APPLIED ECONOMETRIC STUDIES IN AIR QUALITY AND EDUCATION

By

Christopher Khawand

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Economics—Doctor of Philosophy

2016

ABSTRACT

APPLIED ECONOMETRIC STUDIES IN AIR QUALITY AND EDUCATION

By

Christopher Khawand

This dissertation is comprised of three standalone chapters loosely organized around a causal inference theme. It addresses empirical questions in environmental quality and education, while also offering some methodological insight in each chapter.

Chapter 1 is an empirical paper exploring the health effects of air pollution using the large-scale natural experiment generated by U.S. wildfires. In it, I apply a combination of forest fire and atmospheric dispersion models to provide predictions of where wildfire pollution goes after a fire. I use a panel instrumental variables framework to estimate the impact of increasing fine particulate matter pollution (PM2.5) on mortality and perinatal health. Increased short-term exposure to PM2.5 is associated with both increased mortality among elderly and poorer birth outcomes, and toxic metals appear to explain most of the effect.

Chapter 2, co-authored with colleague Wei Lin, is an econometric theory paper focused on the two-sample two-stage least squares estimator (TS2SLS). Here, we aim to fill a gap in the literature on how two-sample instrumental variables estimators behave in finite samples. We show closed-form approximations, simulation evidence, and empirical examples of how the estimator behaves. We offer recommendations for econometric practitioners using two-sample estimation.

Finally, Chapter 3 attempts to answer the question of whether there are ability peer effects in high school classrooms and how strong they are across the ability distribution. I leverage a series of placebo tests to validate whether the estimated peer effects justifiably represent causal effects. I provide a closed-form expression for the placebo estimates, showing that they are directly proportional to the amount of bias in their corresponding peer effect estimates. I find largely plausible peer effects nonlinear in ability, with high-performing students having the greatest positive impact on classrooms' overall performance.

To Lydia "Pikachu-Yoshi-Kitty-Puppy-Raccoon" Khawand

ACKNOWLEDGEMENTS

This dissertation represents five years of hard work and a slew of mistakes overcome. Here's a list of people I'd like to thank for, in one way or another, keeping those mistakes from sinking the ship:

Michael Bates, Julie Harris, and Dan Litwok—with whom I shared comradery, foosball, and MSU Dairy Store ice cream. We have all gazed into the ancient chicken bone in the 1st-year grad student office, and it has gazed into us.

My committee—for their kindness and patience through wild goose chases and long spells of no progress. Gary Solon taught me the fundamental causal inference skills that have sparked an exciting start to my career; Soren Anderson was truly a second advisor, giving me an incredible amount of detailed feedback and morale boosts when things were not going well; and Jeff Wooldridge bestowed upon me nearly all of the useful econometric knowledge that I have and an unwavering committment to finding econometric truths (even if they tend to mostly be in asymptotia).

Lastly, my wife Jennifer and daughter Lydia—they are the Chapter 0 of this dissertation, my life's work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1. Air Quality, Perinatal Health, and Mortality: Causal Evidence from Wildfires

1 Introduction

In the past 40 years, ambient air quality regulation has grown in response to the burgeoning evidence of the public health costs of air pollution. The Clean Air Act Amendments of 1970, the establishment of the Environmental Protection Agency (EPA), and subsequent refinements of air quality standards have all contributed to general downward trends in pollution levels. While the health benefits of air quality improvement are uncontroversial at the highest margins of pollutant levels, the important question remains whether additional reductions will also yield health benefits, and whether those health benefits exceed the marginal costs of abatement. Because pollutants are not randomly assigned and may be correlated with other determinants of health outcomes, an important scientific challenge has been to develop research designs that provide precise, unbiased, and population-representative estimates of air pollution's effects.

In this paper, I exploit quasi-random shocks to ambient fine particulate matter (PM2.5) concentrations generated by large wildfires across the United States to estimate effects on mortality and infant health outcomes. Wildfires are uncontrolled fires primarily occurring in remote wilderness areas, but cause significant variation in urban particulate levels through mechanisms that are plausibly unrelated to non-pollution determinants of health. First, I quantify the effect that wildfires have on air quality by applying a sequence of specialized emissions and dispersion models to historical fire data to generate measures of wildfire pollution for the continental U.S. over time. Then, I estimate the effects of short-term and *in utero* exposures on adult mortality and infant health outcomes, using modeled wildfire PM2.5 as an instrumental variable for station-observed PM2.5. I use extensive pollution monitoring data—spanning 60 PM2.5 subspecies and 18 criteria pollutant and organic gases—to decompose the shock to air quality represented by the wildfire PM2.5 instrument and present a methodology for assessing potential bias from omitted pollutants.

The U.S. Environmental Protection Agency (EPA) has identified six airborne "criteria" pollutants to be regulated under the Clean Air Act that are generally considered harmful to public health: particle pollution (PM2.5 and PM10), carbon monoxide (CO), nitrogen dioxide (NO2), ozone (O3), sulfur dioxide (SO2), and lead (Pb). This study estimates the effects of PM2.5 generated by wildfires and tests whether these estimates potentially reflect the effects of other criteria pollutants. Fine particulate matter, defined as particulate matter less than 2.5 micrometers in diameter (PM2.5), is considered the most dangerous because of its ability to penetrate deep into the human lung and sometimes enter the bloodstream.

I make several contributions to the literature on both the health effects of air pollution and air quality effects of wildfires. First, I systematically assess the impact of large wildfires ($\geq$ 1000 acres) on ground-level air quality in the continental United States. I combine estimates of source emissions with an atmospheric model to retrospectively forecast the spatial distribution of wildfire-related pollutant concentrations in the period following a historical fire event, using the resulting data to predict pollutant concentrations at air pollution monitors. For 78 pollutants, I estimate a set of lower bounds for the percentage of each pollutant's average ambient concentration that is attributable to wildfires; notably, wildfires contribute at least 15% of ambient aggregate PM2.5, 5% of PM10, 5% of O3, and large fractions (15%-35%) of several dangerous metals bound to fine particulates, including arsenic, lead, mercury, nickel, and cadmium. In addition to the aggregate quantities of particulate pollutants they generate, wildfires cycle metallic and other highly toxic industrial emissions previously deposited into wildland vegetation and soils back into the atmosphere, resulting in new ground-level exposures in population centers. These findings underscore the potential significance of wildfires' contribution to public health and that socially optimal fire management policies must take wildfires' health costs from worsened air quality into account. Furthermore, 75% of geographic exposure to wildfire PM2.5 occurs outside of the state of origin, raising the possibility that wildfire management policy in the U.S. may be inefficient due to inter-state spillovers.

Next, I estimate the effect of average monthly PM2.5 exposure on county-level mortality rates

for 2004-2010 via instrumental variables, controlling for weather variables and stringent sets of region-specific fixed effects. Short-term exposure to PM2.5 is associated with mortality with magnitudes consistent with prior literature, and the dose response is approximately linear below regulatory limits of PM2.5. A transitory $10\mu gm^{-3}$ increase in a county's average monthly PM2.5 (approximately a doubling of average ambient concentrations in the sample period) is associated with one additional death per 100,000 individuals. These effects are largely driven by cardiovascular and respiratory fatalities, but PM2.5 is also associated with general disease-related causes of death. Nearly all short-term PM2.5-related deaths are of individuals over age 65, and women are twice as susceptible as men. Because wildfires also emit large quantities of several gaseous pollutants, I attempt to control for potentially correlated pollutants by including a set of comparably modeled controls for NO2, SO2, NH3, and VOC gases. Because of the complex chemical and environmental interactions underlying O3 production and the corresponding difficulty of predicting O3 concentrations from wildfires using the same set of pollution models, I am unable to decisively rule out confounding effects from O3. Based on best available estimates from other studies, I estimate that this effect is on the order of 35% of the estimated effect of PM2.5. Finally, I estimate the effects of prenatal exposure to PM2.5 on premature birth rates, birth weight, and sex ratios, finding small but statistically significant harmful effects. I also find marginally insignificant evidence for negative effects on the fraction of males in a birth cohort, and my estimates are consistent with higher susceptibility of male fetuses to death from pollution shocks estimated in Sanders and Stoecker (2011).

Controlling for non-PM2.5 emissions from wildfires results in a different composition of PM2.5 that more heavily favors toxic species, and I find larger effects in the presence of higher fractions of metals and lower fractions of non-metal particulates. While the intrusive quality of PM2.5 is the basis for the proposed dangers of PM2.5, there is wide heterogeneity in the chemical composition of PM2.5 and some evidence of heterogeneous effects, but relatively little understood about the relative toxicities of individual substances (Bell 2012). PM2.5 is composed of a wide range of substances, including elemental carbon (EC), organic carbon (OC), nitrates (NO3-), sulfates (SO42-),

and metals bound to particulates (such as mercury and lead). Some of these are formed or released directly from an emission source (commonly EC, OC, and metals) and others are formed through chemical reactions in the atmosphere (e.g., nitrates and sulfates). Composition also varies widely by region and over time from cross-sectional differences and seasonal differences within regions (Franklin et al. 2008). This heterogeneity presents problems for the effective regulation of particulate levels, as small exposures of highly toxic species are potentially as dangerous as large exposures of EC, OC, or other species that account for most of PM2.5 mass. Relatedly, it presents statistical challenges for interpeting estimated effects for PM2.5. When controls for non-PM2.5 pollutants from wildfires are included, estimated effects on mortality increase by over two times. For infant health, effects approximately double for prematurity, gestational age, and average birth weight. I interpret this pattern of estimates as evidence that the conditional mixture of PM2.5 identified has increased toxicity that exceeds any reduction in upward bias accomplished by adding controls.

This work attempts to make new methodological and evidentiary contributions to the already-large and diverse economic literature on the health effects of pollution. For short-term health outcomes, panel studies and regional natural experiment studies are two popular research designs. The widely-accepted truism motivating most of the contemporary air pollution literature is that pollution exposure is non-randomly assigned and systematically related to other determinants of health outcomes. Panel studies, such as Currie and Neidell (2005), attempt to address this non-random assignment through exploiting narrow variation through stringent fixed effects. Natural experiment studies try to provide a source of quasi-random assignment by isolating the variation they use to a particular type of pollution-generating (or reducing) event. Strategies have included exploiting the timing of the Clean Air Act of 1970 to predict relatively sudden decreases in particulate concentrations (Chay and Greenstone 2003); changes in daily airport traffic congestion in California caused by weather in other major airports (Schlenker and Walker 2011); weekly panel variation in automobile traffic to identify the effects of carbon monoxide, ozone, and particulate matter on infant mortality rates (Knittel, Miller, and Sanders 2011); and temperature inversions

in Mexico City (Arceo-Gomez et al. 2011). Currie et al. (2013) provide an extensive survey of both types of papers exploring the effects of early-life exposure to pollution, finding a general consensus that airborne pollutants are associated with infant mortality, premature birth, and low birth weight. Papers applying natural experiments to adult mortality have been more infrequent. Chay et al. (2003) uses the timing of the Clean Air Act of 1970, finding insignificant effects on adult and elderly mortality. Pope et al. (2007) use an 8-month national strike of copper smelter workers to estimate the effect of sulfate particulate reductions, finding a 2.5% reduction in mortality over the strike period.

Several papers have attempted to estimate health effects of major wildfire events, implicitly taking their exposure measures as proxies for pollution shocks. Jayachandran (2009) examines sharp increases in particulate pollution from an intense wildfire season in Indonesia in 1997, tracking spatial and temporal variation in pollution from wildfires using satellite-based measures of particulate levels. She finds evidence that prenatal smoke exposures during that period caused a substantial increase in early-life mortality, on the order of a 20 percent increase in the under-age-three mortality rate. Breton, Park, and Wu (2011) estimate that prenatal exposure to high PM2.5 concentrations from a week-long wildfire event in California was associated with an 18g decrease in mean infant birth weight in comparison to counties unaffected by the fires.

Few studies have used modeled exposures from large emission events based on atmospheric transport models, and none have used exposures in tandem with monitoring data to predict health outcomes.[1] Rappold et al. (2012) use modeled wildfire exposures in North Carolina to assess increases in asthma and congestive heart failure risks with reduced-form Poisson regressions. I fill a gap in the literature by incorporating developments in emissions and atmospheric transport modeling and taking advantage of substantial increases in computational power made over the last decade. I unite quasi-random variation in pollution levels predicted from wildfire and at-

[1] A class of study distinct from this one combines modeled exposures with pre-existing estimates of health risks to determine population-wide impacts. For example, Caiazzo et al. (2013) use the Community Multiscale Air Quality (CMAQ) model combined with the U.S. National Emissions Inventory for 2005 to create an annual predicted map of average pollution concentrations, and interpret this as a measure of long-term pollution exposure. Also, several studies use observed changes in particulate measurements and only employ "backward trajectory" calculations to indirectly verify that large changes are due to a specific event, such as wildfires or dust episodes.

mospheric models with observed pollution levels in a panel econometrics framework to estimate health effects, providing a methodology that bridges some of the long-standing gaps between the atmospheric science, epidemiology, and economics literatures on air quality.

2 Data

2.1 Modeled Wildfire Air Pollution

Combining historical wildfire event data and meteorology with scientifically relevant fire and atmospheric transport models, I generate a high-resolution, gridded daily measure of wildfire pollution for the continental U.S. (CONUS) domain. The measure represents a retrospective forecast of where pollution from documented fire events would be likely to have traveled given what is known about atmospheric behavior during and after the fire. To this end, I use the BlueSky Framework software package, which integrates several existing models of emissions and transport processes into a unified process.

2.1.1 Wildfire Data

State and federal agencies responsible for wildfire management keep records on the location, size, and timing of wildfire events. Fire events larger than 1,000 acres are gathered from the Fire Protection Agency (FPA) Fire Occurrence Database (FOD), an interagency collection of fire event reports updated for accuracy and cleaned for duplicates using methods described in Short (2013). The fire event characteristics drawn from this database for modeling are the latitude and longitude point data of the fire, date and time the fire was detected, area of the fire burned in acres, and the date and time at which a fire agency declared it contained. For an available subset of federal fires, I draw the date and time at which a fire agency declared the fire extinguished from a U.S. Geological Survey database of fires reported by the six major federal agencies tasked with managing wildfires. If any of the values except for the containment or extinguish dates are missing, the fire is omitted. Where time of extinguishment data are missing, I empirically estimate the total burn time using

a regression model with categorical dummies for the fire area and the duration from start to containment as predictors, adjusting for region-specific unobserved effects and seasonal effects; where both containment and extinguishment dates are missing, I use the same model without containment time to predict burn duration. The methodology, rationale, and results of the total burn time estimation procedure are described in Appendix A.1. Figure 1 shows the total number of acres burned in fires larger than 1,000 acres mapped by state for 2000-2010, as calculated from the FPA FOD database. The majority of area burned is concentrated in the West, Northwest, and Southwestern United states, with a decreasing eastward and strongly decreasing Northeastward pattern. These large fires constitute over 85 percent of area burned in the United States over this period. Fires smaller than 1,000 acres are a significantly larger percentage of area burned for Northeastern and Central states, but are not included because of high computational costs of the modeling process relative to their small total emissions contributions compared to larger fires.

2.1.2 Description of Wildfire Emissions and Air Pollution Modeling

Wildfires can be started by lightning strikes or direct sunlight when highly flammable fuels (e.g., forest underbrush) endure an extended dry period. Wildfires are also caused by human errors, such as escaped campfires, car accidents, or downed power lines. Occasionally, they are intentionally set by arsonists. Fires are also intentionally set by fire management agencies to preemptively burn fuels for naturally-occurring fires, among other functions. Wildfire incidence peaks in mid-to-late summer, but has varying seasonal peaks by region. The majority of large wildfire events (over 1,000 acres in size) occur in the Western and Northwestern United States.

There are several phenomena which contribute variation to the amount of wildfire-generated pollution at a given point in time in space. Broadly, these are the characteristics of the fire and the meteorological conditions at the time of and shortly after the fire event. The duration of the fire is a function of time till detection, containment efforts, and the containment difficulty of the fire. Besides its role in promoting the rate of spread and ultimate size of a fire, the fuel cover determines the volume and chemical composition of emissions from the fire per unit of area burned. Wildfires'

dominant emissions by mass are PM10, PM2.5, CO, and NOx. In addition to PM2.5 generated by biomass burning, such as Organic Carbons (OC), wildfires release minerals and metals which accumulate in forest soils and vegetation from atmospheric deposition. Nearby historical industrial activity is strongly related to the amount of lead and mercury re-released by fires into the atmosphere, with these re-emissions representing a significant fraction of atmospheric concentrations.

Once generated, emissions travel upward at varying speeds depending on a variety of factors, resulting in a heterogeneous vertical distribution of pollutants in a fire. This vertical distribution then interacts with ambient pressure and wind conditions which result in airborne transport of emissions downwind. Emitted particles (and gases) interact with weather conditions heterogeneously, resulting in relative downwind changes in concentrations that vary by pollutant. Dry deposition is a set of processes by which pollutant concentrations decrease through contact with surfaces, which include gravitational settling and interception (collision with trees, buildings, etc.). Wet deposition is a set of processes by which atmospheric hydrometeors (e.g., precipitation) absorb particles.

