollution data and test scores from over 10,000 U.S. school districts to estimate the relationship between air pollution and test scores. To deal with potential endogeneity we instrument for air quality using (i) year-to-year coal production variation and (ii) a shift-share instrument that interacts fuel shares used for nearby power production with national growth rates. We find that each one-unit increase in particulate pollution reduces test scores by 0.02 standard deviations. Our findings indicate that declines in particulate pollution exposure raised test scores and reduced the black-white test score gap by 0.06 and 0.01 standard deviations, respectively.

Michael Gilraine
Department of Economics
New York University
19 West 4th Street
New York, NY 10012
and NBER
mike.gilraine@nyu.edu

Angela Zheng
Department of Economics
McMaster University
Kenneth Taylor Hall
1280 Main St W
Hamilton, Onta L8S 4E8
Canada
zhenga17@mcmaster.ca

# 1 Introduction

Measuring the external costs of air pollution has been the subject of intense scrutiny – naturally so, given that a full accounting is necessary to implement policies that induce polluters to internalize the society-wide costs of their emissions. Recent research has highlighted the substantial costs of pollution in terms of health, worker productivity, crime, and decision-making (Aguilar-Gomez et al., 2022). Of particular note, air pollution can also impact student learning (Ebenstein et al., 2016; Persico and Venator, 2021; Gilraine, 2020; Duque and Gilraine, 2020; Heissel et al., 2020; Mullen et al., 2020; Marcotte, 2017); an effect that is likely to be of first-order importance given the link between human capital and economic growth.

A key environmental story of the past two decades has been the dramatic improvements in air quality in the U.S. (Currie and Walker, 2019). This pollution decline – largely driven by the decline in coal use – may therefore have materially affected student learning. In addition, data indicate that low-income and minority students are likelier to reside in high-pollution areas (Currie et al., 2020), signalling that improved air quality may have also reduced the pervasive test score gaps we observe in education.

This paper investigates the implications of the recent air quality improvements on America's education system. To do so, we gather satellite-based measures of fine particulate matter – measured by PM2.5 concentrations[1] – to gauge the air quality faced by students in the near-universe of school districts in the United States each academic year. These data are then linked to test score data from the Stanford Education Data Archive (Reardon et al., 2021) covering over 10,000 school districts from 2008-09 through 2017-18. We supplement these data with additional data sources that allow us to control for potential biases arising from other determinants of student achievement that are correlated with air quality, such as local economic conditions (Chay and Greenstone, 2003) and weather (Park et al., 2020). Uniquely, we can also directly account for student sorting by using information on the residential locations of over 80 million Americans aged 18-50 to control for the moving rates for each district-year in the country. An OLS regression indicates that each microgram per cubic meter ($\mu$g/m$^3$) increase in PM2.5 concentrations is associated with a 0.0035 standard deviation decline in student test scores.

To the extent that our detailed controls are insufficiently rich, however, concerns may remain that there exists some unobservable determinant of student achievement that is correlated with air quality. In addition, air quality is likely to be measured with error (Diao et al., 2019; Richmond-Bryant and Long, 2020), attenuating our estimates. We therefore turn to an instrumental variable approach to deal with these issues. Specifically, we develop two empirical strategies based on instrumental variables that leverage variation coming from nearby power plants, which generate roughly thirty percent of particulate matter pollution in the U.S. (in 2019) (McDuffie et al., 2021).

[1] We use fine particulate matter (PM2.5) as our measure of air quality because the scientific literature highlights that these particles are particularly detrimental to health because their small size allows them to penetrate lung tissue and get into the bloodstream (CDC, 2019). Numerous studies have documented that elevated PM2.5 levels can irritate the throat and lungs exacerbating cardiovascular and respiratory disease, leading to increased hospitalizations and mortality (American Lung Association, 2020; Deryugina et al., 2021).

First, we develop an instrument that exploits variation in the proximity of districts to coal power plants and year-to-year fluctuations in power production at these plants, as in Duque and Gilraine (2020). Specifically, we instrument for air quality using the yearly coal-based energy production occurring within 60km of a school district. The choice of 60km is informed by prior research which finds a strong relationship between coal power production and PM2.5 concentrations only for locations within roughly 40-60km of the plant (Levy et al., 2002; Clay et al., 2016). We confirm this relationship, showing that increased power production within 60km of a school district significantly raises PM2.5 concentrations and correspondingly lowers student test scores; increased power production located further than 60km away has limited impacts on air quality and no effect on test scores. Our IV estimates indicate that each microgram per cubic meter ($\mu$g/m$^3$) increase in PM2.5 concentrations causes a 0.02 standard deviation decline in test scores. These estimates are larger in magnitude compared to our OLS estimates, as expected given we suspect OLS regressions will feature severe attenuation bias in this context.

Second, we implement a shift-share design that uses national changes in fuels used for power production by interacting pre-existing shares for four fuel sources (coal, oil, gas, and renewables) nearby a district with annual aggregate growth in each source to instrument for the pollution faced by a school district. We motivate our shift-share instrument by highlighting the dramatic shift from coal to natural gas as the primary fuel used for power production during our time period; a motivation supported by the Rotemberg weights in our shift-share design (Goldsmith-Pinkham et al., 2020). Estimates from our shift-share instrument are near-identical to those from the first instrumental variable strategy and indicate that each microgram per cubic meter ($\mu$g/m$^3$) increase in PM2.5 concentrations causes a 0.02 standard deviation decline in test scores.

Our shift-share instrument mirrors the few industries exposed to aggregate shocks setting considered in Goldsmith-Pinkham et al. (2020). Following Goldsmith-Pinkham et al. (2020), we therefore validate our shift-share design by recasting it as a difference-in-differences design. To do so, we focus on coal – the fuel that contributes the most to particulate pollution – and divide school districts into those with high and low nearby coal production in the pre-period. We then use the sharp 16% drop in aggregate coal production that occurred in 2011-12 as our event,[2] which should affect high-coal districts more than low-coal districts. We show that high- and low-coal districts (as defined using the pre-period) have similar trends in terms of both air quality and test scores leading up to 2011-12, providing support for the parallel trends assumption underlying our difference-in-differences design – and our shift-share instrument more generally – that test scores in high- and low-coal districts would trend similarly after 2011-12 in the absence of the national decline in coal use. After 2011-12, however, these trends diverge with high-coal school districts seeing sharp improvements in air quality and test scores.

While both of our empirical strategies leverage variation in air quality coming from energy production, the identifying source of variation differs. In the first strategy, identification comes from year-to-year changes in coal-based energy production, while in our second strategy

[2] The dramatic shift from coal to natural gas was driven by the shale revolution which caused a pronounced fall in natural gas prices, leading to a switch away from coal to natural gas (Federal Energy Regulatory Commission, 2012).

identification comes from exogenous exposure to national changes in the fuels used for energy production. Irrespective of which empirical strategy we use, our estimates yield a similar result in that each $\mu$g/m$^3$ increase in PM2.5 decreases test scores by 0.02 standard deviations. These estimates are robust to including additional controls to account for student sorting, local economic conditions, and weather.

We next tie our findings to how the dramatic improvements in U.S. air quality have impacted the country's educational performance. We calculate that the average student saw their PM2.5 exposure drop a full 3$\mu$g/m$^3$ over the last two decades. Our estimates therefore indicate that this drop in pollution exposure raised test scores nationwide by a full 0.06 standard deviations. This is a economically meaningful effect; in comparison, an oft-cited nationwide policy that releases the bottom 5% of teachers according to value-added would only raise test scores by 0.02-0.025 standard deviations (Gilraine et al., 2020).

We also investigate how the drop in pollution exposure impacted equity in the education system. We find that the black-white PM2.5 exposure gap declined by 0.50$\mu$g/m$^3$ over the last two decades, lowering the black-white test score gap by 0.01 standard deviations. Overall our findings indicate that improvements in air quality over the past two decades have raised nationwide test scores and improved equity, although further gains are possible. For instance, the elimination of black-white differences in particulate exposure would further close the black-white test score gap by 0.024 standard deviations.

The rest of the paper is organized as follows: The next section describes the two instrumental variable strategies that we use. Section 3 then introduces the data and Section 4 presents our results. Section 5 discusses the implication of these results and concludes.

## 2 Empirical Framework

Our goal is to relate particulate matter measured by PM2.5 concentrations to test scores. To do so, we start with the following OLS model:

$$y_{s,d,c,t} = \alpha + \beta \mathit{PM2.5}_{d,t} + \gamma X_{s,d,c,t} + \eta W_{d,t} + \omega_s + \theta_d + \phi_c + \nu_t + \epsilon_{s,d,c,t}\,, \tag{1}$$

where $y_{s,d,c,t}$ is the test score for subject $s$ in district $d$ for cohort $c$ during year $t$. $\mathit{PM2.5}_{d,t}$ is the nine-month average of PM2.5 leading up to the school testing month for district $d$ in year $t$, which we refer to as the school-year average PM2.5. We also incorporate various district-cohort-year demographics, $X_{s,d,c,t}$, that could be related to testing performance, including: cohort lagged test scores, percent of students with special needs, percent English Language Learners, racial composition, enrollment, percent tested, percent economically disadvantaged, and the gender composition of tested students.[3] As shown by Currie et al. (2020), changes in pollution levels may induce changes in sorting patterns. To control for this, $X_{s,d,c,t}$ also includes the percentage of individuals aged 18 to 50 moving in and out of district $d$ during year $t$. In addition, weather or local economic activity can affect both test scores (Park et al., 2020) and pollution. We

[3] Unfortunately, percent of students with special needs and percent English Language Learners are only available at the district-year level.

therefore also include controls for local economic conditions in $X_{s,d,c,t}$ and include the control vector $W_{d,t}$ which contains controls for temperature and precipitation. We also include subject ($\omega_s$), district ($\theta_d$), cohort ($\phi_c$), and year ($\nu_t$) fixed effects. Standard errors are clustered at the district level.