I utilize a sequence of wildfire models that exploit several facets of these wildfire emissions and pollution transport processes to predict the contribution to PM2.5 levels from wildfires. The computational workflow is explicitly described in Appendix Section A.2.1; Figure 2 depicts the workflow visually. Historical fire events are input into the BlueSky Framework, where fuel loadings, fuel consumption, emissions, and vertical plume rise are estimated; these are fed as emission sources into HYSPLIT, which calculates the concentations' trajectory and dispersion from each source; hourly spatial concentration estimates are calculated at an approximately $1600 km^2$ resolution (approximately 5,000 unique points in the continental US); then, the HYSPLIT predicted concentrations are sampled at pollution monitoring station locations and averaged by county and month to create a monthly panel of county averages of wildfire pollution.

2.1.3 Modeling Tools

The interface between wildfire management and air quality standards has prompted extensive development of tools in the last two decades to appraise the downwind impacts of wildfires. Begin-

ning in 2003, the National Oceanic and Atmospheric Administration (NOAA) developed and implemented the Smoke Forecasting System (SFS) to provide operational forecasts of wildfire PM2.5 (Rolph et al. 2008). A central tool in the NOAA SFS is the BlueSky Framework (BSF), a modeling framework which connects independently developed models of fuel loading, fire consumption, fire emissions, and atmospheric transport (Larkin et al. 2009). The BSF has also been used in development of regional forecasting systems in the Pacific Northwest (O'Neill et al. 2009). The BSF readily accommodates several popular models of each component of the modeling process.

The Fuel Characteristic Classification System (FCCS) is a 1km-resolution spatial map of fuelbed types across the continental United States developed from a combination of fuel photo series, scientific literature, satellite imagery, and expert opinions (Ottmar et al. 2007). CONSUME 3.0 predicts how the amount of fuel consumption for a given fire event divides between flaming, smouldering, and residual phases, each of which have unique contributions to emissions due to differences in combustion efficiency (Prichard et al. 2005). The Fire Emissions Production Simulator (FEPS) is a software module that simulates emission production and plume buoyancy based on a provided consumption profile (Anderson et al. 2004). FEPS is capable of fuel consumption calculations, but this functionality is replaced by CONSUME 3.0 in this modeling process. These three modules have all been used, via the BSF, in the development of national fire emissions inventories since 2008. Lastly, the Hybrid Single-Particle Lagrangian Integrated Trajectory model (HYSPLIT) is a system which uses gridded meteorological data to simulate airmass trajectories, dispersion of concentrations from pollutant plumes, and deposition processes (Draxler and Hess 1997). In addition to being used in the NOAA SFS, HYPSLIT has been used in hundreds of applications, such as modeling fallout dispersion from the Fukushima Daichii nuclear disaster (Draxler et al. 2013), African dust transport to the Iberian peninsula (Escudero et al. 2006), and dispersion of particulate heavy metals from industrial emission sources in Spain (Chen et al. 2013).

2.2 Birth and Mortality Data

Data on the population of births, linked infant deaths, and mortality events in the United States for 2004-2010 come from the U.S. Center for Disease Control's (CDC) National Center for Health Statistics' (NCHS) National Vital Statistics System (NVSS). Data sets contain all non-identifying information recorded on birth and death certificates. Each birth record contains the year and month of the birth event in addition to important perinatal health outcomes, such as birthweight, Apgar scores, estimated gestation, birth complications, and characteristics of the mother and father of the child. Table 3 summarizes these outcomes by gestational category (full-term and pre-term). The mortality data contain individual death records, which include the year and month, county, cause of death, and characteristics of the deceased individual (race, gender, and education). For 2005 and beyond, county identifiers are censored for all counties with fewer than 100,000 individuals.

Causes of death are coded into 39 groups, in accordance with the latest classifications of the International Statistical Classification of Diseases and Related Health Problems (ICD-10). County-by-month mortality rates for each cause are calculated by summing counts from the 34 categories causes of death, including cancers, heart failure, respiratory disease, and other diseases and dividing by a population measure. The population estimates used to calculate rates per 100,000 individuals are from the CDC NCHS Bridged-Race Population Estimates, a set of annual intercensal county population estimates with breakdowns by sex, age, and race. I generate an "all-cause" rate from all non-external, non-accidental causes of death for the general population, and by gender, and infant, child, and 10-year age groups. Table 2 reports summary statistics for mortality rates in the sample.

2.3 Ambient Air Pollution and Weather Data

Daily average monitoring station observations of pollutant levels are gathered from the U.S. Environmental Protection Agency's Air Quality System (AQS), a centralized database of pollutant measurements from state and federal monitors. The geographic and temporal distribution of mea-

surements varies widely by pollutant. The PM2.5 Chemical Speciation Network provides measurements of PM2.5 subspecies of interest, such as metals and nitrates. Some stations collect data at weekly rather than daily frequency. For county-months with missing station-days, I use the average of nonmissing observations by first averaging to monthly station observations, and then averaging station-month values to county-month values. County-months with no station observations are excluded from the sample. For birth and death outcomes, I define the mother's and decedent's county of residences, respectively, as the aggregate geographic units for calculating pollution exposure. For local weather measures, I use data from the North America Land Data Assimilation System on average monthly daily maximum and minimum air temperatures and monthly precipitation quantities for each U.S. county. These data were drawn from the CDC WONDER database. This data source is distinct from the meteorological reanalysis data used as inputs into the pollution transport model.

3 Econometric Approach

3.1 Statistical Model

I consider the following linear model of health outcome y_{it} with a $K \times 1$ vector of endogenous variables representing pollution levels, P_{it}, and a set of unobserved effects:

$$y_{it} = P_{it}\beta + R_{it}\psi + \alpha_i + g_{it}(t) + \varepsilon_{it} \tag{1}$$

$$P_{kit} = z_{it}\gamma_k + R_{it}\psi_k^f + \eta_{ki} + f_{ki}(t) + v_{kit} \tag{2}$$

$$g_{it}(t) = c_{i,a(t)} + s_{i,m(t)} + \tau_i\omega(t) \tag{3}$$

$$f_{ki}(t) = c^f_{ki,a(t)} + s^f_{ki,m(t)} + \tau^f_{ki}\omega(t) \tag{4}$$

Equation (1) shows the relationship between the health outcome (e.g., mean birthweight) y_{it} and pollutants P_{it} for county i in month t. R_{it} is a set of time-varying county characteristics, α_i represents a county fixed effect, and $g_i(t)$ generally represents time-varying unobserved heterogeneity. ε_{it} is an idiosyncratic error term which may generally be correlated with P_{it}. Equation (2) represents the first stage relationship between pollutant k and the vector of at least K excluded modeled wildfire pollution instruments, z_{it}, with a set of fixed effects η_{ki} and $f_{ki}(t)$ matching those in equation (1). Equation (3) defines $g_i(t)$ as the sum of sets of region-year fixed effects $c_{i,a(t)}$, region-month (seasonal) effects $s_{i,m(t)}$, and arbitrary regional time trends $\tau_i\omega(t)$. $a(t)$ and $m(t)$ are functions which convert the global time index t to the correct calendar year (e.g., 2004) and calendar month (e.g., July) indices. "Region" can generally refer to any geographic unit which hierarchically nests counties, including counties, states, and NCDC climate regions. Equation (4) defines $f_{ki}(t)$ in parallel to (3) for P_{kit} (except naturally requiring that fixed effects vary by pollutant k).

Region-year fixed effects account for annual trends in the health outcome including those driven by changes in pollution from sources other than wildfires. This includes region-specific climatological changes and regulatory responses to wildfire incidence or pollution, which might simultaneously affect both wildfire incidence and health outcomes. Region-month fixed effects account for unobserved persistent seasonal differences between regions, such as weather patterns that drive seasonality in wildfire incidence and health outcomes. Including fixed effects increases the plausibility of the assumption that the instrument is exogenous in equation 1; namely, that $E[\varepsilon_{it}|z_{it}, R_{it}, \alpha_i, g_{it}(t)] = 0$.

3.2 Identification

The structural model of atmospheric transport represented by HYPSLIT seamlessly combines emission inputs, trajectory and dispersion calculations, and pollutant removal from the atmosphere through deposition processes to form a single, powerful instrument in the form of a predicted concentration. The dominant source of variation in simulated pollution concentrations using the

HYSPLIT-based modeling framework is the common movement of air parcels (i.e., wind). However, fuel loadings, wet deposition, and dry deposition generate some independent variation among pollutant types that can separately identify their effects. The possibility of separate identification of pollutants breaks down as the pollutants become more similar in the ways that HYSPLIT is able to distinguish them; modeled concentrations of similar pollutants are highly collinear. A corollary of this is that even a perfectly calibrated pollutant instrument will also proxy for the effects of its unmodeled close chemical neighbors, potentially causing bias in estimates of the effect of a specific pollutant species. In the modeling framework used here, variation in downwind wildfire PM2.5 independent from other wildfire pollutants is identified primarily by differences in fuel composition at the wildfire and deposition rates between PM2.5 and gases. Interpretation of the estimated effects is complicated by heterogeneous effects, especially those driven by the chemical composition of the PM2.5 that is statistically identified; this complication is examined in Section 4.1.2. These problems hold true for nearly any attempt to identify the effects of PM2.5.

Previous studies have similary exploited atmospheric phenomena and pollutant characteristics through regression interactions. For example, Schlenker and Walker (2011) interact airport taxi time with wind speed to separately identify CO and NO2, which may be explained by differing dry deposition rates between CO and NO2. NO2 has a higher deposition velocity than CO. Assuming a fixed emission ratio of CO to NO2, higher wind speeds will carry parcels of both pollutants equally far but deposit more NO2 than CO, resulting in an increasing ratio of CO to NO2 in distance from the airport. An alternative explanation they supply is that higher wind speeds change the composition of emissions from airplane engines to be more NO2-heavy. Both deposition differences across pollutants and differences in emissions ratios for specific events would be captured by HYSPLIT's deposition modeling process, with the practical drawback that one must be specific about deposition characteristics and emissions quantities in HYSPLIT's setup.

3.3 Testing and Controlling for Effects from Multiple Pollutants

3.3.1 Controlling for Multiple Wildfire Pollutants

In the ideal empirical setting, one would have a large enough dataset with measurements of all species of interest with identifying instruments for each species and estimate the effects of multiple endogenous variables using 2SLS or otherwise appropriate IV estimator. In reality, station coverage is limited to fewer than 20% of county-month observations in the sample period, and further limited when overlapping species measurements are required. The generation of strong identifying instruments may be both scientifically constrained by the quality of models and practically constrained by computational power. In lieu of the ideal estimation of all pollutants' coefficients, it is feasible to consistently estimate a single structural parameter of interest (in this case PM2.5's effect) without any concern for the structural parameter values for other pollutants. Generally, evidence for a consistent estimate of the effect of PM2.5 can be established by exhausting potential confounding causal pathways through a combination of control variables and pre-testing for omitted variables.[2]

Under the assumption that estimates using the wildfire pollution instrument will reflect effects causally originating with wildfire events only, the primary risk of confounding comes from omitted pollutants which are correlated with the wildfire instrument. A measure of downwind PM2.5 from a wildfire will be correlated with other pollutants emitted concurrently in the same fire's combustion processes, which will also share at least some of its atmospheric trajectory. For example, wildfires simultaneously emit quantities of PM2.5 and NO2, and their atmospheric destinations are highly correlated. In this framework, the health-effect parameter for PM2.5, β_{PM25}, can be identified either through joint IV estimation of all pollutants, or through single-variable IV estimation of PM2.5 alone with controls for pollutants from the same source. This equivalence is motivated by writing the reduced form for equation 1 as follows, only substituting the endogenous variable representing PM2.5 using the first stage based on a single instrument for wildfire PM2.5

[2]Causal pathways can also be credibly ruled out using evidence from rigorous studies that find no effects of an omitted explanator on the outcome of interest, but I do not do this here.

$$y_{it} = z_{pm25,it}\eta_{pm25} + P_{B,it}\eta_B + \Upsilon_{k,it} + \varepsilon^*_{it} \tag{5}$$

$z_{k,it}$ is defined as pollutant k originating from wildfires. $P_{B,it}$ is the vector of all pollutants excluding PM2.5 originating from all sources. For brevity, define $\Upsilon_{k,it}$ as the composite set of controls and effects and ε^*_{it} as a composite error term for equation 1. Partition each pollutant k into its concentration from wildfires and its concentration from all other sources, defining $P_{B,it} = z_{B,it} + \tilde{P}_{B,it}$. Then,

$$y_{it} = z_{pm25,it}\eta_{pm25} + z_{B,it}\eta^{wf}_B + \Upsilon_{k,it} + \tilde{P}_{B,it}\tilde{\eta}_B + \varepsilon^*_{it} \tag{6}$$

Because wildfire PM2.5 in part shares common emission and transport processes with other pollutants from fires, PM2.5 and other wildfire pollution are correlated: $E[z_{B,it}\,|\,z_{pm25,it}, \Upsilon_{k,it}] \neq 0$. Uncontroversially, $E[\tilde{P}_{B,it}\,|\,z_{pm25,it}, \Upsilon_{k,it}] = 0$ is a core assumption for the validity of the instrument; wildfire PM2.5 must be orthogonal to any pollutants in B from all non-wildfire sources. The reduced-form regression of y on z_{pm25} will be inconsistent for η_{pm25}. However, $z_{B,it}$ is observed by virtue of the same modeling process that generates z_{pm25}, and the reduced-form regression of y on z_{pm25} and z_B produces a consistent estimate for η_{pm25}. Correspondingly, provided the other key assumptions for the consistency of IV are met, IV estimation of y on P_{pm25} and z_B with (z_{pm25}, z_B) as instruments is consistent for β_{pm25}. While both the joint IV estimation and the single-variable IV procedures will be consistent for β_{PM25}, single-variable IV is far more feasible to implement; it only requires station observations of PM2.5, an instrument for PM2.5, and adequate controls for correlated pollutants. In some cases, modeled pollutants may be sufficient as controls but not sufficient as identifying instruments; joint IV estimation of PM2.5 and NO2 with a strong instrument for PM2.5 and weak instrument for NO2 may result in an inferior estimate for PM2.5 compared to the corresponding consistent one-variable IV estimate for PM2.5.

An alternative solution to creating an instrument or proxy is to use the endogenous measure of the omitted variable as a control, but in the pollution setting this is not always feasible. First, measurement coverage for each pollutant species incompletely overlaps both across stations and

time. Second, while the quality of station measurements for a particular species might be sufficient for determining whether wildfires have an impact on concentrations of that species in a station-by-station analysis, they may not be appropriate measures of concentrations for aggregate geographic regions used to measure health outcomes (i.e., counties in this paper). Relatedly, to the extent that station measurements (whether due to direct station mismeasurement or spatial error) fail to capture variation from wildfires due to measurement error, the control would fail to account for the influence of the omitted variable. The estimates in this paper instead use the equivalent of a proxy for pollutants emitted from wildfires as controls, thereby reducing or removing their confounding role.

3.3.2 Pre-testing for Omitted Pollutants

It is possible to meaningfully pre-test for potential omitted variables provided there are observations containing values of both the omitted variable and the instrument. Sufficient power in the test obviates the need to develop instruments or controls for the omitted variable if the test is negative. The test is to run a pseudo-first-stage regression of the suspected omitted variable on the current set of instruments and controls and checking whether the current instruments are jointly significant predictors of the proposed omitted variable. In practice, a researcher may not be able to develop an adequate identifying instrument or proxy for the potential omitted variable, and she might not be able to directly control for measures of the omitted variable without losing sample size (or relying on imputation methods). The creation of new instruments or proxies for new pollutant species is constrained practically by computational requirements and development time for accurate emission factors and deposition parameters. Separate identification of pollutants is also statistically limited by the mechanical richness of the modeling process. As separate identification of pollutants in the modeling process used here is driven by differential emissions and deposition behavior, pollutants with very similar emission and deposition properties will be weakly identifiable unless some part of the modeling process is upgraded to exploit other differences in characteristics not accounted for by HYSPLIT (e.g., buoyancy, aerodynamic, or photochemical properties).

To illustrate, consider a simple two-pollutant example with Pollutant A and Pollutant B and a wildfire-generated measure of Pollutant A as an identifying instrument. Assume we have a prior belief that Pollutant B causes mortality. If Pollutant B is positively correlated with wildfire-generated Pollutant A, then an instrumental variables regression of mortality on Pollutant A with wildfire-generated Pollutant A as an instrument and no control for Pollutant B will be biased upward due to the confounding effect of Pollutant B. Hence, a pseudo-first-stage regression of Pollutant B observations on wildfire-generated Pollutant A which produces a significantly positive coefficient on wildfire-generated Pollutant A is interpreted as evidence of this upward bias (in context of the prior belief that Pollutant B has an effect on mortality).