**Econometric Concerns:** There are two major econometric concerns with the estimation of equation (1). First, PM2.5 is measured with error (see Diao et al. (2019); Richmond-Bryant and Long (2020)). The key reason for this measurement error is that PM2.5 monitors are sparse, with only around 1,100 air monitors covering the entire United States. To estimate PM2.5 for areas without monitors, researchers use state-of-the-art models that combine satellite retrievals of aerosol optical depth, chemical transport modeling, and ground-based measurements (see Section 3 for more details). These models, however, only explain about 70 percent of variation in PM2.5 concentration (Van Donkelaar et al., 2019), indicating that there is likely substantial measurement error, which will lead to OLS estimates being attenuated.

Second, there may be unobservable time-varying local characteristics that influence both pollution and student performance. For example, increased local economic activity may draw in new high-performing students or raise current students' performance through an income effect (Dahl and Lochner, 2012) and simultaneously increase pollution. While the inclusion of local economic and sorting controls should help, concerns may remain. To deal with these concerns we use an instrumental variables approach that leverages differential exposure to fuel types used in nearby power production, detailed below.

## 2.1 Instrumental Variable Strategy 1: Year-to-Year Coal Power Production

Our first strategy mirrors Duque and Gilraine (2020) and exploits variation in the proximity of districts to coal power plants and year-to-year fluctuations in power production at these plants. We implement this strategy by instrumenting for a district's PM2.5 exposure with the yearly power production in nearby coal plants. Specifically, for each district we calculate the total coal power production occurring within 20 km bins of a district's centroid, up to a maximum of 60 km. Our distance choice is informed by prior research which finds a strong relationship between coal power production and PM2.5 concentrations only for locations within roughly 40-60 km of the plant, but not beyond (Levy et al., 2002; Clay et al., 2016). We then instrument for PM2.5 using the total coal power production occurring within these three separate distance bins.

We therefore estimate the following two-stage least squares system:

$$PM2.5_{d,t} = \alpha + \sum_i \xi_i \text{coal}_{i,d,t} + \gamma X_{s,d,c,t} + \eta W_{d,t} + \omega_s + \theta_d + \phi_c + \nu_t + \epsilon_{s,d,c,t} \tag{2}$$

$$y_{s,d,c,t} = \alpha + \upsilon \widehat{PM2.5}_{d,t} + \gamma X_{s,d,c,t} + \eta W_{d,t} + \omega_s + \theta_d + \phi_c + \nu_t + \epsilon_{s,d,c,t}\,, \tag{3}$$

where $\text{coal}_{i,d,t}$ is the amount of coal production (in one million MwH units) within $i \in \{$0-20, 20-40, 40-60$\}$ km of district $d$ during school year $t$, and all other variable are defined in Equation (1). The coefficients of interest in the first-stage, $\xi_i$, represent how much PM2.5 increases when coal power production increases by one million MwH. Intuitively, we expect that $\xi_i$ should be strictly positive and decrease as $i$ gets larger: coal production increases district pollution levels but its effect falls as that production takes place further from the district. In the second-stage, the coefficient of interest is $\upsilon$ which indicates the decline in test scores when $PM_{2.5}$ exposure increases by one microgram per cubic meter ($\mu$g/m$^3$).

## 2.2 Instrumental Variable Strategy 2: Shift-Share Instrument

The composition of energy production in the United States has changed substantially over the past decade with the dominant fuel shifting from coal to natural gas – see Figure A.1 for a visualization of these national trends. The academic year 2011-12 in particular marked an exceptional shift with coal use dropping a dramatic sixteen percent that year alone to a thirty-year low (US Energy Information Administration, 2021a; Federal Energy Regulatory Commission, 2012). Given that natural gas emits substantially less air particulates (US Energy Information Administration, 2021b), the rapid shift from coal to natural gas should cause pronounced improvement in air quality for districts in areas that used to rely on coal for energy production. This motivates our shift-share instrument that leverages this change in fuels used for power production by interacting pre-existing shares for four fuel sources (coal, oil, gas, and renewables)[4] nearby a district with annual aggregate growth in each source.

Formally, we construct our shift-share instrument by interacting 2004-05 fuel shares across the four fuel sources (coal, oil, gas, and renewables) with annual aggregate growth in each source. The two-stage estimation is:

$$PM2.5_{d,t} = \alpha + \sum_f \delta_{2005,f,d}\Gamma_{t,f} + \gamma X_{s,d,c,2005} \cdot \nu_t + \eta W_{d,t} + \omega_s + \theta_d + \phi_c + \nu_t + \epsilon_{s,d,c,t} \tag{4}$$

$$y_{s,d,c,t} = \alpha + \rho \widehat{PM2.5}_{d,t} + \gamma X_{s,d,c,2005} \cdot \nu_t + \eta W_{d,t} + \omega_s + \theta_d + \phi_c + \nu_t + \epsilon_{s,d,c,t} \tag{5}$$

where $\delta_{2005,f,d}$ is the share of district $d$'s 2004-05 fuel production within 40km of its centroid from source $f \in \{\textit{coal, gas, oil, renewables}\}$, $\Gamma_{t,f}$ is the growth rate of fuel $f$ in year $t$, and all other variable are defined in equation (1). Following Goldsmith-Pinkham et al. (2020), we construct $X_{s,d,c,2005} \cdot \nu_t$ by interacting our district-level controls in 2004-05 with year fixed effects since controls measured after shifts in fuel sources can cause bias as these shifts may also affect the district covariates themselves.[5] The regressions are weighted by the total aggregate power production in 2004-05 within 40km. The shift-share instrument can be thought of as using the fuel shares to measure the extent to which a district is exposed to aggregate changes in fuel production.

[4] These four fuel sources covered more than 98 percent of energy production in the United States in 2009 (US Energy Information Administration, 2021a). Note that we classify nuclear energy as renewable since nuclear reactors do not emit any air pollution.

[5] There are two exceptions to this. First, we control for lagged cohort test scores so that our outcome can be interpreted as test score growth. Second, we use time-varying weather controls as we would not expect shifts in fuel sources to impact weather (at least locally and in the relatively short-term) and so the weather controls are unaffected by the policy change negating the reason controls are fixed in the shift-share design.

# 3 Data

This section provides a brief overview of the datasets that we use. Appendix B provides more detail.

## 3.1 School District Performance and Demographics

School district performance data is from the Stanford Education Data Archive (SEDA) for the school years 2008-09 to 2017-18 (Reardon et al., 2021). This dataset constructs a national scale of district performance using proficiency rates from reading and mathematics state assessment exams. Since state assessments may differ in terms of their proficiency standards, Reardon et al. (2021) bring in test score results from the nationwide National Assessment of Educational Progress to place district performance across different states on the same scale. The test scores are reported in standardized deviations relative to the average performance of a national reference cohort for the same subject and grade.[6] In our analysis we stack the math and reading test score data and so our findings indicate the average impact on math and reading scores.

SEDA also contains information on district-grade-year demographics such as total enrollment, number of students tested, gender and racial shares, and the percentage of special needs students and English learner students. In the shift-share design, we use data from the National Center for Education Statistics to gather these same demographics.

We add to the student testing dataset information on whether each state tested students in the fall, spring, or year-round. While testing is usually done in the spring, in the early years of the sample, a few states tested in the fall or year-round (see Table A.1). However, these states switched to spring testing eventually when they adopted Common Core standards. For our main analyses we restrict our sample to states who test in the spring so that we can focus on exposure to PM2.5 during the academic year. We have confirmed that the vast majority of testing in these states occurs during May.[7] We show later in robustness exercises that the inclusion of non-spring testing states does not affect our results.

## 3.2 Pollution

Our pollution data come from Van Donkelaar et al. (2019) who produce monthly PM2.5 concentrations at a 0.01 degree by 0.01 degree resolution (roughly 1.1km by 1.1km at the equator) for the United States (excluding Hawaii) from 2008-2018. We assign each cell in these gridded raster data to a school district using its centroid.[8] The average PM2.5 experienced by a school district during a given month is then the average PM2.5 across all cells located in the school district. In our main results, we take the average of these monthly PM2.5 levels during

[6] See Fahle et al. (2021) for the technical documentation.

[7] While states often have wide testing windows, student testing almost always occurs in May. For example, during our sample period California's testing window ran from the start of March to the start of June (e.g., 2014-15 testing window was March 4 - June 4). The overwhelming majority of schools, however, conducted their tests in May.

[8] District locations are determined using the shapefile provided by Reardon et al. (2021).

the academic year (September to May) as the air pollution that students have experienced.

These monthly PM2.5 concentration estimates come from state-of-the-art environmental science research that combines spatially-continuous measurements of pollution from satellites (i.e., aerosol optical depth) with other observable pollution correlates such as emissions inventories, chemical transport models, land use characteristics, and weather patterns (Di et al., 2016; Hammer et al., 2020). The basic idea in these papers is to use model selection techniques to build a predictive model of PM2.5 concentrations by correlating EPA monitor data with the pollution correlates listed above. The chosen model then predicts pollution in regions without air monitor data.