This test for omitted variables holds under one additional assumption: the direction, but not necessarily the magnitude, of the average partial effect of the instrument is the same between the samples used for testing and estimation. If the instrument is monotonically related to the endogenous variable of interest for the population, this assumption is satisfied. For the relationship between wildfire-generated pollution to observed pollution, these assumptions are likely to hold. While there may be first-stage heterogeneous effects of the modeled wildfire-generated pollution (either due to heterogeneous modeling error or because of true heterogeneity in the world due to chemistry or other processes), I assume that effects are bounded by zero. With the exception of a few highly reactive pollutants and/or pollutants with low atmospheric quantities, wildfire pollution can be generally expected to homogeneously weakly increase (or decrease) each pollutant type across geographic location and time. Let superscript A and superscript B denote that the variable is drawn from estimation sample's and testing sample's subpopulations, respectively. The assumption can be written as

$$\frac{\partial E(P_{it}^{A}|z_{it}^{A},R_{it}^{A},\eta_{ki}^{A},f_{ki}^{A}(t))}{\partial z_{it}^{A}}>0 \iff \frac{\partial E(P_{it}^{B}|z_{it}^{B},R_{it}^{B},\eta_{ki}^{B},f_{ki}^{B}(t))}{\partial z_{it}^{B}}>0. \tag{7}$$

In the linear case, this simply translates to the pseudo-first-stage coefficients having the same direction in each sample (i.e., $\gamma_k^A > 0 \iff \gamma_k^B > 0$). The corresponding hypothesis test is of $H_0 : \gamma_k^A = \gamma_k^B = 0$, $H_A : \gamma_k^A \neq 0$ using the t-test of $H_0 : \gamma_k^B = 0$ from the regression using the testing sample

B. Because of sampling error, failure to reject the null does not rule out omitted pollutants, but the estimate's confidence interval can be informative of the largest effect that is statistically supported by the given estimate. The true coefficient in the testing sample could be substantively smaller than the coefficient in the estimation sample, in which case the confidence interval bound may be misleadingly low. A more stringent assumption, which would imply (7), is similar to the necessary assumption for the consistency of two-sample IV estimators: $\gamma_k^A = \gamma_k^B$. This assumption permits a more literal interpretation of the coefficients and confidence intervals when the omitted variables test is conducted with a set of observations that is not identical to that being used to estimate the equation of interest. I perform and interpret this test for criteria and organic gases in Section 4.1.1.

4 Results

4.1 Wildfires' Effect on Ambient Air Quality

4.1.1 First Stage: Wildfires' Effect on Ambient Concentrations of Pollutants

Wildfires have a considerable impact on urban air quality, and noticeably and dangerously so for larger wildfires close to urban centers. The wildfire PM2.5 instrument is a strong predictor of PM2.5, but also captures some of the relationship between wildfires and other criteria pollutants. For each pollutant, I regress the county-monthly average of its station values on the county-monthly average of the wildfire PM2.5 instrument (sampled at the station sites), and I control for county, state-year, and state-month fixed effects, and quadratics of average minimum temperature, maximum temperature, and precipitation. I measure the average contribution by wildfires for each pollutant's concentration in the estimation sample by calculating its partial fitted value $z_{it}\hat{\gamma}^B$, and calculate the percentage of all concentrations of that pollutant attributable to the instrument by dividing by the average measured concentration. These percentages can be interpreted as lower bounds of the amount of each pollutant attributable to wildfires in the CONUS. I repeat this procedure controlling for estimates of NO2, SO2, NH3, and organic (VOC) gases from wildfires and assess how a unit increase in the wildfire instrument predicts downwind concentrations of criteria

gases, organic gases, and PM2.5 subspecies. In another specification, I control for only wildfire NO2 and SO2.

Panel A in Tables 4, 6, and 5 shows the estimated regression coefficients and percentage of average ambient concentrations contributed by wildfires for criteria pollutants, non-metallic PM2.5, and five of the most toxic PM2.5 species. Appendix Tables 16 through 20 repeat this exercise for all other metallic PM2.5 species. Under the assumption that the estimated coefficients reflect purely causal relationships, the maximum of the wildfire percentage of ambient concentration across different control pollutant specifications can be interpreted as an estimated lower bound on the true percentage of ambient concentrations caused by wildfires. Assuming the station sets are representative of the U.S., the instrument predicts nearly 15% of PM2.5 levels and 5% of PM10 levels. Controlling for non-PM2.5 species alters the distribution of pollutants predicted by the instrument, which has significant implications for health effects estimates. Panel B of the pollution regression tables report the estimated coefficients for the regression with controls for other pollutants. The wildfire instrument ceases to be a statistically (and chemically) significant predictor of PM10, while still predicting 15% of PM2.5 mass.

Interpreting hypothesis tests for these estimates as the omitted variables test described in Section 3.3 for IV estimates with no controls for other pollutants, we expect the effect identified by the wildfire PM2.5 instrument to be biased upward by any health effects of non-PM2.5 pollutants that are significantly associated with the PM2.5 instrument. Hence, PM10 and two criteria gases, O3 and NO2, are possible confounders, though the contributions predicted by the instrument for these pollutants are only 4.7%, 3%, and 5.7% of ambient levels. Organic gases are insignificantly predicted, both statistically and in magnitude. Because of sampling error, this test does not rule out that other pollutants with statistically insignificant coefficients may still confound estimates, especially if their 95% confidence interval upper bound is a quantity that could have meaningful health effects. For example, benzene is insignificantly predicted at 6% of total concentrations, but its confidence interval upper bound is 16.2% of benzene, which is arguably a quantity that could have a marginal health impact. Benzene concentrations may only be poorly detected statistically;

short-lived organic gases, such as m-xylene and toluene (8 to 48 hours, Prinn et al. 1987) show neither statistically nor substantively significant effects, while benzene has a comparatively long atmospheric lifetime (2 weeks to 2 months).

Wildfires have the unique property of inducing changes in PM2.5 almost uniformly across both highly and lightly polluted areas. This property is favorable to estimating population-representative effects, since an area's non-wildfire pollution levels drives nonlinear dose response and might also be correlated with effect heterogeneity due to other factors (e.g., highly-polluted areas also have low-income individuals who are more vulnerable to pollution shocks). Figure 4a shows a quantile-quantile plot of all PM2.5 against the estimated implied counterfactual PM2.5 (a world with no wildfire PM2.5), with each point representing the numerical values at which the same quantile occurs in each distribution. The quantile relationships are approximately parallel to the line of distributional equivalence and shifted upward, suggesting that wildfire PM2.5 largely preserves the shape of the distribution of PM2.5 and only shifts the mean. For comparison, Figure 4b shows a comparable quantile-quantile plot when the counterfactual is estimated using the same set of fixed effects and station observations (i.e., a pure panel data approach) instead of fixed effects-IV, revealing a considerably different distribution of margins of change for PM2.5 driven mostly by left- and right-tail behavior.

4.1.2 PM2.5 Chemical Composition Identified by Wildfire Instrument

The types and quantities of PM2.5 predicted by the instrument significantly change when non-PM2.5 controls are included. Section 4.2.2 outlines an argument for how this changes the interpretation of health effects estimates because of changes in the level of toxicity per unit PM2.5. While the total mass of PM2.5 predicted by the instrument only decreases by 10%, the fractions of subspecies groups change significantly. In the non-metallic category, Organic Carbons decrease in concentration by 60-75% per unit wildfire PM2.5, Elemental Carbons by 40-50%, and hydrogen PM2.5 by 70-85%. Bromine PM2.5 increases by 100%, and nitrates by 50%, while the influence of sulfates stays approximately the same. Several metallic PM2.5 species become more strongly

represented per unit of wildfire PM2.5 by at least 50%: Arsenic, Lead, Nickel, Mercury, Cadmium, Barium, Cesium, Cobalt, Gallium, Lanthanum, Selenium, Niobium, and Rubidium. The estimated fraction of atmospheric mercury PM2.5 attributable to wildfires becomes approximately 30 percent, parallel to the fraction established in an inventory of mercury wildfire emissions in the U.S. (Wiedenmyer and Friedli 2007). Predicted arsenic increases by a factor of nearly 30, now accounting for 19 percent of ambient arsenic concentrations. Lead, Nickel, and Cadmium are also all significantly enhanced per unit wildfire PM2.5.

The speciated PM2.5 data present a fairly complete picture of PM2.5 in the U.S. Over 81% of average PM2.5 concentration is accounted for by the subspecies I model. The remaining unexplained PM2.5 concentration may be due to known PM2.5 species which I measure imperfectly or not at all (such as sea salt and dust) and differences in mean concentrations between the PM2.5 Speciation network and general PM2.5 station samples. Moreover, heterogeneous coefficients between testing and estimation samples are not likely to drive most of the results. The number of observations measuring total PM2.5 exceeds the number of observations measuring individual species by 20,000 to 30,000, driven mostly by spatial variation in station coverage. Despite the disparity in spatial sampling, the instrument's estimated effect on total PM2.5 concentration is closely matched by the sum of coefficients for individual PM2.5 species (in the no-controls case, a less than 1% difference). This suggests that any between-sample differences in the relationship between the wildfire instrument and pollutants are mean-zero across PM2.5 subspecies.

Some coefficients for metallic species are negative. The causal interpretation for negative coefficients is that something in the pollutant plume causes a chemical reaction that removes quantities of another species or its precursors (e.g., through oxidation or binding). Many metal PM2.5 species, including mercury, are defined as the metal bound to other airborne particles, such as black carbon (soot). Chemical reactions with wildfire emissions may change such metals back to their gaseous phases, or additional substances may bind to and change the particle to a larger size class. Another possibility is that the relationship is not causal. The PM2.5 instrument is generated using a set of emissions factors for all PM2.5. If there is geographic heterogeneity of subspecies emissions

(e.g., aluminum, silicon, and other metals) that is negatively correlated with the total amount of PM2.5 emitted, high downwind PM2.5 values will also be negatively correlated with those metals. The final possibility is that stations' measurement methods may have some systematic measurement error for subspecies measurements that varies with the amount of other substances in the air.

4.2 Short-term Effects on Mortality

4.2.1 Short-term Effects of PM2.5 on All-Cause Mortality

Panel A of Table 8 reports the 2SLS estimates of the effect of average monthly PM2.5 on monthly all-cause mortality rates using wildfire PM2.5 as an instrument, each column reporting a specification with a different set of fixed effects. Estimates range from 0.67 to 1.05 additional deaths per 100,000 people per monthly $10\mu gm^{-3}$ increase in PM2.5. Panel D reports OLS estimates with the same fixed effects and weather controls as the 2SLS estimates; they are insignificant and sharply estimated close to zero, reflecting the important role of exposure measurement error and omitted variables causing downward bias. Estimated effects using 2SLS increase with the inclusion of more stringent region-specific fixed effects, providing some evidence of region-specific confounders to wildfire PM2.5 such as unobserved seasonal weather factors or endogenous annual policy responses to poor air quality or high wildfire activity. It is also partially explainable by changes in the finite-sample bias of the 2SLS estimator across specifications because of relative changes in the ratio of endogeneity in PM2.5 to the strength of the first-stage relationship (see Appendix A.5); however, confidence intervals based on the inverted Anderson-Rubin test statistic (Anderson and Rubin 1949; Finlay and Magnusson 2009) are very close to the conventional asymptotic confidence intervals, which is evidence against any meaningful bias from weak instruments. Finally, these changes can be attributable to changes in the PM2.5 composition identified by wildfire PM2.5, since different fixed effects may remove certain correspondingly invariant characteristics of wildfire PM2.5. The effect size in column 4 translates to approximately 39,230

premature deaths per year in the U.S. due to monthly exposure to PM2.5, based on the 2010 U.S. population and assuming the sample average PM2.5 of $10.6 \mu g m^{-3}$ is representative of the entire U.S. I find evidence that many of these deaths are driven by forward displacement of mortality within six months in Section 4.2.5.

OLS estimates may be downward-biased because of some combination of correlated unobservables not removed by fixed effects or measurement errors (potentially worsened by fixed effects). The traditional culprits for bias, such as residential sorting, seasonality, and coincidental trends presumably have their influence removed by the stringent fixed effects imposed in each specification. The identifying variation for the OLS estimates remaining is based on within-region, within-year, within-season comparisons, with variation likely to be driven by the totality of incidental variations in PM2.5 emissions and weather patterns. Co-emission would bias estimates upward, as changes in PM2.5 emissions would likely be accompanied by changes in other pollutants. On the other hand, the activities underlying emissions of PM2.5 are likely correlated with several time-varying economic and health behavior processes, including changes in traffic, smoking and drug use, short-term health inputs, physical activity, and stressful events.

More likely is that measurement error plays a significant role in shrinking both OLS and 2SLS estimates toward zero, though the 2SLS estimate corrects this measurement error to the extent that both PM2.5 and the wildfire PM2.5 are characterized by classical measurement error. County-level averages of PM2.5 and the wildfire PM2.5 instrument are calculated from raw averages of measurements at the sites of pollution monitors, which are not always spatially representative. In the traditional errors-in-variables setup, nonzero correlation between the true value of the regressor and the measurement error (i.e., non-classical error) has different implications for bias (expression in Appendix A.4). In the case of negative correlation large enough relative to the signal value of the mismeasured regressor, the coefficient estimate can also reverse sign. Stations tend to be located in more densely populated and plausibly more polluted areas. More densely populated areas have higher pollution but would have their aggregate exposures well-measured by local station observations. Less-densely populated and less-polluted areas will use information on PM2.5 from more

highly-populated areas, resulting in overestimation of PM2.5 levels. The combination of these two factors may result in a negative correlation between the measurement error and PM2.5 levels.

Table 11 reports estimates by age group, revealing that the observed aggregate effects are primarily driven by the three age groups over age 65. Elderly individuals are more likely to be living at vulnerable health margins, and thus are more susceptible to a relatively short-term shock to pollution cause a life-threatening health complication. Also (not reported in tables), the estimated effect is twice as large for women as it is for men. Similarly, Chen et al. (2005) find a higher increased relative risk for females for fatal heart disease and Kunzli et al. (2005) for atherosclerosis from PM2.5 exposure.

4.2.2 Heterogeneous Effects of PM2.5 by Chemical Composition

The inclusion of any of the controls for other pollutants results in a sharp increase in the estimated effect of PM2.5 on all-cause mortality by about two and a half times (Panels B and C, Table 8). In tandem with the distinctive changes in composition across the specifications observed in Section 4.1.2, the increase in mortality estimates with additional controls suggests that PM2.5 has heterogeneous effects that depend on its underlying chemical composition. Because of changes in the toxicological properties of the PM2.5 whose effects are being measured, the interpretation of changes in effect estimates across different identification strategies is potentially ambiguous, even when the regions and emissions sources being studied are identical across estimation methods. In a homogeneous-effects world, a pollutant A's health effect estimate is biased upward by the effect of harmful pollutant B co-emitted from wildfires, implying that including controls for pollutant B would make the estimated effect of pollutant A smaller in expectation. This property does not always hold if there are heterogeneous effects from chemical composition. Specifically, heterogeneous chemical composition may result in some controls removing the statistical influence of some subspecies of pollutant A in favor of more harmful ones. A $1 \mu gm^{-3}$ increase in ambient PM2.5 induced by general wildfire PM2.5 will have a smaller marginal health impact than a $1 \mu gm^{-3}$ increase of wildfire-emitted PM2.5 subspecies with above-average toxicity. Control-

ling for another wildfire pollutant eliminates variation from emissions and atmospheric trajectory components common between the control pollutant and PM2.5 in the identification of the PM2.5 coefficient, resulting in greater weight on identification idiosyncratic to the fuel type at the fire and deposition behavior. For example, Organic Carbons (OC), common byproducts of primary combustion, are a major constituent of wildfire emissions by mass across all wildfire fuel types and would thus have a large part of their influence removed by including any other wildfire pollutant controls due to their commonality to all fires.[3]

Because including controls changes the breakdown of PM2.5 that is identifying the effect to favor relatively more mass from highly toxic species (as demonstrated in Section 4.1.2), we can not unequivocally expect including controls to have a net downward effect on the magnitude of health effects estimates. Hence, the increase in mortality estimates is prima facie evidence of large compositional effects in PM2.5. As shown in Section 4.1.2, the specifications in Panel B and Panel C reflect changes in PM2.5 composed of greater proportions of metals and nitrates than the no-control specifications in Panel A. A common finding in the epidemiological and medical literature is that PM2.5 effects are higher in the presence of metallic PM2.5 subspecies. Bell (2012) finds 15% larger PM2.5 effect estimates for cardiovascular and respiratory morbidity when ambient Nickel (N) is elevated. In a study of rats, Pozzi et al. (2003) find evidence that inflammatory response from particulates is driven by contaminants adsorpted onto particles by comparing inflammatory responses between exposures to urban-sampled particulate matter and pure black carbon.