These data have very good in-sample fit as they match the "true" PM2.5 concentrations (as measured by the EPA air monitors) very well. That said, the measures are not perfect: Van Donkelaar et al. (2019) conduct a cross-validation exercise which withholds a random 10 percent of the EPA monitors from their sample and find that only 70 percent of the observed variation in PM2.5 is explained by their model. The significant measurement error is likely to materially attenuate OLS estimates of pollution on test scores, motivating the instrumental variable approach we take.[9]

## 3.3 Energy Production

Data on plant-level energy production by fuel source are from the Energy Information Administration (EIA) form 923. The EIA form is a mandatory report for all electric power plants connected to the electrical grid that have a total generating capacity of one megawatt or more. The survey contains information on monthly power production at the plant level, which we aggregate up to plant level production for each school year from 2008-2009 to 2017-2018. (School years are defined as running from September to May). We then merge each plant's latitude and longitude from EIA form 860 onto these data.

## 3.4 Controls

**Weather:** One salient concern is that weather can directly influence pollution and student performance (Park et al., 2020). We therefore construct weather controls using data from the National Oceanic and Atmospheric Administration's Daily Global Historical Climatology Network, which includes daily station-level data for thousands of weather stations across the United States. We restrict our data to stations reporting valid readings for at least 95 percent of school days[10] and impute the small proportion of missing daily observations using the nearest station. Our weather data include approximately 5,400 weather stations that record daily temperature data and 6,500 stations that register precipitation data for the school years 2008-09 to 2017-18. Each school district is assigned to the weather station nearest to its centroid, resulting in average distances of 11.9 and 10.7 miles between a district's centroid and the nearest

[9] In addition to classical measurement error, Fowlie et al. (2019) find evidence that satellite-derived PM2.5 estimates are biased downward for high PM2.5 concentrations. In our context, this will bias our estimates downward as the shifts in PM2.5 caused by reduced energy production will be more pronounced than we estimate.

[10] We define school days in our data as any weekday from September 1 to May 31.

temperature and precipitation station, respectively.

We use the matched station-district data to construct flexible controls for the weather experienced by students. For each academic year, we construct means of daily temperature highs and lows as well as counts of the number of days falling within six temperature bins to capture extreme temperature events. Similarly, for each academic year we control for mean precipitation and counts of days with precipitation falling into two bins, accounting for heavy levels of rainfall. Similar controls are also constructed for snowfall, given the possibility that snow may lead to school closures that can independently affect student achievement (Goodman, 2014). (More detail on the exact weather controls is provided in Table A.8.)

**Moving Controls:** We use records on residential address history from the data company Infutor, which tracks individual addresses through credit history information. The addresses in Infutor contain street address, city, and zip code, which we geocode to get latitude and longitude. In our sample we focus on individuals aged between 18 - 50, whose address we can match to a school district using district shapefiles from Reardon et al. (2021). We have roughly 84 million unique individuals in the dataset. We then construct the move-in rate by calculating the fraction of residents of district $d$ who moved in during year $t$, and the move-out rate as the fraction of residents of district $d$ who left at year $t$. Note that the move rates are calculated by academic year, so that individuals who moved in September or later of year $t$ are coded as moving in the academic year ending in $t + 1$.

**Local Economic Controls:** Lastly, we bring in local economic controls at the district level from the American Community Survey. Our controls include percentage employed, percentage in the labor force, percentage employed in the utilities and manufacturing sector, percentage with a bachelor's degree or higher, and percentage of single mother households, which may change in response to shifting energy demand.

Our data consists of 11,476 unique districts over the academic years 2008-09 through 2017-18. Table B.1 in the Appendix presents summary statistics for our sample. The average district has a school-year mean PM2.5 of 7.19 (to put this in perspective, the CDC considers 12 $\mu g/m^3$ to be the acceptable annual level). The average school district consists of 73% white students, with Hispanics being the largest minority group at 13%. Roughly 4% of students are learning English as a second language, and 14% are classified as special needs. Close to half of students are eligible for a free or a reduced lunch. The average move-in and move-out rates for a district are both around 4 percent.

## 4 Results

We start by running the OLS regression described in Equation (1), with Panel A of Table A.2 presenting the results across four columns with varying sets of controls. Our preferred specification is in Column (4), which includes all controls. Looking at the first row, we see that air particulate matter decreases test scores: A one microgram per cubic meter ($\mu$g/m$^3$) increase in PM2.5 concentration is associated with a 0.0035 standard deviation lower test score.

As previously discussed, however, we suspect that OLS estimates are materially attenuated due to measurement error in the PM2.5 variable. Given this, we turn to results from our two IV strategies.

### 4.1 Instrumental Variable Strategy I: Year-to-Year Coal Power Production

First, we implement the IV strategy that leverages year-to-year variation in coal production as described in Section 2.1. Before running the IV regression, we provide some visual evidence of the relationships between coal power production, PM2.5, and test scores that underlie the IV regression.

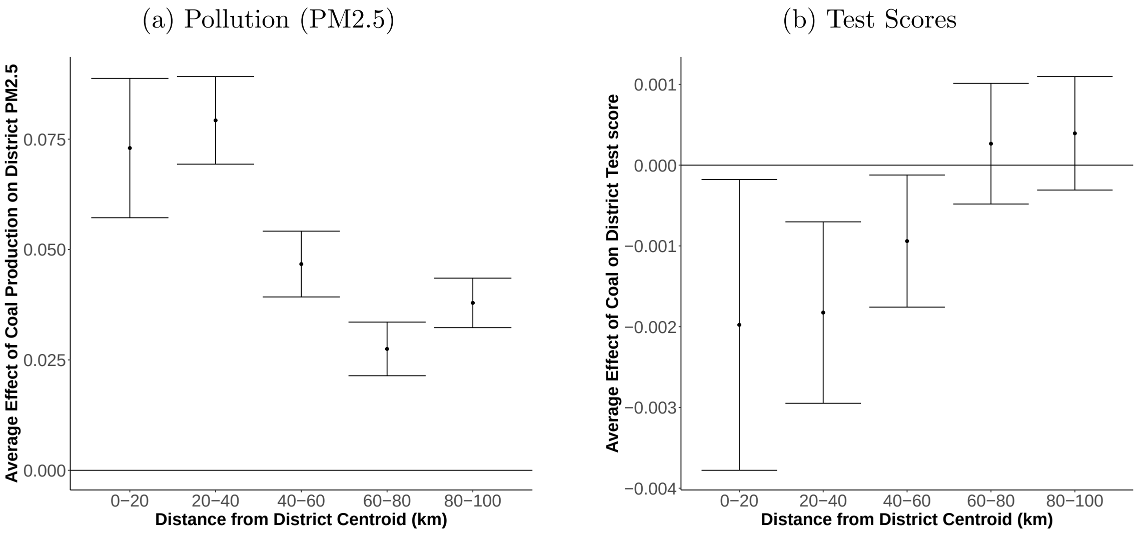
Figure 1(a) plots the coefficients for our first-stage regression of PM2.5 on coal production given by equation (2) for each 20km distance bin. Each coefficient represents the effect of a one million MwH increase in coal production among plants in that distance bin on average school-year PM2.5. As expected, coal-based power production that occurs close to the district centroid has the greatest effect on air pollution. For plants within 40km, a one million Mwh increase in production causes a 0.075 increase in the district's PM2.5 concentration. The relationship between power production and PM2.5 then declines by roughly half for plants 40-60km away and by two-thirds for plants beyond 60km.

Next, Figure 1(b) presents results of a reduced form regression of test scores on coal production. These regressions are identical to those in Figure 1(a), but replace PM2.5 levels with test scores. The point estimates in 1(b) indicate that increased coal production nearby a district decreases that district's test scores. Indeed, a similar relationship between distance and the impact of coal production that we observed in Figure 1(a) reveals itself: a one million MwH increase in production plants within 40km causes a 0.0019 standard deviation decline in student test scores. The relationship between power production and test scores then declines by roughly half for plants 40-60km away and disappears for plants beyond 60km.

Taken together, the two figures highlight that proximity to coal production is a strong predictor of a district's PM2.5 levels and test scores. Notably, there is a large drop in the effect of coal-based power production on PM2.5 pollution levels (and test scores) at distances beyond 40-60km. These findings corroborate prior work showing that PM2.5 levels drop significantly outside of a 40-60km distance from coal plants (Levy et al., 2002; Clay et al., 2016). Our reduced form findings then show a similar relationship whereby the impact of coal-based power production on test scores dissipates past 40-60km, in line with the point at which the impact of coal use on district PM2.5 levels also fall. The fact that the impact of coal power production on PM2.5 and test scores fall in tandem as we look at districts further away from the coal plant implies that our IV estimates are nearly identical for our three IVs that use coal-based production in 20km bins up to 60km. Table A.3 confirms that this is indeed the case.

We next run the IV regression described by equations (2) and (3), with results reported in Panel A of Table (1). Taking our preferred estimates in Column (4) with all controls, the IV results indicate that a one microgram per cubic meter ($\mu$g/m$^3$) increase in PM2.5 concentration reduces district test scores by 0.021 standard deviations. The estimate is statistically significant and larger in magnitude compared to the OLS estimate, in line with the IV estimate correcting

Figure 1: Distance to Coal Plants


(a) Pollution (PM2.5)

(b) Test Scores

Notes: Figure 1(a) plots the coefficients for our first-stage regression of PM2.5 on coal production for each 20km distance bin up to 100km. Each coefficient represents the effect of a one million MwH increase in coal production on average PM2.5 during the academic year. The figure therefore visualizes the first-stage of the IV regression given by equation (2). Figure 1(b) then plots the coefficients for a reduced form regression of test scores on coal production for each 20km distance bin up to 100km. The horizontal line in each figure represents a point estimate of zero, while the whiskers around each point estimate represent 95% confidence intervals with standard errors clustered at the district level. The specification used matches those of Column (1) in Table 1, with only controls for student covariates included. A full list of controls is available in Table A.8.

for attenuation bias.