There are also some shifts in predictions of criteria and organic gases depending on the set of pollutant controls, suggesting a potential role of changing correlations with omitted pollutants driving the increase in mortality effects. However, the loss of PM2.5 mass from carbons and gain from metals is roughly stable across estimates using different pollutant control groups; the changes in estimated mass contributions by the instrument to these gaseous species varies widely with control groups; and estimates are relatively stable across control group sets after the first

[3] Also, the deposition parameters chosen for PM2.5 place relatively more weight on PM2.5 species whose deposition characteristics mimic the chosen parameters most closely. This has ambiguous estimation consequences without further investigation of the distribution of emission deposition characteristics across PM2.5 subspecies.

control pollutant is included. While this analysis is not a substitute for joint IV estimation of all pollutants, this is evidence that most of the increases in effects are driven by a set of PM2.5 species and not from confounding by simultaneous wildfire emissions of criteria and organic gases. Despite attempts to control for O3 production by modeling its key precursor NO2, the instrument predicts approximately 1ppb of O3 per $0.1\mu gm^{-3}$ of PM2.5 predicted by the instrument across specifications. Bell et al. (2004) find a 0.52% increase in daily mortality per 10ppb increase in the previous week's O3; if this effect were true and the base mortality rate is 67.6 deaths per 100,000, then 0.35 of the 1.04 deaths per $10\mu gm^{-3}$ of PM2.5 estimated with the wildfire PM2.5 instrument and no controls are attributable to bias from O3. The estimate for O3-related bias is comparable for the effect with all non-PM2.5 controls, but relatively smaller (0.35 of 2.68 deaths).

4.2.3 Nonlinear Effects of PM2.5

Using a control function approach to estimate nonlinear dose response of short-term mortality, I find that the marginal effect of PM2.5 slightly declines at low concentrations (less than $5\mu gm^{-3}$) and becomes approximately linear. Previous studies using multi-city time series analyses examining short-term PM2.5 dose response have also found a roughly linear relationship below the NAAQS concentration level of $25\mu gm^{-3}$ for all-cause mortality (Schwartz, Laden, and Zanobetti 2002; Stieb et al. 2008); Daniels et al. (2000) additionally finds approximate linearity in PM10 for all-cause mortality. Piecewise regressions for an endogenous variable can be easily estimated via control function methods without the need to develop additional identifying instruments. In the control function procedure, the first-stage regression is identical to the conventional IV first stage, but the residuals from that regression are generated and used as a control variable in a regression of the outcome on the endogenous variable. Results can be made further robust to endogeneity by controlling for corresponding nonlinear functions of the control function residual, accounting for changing correlation with the error term across the support of the endogenous variable. The only other required assumptions are mean independence of the instrument from the structural error and that the distribution of the first stage is correctly specified; in the case of wildfire pollution, the en-

dogenous explanatory variable is continuous and its relationship to the instrument is conceptually linear. Figure 5 is a graph of the fitted values and 95% confidence interval of a spline regression dividing average monthly PM2.5 concentrations into splines by decile (denoted by vertical bars), controlling for the linear control function residual.

4.2.4 Short-Term Effects of PM2.5 by Cause of Death

I estimate effects on mortality rates by broad cause-of-death categories using specification (4) from Table 8, and report the results in Table 9. Unsurprisingly, the fatal effects of PM2.5 manifest most strongly through cardiovascular and respiratory causes, consistent with prior literature. A $10 \mu gm^{-3}$ increase in average monthly PM2.5 is associated with additional deaths from ischemic heart disease (0.26 additional deaths per 100,000), cerebrovascular disorders (0.17), influenza and pneumonia (0.15), and chronic lower respiratory disease (0.19). PM2.5 also has an impact (0.16) on deaths in the ICD-10's broad "Other Diseases" category, suggesting that PM2.5 exposures either lead to complications for already-vulnerable individuals or also cause cardiovascular and respiratory-related deaths for individuals whose cause of death is coded in accordance with the presence of another major health condition.

These wide-ranging effects are supported by the medical literature, which generally finds various undesirable immune system and other bodily responses to fine particulates. Proposed pathophysiological pathways for short-term effects to exposure of PM reviewed in Brook et al. (2010) and Pope et al. (2003) include the production of proinflammatory cytokines that create a systemic inflammatory response affecting bodily areas outside the lungs (also in van Eeden et al. 2001), systemic oxidative stress, changes in coagulation, changes in blood pressure, impaired vascular function, and increased heart rate variability. Brook et al. (2010) cite some conflicting evidence on the effects of particulates on biomarkers for these pathways, likely due to heterogeneity in chemical composition and exposure duration and intensity, but nevertheless reveal a common association between PM2.5 and important biomarkers related elevated risks of cardiovascular and respiratory morbidity. Specific studies have also specifically tied certain types of morbidity to particulate pol-

lution, such as pneumonia (Zeikloff et al. 2002; Zeikloff et al. 2003) and chronic obstructive pulmonary disease (MacNee and Donaldson 2003). There are also research findings which associate subspecies with certain respiratory and cardiovascular health effects. Dye et al. (2001) find pulmonary injury in rats after exposure to PM2.5 subcomponents, with suggestive evidence of the high pulmonary toxicity of metal particulates, while Huang and Ghio (2006) implicate arsenic, mercury, and nickel exposure as causes for anemia, tachycardia, and increased blood pressure.

The inclusion of the wildfire non-PM2.5 pollution controls show the corresponding increases of toxicity of implied changes in PM2.5 across these dominant causes of death. Effects per unit mass PM2.5 on increase by factors of approximately 1.8 for ischemic heart disease and cerebrovascular deaths and 2.3 for chronic lower respiratory deaths, while increasing by a factor of 3 for influenza/pneumonia and other disease-related deaths (though individually remain within sampling error of the no-control effect sizes). Assuming these accurately represent the comparative magnitudes of true effects and that changes in identified PM2.5 composition explain most of the estimated increase in mortality per unit mass, this implies greater toxicity of PM2.5 metals for respiratory and general illnesses relative to cardiovascular illnesses. One explanation is that metals interfere with antimicrobial processes in the lungs, thereby raising the risk and severity of infection. Systemic inflammatory response may also inhibit the body's ability to fight infections outside the lungs.

As a sensitivity check, I estimate whether wildfire-instrumented PM2.5 has an impact on external causes of death (Table 10), with rationale comparable to Heutel and Ruhm (2013): if effect estimates are driven by confounding variation from seasonal or trending factors related to both wildfires and mortality, then external causes of death physiologically unrelated to wildfires might show an effect. I consider 5 outcome groups as classified by the ICD-10: deaths from motor vehicle accidents, accidents, suicides, assaults/homicides, and from "all other external and unspecified causes." Motor vehicle accidents may regardless be affected by wildfires in extreme cases, as wildfires near major roadways can rapidly impede visibility causing massive, multi-vehicle accidents (Collins et al. 2009). ICD-10's "all other external and unspecified" category contains deaths due

to fire exposure and acute smoke inhalation, which would reflect the deaths of firefighters, rural residents, campers, hikers, and other individuals who may be trapped in the vicinity of a wildfire. However, neither of these show any relationship to wildfire smoke, which is some evidence that wildfire pollution exposure is driven by fires distant enough to not have potential direct effects of fire events themselves (e.g., stress caused by imminent danger or property damage). I also find no relationship to suicides and homicides. With no wildfire pollution controls, I find a moderate, marginally statistically significant positive effect on the deaths under the "other unspecified accidents or adverse effects" category, which includes all deaths due to complications related to surgery or medication. This result may be explained by expected increase in the frequency of medical care being administered for increased rates of morbidity due to pollution.

4.2.5 Lagged and Lead Short-Term Associations with PM2.5

Estimating causal associations of air pollution with health outcomes is complicated by a wide range of potential intertemporal relationships between outcome and regressor, both causal and non-causal. There are three reasons to expect lagged pollution values to have negative effects: forward displacement of deaths, depletion of wildfire fuel stocks combined with contemporaneous measurement error, and denominator error in population rates due to annual population measures. In Table 12, I report reduced form estimates of lead, lagged, and both lead and lagged effects of the instrument on all-cause mortality, as well as the joint F-statistic of lead/lagged coefficients. I find evidence of forward displacement and generally violations of the strict exogeneity assumption for fixed effects estimators.

Pollution exposure causes forward displacement of an event if it causes the relocation of an event that otherwise would have occurred to an earlier time period. Schlenker and Walker (2011) argue that welfare impacts of air pollution through morbidity would be be overestimated if forward displacement occurs and is not taken into account (but they test for and find no evidence of forward displacement of hospitalizations). Unless there is a value on postponing a particular outcome, the only negative impact pollution exposure would have on welfare is through events that counterfac-

tually would not have existed if not for the exposure. Since everybody dies[4], welfare effects of pollution-induced mortality can only be measured through the average change in life expectancy. Short-term pollution exposures may primarily only affect those who would otherwise die within a few months, but forward displacement of mortality in this sense is still economically meaningful as long as individuals place positive value on an additional month of life, though one might expect that such value is lower than that of a healthy working individual. The estimates in the second column reveal significant forward displacement.

If wildfire smoke is measured with substantial error, part of the error term of observed pollution is a function of the true level of wildfire pollution, which may in turn be predicted by past (or future) wildfire pollution due to fuel stock dynamics. A large wildfire may burn fuels accumulated over long periods that are not immediately replaced. Wildfires in the near future in the same area are well-situated to affect the same downwind areas as the past large wildfire, but likely to have smaller sizes and shorter durations. In turn, high concentrations in the past predict low concentrations in the present, which would result in lower present mortality.

Lastly, error in the population measure used to calculate mortality rates may cause a lagged negative relationship between mortality and pollution to appear. I measure mortality rates using annual intercensal estimates of population, but measure mortality effects with monthly frequency. Holding changes due to births and migrations fixed, if contemporaneous pollution causes deaths in one month, then the following month's population count is too high, resulting in a measured mortality rate lower than the true rate. The measured rate is hence negatively correlated with the previous month's pollution, generating downward bias in estimates of lagged effects.

Jointly significant lead and lagged effects are interpreted as evidence of violation of the strict exogeneity assumption needed for large-N (number of cross-sectional observations) consistency of fixed effects estimators with a small number of time periods. The inconsistency has bounds shrinking at a rate proportional to the number of time periods (Wooldridge 2010), which in this case is 84 months. In all three specifications I find evidence that strict exogeneity is violated.

[4]I was unable to find a citation for this.

Lagged and lead effects may also be indicative of shocks correlated with the regressor that affect multiple time periods and the outcome variable. In the wildfire setting, this may be weather or climatological variables not adequately captured by temperature, precipitation, and annual and seasonal regional fixed effects.

4.3 Effects on Infant Health

While infants at the most vulnerable health margins may be more likely to die from pollution shocks, the larger population of surviving infants may have their health after birth and subsequent quality of life impacted by *in utero* pollution exposure. Table 13 reports IV estimates for average exposures over the 9 months preceding birth and 4 months preceding birth. Prenatal exposure to PM2.5 has a strong effect on premature births, with effects concentrated in the 4 months leading up to birth. A $10\mu gm^{-3}$ increase in PM2.5 over the gestational period is associated with a 2.6 percentage point increase in the number of premature births and an average decrease in gestational age of 0.23 weeks. There are also negative, but not statistically significant effects on average birth weight, amounting to a $19g$ decrease per $10\mu gm^{-3}$ increase in PM2.5.

As with the mortality outcomes, controlling for NO2 and SO2 strengthens effects, testament to the increased relative toxicity of a unit change in PM2.5; in the final 4 months before birth, a $10\mu gm^{-3}$ increase in PM2.5 lowers average birth weights by $31g$, but there is no significant increase in the likelihood of low birth weight. If the increased toxicity also would result in increased fetal attrition (weakly suggested by the increase in the effect on percentage of female births), then this effect is likely to be occurring for healthier neonates. Alternatively, the effect could be driven by additional growth losses for neonates who regardless of exposure would have been low birth weight. There are at least four classes of physiological mechanisms which may explain the observed negative associations with birth weight: intrauterine growth restriction, fetal genetic or epigenetic changes, pollutant-DNA adducts, and premature birth (Slama et al. 2008). Prematurity may be highly correlated with any of the other mechanisms, or the increased rates of prematurity alone could be driving most of the effect.

The complex interaction of birth timing, overlapping exposures between birth cohorts, and strict exogeneity requirements for fixed effects estimators are possible hazards to identifying meaningful effects of in utero exposure. Because these exposure estimates are framed relative to the birth month, and not the month of conception, substantial harmful effects may be attributable to displacement of unhealthy births from future cohorts into current ones via decreases in gestational age. In the same vein, I expect that displacement due to premature births (and fetal deaths) caused by PM2.5 exposure will cause bias in the opposite direction due to cohort composition effects, as infants with worse health outcomes are deselected from a birth cohort and displaced into earlier cohorts (or completely removed the sample due to fetal death). Exposure timing varies even for births within the same month (by as much as 30 days), resulting in a mixture of true exposure effects estimated in each exposure window. More complicatedly, if the error is not strictly exogenous conditional on the exposure measures and controls, then exposure windows with a mixture of true exposure period and non-exposure periods will reflect a mixture of exposure effects and strict exogeneity violations (i.e., feedback between the dependent variable and lead/lagged values of the regressor).

Attrition from fetal deaths is likely to cause downward bias in the magnitude of these estimates. One key piece of evidence for fetal attrition is the large, albeit imprecisely estimated, effect of average exposure on the sex ratio: each $1 \mu g m^{-3}$ increase in average PM2.5 exposure over nine months before birth raises the percentage of female births by 0.2 percentage points with no non-PM2.5 controls and 0.4 percentage points with NO2 and SO2 controls. This magnitude is comparable to the findings of Sanders and Stoecker (2011) for Total Suspended Particulates (TSPs), which are all particles less than $100 \mu m$. Limited monitor coverage at the time of Clean Air Act makes it impossible to ascertain the effects TSP reductions had on fine particulates. Using rough conversion factors (based on ratios of means in the AQS data) for TSPs to PM10 of 0.55, and PM10 to PM2.5 of 0.6, a one-unit change in TSPs corresponds to a 0.33 unit change in PM2.5, translating the estimate to 0.067 percentage points per unit change in TSP compared to Sanders and Stoecker's (2011) 0.088.

This pattern of results is comparable to Bharadwaj and Eberhard's (2008) estimates of the

effects of PM10 in Santiago, Chile on birth outcomes, but with smaller magnitudes. They estimate a $125g$ effect on birth weight per $17.57\mu gm^{-3}$ (one standard deviation) increase in PM10 pollution 1-16 weeks before birth, whereas I estimate a substantially smaller effect of $32g$ for a comparable change in PM2.5 (again using a conversion factor of $PM10 = 0.6 \times PM2.5$). Besides differences in toxicity between PM10 and PM2.5 (which we regardless might expect to make the difference smaller), this large difference is likely to be driven by some combination of nonlinear effects due to the substantially higher pollution levels in their sample period and the effects of omitted pollutants that also significantly decrease with rainfall. Average PM10 in the U.S. sample period is $18\mu gm^{-3}$ compared to $76\mu gm^{-3}$ in the Santiago sample, and any increasing dose response would be reflected. The rainfall instrument is likely to be strongly associated with decreases in non-PM10 pollutants relative to its association with PM10. While the wildfire instrument does predict some non-PM2.5 pollution levels, this contribution (and thus potential upward bias in estimates' magnitudes) is constrained by the wildfire instrument's dependence on wildfire-specific PM2.5 emissions and PM2.5-specific deposition parameters, compared to the broad and relatively less PM-heavy distribution of pollutants from all industrial sources in or near Santiago. Lastly, because they identify their pollution changes through rainfall, they also identify effects on health outcomes through the associated changes in water pollution generated thru deposition; deposited pollutants run off into water and food supplies and are exposed to individuals through consumption and skin contact. This can bias their estimates either way, depending on whether the pollutants are more harmful after deposition or in the air. In contrast, I control for local rainfall, which will generally account for the aggregate effect of deposited pollutants that could affect health outcomes through the water supply. If deposition occurs in watersheds outside of the area that affect the area's water supply and precipitation differs significantly between the two areas, then the airborne pollutant estimated effects may still pick up effects from associated changes in water supply quality.

5 Wildfire Externalities and Current Management Policy

Wildfires induce significant changes in PM2.5 concentrations over long distances, with polluted air parcels crossing intranational and international boundaries. Assuming that monitoring stations are representative of a state's overall exposure to wildfire pollution, I calculate the fraction of wildfire PM2.5-months that occur outside the state of the wildfire, finding that over 75% of geographic exposure to PM2.5 from large wildfire events in the continental U.S. occurs in states other than the state of origin. Table 7 reports the percentage of modeled wildfire PM2.5 exposure that occurs outside of each state of the wildfire occurrence as an approximation of the intensity of inter-state pollution externalities from wildfires. Because of the implied externalities, wildfire management is subject to the classic tradeoff between inefficient local management behavior and potentially inefficient centralized, uniform policies for environmental goods. To the extent that local jurisdictions in charge of wildfires (e.g., state fire agencies) are individual actors and ignore inter-state pollution spillovers in making fire management decisions, then they will tend to under-suppress wildfire activity or engage in more aggressive prescribed burning for other local benefits. The structure of wildfire management in the U.S. is a complicated mixture of many agencies acting individually and collaborating at multiple levels of government, while the Clean Air Act does not penalize states for pollution from naturally-occurring wildfires. Hence, it is unclear whether current wildfire management efforts properly account for the welfare effects from poor air quality.