**Validity:** Our instrument leverages year-to-year production variation for identification. The year-to-year production variation guards against bias coming from individuals sorting into more/less polluted areas that are not captured by our moving controls. Specifically, since production changes every year, bias coming from sorting must be driven by students moving districts yearly in response to the yearly production changes which seems unlikely. In addition, the fact that our IV estimates are similar when coal production in each distance bin – namely production within 0-20km, 20-40km, and 40-60km of the coal plant – are used separately as instruments alleviates concerns over the economic impacts of plant downsizing or closures driving our result. Specifically, we would expect the local economic impacts of plant closures (e.g., via job losses) to be concentrated in the region close to the plant (i.e., the 0-20km bin). In Table A.3, however, we observe similar estimates regardless of whether the 0-20km or 40-60km production instruments are used.

## 4.2 Instrumental Variable Strategy II: Shift-Share Instrument

Our second empirical strategy is a shift-share instrument that leverages the differential exposure of districts to national changes in fuels used for power production. We start by describing the key identifying variation that underlies the shift-share instrument; namely, the drastic shift from coal to natural gas. Figure A.1 shows aggregate power production from coal and natural gas over our time period and indicates that there is a pronounced decline in coal-based power production with a corresponding rise in natural gas power production over time, with this shift becoming especially pronounced starting in the 2011-12 school year.

Our shift-share instrument takes full advantage of these national changes in the mix of fuels used for power production by comparing school districts with different exposures to various fuels used for nearby power production. To do so, we use the interaction between a district's pre-existing fuel shares (measured in 2004-05) and annual national growth in each fuel source to instrument for the pollution faced by a school district as described by equations (4) and (5). Panel B of Table 1 reports the results.[11] Our preferred coefficient in Column (4) is -0.017, which is similar to our estimate from the first empirical design (-0.021). Taken together, our two IV strategies indicate that each $\mu$g/m$^3$ of PM2.5 concentration reduce test scores by roughly 0.02 of a standard deviation.

**Validity:** Recent work by Borusyak et al. (2021) and Goldsmith-Pinkham et al. (2020) highlight how to validate the shift-share research design. We motivated our shift-share instrument by emphasizing the dramatic shift from coal to natural gas. Following Goldsmith-Pinkham et al. (2020), we confirm that this motivation is borne out empirically by calculating the Rotemberg weights for our four fuels sources. We find that exposure to nearby coal energy production is the most important source of variation in the shift-share instrument, receiving a Rotemberg weight of 0.50. Gas and oil-based energy production are also important sources of variation and have Rotemberg weights of 0.21 and 0.27, respectively. The renewable energy sector provides little identifying variation here – as we would expect as it is emissions-free – receiving a Rotemberg weight of only 0.02. Given that our motivation is confirmed in the data, our shift-share setup resembles the few industries exposed to a common shock setting considered in Goldsmith-Pinkham et al. (2020). The validity of our instrument therefore relies on the assumption that the differential effect of higher exposure of one power producing industry (compared to another) only affects the change in test scores through air pollution, and not through any potential confounding channel.

We follow Goldsmith-Pinkham et al. (2020) and test the validity of our identifying assumption in two ways. First, we investigate the correlation between districts' initial exposure shares (constructed using 2004-05 electricity production) and the characteristics of those districts in that initial period. To control for regional variation in fuel mixes, correlation coefficients are calculated within states and then averaged across all states. These correlations are reported in Table A.4. Importantly, we find limited correlation between 2004-05 fuel shares and the moving controls as well as the local economic controls. This is reassuring as it shows the initial fuel shares that we use to construct our instrument are not correlated with factors that predict changes in test scores, which is encouraging that omitted variables are not biasing estimation. In addition, we note that our point estimates are quite stable across the different sets of controls we use (see columns (1)-(4) in Table 1), which is comforting as movements in point estimates

[11] Compared to our first IV design displayed in Panel A, our shift-share design has roughly 90,000 fewer observations. The cause of the drop in observations is twofold: (i) district characteristics are missing in 2004-05 (needed to construct controls), and (ii) there is no power production occurring within 40km of the district's centroid in 2004-05. Roughly three-quarters of the observation loss comes from the latter cause.

Table 1: Main Results

| **Outcome:** Standardized Test Scores | | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| **Panel A:** Empirical Strategy I: Distance and Production Variation | | | | |
| IV Estimate | $-0.0198^{***}$ | $-0.0198^{***}$ | $-0.0204^{***}$ | $-0.0206^{***}$ |
| ($\mu$g/m$^3$) | (0.0042) | (0.0042) | (0.0045) | (0.0045) |
| First-Stage F-stat | 203.92 | 202.15 | 191.48 | 191.39 |
| Observations | 701,199 | 694,257 | 694,257 | 694,257 |
| **Panel B:** Empirical Strategy II: Shift-Share Instrument | | | | |
| IV PM2.5 Estimate | $-0.0176^{***}$ | $-0.0206^{***}$ | $-0.0181^{**}$ | $-0.0165^{**}$ |
| ($\mu$g/m$^3$) | (0.0050) | (0.0064) | (0.0071) | (0.0069) |
| First-Stage F-stat | 662.08 | 408.31 | 377.40 | 390.53 |
| Observations | 607,482 | 604,232 | 604,232 | 604,193 |
| **Controls Used:** | | | | |
| Student Covariates | Yes | Yes | Yes | Yes |
| Local Economic Controls | No | Yes | Yes | Yes |
| Weather Controls | No | No | Yes | Yes |
| Sorting Controls | No | No | No | Yes |

Notes: This table reports results from our two empirical strategies. Test scores are measured in standard deviations, while PM2.5 is measured in micrograms per cubic meter ($\mu$g/m$^3$). Panel A displays the point estimates from our first empirical methodology leveraging year-to-year production variation in nearby coal plants. In particular, we use yearly coal-based power production within different distance bins as instruments for PM2.5, as described by equations (2) and (3). Panel B reports results from our second empirical methodology which uses a shift-share instrument. Specifically, we use the interaction between pre-existing exposure to different power production interacted with national growth rates as our instrument, which is described in equations (4) and (5). The 'First-Stage F-stat' in both panels displays the Kleibergen-Paap F statistic to assess the statistical significance of the instruments' ability to predict PM2.5. All regressions include subject, district, cohort, and year fixed effects. The control variables used for each set of controls is given in Table A.8. Standard errors are clustered at the district level. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

when conditioning on observable confounders suggest the potential importance of unobserved confounders (Altonji et al., 2005).

Second, we use a difference-in-differences design to explore whether there were parallel trends before the shock to aggregate fuel composition. If parallel trends hold then it is likely that the common shock to national fuel mixes caused the change in test scores, rather than pre-existing differences between districts with different fuel shares. To conduct the difference-in-differences, we compare districts that are exposed to high relative to low coal production before and after the large decline in aggregate coal production in 2011-12. Specifically, we classify districts into two groups – low and high – based on 2004-05 coal production that took place within 40km of a district's centroid. We focus only on districts that had some positive coal production within that distance range[12] and define "low" and "high" coal districts based on whether their exposure to coal production in 2004-05 was above or below the mean.

Figure 2(a) displays the PM2.5 concentrations among low- (dashed line) and high-coal (solid line) districts over time. As expected, there is a large drop in PM2.5 concentration in 2011-12 for high- relative to low-coal districts, which coincides with the significant decrease in U.S. coal production that year. We therefore conduct the following regression using data from 2008-09 to 2015-16 to compare test scores in high- relative to low-coal districts relative to a reference year (which we set as 2010-11):

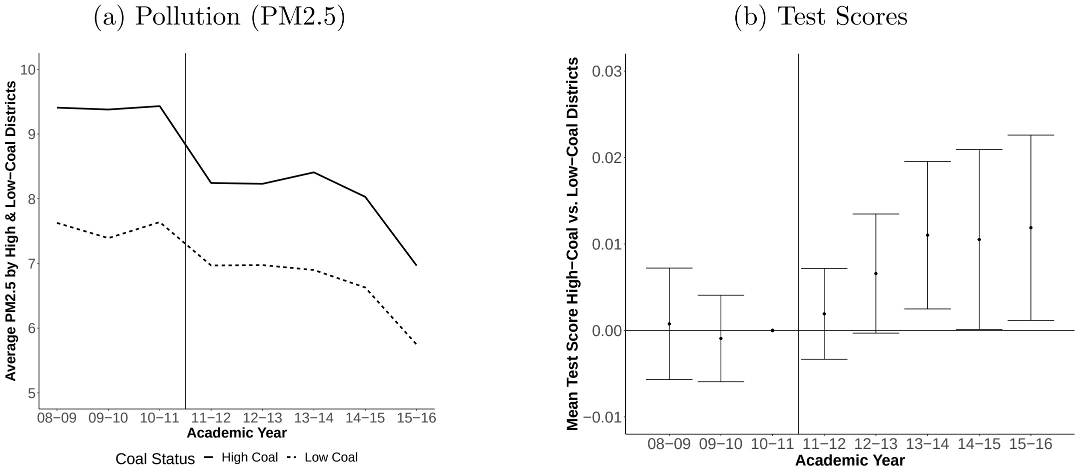
$$y_{s,d,c,t} = \alpha + \sum_{j \neq 2010-11} \beta_j High_d * \mathbb{1}\{year{=}j\}_t + \gamma X_{s,d,c,t} + \eta W_{d,t} + \omega_s + \theta_d + \phi_c + \nu_t + \epsilon_{s,d,c,t}\,, \quad (6)$$

where $High_d$ is an indicator for being a high-coal district, $\mathbb{1}\{year{=}j\}$ is a year indicator, and all other variables are defined in equation (1).