While air quality externalities largely make wildfire abatement a national environmental good, it is uncertain whether fire policy would strongly improve with greater centralization. Banzhaf and Chupp (2011) show for the U.S. electricity sector that a uniform federal pollution abatement policy has better welfare implications than decentralized state policies because the inter-state spillovers addressed by a uniform policy are relatively more important than the between-state heterogeneity of benefits addressed by decentralized policies. They argue that relatively inelastic marginal cost of abatement in the relevant region of the uniform policy results in smaller distortions from ignoring between-state heterogeneity of marginal benefits. Wildfires are characterized by large inter-state

spillovers, but the concavity or convexity properties of the marginal costs of abatement are unclear, as are their true marginal damages. Wildfire management has two dimensions of abatement: pre-fire measures, such as prescribed burning and fuel clearing, and suppression efforts. Marginal costs of suppression efforts are relatively easy to measure; for example, Donovan (2006) finds a convex marginal cost function for the number of contract-based firefighting crews hired in a season. Regardless, all abatement measures may have strong heterogeneity and uncertainty in marginal benefits and costs associated with them. Prescribed wildfires themselves generate pollution and some ecological hazards because of their artificial timing (Knapp et al. 2009). Naturally-occurring wildfires have ecological benefits, such as biodiversity and better disease regulation, which may potentially counterbalance the marginal benefits of improved air quality (e.g., Keane and Karau 2010). Even suppression's benefits cannot be well-accounted for, as aggressive suppression can lead to higher likelihood and intensity of future fires by altering the nature of fuel accumulation (Yoder 2004).

Despite federal guidelines governing fire suppression attempts in the interest of protecting public health (Fire Executive Council 2009), the incentives facing the agencies making fire management choices are vague relative to the regulation of agents generating industrial air pollution. The Clean Air Act distinguishes between "unplanned" and "planned" fire, only penalizing states for the pollution generated by planned fire (i.e., prescribed burns), resulting in the adverse health effects of natural wildfires not being inherently taken into account by air quality regulations (Engel and Reeves 2011). There are federal directives and funding for wildfire management, with $3.9 billion allocated for FY2014 (Bracmort 2013). Decision-making regarding suppression and prescribed burning is not federally-determined, however. Currently, fire management in the U.S. predominantly falls upon five federal agencies for fires over 1,000 acres[5] and individual state, county, and local agencies, with frequent interagency collaboration. For the fires in the sample period, 36% were reported by state, county, and local agencies, while the remainder were federally-reported. The Forest Service and and Bureau of Land Management reported the majority of the remaining

[5]These are the the Bureau of Land Management (BLM), Bureau of Indian Affairs (BIA), the U.S. Forest Service (USFS), Fish and Wildfire Service (FWS), and National Park Service (NPS) for 99% of federally-reported fires.

fires. There are ambiguities regarding which agency is responsible for suppression decisions; for example, the agency making the report does not always commit all of the resources tasked with managing the fire, and multiple agencies may report the same fire but only one record is retained in the FPA fire database.

6 Conclusion

This study uses new tools to measure the health externality costs of both industrial and natural sources of air pollution and provides estimates for the effects of fine particulate matter on mortality and infant health. To my knowledge, it is the first to synthesize historical emissions, atmospheric transport models, and ground-level monitoring data at a large scale to estimate the distribution of environmental pollutants and their health effects in the United States. Its design provides spatially and temporally smooth measures of pollution shocks, and the ability to construct a full emissions-to-destination modeling process provides a large degree of customizability and control over the variation used to identify changes in air quality. The choice of wildfires as emissions source results in geographically wide-reaching variation in particulate levels, inducing both small and large shocks to highly polluted and relatively unpolluted areas. The findings of effects on short-term mortality and infant health contribute to the body of evidence supporting that PM2.5, and generally air quality, has important impacts on human health. They also highlight the importance of fire management as an important public health issue.

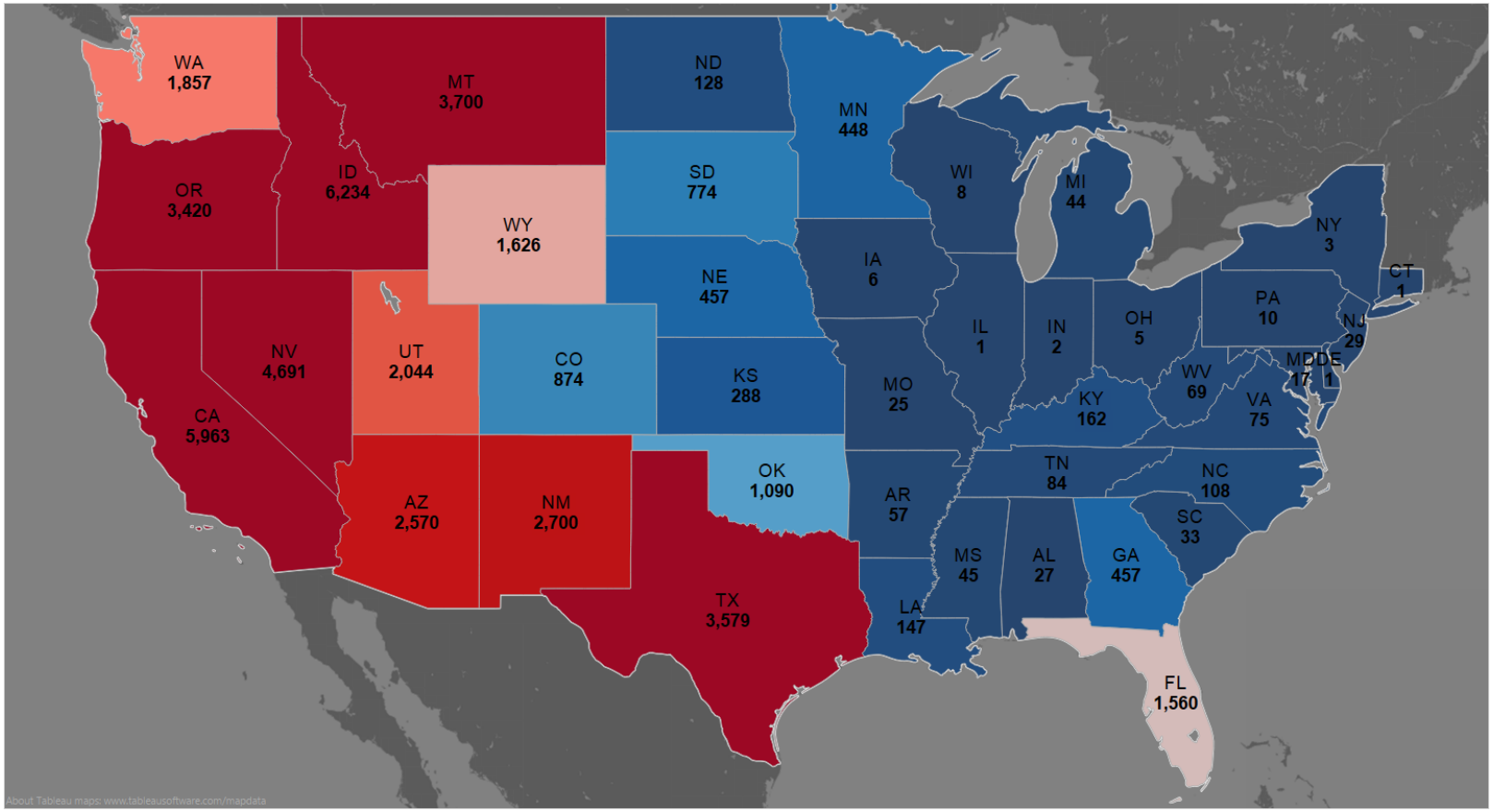
As might be expected with a new source of data, there are several statistical issues which must be addressed to fully realize the potential of wildfires to identify useful, policy-relevant health effects parameters. Incorrect exposure measurements in both space and time create potentially serious measurement error problems which are only partially alleviated by instrumental variables techniques. Imperfect monitoring coverage results in measurement error of exposures both within and between geographic units. Spatial measurement error can be alleviated through more comprehensive measures of ambient pollution, generated through a combination of interpolation of data points, remote sensing data, and two-sample instrumental variables estimation techniques

(Khawand 2014). Two-sample IV techniques can also be used to include geographic regions with no monitoring coverage in estimating health effects, resulting in estimated average effects more representative of the U.S. population. In the short run, this study can be improved upon through developing richer model inputs from higher-quality data products that require substantially greater computational input to implement. Satellite products for fire detection allow wildfire burn dynamics to be better parsed out in space and time. Higher-resolution meteorological products can be used to better capture short-range dispersion patterns, which in turn require more intensive geographic sampling schemes to properly translate to aggregate concentrations.

The modeling of wildfires' air quality impact itself also stands to be significantly improved. The relationship between the wildfire pollution forecasts and actual pollution levels, while intuitively seeming to be relatively uncomplicated, is subject to situation-specific measurement errors due to the complex interaction among fire, fuel, and meteorological data inputs and modeling assumptions. Modeling errors may occur due to unmodeled heterogeneity at the source or between the source and the destination. A richer exploration of heterogeneous source-receptor relationships is needed to understand where modeling errors may result in putting undue weight on health effects in certain areas or discarding useful variation in others. Extensive further work, particularly in collaboration with scientists in the wildfire community, is required to improve the realism and predictive power of the wildfire pollution simulation.

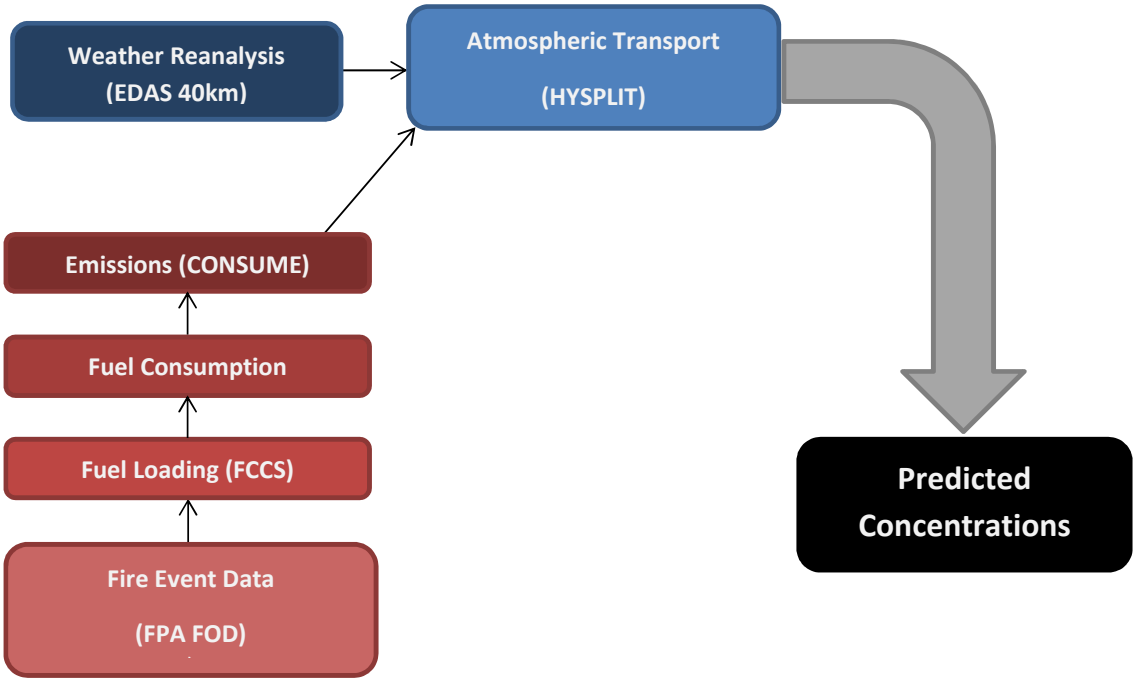
APPENDIX

Figure 1: Number of Acres Burned (Thousands) for All Fires Greater than 1,000 Acres, 2000-2010



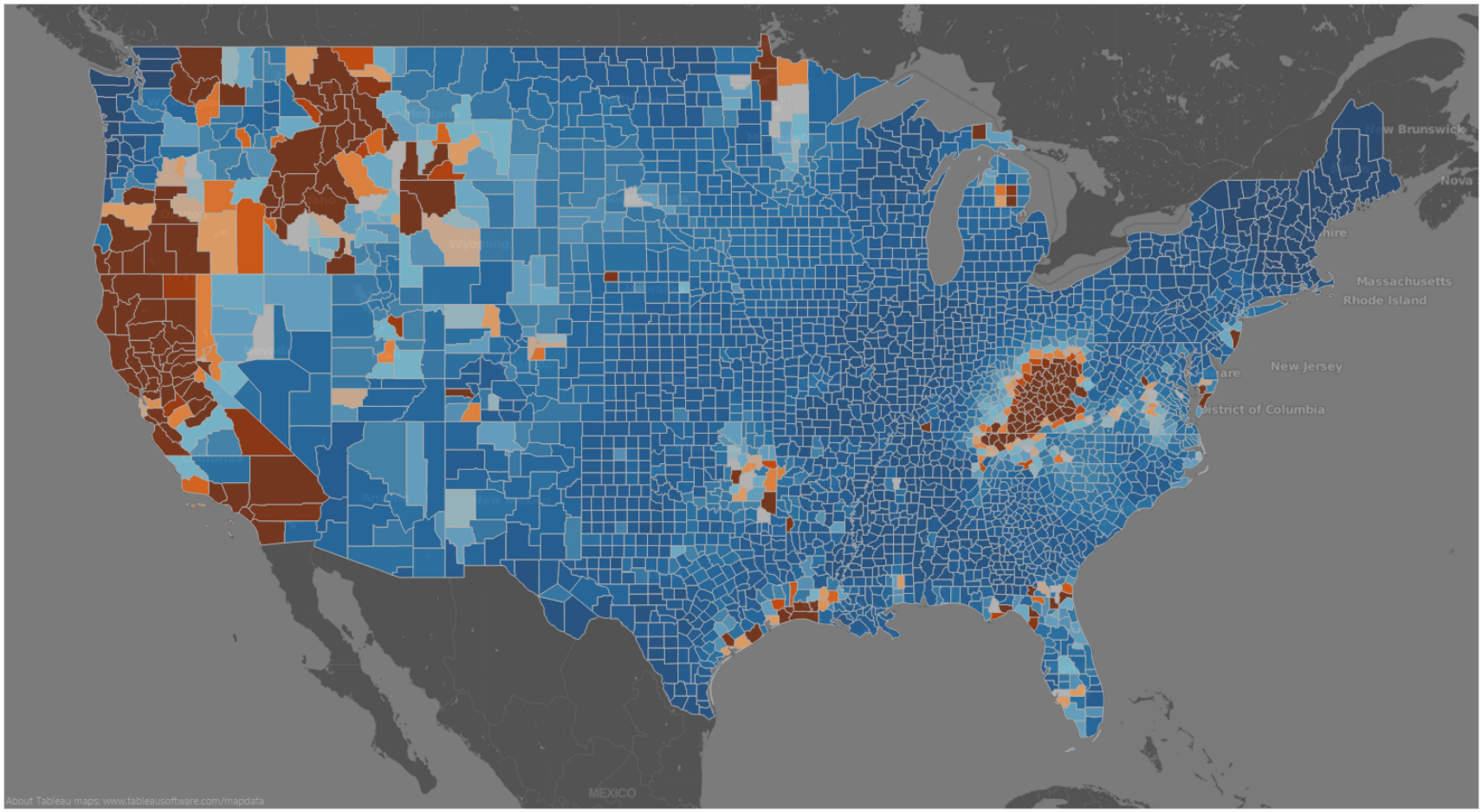
Map shows the number of acres (in thousands) for all 1,000 acre or greater fires in the US from 2000 to 2010 by state, ranging from red (most area burned) to blue (least area burned).

Figure 2: Wildfire Air Pollution Modeling - BlueSky Framework Workflow



Flow chart depicting the modeling workflow to produce pollution concentration outputs from ingestion of fire data to output by the HYSPLIT model.]