We plot the $\beta_j$ coefficients for each year in Figure 2(b). In the years leading up to 2011-12, we see that test score differences between high- and low-coal districts are stable. This provides support for the parallel trends assumption underlying our difference-in-differences design – and our shift-share instrument more generally – that test scores in high- and low-coal districts would trend similarly after 2011-12 in the absence of the national decline in coal use. After 2011-12, we see large increases in test scores among high- relative to low-coal districts, mirroring the relative decline in PM2.5 concentrations in the high-coal districts.

[12] Effectively this restriction eliminates Western states that have never relied on coal for electricity and leaves us with roughly 5,100 districts.

Figure 2: Difference-in-Differences


(a) Pollution (PM2.5)

(b) Test Scores

Notes: The left-hand figure shows the average PM2.5 levels for high-coal districts (solid line) and low-coal districts (dashed line) from 2008-09 to 2015-16. The right-hand figure then shows the average test performance of high-coal districts relative to low-coal districts as described in equation (6). The coefficient for the academic year 2010-11 is normalized to zero and whiskers represent 95% confidence intervals with standard errors clustered at the district level. Controls for student covariates along with subject, district, cohort, and year fixed effects are included.

### 4.3 Robustness

We perform three additional robustness checks to ensure that our results are invariant to several choices we made. First, in Panel A of Table A.5 we re-run our shift-share specification using fuel shares from energy production in 2000-01. This check alleviates concerns about whether choosing 2004-05 to construct fuel shares is driving our results. Table A.5 shows that we find similar results, with our preferred specification yielding a point estimate of -0.021 (compared to -0.017 when 2004-05 shares were used).

Second, our shift-share instrument was constructed using power production occurring within 40km of the school district. Panel B of Table A.5 shows the results if we change this distance to within 60km of the district centroid. The results are quite similar: using all controls the estimate is -0.0168 as compared to -0.0165 in the baseline shift-share instrument.

Third, in Table A.6 we use the entire sample of districts including those who did not test during the spring. Column (1) runs our year-to-year production variation instrument while Column (2) and (3) use the shift-share instrument with 2004-05 and 2000-01 shares, respectively. Our estimates are similar to our main results, indicating that the exclusion of non-spring testing states are not causing selection issues.

## 5 Discussion and Conclusion

Our two empirical strategies indicate that each $\mu$g/m$^3$ reduction in PM2.5 increases test scores by 0.02 standard deviations. Panel A of Table 2 places these estimates into perspective by reporting the PM2.5 exposure an average student in the United States faced in 2002-03, 2010-11, and 2018-19. We see a large decline in pollution exposure over the last two decades:

Table 2: Pollution Exposure and Test Score Gaps Over Time

| Academic Year | 2002-03 | 2010-11 | 2018-19 | Change from 2002-03 to 2018-19 |
|---|---|---|---|---|
| Panel A. Mean PM2.5 Exposure | | | | |
| Average PM2.5 Exposure | 10.47 | 8.72 | 7.47 | -3.00 |
| Panel B. Black-White Exposure Gap | | | | |
| Black PM2.5 Exposure | 11.51 | 9.60 | 8.10 | -3.41 |
| White PM2.5 Exposure | 9.83 | 8.32 | 6.92 | -2.91 |
| **Black-White PM2.5 Gap** | **1.68** | **1.28** | **1.18** | **-0.50** |
| Panel C. Black-White Test Score Gap | | | | |
| Black Mean Test Score | -0.61 | -0.51 | -0.47 | 0.14 |
| White Mean Test Score | 0.29 | 0.28 | 0.25 | -0.04 |
| **Black-White Test Score Gap** | **0.90** | **0.79** | **0.72** | **-0.18** |

Notes: This table reports average PM2.5 exposure nationwide over time, with Panel B splitting these by race. We construct PM2.5 exposures by taking the grade 3-8 enrollment-weighted district PM2.5 measures for a given year, with enrollment data coming from the National Center for Education Statistics (2022). Panel C then displays test scores by race, with test scores data coming from National Assessment of Educational Progress (2022). To construct mean test scores by race, we use fourth grade test scores from both mathematics and reading and standardize these test scores using the standard deviation across all students for a given test-year combination. These standardized subject-specific scores are then averaged.

the average student saw their PM2.5 exposure drop a full $3\mu\text{g/m}^3$ from 2002-03 to 2018-19. Our estimates therefore indicate that air quality improvements raised test scores nationwide by a full 0.06 standard deviations. This nationwide test score impact is large; in comparison a nationwide policy that releases the bottom 5% of teachers according to value-added would only raise test scores by 0.02-0.025 standard deviations (Gilraine et al., 2020).

Given that the decline in pollution exposure over the past two decades has disproportionately benefited areas where African Americans reside (Currie et al., 2020), we also expect impacts on test score equity. Panel B reports PM2.5 exposure for the average white and black student in the United States. From 2002-03 to 2018-19 the black-white PM2.5 exposure gap declined by $0.50\mu\text{g/m}^3$. This suggests that changes in pollution exposure decreased the black-white test score gap by 0.01 standard deviations, or 6 percent of the 0.18 standard deviations decline in the black-white test score gap over this period. These estimates assume that the impact of pollution on test scores are homogeneous across racial groups. We investigate this assumption in Table A.7, which presents our IV estimates for districts in the top tercile of percentage of Black students (Column 1) and the lowest tercile (Column 2). The estimates from both sets of districts are similar, suggesting that there is no significant difference in how pollution affects test scores by the proportion of Black students. Thereofre, our estimates of the decline in air pollution on the Black-white test score gap are unlikely to be confounded by heterogeneous effects.

This paper analyzed the impact of recent air quality improvement on American education. Using instrumental variables that leverage exposure to nearby power production, we find that air pollution significantly lowers student test scores. Given that, the large $3\mu\text{g/m}^3$ drop in PM2.5 concentrations experienced by the average student materially raised test scores nationwide. Substantial improvements to student performance and equity through cleaner air are still possible, however. For example, decreasing average PM2.5 to that of the first quartile district would raise nationwide test scores by 0.036 and completely eliminating black-white differences in particulate exposure would further decrease the black-white test score gap by 0.024 standard deviations.

## References

Aguilar-Gomez, Sandra, Holt Dwyer, Joshua S. Graff Zivin, and Matthew J. Neidell (2022), "This is air: The "non-health" effects of air pollution." Working Paper 29848, National Bureau of Economic Research, URL https://www.nber.org/papers/w29848.

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber (2005), "Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools." *Journal of Political Economy*, 113, 151–184.

American Lung Association (2020), "Particle pollution." https://www.lung.org/clean-air/outdoors/what-makes-air-unhealthy/particle-pollution.

Borusyak, Kirill, Peter Hull, and Xavier Jaravel (2021), "Quasi-Experimental Shift-Share Research Designs." *Review of Economic Studies*, 89, 181–213.

CDC (2019), "Particle pollution." https://www.cdc.gov/air/particulatematter.html.

Chay, Kenneth Y. and Michael Greenstone (2003), "The impact of air pollution on infant mortality: Evidence from geographic variation in pollution shocks induced by a recession." *Quarterly Journal of Economics*, 118, 1121–1167.

Clay, Karen, Joshua Lewis, and Edson Severnini (2016), "Canary in a coal mine: Infant mortality, property values, and tradeoffs associated with mid-20th century air pollution." Working Paper 22155, National Bureau of Economic Research, URL http://www.nber.org/papers/w22155.

Currie, Janet, John Voorheis, and Reed Walker (2020), "What caused racial disparities in particulate exposure to fall? New evidence from the Clean Air Act and satellite-based measures of air quality." Working Paper 26659, National Bureau of Economic Research, URL http://www.nber.org/papers/w26659.

Currie, Janet and Reed Walker (2019), "What do economists have to say about the Clean Air Act 50 years after the establishment of the Environmental Protection Agency?" *Journal of Economic Perspectives*, 33, 3–26.

Dahl, Gordon B. and Lance Lochner (2012), "The impact of family income on child achievement: Evidence from the earned income tax credit." *American Economic Review*, 102, 1927–56.

Deryugina, Tatyana, Nolan Miller, David Molitor, and Julian Reif (2021), "Geographic and socioeconomic heterogeneity in the benefits of reducing air pollution in the United States." *Environmental and Energy Policy and the Economy*, 2, 157–189.

Di, Qian, Itai Kloog, Petros Koutrakis, Alexei Lyapustin, Yujie Wang, and Joel Schwartz (2016), "Assessing $\mathrm{PM}_{2.5}$ exposures with high spatiotemporal resolution across the continental United States." *Environmental Science & Technology*, 50, 4712–4721.

Diao, Minghui, Tracey Holloway, Seohyun Choi, Susan M. O'Neill, Mohammad Z. Al-Hamdan, Aaron Van Donkelaar, Randall V. Martin, Xiaomeng Jin, Arlene M. Fiore, Daven K. Henze, Forrest Lacey, Patrick L. Kinney, Frank Freedman, Narasimhan K. Larkin, Yufei Zou, James T. Kelly, and Ambarish Vaidyanathan (2019), "Methods, availability, and applications of PM2.5 exposure estimates derived from ground measurements, satellite, and atmospheric models." *Journal of the Air & Waste Management Association*, 69, 1391–1414.

Duque, Valentina and Michael Gilraine (2020), "Coal use and student performance." Technical Report 251, Annenberg Institute at Brown University, URL http://www.edworkingpapers.com/ai20-251.

Ebenstein, Avraham, Victor Lavy, and Sefi Roth (2016), "The long-run economic consequences of high-stakes examinations: Evidence from transitory variation in pollution." *American Economic Journal: Applied Economics*, 8, 36–65.

Fahle, Erin M., Belen Chavez, Demetra Kalogrides, Benjamin R. Shear, Sean F. Reardon, and Andrew D. Ho (2021), "Stanford education data archive technical documentation version 4.1 June 2021." URL https://stacks.stanford.edu/file/druid:db586ns4974/seda_documentation_4.1.pdf.

Federal Energy Regulatory Commission (2012), "2012 state of the markets report." Retrieved from https://www.ferc.gov/sites/default/files/2020-05/2012-som-final.pdf.

Fowlie, Meredith, Edward Rubin, and Reed Walker (2019), "Bringing satellite-based air quality estimates down to earth." In *AEA Papers and Proceedings*, volume 109, 283–88.

Gilraine, Michael (2020), "Air filters, pollution and student achievement." Technical Report 188, Annenberg Institute at Brown University, URL http://www.edworkingpapers.com/ai20-188.

Gilraine, Michael, Jiaying Gu, and Robert McMillan (2020), "A new method for estimating teacher value-added." Working Paper 27094, National Bureau of Economic Research, URL http://www.nber.org/papers/w27094.

Goldsmith-Pinkham, Paul, Isaac Sorkin, and Henry Swift (2020), "Bartik instruments: What, when, why, and how." *American Economic Review*, 110, 2586–2624.

Goodman, Joshua (2014), "Flaking out: Student absences and snow days as disruptions of instructional time." Working Paper 20221, National Bureau of Economic Research.

Hammer, Melanie S., Aaron van Donkelaar, Chi Li, Alexei Lyapustin, Andrew M. Sayer, N. Christina Hsu, Robert C. Levy, Michael J. Garay, Olga V. Kalashnikova, Ralph A. Kahn, Michael Brauer, Joshua S. Apte, Daven K. Henze, Li Zhang, Qiang Zhang, Bonne Ford, Jeffrey R. Pierce, and Randall V. Martin (2020), "Global estimates and long-term trends of fine particulate matter concentrations (1998–2018)." *Environmental Science & Technology*, 54, 7879–7890.

Heissel, Jennifer A., Claudia Persico, and David Simon (2020), "Does pollution drive achievement? The effect of traffic pollution on academic performance." *Journal of Human Resources*, 1218–9903R2.

Levy, Jonathan I., John D. Spengler, Dennis Hlinka, David Sullivan, and Dennis Moon (2002), "Using CALPUFF to evaluate the impacts of power plant emissions in Illinois: Model sensitivity and implications." *Atmospheric Environment*, 36, 1063–1075.

Marcotte, Dave E. (2017), "Something in the air? Air quality and children's educational outcomes." *Economics of Education Review*, 56, 141–151.

McDuffie, Erin E., Randall V. Martin, Joseph V. Spadaro, Richard Burnett, Steven J. Smith, Patrick O'Rourke, Melanie S. Hammer, Aaron van Donkelaar, Liam Bindle, Viral Shah, Lyatt Jaeglé, Gan Luo, Fangqun Yu, Jamiu A. Adeniran, Jintai Lin, and Michael Brauer (2021), "Source sector and fuel contributions to ambient PM2.5 and attributable mortality across multiple spatial scales." *Nature communications*, 12, 3594.

Mullen, Casey, Sara E Grineski, Timothy W Collins, and Daniel L. Mendoza (2020), "Effects of PM2. 5 on third grade students' proficiency in math and English language arts." *International Journal of Environmental Research and Public Health*, 17, 6931.

National Assessment of Educational Progress (2022), "NAEP data explorer." Retrieved from https://nces.ed.gov/nationsreportcard/data/.

National Center for Education Statistics (2022), "Common core of data." Retrieved from https://nces.ed.gov/ccd/.

Park, R. Jisung, Joshua Goodman, Michael Hurwitz, and Jonathan Smith (2020), "Heat and learning." *American Economic Journal: Economic Policy*, 12, 306–39.

Persico, Claudia L. and Joanna Venator (2021), "The effects of local industrial pollution on students and schools." *Journal of Human Resources*, 56, 406–445.

Reardon, S. F., A. D. Ho, B. R. Shear, E. M. Fahle, D. Kalogrides, H. Jang, and B. Chavez (2021), "Stanford education data archive (version 4.1)." Retrieved from http://purl.stanford.edu/db586ns4974.

Richmond-Bryant, Jennifer and Thomas C. Long (2020), "Influence of exposure measurement errors on results from epidemiologic studies of different designs." *Journal of Exposure Science & Environmental Epidemiology*, 30, 420–429.

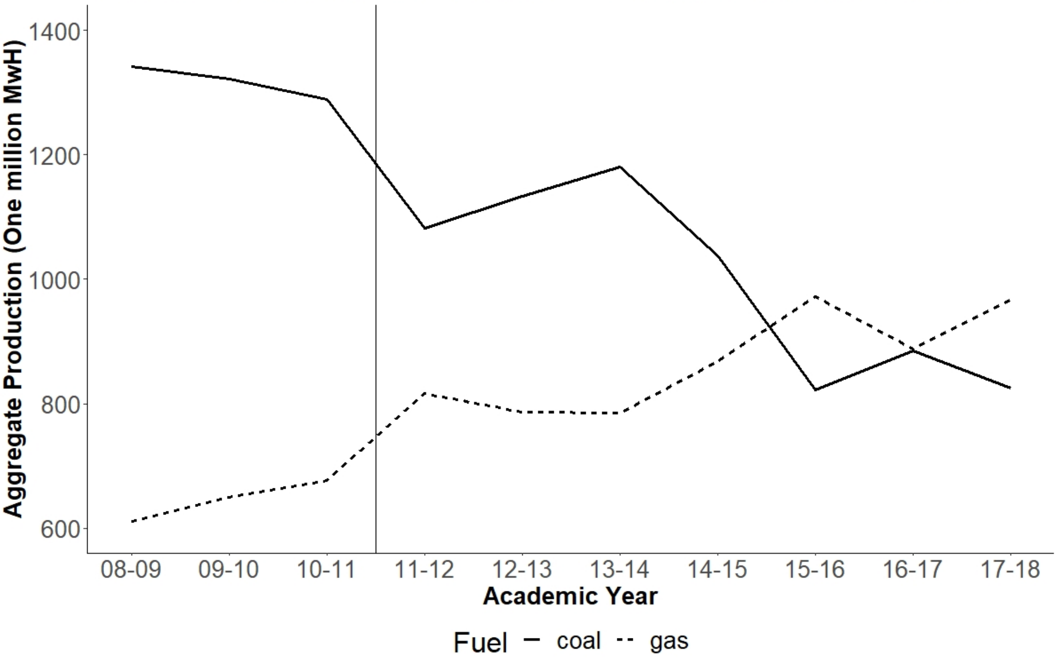
US Energy Information Administration (2021a), "Form eia-923 detailed data with previous form data (eia-906/920)." Retrieved from https://www.eia.gov/energyexplained/natural-gas/natural-gas-and-the-environment.php.

US Energy Information Administration (2021b), "Natural gas explained." Retrieved from https://www.eia.gov/energyexplained/natural-gas/natural-gas-and-the-environment.php.

Van Donkelaar, Aaron, Randall V. Martin, Chi Li, and Richard T. Burnett (2019), "Regional estimates of chemical composition of fine particulate matter using a combined geoscience-statistical method with information from satellites, models, and monitors." *Environmental Science & Technology*, 53, 2595–2611.

# A Appendix Figures and Tables

Figure A.1: Aggregate U.S. Coal and Gas Electricity Production over Time



Notes: This figure plots total coal (solid line) and natural gas (dashed line) electricity production in the United States from academic years 2008-09 to 2017-18. Academic years are defined as running from September-May. The x-axis is academic year and the y-axis is aggregate production of each fuel in units of one million MwH. Source: US Energy Information Administration (2021a).

Table A.1: States that administered standardized testing in the Fall or Year-round.

| State<br>(1) | Testing Calendar<br>(2) | Year of Switch to Spring Testing<br>(3) |
|---|---|---|
| Delaware | Year round | 2014-2015 |
| Idaho | Year round | 2014-2015 |
| Iowa | Year round | 2014-2015 |
| Maine | Fall | 2014-2015 |
| Michigan | Fall | 2014-2015 |
| New Hampshire | Fall | 2014-2015 |
| North Dakota | Fall | 2014-2015 |
| Oregon | Year round | 2013-2014 |
| Rhode Island | Fall | 2015-2016 |
| Vermont | Fall | 2014-2015 |
| Wisconsin | Fall | 2014-2015 |

Notes: This table lists states that conducted standardized testing in the Fall or Year round. Column (1) lists the state, Column (2) lists the testing calendar and Column (3) lists the academic year that the state switched to testing in the spring. *Source:* Author's own calculations from news articles. See Appendix B.