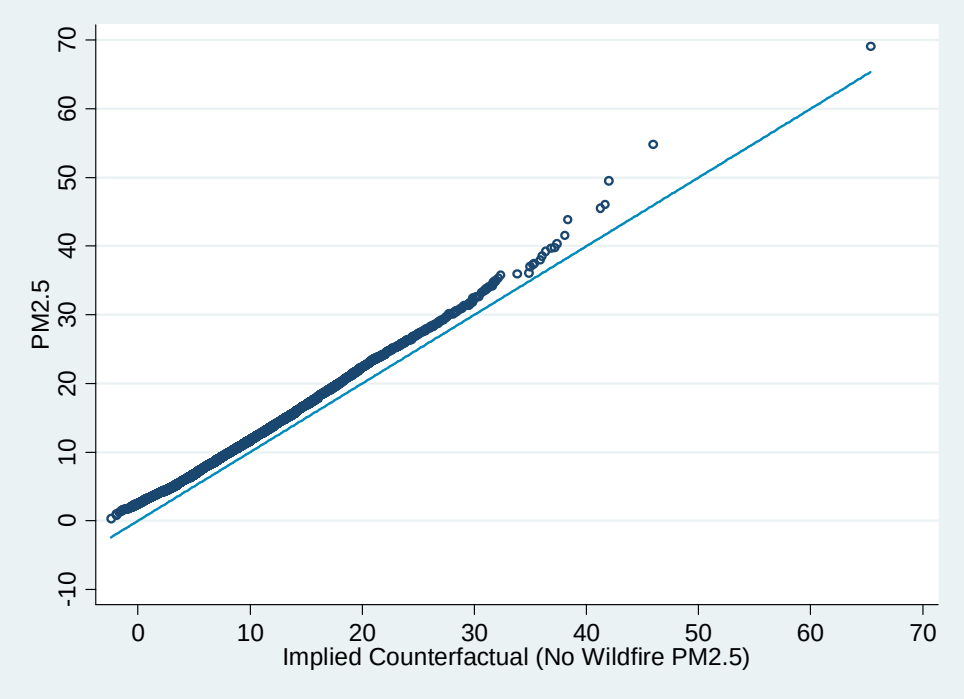
Figure 3: Average Raw Wildfire PM2.5 Output by County, CONUS, 2004-2010



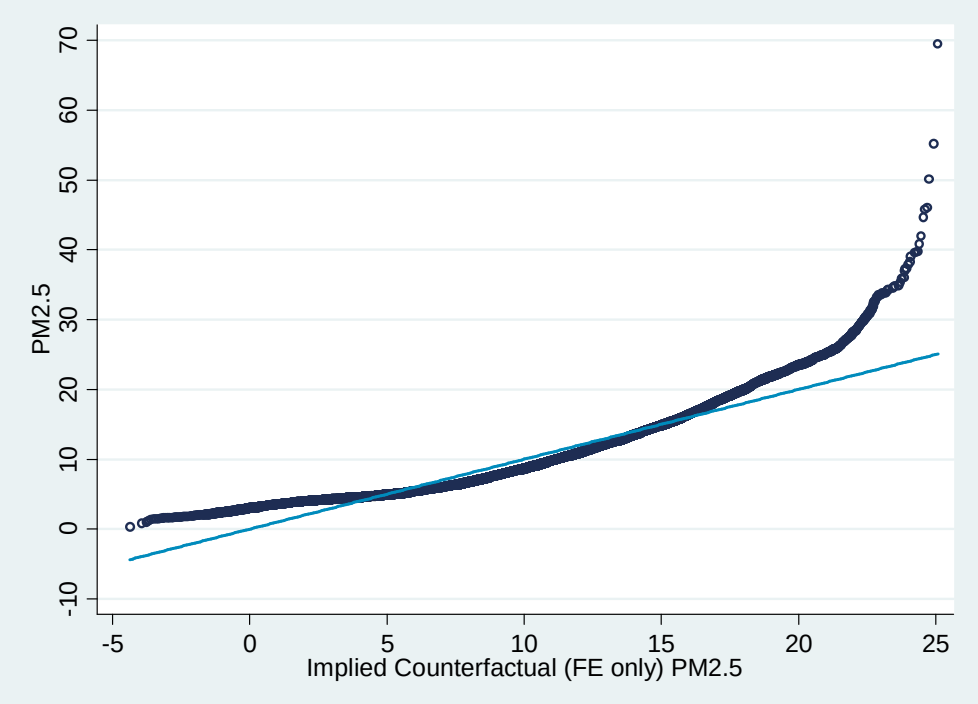
Map of untransformed average PM2.5 concentrations by U.S. county for 2004-2010 sample period. Dark blue values represent low concentrations and brown values high concentrations (e.g, California has high concentrations, while Maine has low concentrations).

Figure 4: Quantile-Quantile Plots of PM2.5 versus Counterfactuals

(a) PM2.5 (with Wildfire PM2.5) versus Estimated Counterfactual PM2.5 (No Wildfire PM2.5)

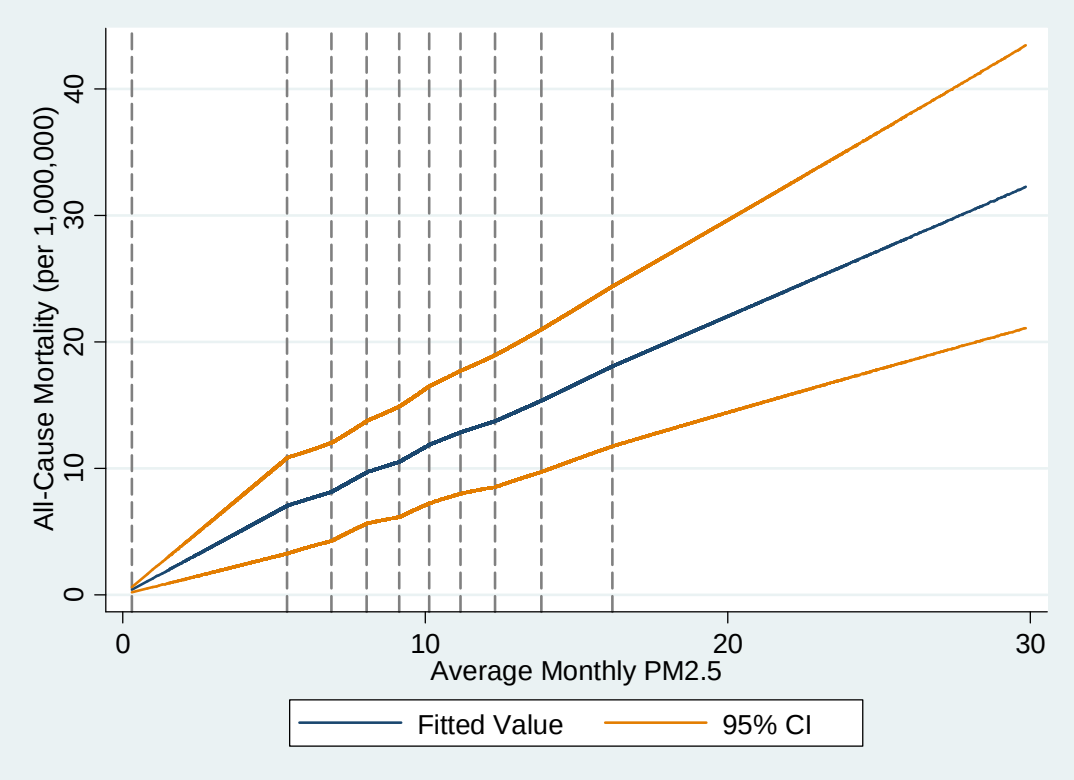


(b) PM2.5 versus Counterfactual PM2.5 Estimated with Fixed Effects



Sub-figure A plots PM2.5 against the counterfactual estimated using the wildfire PM2.5 instrument; sub-figure B plots it against the counterfactual as estimated using state-year, state-month, and county fixed effects. Each point on the plot represents the values in each distribution at which the quantiles are equivalent.

Figure 5: Spline Control Function Regression of All-Cause Mortality on PM2.5, by Decile



This is a plot of the estimated effect of average monthly PM2.5 estimated by splines in deciles of average monthly PM2.5 conditional on a linear control function residual using wildfire PM2.5 as the excluded instrument.

Table 1: AQS PM2.5 and NLDAS Weather Descriptive Statistics

	Mean	Std. Dev	5th Pctile.	Median	95th Pctile.	N
Average Daily Pollution and Weather Measures						
PM2.5 Concentration (ug/m3)	10.60	4.37	4.50	10.13	18.47	37,259
Precipitation (mm)	2.72	2.04	0.25	2.35	6.45	171,014
Maximum Air Temperature (F)	63.82	19.43	29.13	65.96	91.02	171,014
Minimum Air Temperature (F)	46.09	17.23	16.88	46.49	72.48	171,014

County-level descriptive statistics for 2004-2010 across all U.S. counties. PM2.5 concentration is only available for county-months with monitoring data.

Table 2: Monthly, County-Level Mortality Rate (per 100,000) by Subgroup from U.S. Death Certificates, 2004-2010

	Mean	Std. Dev	5th Pctile.	Median	95th Pctile.	N
All individuals	78.58	33.62	36.38	73.73	136.38	171,182
< 1 year	54.67	222.98	0.00	0.00	278.55	170,782
01 to 04	1.56	18.77	0.00	0.00	2.15	170,787
05 to 14	0.69	6.50	0.00	0.00	1.52	170,796
15 to 24	1.71	10.65	0.00	0.00	7.56	170,855
25 to 34	4.18	18.92	0.00	0.00	22.82	170,836
35 to 44	11.52	27.14	0.00	0.00	50.48	170,837
45 to 54	31.70	41.41	0.00	24.25	98.14	170,847
55 to 64	73.93	70.76	0.00	64.44	188.95	170,834
65 to 74	173.19	128.88	0.00	159.49	386.10	170,820
75 to 84	431.85	255.33	0.00	410.98	833.33	170,831
85+	1247.40	747.52	0.00	1173.70	2395.20	170,823
All Ages, Male	76.67	42.35	24.85	70.17	147.49	171,075
All Ages, Female	79.28	43.25	26.26	72.74	152.65	170,917
County Population	110,000	350,000	3,394	28,494	490,000	171,182

County-level mortality rates calculated using U.S. death certificate date for 2004-2010. All figures are scaled per 100,000 county population.

Table 3: Monthly, County-Level Mean Birth Outcomes and Rates for Birth Cohorts from U.S. Birth Certificates, 2004-2010

	Mean	Std. Dev	5th Pctile.	Median	95th Pctile.
All Births					
# Births	113	391	3	28	476
Avg. Birth Weight (g)	3268	197	2966	3274	3556
% Male	48.8%	15.1%	25.0%	49.1%	71.4%
% Low Birth Weight	8.1%	8.3%	0.0%	7.1%	21.4%
APGAR 0-3	0.5%	2.3%	0.0%	0.0%	2.9%
APGAR 4-6	1.4%	3.9%	0.0%	0.0%	6.7%
APGAR 7-8	13.0%	14.4%	0.0%	9.4%	40.0%
APGAR 9-10	82.7%	19.2%	46.0%	88.0%	100.0%
APGAR Unknown	2.4%	13.7%	0.0%	0.0%	3.8%
Preterm Birth	12.6%	10.4%	0.0%	11.7%	29.2%
Full-Term Birth	87.3%	10.5%	70.5%	88.1%	100.0%
Gest. Age Unknown	0.2%	1.3%	0.0%	0.0%	0.4%
N = 259,471					
Full-Term Births Only (< 37 wks.)					
# Births	98	339	2	24	417
Avg. Birth Weight (g)	3368	167	3108	3371	3619
% Male	49.1%	15.9%	25.0%	49.8%	75.0%
% Low Birth Weight	3.2%	5.5%	0.0%	1.8%	11.9%
APGAR 0-3	0.2%	1.6%	0.0%	0.0%	0.9%
APGAR 4-6	1.0%	3.5%	0.0%	0.0%	5.0%
APGAR 7-8	11.5%	14.5%	0.0%	7.7%	40.0%
APGAR 9-10	85.0%	19.4%	50.0%	90.6%	100.0%
APGAR Unknown	2.3%	13.7%	0.0%	0.0%	3.1%
N = 258,748					
Pre-Term Births Only (≥ 37 wks.)					
# Births	14	48	0	3	60
Avg. Birth Weight (g)	2570	481	1778	2574	3323
% Male	46.4%	29.2%	0.0%	48.5%	100.0%
% Low Birth Weight	41.3%	29.4%	0.0%	42.3%	100.0%
APGAR 0-3	2.8%	9.8%	0.0%	0.0%	16.7%
APGAR 4-6	4.2%	12.1%	0.0%	0.0%	25.0%
APGAR 7-8	22.3%	25.8%	0.0%	16.7%	100.0%
APGAR 9-10	68.1%	30.0%	0.0%	74.1%	100.0%
APGAR Unknown	2.6%	14.5%	0.0%	0.0%	4.7%
N = 216,452					

County-level birth outcome descriptive statistics derived from U.S. birth certificate data for 2004-2010.

Table 4: First Stage Regression of PM2.5 and Regressions of Criteria Pollutants on Wildfire Instrument

	Fine Particulate (PM2.5)	**Coarse Particulate (PM10)**	**Carbon Monoxide (CO)**	**Sulfur Dioxide (SO2)**	**Nitric Oxide (NO)**	**Nitrogen Dioxide (NO2)**	**Ozone (O3)**
				Panel A: OLS			
Coefficient	**1.1e-01*** **	**5.5e-02** **	-1.2e-04	2.9e-03	-2.2e-02	**2.2e-02*** **	**1.1e-04*** **
	(1.1e-02)	(2.4e-02)	(3.2e-04)	(3.4e-03)	(1.5e-02)	(5.8e-03)	(1.1e-05)
% Wildfire	15.3%	4.7%	-0.4%	1.5%	-4.8%	3.0%	5.7%
95% CI Upper	*18.5%*	*8.7%*	*1.6%*	*4.8%*	*1.6%*	*4.6%*	*6.9%*
95% CI Lower	*12.2%*	*0.7%*	*-2.4%*	*-1.9%*	*-11.2%*	*1.4%*	*4.6%*
				Panel B: OLS with Wildfire NO2, SO2 Controls			
Coefficient	**1.1e-01*** **	4.7e-02	6.1e-04	4.0e-03	-1.4e-02	5.5e-03	**9.2e-05*** **
	(1.2e-02)	(2.9e-02)	(3.7e-04)	(4.9e-03)	(2.1e-02)	(8.0e-03)	(1.8e-05)
% Wildfire	16.1%	4.0%	2.0%	2.0%	-3.1%	0.8%	4.8%
95% CI Upper	*19.5%*	*9.0%*	*4.3%*	*6.8%*	*5.8%*	*2.9%*	*6.6%*
95% CI Lower	*12.8%*	*-0.9%*	*-0.4%*	*-2.8%*	*-12.1%*	*-1.4%*	*3.0%*
				Panel C: OLS with Wildfire NO2, SO2, NH3, VOC Controls			
Coefficient	**1.0e-01*** **	3.3e-02	5.6e-04	4.6e-03	8.5e-03	3.8e-03	**9.7e-05*** **
	(1.1e-02)	(2.7e-02)	(3.7e-04)	(4.8e-03)	(1.8e-02)	(8.6e-03)	(1.9e-05)
% Wildfire	14.5%	2.8%	1.8%	2.3%	1.8%	0.5%	5.1%
95% CI Upper	*17.5%*	*7.3%*	*4.1%*	*7.1%*	*9.6%*	*2.9%*	*7.0%*
95% CI Lower	*11.5%*	*-1.7%*	*-0.5%*	*-2.4%*	*-6.0%*	*-1.8%*	*3.1%*
Mean Conc.	1.10E+01	1.80E+01	4.60E-01	2.90E+00	7.00E+00	1.10E+01	3.00E-02
N	36752	14706	11955	14719	10665	13119	26063

Coefficients are for a 1µgm-3 change in PM2.5. Units are ppb for SO2, NO, and NO2 and ppm for O3 and CO. "% Wildfire" is calculated as the overall quantity of pollutant predicted by the instrument divided by the mean concentration times 100%. Standard errors clustered at state-year level are in parentheses. Significance stars represent p < 0.1 (*), p < .05 (**), p < .01 (***).

Table 5: Regressions of Highly Toxic PM2.5 Subspecies on Wildfire PM2.5

	Arsenic	**Mercury**	**Lead**	**Nickel**	**Cadmium**
			Panel A: OLS		
Coefficient	2.1e-07	**-1.6e-05*** **	**2.0e-05** **	4.7e-06	9.5e-06
	(1.5e-06)	(4.1e-06)	(8.6e-06)	(3.2e-06)	(6.5e-06)
% Wildfire	0.4%	-24.0%	10.6%	6.7%	8.4%
95% CI Upper	*6.8%*	*-12.1%*	*19.7%*	*15.5%*	*19.8%*
95% CI Lower	*-5.9%*	*-35.9%*	*1.5%*	*-2.2%*	*-3.0%*
			Panel B: OLS with Wildfire NO2, SO2 Controls		
Coefficient	**8.7e-06*** **	**2.4e-05*** **	**2.6e-05***	8.5e-06	**3.8e-05*** **
	(2.3e-06)	(6.3e-06)	(1.4e-05)	(6.3e-06)	(9.4e-06)
% Wildfire	18.6%	35.5%	14.1%	12.1%	33.9%
95% CI Upper	*28.0%*	*53.6%*	*28.6%*	*29.8%*	*50.2%*
95% CI Lower	*9.1%*	*17.4%*	*-0.5%*	*-5.6%*	*17.6%*
			Panel C: OLS with Wildfire NO2, SO2, NH3, VOC Controls		
Coefficient	**9.1e-06*** **	**2.2e-05*** **	**3.3e-05** **	**1.1e-05***	**3.5e-05*** **
	(2.4e-06)	(6.2e-06)	(1.3e-05)	(6.3e-06)	(9.1e-06)
% Wildfire	19.4%	31.7%	17.9%	16.3%	30.7%
95% CI Upper	*29.3%*	*49.6%*	*32.2%*	*33.9%*	*46.5%*
95% CI Lower	*9.6%*	*13.9%*	*3.6%*	*-1.2%*	*14.9%*
Mean Concentration	7.00E-04	1.10E-03	2.80E-03	1.00E-03	1.70E-03
N	15,566	8,300	15,624	15,624	10,439

Coefficients are for a 1-unit change in the wildfire PM2.5 instrument. Units are in µgm-3 for all PM2.5. "% Wildfire" is calculated as the overall quantity of pollutant predicted by the instrument divided by the mean concentration times 100%. Standard errors clustered at state-year level are in parentheses. Significance stars represent $p < 0.1$ (*), $p < .05$ (**), $p < .01$ (***).

Table 6: Regressions of Non-Metallic PM2.5 Subspecies on Wildfire PM2.5

	Organic Carbon (OC)	**Elemental Carbon (EC)**	**Hydrogen**	**Chloride**	**Bromine**	**Sulfur**	**Nitrite**	**Soil**	**Sulfate**	**Nitrate**
					Panel A: OLS					
Coefficient	**2.3e-02*** **	**3.0e-03*** **	**2.9e-03*** **	-2.6e-04	**1.9e-05*** **	**9.7e-03*** **	**1.1e-04** **	-2.3e-04	**2.7e-02*** **	**2.1e-02*** **
	(4.8e-03)	(8.4e-04)	(4.9e-04)	(4.8e-04)	(3.7e-06)	(1.3e-03)	(4.5e-05)	(1.8e-03)	(3.8e-03)	(3.1e-03)
% Wildfire	25.3%	13.7%	15.9%	-4.9%	10.4%	17.9%	10.8%	-0.5%	17.4%	28.2%
95% CI Upper	*35.9%*	*21.2%*	*21.2%*	*12.9%*	*14.4%*	*22.5%*	*19.2%*	*6.7%*	*22.1%*	*36.4%*
95% CI Lower	*14.8%*	*6.3%*	*10.6%*	*-22.8%*	*6.4%*	*13.3%*	*2.4%*	*-7.7%*	*12.6%*	*20.1%*
					Panel B: OLS with Wildfire NO2, SO2 Controls					
Coefficient	5.6e-03	1.5e-03	6.7e-04	-1.4e-03	**3.8e-05*** **	**1.1e-02*** **	4.4e-05	8.7e-04	**3.2e-02*** **	**3.2e-02*** **
	(4.7e-03)	(1.7e-03)	(6.7e-04)	(9.8e-04)	(6.1e-06)	(1.5e-03)	(4.4e-05)	(2.3e-03)	(4.6e-03)	(4.5e-03)
% Wildfire	6.3%	6.8%	3.7%	-26.8%	20.8%	20.0%	4.2%	1.8%	20.6%	42.4%
95% CI Upper	*16.7%*	*21.8%*	*10.9%*	*9.7%*	*27.4%*	*25.6%*	*12.5%*	*11.2%*	*26.2%*	*54.2%*
95% CI Lower	*-4.1%*	*-8.2%*	*-3.6%*	*-63.2%*	*14.2%*	*14.3%*	*-4.1%*	*-7.7%*	*14.9%*	*30.6%*
					Panel C: OLS with Wildfire NO2, SO2, NH3, VOC Controls					
Coefficient	**9.6e-03***	1.7e-03	4.1e-04	-9.9e-04	**3.6e-05*** **	**8.6e-03*** **	5.7e-05	2.1e-03	**2.5e-02*** **	**3.2e-02*** **
	(5.3e-03)	(1.5e-03)	(6.7e-04)	(9.7e-04)	(5.4e-06)	(1.7e-03)	(4.6e-05)	(2.5e-03)	(5.0e-03)	(4.1e-03)
% Wildfire	10.7%	7.7%	2.3%	-18.8%	19.9%	15.9%	5.4%	4.3%	15.9%	42.6%
95% CI Upper	*22.3%*	*20.6%*	*9.5%*	*17.5%*	*25.7%*	*22.1%*	*14.1%*	*14.2%*	*22.2%*	*53.4%*
95% CI Lower	*-0.8%*	*-5.3%*	*-5.0%*	*-55.0%*	*14.1%*	*9.7%*	*-3.3%*	*-5.6%*	*9.7%*	*31.7%*
Mean Concentration	1.30E+00	3.20E-01	2.60E-01	7.60E-02	2.70E-03	8.00E-01	1.50E-02	7.10E-01	2.30E+00	1.10E+00
N	6,477	6,469	6,281	6,359	15,481	15,561	6,299	6,281	15,628	15,378

Coefficients are for a 1-unit change in the wildfire PM2.5 instrument. Units are in µgm-3 for all PM2.5. "% Wildfire" is calculated as the overall quantity of pollutant predicted by the instrument divided by the mean concentration times 100%. Standard errors clustered at state-year level are in parentheses. Significance stars represent p < 0.1 (*), p < .05 (**), p < .01 (***).

Table 7: Percentage of Wildfire PM2.5 Exposure Outside of the State of Origin

AL	AR	AZ	CA	CO	CT	DE	FL	GA	IA	ID	IL	IN	KS	KY	LA
72.2%	72.1%	77.9%	80.3%	56.9%	7.3%	--	83.4%	86.7%	--	74.1%	83.6%	74.5%	92.5%	82.0%	73.7%
MA	**MD**	**ME**	**MI**	**MN**	**MO**	**MS**	**MT**	**NC**	**ND**	**NE**	**NH**	**NJ**	**NM**	**NV**	**NY**
--	96.5%	--	84.8%	94.3%	69.8%	58.5%	75.7%	86.1%	75.1%	90.6%	--	65.8%	69.8%	73.7%	68.0%
OH	**OK**	**OR**	**PA**	**RI**	**SC**	**SD**	**TN**	**TX**	**UT**	**VA**	**VT**	**WA**	**WI**	**WV**	**WY**
80.4%	61.7%	72.0%	78.4%	--	58.0%	50.2%	75.2%	92.3%	65.2%	78.6%	--	76.2%	--	--	54.5%

Each cell represents the fraction of raw average wildfire PM2.5 unit-months that occurs within the wildfire's state of origin. Empty cells indicate states with no wildfires larger than 1,000 acres in the sample period.

Table 8: IV Estimates: PM2.5 Effects on All-Cause Mortality (by Fixed-Effects Specification)

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: 2SLS						
Avg. PM2.5 ($10\mu gm^{-3}$)	**0.671*** **	**0.806*** **	**0.881*** **	**1.041*** **	**1.049*** **	**0.926*** **
	(0.121)	(0.126)	(0.123)	(0.199)	(0.125)	(0.129)
First-Stage F-Statistic	67.791	80.118	79.523	79.078	82.994	76.378
First-Stage Partial R^2	0.03	0.025	0.025	0.024	0.025	0.028
Panel B: 2SLS - Wildfire NO2, SO2 Controls						
Avg. PM2.5 ($10\mu gm^{-3}$)	**1.590** **	**1.965*** **	**2.067*** **	**2.350*** **	**2.419*** **	**2.875*** **
	-0.701	(0.670)	(0.695)	(0.737)	(0.736)	(1.015)
First-Stage F-Statistic	10.102	17.334	16.737	19.645	20.799	13.569
First-Stage Partial R^2	0.003	0.004	0.004	0.004	0.004	0.003
Panel C: 2SLS -Wildfire NO2, SO2, NH3, VOC Controls						
Avg. PM2.5 ($10\mu gm^{-3}$)	**1.791** **	**2.269*** **	**2.365*** **	**2.680*** **	**2.779*** **	**3.175*** **
	(0.780)	(0.802)	(0.823)	(0.870)	(0.874)	(1.145)
First-Stage F-Statistic	9.123	14.695	14.556	17.441	18.459	12.713
First-Stage Partial R^2	0.003	0.003	0.003	0.003	0.003	0.003
Panel D: OLS						
Avg. PM2.5 ($10\mu gm^{-3}$)	***-0.043****	-0.017	-0.022	-0.013	-0.009	0.001
	(0.024)	(0.025)	(0.025)	(0.024)	(0.025)	(0.026)
Fixed Effects						
Year	Y	Y	--	--	--	--
Month	Y	--	--	--	--	--
County	Y	Y	Y	Y	--	--
County-Month	N	N	N	N	N	Y
County-Year	N	N	N	N	Y	N
Climate Region-Month	N	Y	Y	--	--	--
State-Year	N	N	Y	Y	--	Y
State-Month	N	N	N	Y	Y	--

N = 36,752. Coefficients are effects for mortality rate per 100,000 population for a 10µgm-3 change in PM2.5. Standard errors clustered at state-year level are in parentheses. Significance stars represent p < 0.1 (*), p < .05 (**), p < .01 (***).

Table 9: IV Estimates: PM2.5 Effects on Mortality (by Cause)

All-Cause	All-Cause (ln(rate))	Ischemic Heart Disease	Other Heart Disease	Cerebrovascular	Influenza & Pneumonia	Chronic Lower Respiratory	ICD-10 "All Other (Residual)"
			Panel A: 2SLS				
1.041* **	**0.013*** **	**0.258*** **	0.066	**0.166*** **	**0.146*** **	**0.194*** **	**0.163** **
(0.233)	(0.003)	(0.073)	(0.041)	(0.043)	(0.044)	(0.048)	(0.070)
			Panel B: 2SLS with Wildfire NO2, SO2, NH3, VOC Controls				
2.680* **	**0.033*** **	**0.472** **	**0.257***	**0.292** **	**0.458*** **	**0.454*** **	**0.529** **
(0.885)	(0.012)	(0.225)	(0.133)	(0.134)	(0.167)	(0.169)	(0.230)
			Panel C: OLS				
-0.013	0.000	-0.010	0.007	***-0.015*** **	***-0.009*** **	-0.003	-0.003
(0.024)	(0.000)	(0.011)	(0.007)	(0.006)	(0.004)	(0.006)	(0.009)
			Outcome Means (monthly, per 100,000)				
67.64	4.16	11.97	5.53	4.20	1.67	4.20	11.92

N = 36,752

Coefficients are effects for mortality rate per 100,000 population for a 10μgm-3 change in PM2.5. Standard errors clustered at state-year level are in parentheses. Significance stars represent $p < 0.1$ (*), $p < .05$ (**), $p < .01$ (***).

Table 10: IV Estimates: PM2.5 (Non-)Effects on Mortality from External Causes

Motor Vehicle Accidents	**Other Unspecified & Adverse Effects**	**Homicides**	**Suicides**	**Other External Causes**
		Panel A: 2SLS		
0.020	**0.083****	-0.031	0.001	0.008
(0.026)	(0.040)	(0.021)	(0.009)	(0.007)
		Panel B: 2SLS with Wildfire NO2, SO2, NH3, VOC Controls		
0.067	0.060	-0.063	-0.007	0.006
(0.076)	(0.111)	(0.060)	(0.024)	(0.020)
		Panel C: OLS		
-0.001	-0.003	0.007**	-0.000	-0.000
(0.004)	(0.005)	(0.003)	(0.002)	(0.001)
		Outcome Means (monthly, per 100,000)		
1.35	2.53	1.10	0.42	0.18

N = 36,752

Coefficients are effects for mortality rate per 100,000 population for a 10μgm^{-3} change in PM2.5. Standard errors clustered at state-year level are in parentheses. Significance stars represent $p < 0.1$ (*), $p < .05$ (**), $p < .01$ (***).

Table 11: IV Estimates: PM2.5 Effect on All-Cause Mortality by Age Group

01 to 04	**05 to 14**	**15 to 24**	**25 to 34**	**35 to 44**
		Panel A: 2SLS		
-0.028	-0.043	0.011	***-0.230****	0.032
(0.088)	(0.037)	(0.072)	(0.124)	(0.148)
		Panel C: 2SLS -Wildfire NO2, SO2, NH3, VOC Controls		
-0.014	-0.057	0.069	-0.395	0.042
(0.241)	(0.102)	(0.199)	(0.344)	(0.408)
45 to 54	**55 to 64**	**65 to 74**	**75 to 84**	**85+**
		Panel A: 2SLS		
0.103	0.584	**3.140*** **	**7.009*** **	**27.027*** **
(0.227)	(0.419)	(0.907)	(1.956)	(6.978)
		Panel C: 2SLS -Wildfire NO2, SO2, NH3, VOC Controls		
0.275	1.615	**6.042** **	**14.892** **	**78.988*** **
(0.626)	(1.223)	(2.795)	(6.181)	(26.583)

Coefficients are effects for mortality rate per 100,000 population for a 10µgm-3 change in PM2.5. Standard errors clustered at state-year level are in parentheses. Significance stars represent $p < 0.1$ (*), $p < .05$ (**), $p < .01$ (***).

Table 12: Reduced Form Lead and Lagged Wildfire PM2.5 Effect on All-Cause Mortality

	(1)	(2)	(3)
6 Month Lead		-0.007	0.011
		(0.01)	(0.02)
5 Month Lead		0.009	**-0.035****
		(0.019)	(0.016)
4 Month Lead		**0.046****	**0.046****
		(0.019)	(0.023)
3 Month Lead		0.016	0.008
		(0.018)	(0.022)
2 Month Lead		**-0.035***	-0.009
		(0.019)	(0.019)
1 Month Lead		0.015	-0.018
		(0.019)	(0.023)
Contemp.	**0.077*****	**0.046*****	**0.036***
	(0.018)	(0.018)	(0.022)
1 Month Lag	-0.022		**-0.043****
	(0.016)		(0.018)
2 Month Lag	-0.008		0.016
	(0.016)		(0.019)
3 Month Lag	**-0.049*****		-0.03
	(0.016)		(0.019)
4 Month Lag	-0.021		-0.031
	(0.019)		(0.021)
5 Month Lag	**-0.052*****		**-0.062*****
	(0.019)		(0.022)
6 Month Lag	0.009		0.036
	(0.018)		(0.023)
Joint F-test Leads/Lags (p-value)	0.00015	0.00055	0.00028
N	30,355	30,394	24,096

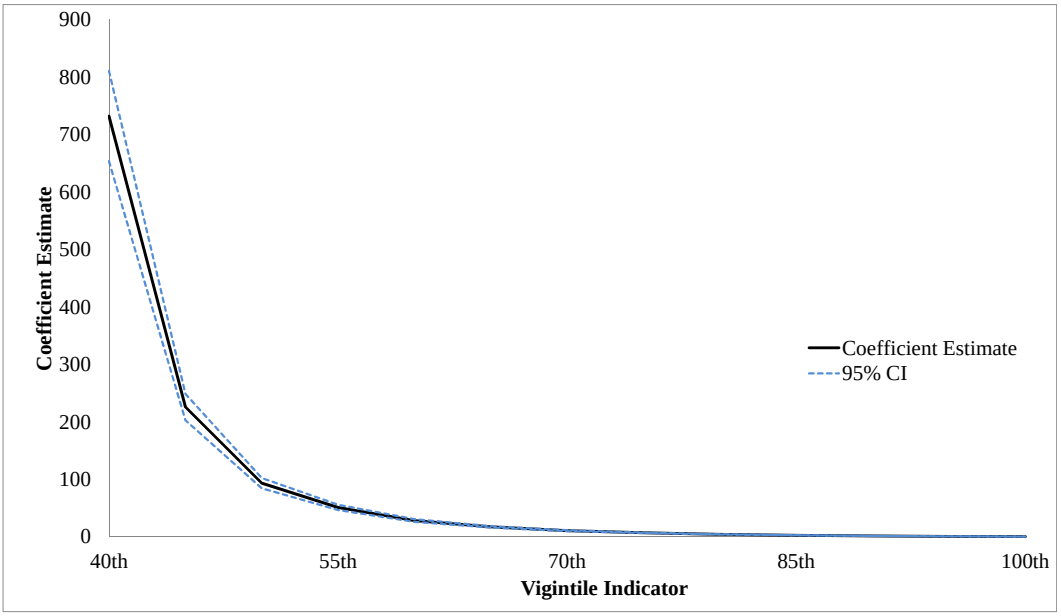
Table 13: IV Estimates: Effect of PM2.5 Exposure for Full Gestation and 16 Weeks Before Birth on Birth Outcomes

	% Female	**Gestational Age (Weeks)**	**% Premature (< 37 Weeks GA)**	**Avg. Birth Weight (g)**	**Low Birth Weight (< 2500g)**	**Low Apgar (< 5)**
			Panel A: 2SLS			
Avg. PM2.5 (9 mo. Before Birth)	**0.00226***	**-0.0228*****	**0.00257*****	-1.9235	0.0090	-0.0003
	(0.00116)	(0.0077)	(0.00083)	(1.7574)	(0.0070)	(0.0005)
1-16 Weeks Before Birth	0.00117	**-0.0238*****	**0.00195*****	-1.7355	0.0031	0.0000
	(0.00086)	(0.0055)	(0.00064)	(1.1570)	(0.0046)	(0.0003)
FS F-stat F = 156.45, Partial R2 = 0.0977, N = 43,585						
			Panel B: 2SLS - Wildfire NO2, SO2 Controls			
Avg. PM2.5 (9 mo. Before Birth)	0.00397	**-0.0464*****	**0.00388****	-4.4164	0.0006	0.0001
	(0.00244)	(0.0164)	(0.00177)	(3.6309)	(0.0015)	(0.0010)
1-16 Weeks Before Birth	0.00107	**-0.039*****	**0.00297*****	**-3.1059****	0.00001	0.0002
	(0.00112)	(0.0078)	(0.00088)	(1.4327)	(0.0006)	(0.0004)

FS F-stat = 37.68, Partial R2 = 0.023, N = 43,585. PM2.5 is denominated in 1ugm-3 and averaged over the specified period. The instrument is wildfire PM2.5 averaged over the same period, and all controls are averaged over the same period.