Table A.2: OLS Regression Results

| **Outcome:** Standardized Test Scores | | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| PM2.5 | −0.0034*** | −0.0035*** | −0.0034*** | −0.0035*** |
| | (0.0005) | (0.0005) | (0.0005) | (0.0005) |
| Lagged Test Score | 0.4759*** | 0.4758*** | 0.4759*** | 0.4759*** |
| | (0.0021) | (0.0021) | (0.0021) | (0.0021) |
| % Black | −0.4180*** | −0.4116*** | −0.4116*** | −0.4114*** |
| | (0.0335) | (0.0335) | (0.0335) | (0.0335) |
| % Hispanic | −0.1775*** | −0.1715*** | −0.1715*** | −0.1752*** |
| | (0.0257) | (0.0257) | (0.0256) | (0.0257) |
| Avg. Max. Temp (F) | | | 0.0006 | 0.0008* |
| | | | (0.0004) | (0.0004) |
| Move-in Rate | | | | −0.0889** |
| | | | | (0.0393) |
| Move-out Rate | | | | −0.0687 |
| | | | | (0.0434) |
| Observations | 701,199 | 694,257 | 694,257 | 694,257 |
| **Controls Used:** | | | | |
| Student Covariates | Yes | Yes | Yes | Yes |
| Local Economic Controls | No | Yes | Yes | Yes |
| Weather Controls | No | No | Yes | Yes |
| Sorting Controls | No | No | No | Yes |

Notes: This table reports the results from an OLS regression of test scores on PM2.5, as described in equation (1). Test scores are measured in standard deviations, while PM2.5 is measured in micrograms per cubic meter ($\mu$g/m$^3$). All regressions include subject, district, cohort, and year fixed effects. The control variables used for each set of controls is given in Table A.8. Standard errors are clustered at the district level. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Table A.3: Instrumental Variable Strategy I: IV Estimates Across Distance Bins

| **Outcome:** Standardized Test Scores | (1) | (2) | (3) |
|---|---|---|---|
| IV Estimate | $-0.023^{**}$ | $-0.021^{***}$ | $-0.019^{**}$ |
| ($\mu$g/m$^3$) | (0.009) | (0.006) | (0.008) |
| | | | |
| First-Stage F-stat | 134.78 | 331.24 | 181.36 |
| Observations | 694,257 | 694,257 | 694,257 |
| **Instrument** | Coal 0-20km | Coal 20-40km | Coal 40-60km |
| **Controls Used:** | | | |
| Student Covariates | Yes | Yes | Yes |
| Local Economic Controls | Yes | Yes | Yes |
| Weather Controls | Yes | Yes | Yes |
| Sorting Controls | Yes | Yes | Yes |

Notes: This table reports results from the two-stage regression defined by equations (2) and (3) when we use only one of our instruments – coal-based production within 0-20km, 20-40km, or 40-60km – rather than all three. Columns (1), (2), and (3) use coal production at a distance of 0-20km, 20-40km, and 40-60km from the district centroid as an instrument, respectively. Test scores are measured in standard deviations, while PM2.5 is measured in micrograms per cubic meter ($\mu$g/m$^3$). The 'First-Stage F-stat' in both panels displays the Kleibergen-Paap F-stat of the first-stage, which gives the F-statistic from a F-test testing the statistical significance of the instruments' ability to predict PM2.5. All regressions include subject, district, cohort, and year fixed effects. The control variables used for each set of controls is given in Table A.8. Standard errors are clustered at the district level. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Table A.4: Correlation between 2004-05 Fuel Shares and 2004-05 Covariates

| | Coal<br>Share<br>(1) | Gas<br>Share<br>(2) | Oil<br>Share<br>(3) | Renewable<br>Share<br>(4) |
|---|---|---|---|---|
| % White | −0.044 | −0.042 | 0.003 | 0.067 |
| % Black | 0.050 | 0.035 | −0.011 | −0.076 |
| % Hispanic | −0.006 | 0.045 | 0.002 | −0.018 |
| % Asian | 0.036 | 0.054 | −0.033 | −0.060 |
| % Free Lunch | −0.045 | −0.032 | 0.052 | 0.047 |
| % Special Needs | −0.007 | −0.034 | 0.017 | 0.032 |
| % English Language | 0.013 | 0.027 | −0.048 | 0.001 |
| Move-in | 0.031 | 0.010 | −0.076 | 0.008 |
| Move-out | 0.062 | 0.066 | −0.049 | −0.103 |
| % Bachelor's | 0.048 | 0.050 | −0.048 | −0.059 |
| % Employed | −0.024 | 0.020 | −0.014 | −0.003 |
| % Single Mother | 0.078 | 0.045 | −0.015 | −0.095 |

Notes: This table lists the correlation between 2004-05 district covariates and the 2004-05 share of total aggregate fuel production. To control for regional variation in fuel mixes, correlation coefficients are calculated within states and then averaged across all states. The four columns are the different types of fuels (coal, gas, oil, renewables) and covariates are listed in the rows. Source: Author's own calculations from the 2005-2009 American Community Survey, US Energy Information Administration (2021a) and Infutor.

Table A.5: Robustness: Shift-Share Instrument

| **Outcome:** Standardized Test Scores | | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| **Panel A:** Shift-Share Instrument, 2000-01 Shares | | | | |
| IV PM2.5 Estimate | $-0.023^{***}$ | $-0.029^{***}$ | $-0.023^{***}$ | $-0.021^{***}$ |
| ($\mu$g/m$^3$) | (0.006) | (0.007) | (0.008) | (0.007) |
| First-Stage F-stat | 396.30 | 309.916 | 269.51 | 293.54 |
| Observations | 584,201 | 583,282 | 583,282 | 583,243 |
| **Panel B:** Shift-Share Instrument: Production within 60km | | | | |
| IV PM2.5 Estimate | $-0.0166^{***}$ | $-0.0213^{***}$ | $-0.0185^{***}$ | $-0.0168^{***}$ |
| ($\mu$g/m$^3$) | (0.0044) | (0.0057) | (0.0063) | (0.0062) |
| First-Stage F-stat | 857.80 | 512.49 | 453.59 | 479.41 |
| Observations | 663,594 | 660,017 | 660,017 | 659,905 |
| **Controls Used:** | | | | |
| Student Covariates | Yes | Yes | Yes | Yes |
| Local Economic Controls | No | Yes | Yes | Yes |
| Weather Controls | No | No | Yes | Yes |
| Sorting Controls | No | No | No | Yes |

Notes: This table reports results for two robustness tests from our shift-share empirical strategy. In Panel A we run our baseline estimation but with shares defined in 2000-01. In Panel B we run our baseline estimation but use production within 60km of a district centroid. Test scores are measured in standard deviations, while PM2.5 is measured in micrograms per cubic meter ($\mu$g/m$^3$). Specifically, we use the interaction between pre-existing exposure to different power production interacted with national growth rates as our instrument, which is described in equations (4) and (5). The 'First-Stage F-stat' in both panels displays the Kleibergen-Paap F-statistic to assess the statistical significance of the instruments' ability to predict PM2.5. All regressions include subject, district, cohort, and year fixed effects. The control variables used for each set of controls is given in Table A.8. Standard errors are clustered at the district level. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Table A.6: Robustness: Including Non-Spring Testing States

| **Outcome:** Standardized Test Scores | | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| IV Estimate | $-0.019^{***}$ | $-0.018^{***}$ | $-0.023^{***}$ |
| ($\mu$g/m$^3$) | (0.005) | (0.006) | (0.007) |
| Observations | 774,627 | 678,795 | 652,580 |
| **Specification** | Coal Plant IV | Shift-Shares 2004-05 | Shift-Shares 2000-01 |
| **Controls Used:** | | | |
| Student Covariates | Yes | Yes | Yes |
| Local Economic Controls | Yes | Yes | Yes |
| Weather Controls | Yes | Yes | Yes |
| Sorting Controls | Yes | Yes | Yes |

Notes: This table reports results using the entire sample of districts, regardless of whether they tested during the spring or fall or year-round. Test scores are measured in standard deviations, while PM2.5 is measured in micrograms per cubic meter ($\mu$g/m$^3$). Column (1) displays the point estimates from our first empirical methodology leveraging year-to-year production variation. In particular, we use yearly coal-based power production within different distance bins as instruments for PM2.5, as described by equations (2) and (3). Column (2) reports results from our second empirical methodology which uses a shift-share instrument. Specifically, we use the interaction between pre-existing exposure from 2004-05 to different power production interacted with national growth rates as our instrument, which is described in equations (4) and (5). Column (3) presents estimates of the shift-share instrument when using fuel shares from 2000-01. All regressions include subject, district, cohort, and year fixed effects. The control variables used for each set of controls is given in Table A.8. Standard errors are clustered at the district level. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Table A.7: Heterogeneity by Race and Income

| **Outcome:** Standardized Test Scores | | | | |
|---|---|---|---|---|
| | High % Black | Low % Black | Low % FRL | High % FRL |
| | (1) | (2) | (3) | (4) |
| **Panel A:** Empirical Strategy I: Distance and Production Variation | | | | |
| IV PM2.5 Estimate | $-0.0219^{***}$ | $-0.0228^{*}$ | $-0.0263^{***}$ | $-0.0119^{*}$ |
| ($\mu$g/m$^3$) | (0.0060) | (0.0124) | (0.0071) | (0.0070) |
| First-Stage F-stat | 103.13 | 28.79 | 73.13 | 67.87 |
| Observations | 261,350 | 233,301 | 266,002 | 250,290 |
| **Panel B:** Empirical Strategy II: Shift-Share Instrument | | | | |
| IV PM2.5 Estimate | $-0.0196^{**}$ | $-0.0150$ | $-0.0106$ | $-0.0284^{***}$ |
| ($\mu$g/m$^3$) | (0.0080) | (0.0170) | (0.0092) | (0.0105) |
| First-Stage F-stat | 223.45 | 71.24 | 155.83 | 164.18 |
| Observations | 241,245 | 205,267 | 245,680 | 212,174 |
| **Controls Used:** | | | | |
| Student Covariates | Yes | Yes | Yes | Yes |
| Local Economic Controls | Yes | Yes | Yes | Yes |
| Weather Controls | Yes | Yes | Yes | Yes |
| Sorting Controls | Yes | Yes | Yes | Yes |

Notes: This table reports results for districts in the top tercile of % of Black students in Column (1) and the bottom tercile in Column (2). Terciles are calculated within states. In Panel A we run our baseline estimation with the instrument using distance and production variation. In Panel B we use the Shift-Share instrument. Test scores are measured in standard deviations, while PM2.5 is measured in micrograms per cubic meter ($\mu$g/m$^3$). Specifically, we use the interaction between pre-existing exposure to different power production interacted with national growth rates as our instrument, which is described in equations (4) and (5). The 'First-Stage F-stat' in both panels displays the Kleibergen-Paap F-statistic to assess the statistical significance of the instruments' ability to predict PM2.5. All regressions include subject, district, cohort, and year fixed effects. The control variables used for each set of controls is given in Table A.8. Standard errors are clustered at the district level. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Table A.8: Control Variables Sets

| **Controls:** | Student Covariates | Sorting Controls | Local Economic Controls | Weather |
|---|---|---|---|---|
| Cohort lagged test scores | X | | | |
| % Tested: male | X | | | |
| % Black | X | | | |
| % Hispanic | X | | | |
| % White | X | | | |
| % Tested | X | | | |
| Total Enrollment | X | | | |
| % English Language Learner | X | | | |
| % Special Needs | X | | | |
| % Free & Reduced Lunch | X | | | |
| Move-in rate | | X | | |
| Move-out rate | | X | | |
| % Employed | | | X | |
| % Labor Force Participation | | | X | |
| % Employed in utility sector | | | X | |
| % Employed in manufacturing sector | | | X | |
| % Bachelor’s or higher | | | X | |
| % Single mother households | | | X | |
| Avg. Rainfall | | | | X |
| Days 1-2” Rainfall | | | | X |
| Days $\geq$ 2” Rainfall | | | | X |
| Avg. Snow | | | | X |
| Days 1-2” Snow | | | | X |
| Days $\geq$ 2” Snow | | | | X |
| Avg. Min. Temp. | | | | X |
| Avg. Max. Temp. | | | | X |
| Days $\geq$ 100F | | | | X |
| Days 90-100F | | | | X |
| Days 80-90F | | | | X |
| Days 10 -20F | | | | X |
| Days 0 -10F | | | | X |
| Days $\leq$ 0F | | | | X |

Notes: This table lists the control variables used in each set of controls. “Student Controls” are from the Reardon et al. (2021) and the National Center for Education Statistics. “Sorting Controls” are from Infutor. “Local Economic Controls” are the ACS estimates from the Reardon et al. (2021). Weather Controls are from the National Oceanic and Atmospheric Administration’s Daily Global Historical Climatology Network and are calculate as the monthly averages from September to May. All regressions include subject, district, cohort, and year fixed effects.

# B Data Appendix

This appendix describes the data we use in detail and how we merge the various data sets together.

**SEDA Data:** School district performance data is from the Stanford Education Data Archive (SEDA) for the school years 2008-09 to 2017-18 (Reardon et al., 2021). Following the recommendation in the accompanying technical documents, we use the research file with 'cohort standardized' test scores (i.e., use the "seda_geodist_long_CS_4.1" research file). We then merge the accompanying district covariate data file (which also includes the local economic controls) on ("seda_cov_geodist_long_4.1"). We drop any district-year observations with missing test score information for that year. These data contain 1,127,781 district-subject-cohort-year observations, covering 11,806 school districts.

**Pollution Data:** Our next data set comes from Van Donkelaar et al. (2019) and contain monthly PM2.5 concentrations at a 0.01 degree by 0.01 degree resolution (roughly 1.1km by 1.1km at the equator) for the United States (excluding Hawaii) from 2008-2018. Specifically, we download the monthly North American Regional Estimates data files (titled V4.NA.02 PM2.5) from September 2008 through May 2018, excluding the months of June, July, and August. These data come in a gridded raster format; we use ArcGIS's 'raster to point' tool to assign each 0.01 degree by 0.01 degree cell to its centroid. We then use a point-in-polygon operation to assign these data to school districts using the district shapefile that accompanies the SEDA data (titled "seda_shapefiles_2019_4.0"). All the monthly PM2.5 readings that are assigned to a school district are then averaged, giving us average PM2.5 concentrations for each district-month. We then merge these data onto the SEDA data. The merge rate is nearly perfect, although we do lose 1.2 % of districts, leaving us with 1,121,041 district-subject-cohort-year observations, covering 11,663 school districts.

**Moving Data:** We then merge in data on move-in and move-out rates at the district-academic year level from Infutor. This dataset was calculated by geocoding residential addresses in Infutor and matching each address to a school district. We then used information on when an individual moved in and moved out of each address to calculate move rates. Our move rates focus on individuals aged 18 to 50. The merge results in 1,120,717 observations with 11,655 unique school districts.

**Energy Production Data:** Data on energy production by fuel source come from the Energy Information Administration EIA-923 form. This data contains information on monthly fuel production from all large U.S. power plants. The fuel types we use are: coal, gas, oil, and renewable (nuclear is classified as renewable). We calculate production for each plant over the academic year, starting in September, from 08-09 to 17-18. Districts are matched to plants based on the distance from the district centroid to the plant's location. Merging in this dataset results in no observations lost as districts that are not close (within 100km) to any power plant are classified as having zero exposure to production.

**Weather Data:** To build our weather controls we use data from the National Oceanic and Atmospheric Administration's Daily Global Historical Climatology Network, which includes daily station-level data for weather stations across the United States. Weather stations in the data often record only precipitation data, so we construct separate datasets for temperature, precipitation, and snow. We define school days in our data as any weekday from September 1 to May 31 and keep weather stations with a valid reading for at least 95 percent of school days following Park et al. (2020). For those few missing days, we impute missing daily observations using the nearest weather station with a valid reading that day.

For temperature, our data cover 5,397 stations with valid temperature readings for at least 95 percent of school days during our time period; the missing daily observations which cover one percent of the data – 109,246 out of 10,534,944 daily observations – are then imputed using the nearest station with a valid temperature reading for that day. School districts are then assigned to their nearest weather station, leaving us with 3,807 distinct weather stations covering the 11,655 school districts in our sample. On average, the nearest weather station is located 11.9 miles from a district's centroid; the first and third quartiles of distance are 5.4 and 14.7 miles, respectively.

The precipitation (snow) data cover 6,575 (4,225) stations with valid precipitation (snow) readings for at least 95 percent of school days during our time period;[13] the missing daily observations which cover one (nine) percent of the data – 150,419 out of 12,834,400 (486,778 out of 5,467,150) daily observations – are then imputed using the nearest station with a valid precipitation (snow) reading for that day. We then assign school districts to their nearest weather station, leaving us with 4,582 (3,247) distinct weather stations covering the 11,655 school districts in our sample. On average, the nearest weather station is located 10.7 (14.9) miles from a district's centroid; the first and third quartiles of distance are 4.4 (5.8) and 13.2 (19.2) miles, respectively.

**Sample Restrictions:** We include lagged test scores as controls in our models, necessitating us to drop our first year of data (2008-09) along with the first tested grade (third grade) to maintain consistent sample sizes across specifications. This leaves us with 805,424 district-subject-cohort-year observations, covering 11,496 school districts for grades 4-8 from 2009-10 through 2017-18. We also drop observations without information on district covariates, resulting in 784,586 observations in our full sample (summary statistics in Column (1) of Table B.1).

Lastly, for our main analysis data set we drop districts in states that do not test in the spring (see Table A.1), dropping 83,387 observations. Our final data contains 701,199 district-subject-cohort-year observations covering 11,419 districts (summary statistics in Column (2) of Table B.1).

[13] For snow, we only require the weather station to have valid readings for at least 75 percent of school days. This decision is necessitated by the fact that many stations do not report snow readings in September or May.

Table B.1: Summary Statistics

| **Summary Statistics** | Full Sample[1]<br>(1) | Analysis Sample[2]<br>(2) | Shift-Share Sample[3]<br>(3) |
|---|---|---|---|
| School-year Average PM2.5 | 7.19 | 7.22 | 7.44 |
| Test score (std) | 0.029 | 0.025 | 0.048 |
| Lagged Test score (std) | 0.030 | 0.028 | 0.05 |
| % Black | 8.6 | 9.1 | 9.5 |
| % Hispanic | 13.6 | 14.2 | 14.1 |
| % Asian | 2.1 | 2.2 | 2.5 |
| % White | 73.3 | 71.8 | 72.0 |
| % Free and Reduced Lunch | 48.9 | 49.7 | 47.9 |
| % English Language Learner | 4.26 | 4.45 | 4.46 |
| % Special Needs | 13.9 | 14 | 14.70 |
| Move-in rate | 0.049 | 0.047 | 0.048 |
| Move-out rate | 0.044 | 0.043 | 0.044 |
| Observations | 784,586 | 701,199 | 607,482 |
| Number of Districts | 11,476 | 11,419 | 9,691 |

[1] Full Sample includes grades 4-8 for school years 2009-10 through 2017-18. (Grade 3 and school-year 2008-09 are excluded from our sample so that we can control for lagged cohort test scores.)

[2] Analysis Sample is the same as Full Sample without states that have spring testing (see Table A.1 for a list of these States).

[3] Shift-Share sample is the same as the Analysis Sample but is restricted to districts with: (i) district covariates in 2004-05, and (ii) some positive power production within 40km in 2004-05.