Figure 6: Piecewise Regression Coefficient Estimates of Daily Station PM2.5 on Raw and Log-transformed Wildfire PM2.5 Model Output, by Vigintile

(a) Raw Wildfire PM2.5 Output



(b) Log-Transformed Wildfire PM2.5 Output

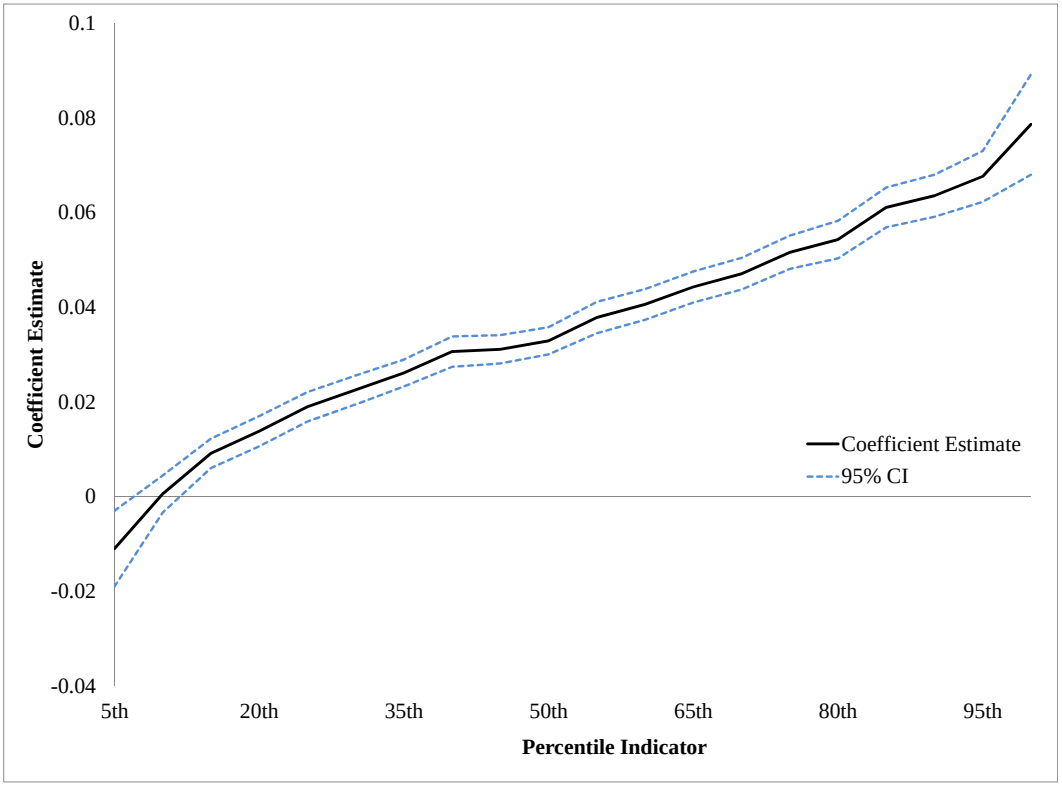


Table 14: Henry's Law Constants and Dry Deposition Velocities for Gaseous Pollutants

Pollutant	Henry's Law Constant	Citation	Dry Deposition Velocity	Citation
Ozone (O3)	9.40E-03		1.25E-02	
Sulfur Dioxide (SO2)	**1.23E+01**		1.50E-02	
Nitrogen Dioxide (NO2	**1.20E-02**		3.60E-03	
Nitric Oxide (NO)	**1.90E-03**		0.00E+00	
Carbon Monoxide (CO)	9.40E-04		3.00E-04	
Methane (CH4)	1.40E-03			
Carbon Dioxide (CO2)	3.40E-02			
Ammonia (NH3)	**6.10E+01**		6.50E-03	
Formaldehyde (HCHO)	3.20E+03		5.00E-03	
Mercury Elemental (Gas	9.30E-02		1.00E-04	
Mercury Reactive Gaseo	1.40E+06		1.00E-03	
Toluene (C6H5CH3)	**1.50E-01**		**0.00E+00**	
Benzene	1.60E-01			
O-Xylene	1.30E-01			
M-Xylene	1.90E-01			
P-Xylene	1.30E-01			

Table 15: Regression of Organic Gases on Wildfire PM2.5

	M/P Xylene	Benzene	Toluene	Ethylbenzene	O-Xylene	Styrene
			Panel A: OLS			
Coefficient	-2.4e-04	8.7e-03	1.5e-02	8.5e-04	1.6e-03	-5.5e-03
	(5.1e-03)	(6.5e-03)	(1.7e-02)	(2.6e-03)	(2.8e-03)	(3.8e-03)
% Wildfire	-0.2%	6.6%	4.6%	1.8%	2.9%	-18.8%
95% CI Upper	7.1%	16.2%	15.2%	12.5%	13.0%	7.2%
95% CI Lower	-7.5%	-3.0%	-6.0%	-9.0%	-7.3%	-44.9%
			Panel B: OLS with Wildfire NO2, SO2 Controls			
Coefficient	1.1e-02	1.3e-02	6.0e-03	-2.0e-03	-4.3e-03	-2.0e-03
	(7.4e-03)	(8.1e-03)	(4.0e-02)	(6.4e-03)	(7.3e-03)	(4.3e-03)
% Wildfire	8.1%	9.6%	1.9%	-4.3%	-7.9%	-7.0%
95% CI Upper	18.7%	21.5%	26.9%	22.1%	18.6%	22.1%
95% CI Lower	-2.5%	-2.4%	-23.1%	-30.7%	-34.5%	-36.1%
			Panel C: OLS with Wildfire NO2, SO2, NH3, VOC Controls			
Coefficient	1.0e-02	**1.7e-02****	1.9e-02	1.7e-03	-6.3e-05	-2.3e-03
	(7.7e-03)	(7.7e-03)	(3.1e-02)	(4.8e-03)	(5.2e-03)	(5.1e-03)
% Wildfire	7.6%	12.9%	6.0%	3.6%	-0.1%	-7.8%
95% CI Upper	18.6%	24.4%	25.3%	23.4%	18.9%	26.8%
95% CI Lower	-3.5%	1.5%	-13.3%	-16.2%	-19.1%	-42.4%
Mean Concentration	2.00E+00	1.90E+00	4.60E+00	6.90E-01	7.90E-01	4.30E-01
N	8044	8924	8653	8580	8364	7761

	Chloroform	Carbon Tetrachloride	Methyl Chloroform	Tetrachloro-ethylene	Trichloro-ethylene	Dichloro-Trifluoroethane
			Panel A: OLS			
Coefficient	-8.6e-05	-1.8e-04	-2.6e-04	-2.4e-04	-5.6e-05	-5.6e-05
	(7.6e-05)	(1.3e-04)	(5.1e-04)	(2.8e-04)	(4.9e-04)	(4.9e-04)
% Wildfire	-4.9%	-2.9%	-6.4%	-4.5%	-2.2%	-2.2%
95% CI Upper	3.6%	1.3%	17.8%	5.8%	36.5%	36.5%
95% CI Lower	-13.3%	-7.1%	-30.6%	-14.7%	-41.0%	-41.0%
			Panel B: OLS with Wildfire NO2, SO2 Controls			
Coefficient	-1.8e-04	-9.6e-05	-4.2e-04	**-1.0e-03***	8.5e-04	8.5e-04
	(1.5e-04)	(1.4e-04)	(8.7e-04)	(5.9e-04)	(6.4e-04)	(6.4e-04)
% Wildfire	-10.5%	-1.5%	-10.1%	-18.7%	34.2%	34.2%
95% CI Upper	6.1%	2.9%	31.2%	3.1%	84.9%	84.9%
95% CI Lower	-27.0%	-6.0%	-51.4%	-40.6%	-16.4%	-16.4%
			Panel C: OLS with Wildfire NO2, SO2, NH3, VOC Controls			
Coefficient	-1.9e-04	**-2.4e-04***	-7.6e-04	-6.3e-04	9.5e-04	9.5e-04
	(1.6e-04)	(1.5e-04)	(1.1e-03)	(6.0e-04)	(7.6e-04)	(7.6e-04)
% Wildfire	-11.0%	-3.8%	-18.3%	-11.9%	38.2%	38.2%
95% CI Upper	6.6%	0.7%	34.1%	10.3%	98.9%	98.9%
95% CI Lower	-28.6%	-8.4%	-70.7%	-34.2%	-22.5%	-22.5%
Mean Concentration	2.50E-02	8.80E-02	6.00E-02	7.60E-02	3.50E-02	3.50E-02
N	8086	7687	7569	8131	8090	8090

Coefficients are for a one-unit change in the wildfire PM2.5 instrument. Units for organic gases are ppbC (parts per billion carbon). "% Wildfire" is calculated as the overall quantity of pollutant predicted by the instrument divided by the mean concentration times 100%. Standard errors clustered at state-year level are in parentheses. Significance stars represent $p < 0.1$ (*), $p < .05$ (**), $p < .01$ (***).

Table 16: Regression of PM2.5 Metal Subspecies on Wildfire PM2.5, Set I

	Aluminum	**Ammonium Ion**	**Ammonium Nitrate**	**Ammonium Sulfate**	**Antimony**	**Barium**	**Calcium**	**Cerium**	**Cesium**	**Chlorine**
Coefficient	**-4.6e-04*** **	**2.1e-02*** **	**9.8e-03*** **	**1.8e-02*** **	-2.4e-05	**7.7e-05*** **	4.2e-05	**6.4e-05** **	**6.6e-05*** **	-1.1e-04
	(1.2e-04)	(2.6e-03)	(2.7e-03)	(3.7e-03)	(1.9e-05)	(2.5e-05)	(8.8e-05)	(2.7e-05)	(1.7e-05)	(1.9e-04)
% Wildfire	-15.9%	24.0%	23.1%	12.3%	-2.3%	9.0%	1.2%	7.0%	9.0%	-5.0%
95% CI Upper	*-7.5%*	*29.9%*	*35.7%*	*17.3%*	*1.3%*	*14.8%*	*6.1%*	*12.7%*	*13.7%*	*11.1%*
95% CI Lower	*-24.3%*	*18.2%*	*10.6%*	*7.3%*	*-5.9%*	*3.2%*	*-3.7%*	*1.2%*	*4.3%*	*-21.1%*
Coefficient	**-6.0e-04*** **	**2.8e-02*** **	**2.0e-02*** **	**2.1e-02*** **	**-5.3e-05***	**1.2e-04*** **	1.1e-04	**2.2e-04*** **	**5.5e-05** **	-5.0e-04
	(1.5e-04)	(3.5e-03)	(5.1e-03)	(6.3e-03)	(2.8e-05)	(4.1e-05)	(1.2e-04)	(3.9e-05)	(2.6e-05)	(3.7e-04)
% Wildfire	-20.8%	32.2%	0.4806919	14.3%	-5.1%	13.6%	3.2%	23.8%	7.4%	-21.8%
95% CI Upper	*-10.5%*	*40.2%*	*0.7209205*	*22.8%*	*0.3%*	*22.9%*	*10.2%*	*32.2%*	*14.4%*	*9.7%*
95% CI Lower	*-31.1%*	*24.2%*	*0.2404632*	*5.8%*	*-10.6%*	*4.3%*	*-3.8%*	*15.4%*	*0.4%*	*-53.4%*
Coefficient	**-5.6e-04*** **	**2.4e-02*** **	**1.8e-02*** **	**1.5e-02** **	-4.0e-05	**1.2e-04*** **	1.2e-04	**2.3e-04*** **	**6.5e-05** **	-3.8e-04
	(1.6e-04)	(3.2e-03)	(4.6e-03)	(6.4e-03)	(3.0e-05)	(4.0e-05)	(1.2e-04)	(4.0e-05)	(3.0e-05)	(3.4e-04)
% Wildfire	-19.3%	27.5%	43.4%	10.2%	-3.9%	14.5%	3.4%	25.4%	8.8%	-16.5%
95% CI Upper	*-8.6%*	*34.7%*	*64.7%*	*18.9%*	*1.8%*	*23.6%*	*10.4%*	*34.0%*	*16.8%*	*12.9%*
95% CI Lower	*-30.1%*	*20.2%*	*22.0%*	*1.5%*	*-9.5%*	*5.5%*	*-3.6%*	*16.8%*	*0.8%*	*-45.9%*
Mean Concentration	4.30E-02	1.30E+00	6.10E-01	2.10E+00	1.60E-02	1.30E-02	5.20E-02	1.40E-02	1.10E-02	3.40E-02
N	15529	10899	6299	6295	10859	10719	15516	10273	10432	15561

Table 17: Regression of PM2.5 Metal Subspecies on Wildfire PM2.5, Set II

	Chromium	**Chromium Vi**	**Cobalt**	**Copper**	**Europium**	**Gallium**	**Gold**	**Hafnium**	**Indium**
Coefficient	-6.1e-06	**6.6e-05*** **	**1.6e-06*** **	5.8e-06	-1.1e-05	2.6e-06	2.0e-06	**4.2e-05*** **	**-1.3e-05** **
	(1.0e-05)	(1.7e-05)	(5.7e-07)	(8.3e-06)	(1.5e-05)	(1.6e-06)	(2.6e-06)	(1.1e-05)	(6.0e-06)
% Wildfire	-5.5%	9.0%	3.2%	2.5%	-3.6%	2.7%	1.3%	8.4%	-2.5%
95% CI Upper	*12.6%*	*13.7%*	*5.3%*	*9.5%*	*5.5%*	*6.1%*	*4.6%*	*12.6%*	*-0.2%*
95% CI Lower	*-23.6%*	*4.3%*	*1.0%*	*-4.5%*	*-12.7%*	*-0.7%*	*-2.0%*	*4.2%*	*-4.7%*
Coefficient	2.5e-06	**5.5e-05** **	**2.8e-06** **	5.4e-06	9.1e-06	-6.4e-07	**-7.3e-06** **	-1.6e-05	-8.6e-06
	(1.8e-05)	(2.6e-05)	(1.2e-06)	(1.6e-05)	(2.6e-05)	(2.3e-06)	(3.4e-06)	(1.2e-05)	(7.8e-06)
% Wildfire	2.2%	7.4%	5.5%	2.3%	2.9%	-0.7%	-4.7%	-3.2%	-1.6%
95% CI Upper	*34.3%*	*14.4%*	*9.9%*	*16.1%*	*19.2%*	*4.0%*	*-0.4%*	*1.6%*	*1.3%*
95% CI Lower	*-29.8%*	*0.4%*	*1.0%*	*-11.5%*	*-13.4%*	*-5.3%*	*-9.1%*	*-7.9%*	*-4.5%*
Coefficient	7.5e-06	**6.5e-05** **	**3.7e-06*** **	5.0e-06	-6.2e-06	**4.4e-06***	-8.3e-07	7.2e-06	-6.9e-06
	(1.8e-05)	(3.0e-05)	(1.2e-06)	(2.0e-05)	(2.9e-05)	(2.6e-06)	(3.9e-06)	(1.2e-05)	(8.0e-06)
% Wildfire	6.7%	8.8%	7.1%	2.1%	-2.0%	4.6%	-0.5%	1.4%	-1.3%
95% CI Upper	*37.9%*	*16.8%*	*11.6%*	*18.7%*	*15.9%*	*9.9%*	*4.3%*	*6.3%*	*1.7%*
95% CI Lower	*-24.4%*	*0.8%*	*2.5%*	*-14.5%*	*-19.9%*	*-0.8%*	*-5.4%*	*-3.4%*	*-4.3%*
Mean Concentration	1.70E-03	1.10E-02	7.80E-04	3.50E-03	4.80E-03	1.50E-03	2.40E-03	7.70E-03	7.90E-03
N	15579	10432	10769	15561	7850	7896	7896	7850	10319

Table 18: Regression of PM2.5 Metal Subspecies on Wildfire PM2.5, Set III

	Iridium	**Iron**	***Lanthanum***	**Magnesium**	**Manganese**	**Molybdenum**	**Niobium**	**Potassium**	**Potassium Ion**
Coefficient	-9.1e-07	-2.0e-04	**1.0e-04*** **	**-1.4e-04*** **	5.2e-06	4.3e-06	**2.5e-06***	**6.4e-04*** **	**5.7e-04*** **
	(3.0e-06)	(1.3e-04)	*(2.5e-05)*	(3.4e-05)	(1.7e-05)	(3.9e-06)	(1.4e-06)	(1.4e-04)	(2.0e-04)
% Wildfire	-0.5%	-4.2%	*11.8%*	-13.4%	2.8%	2.1%	2.0%	15.1%	13.9%
95% CI Upper	*2.7%*	*1.1%*	*17.6%*	*-7.0%*	*20.8%*	*5.8%*	*4.2%*	*21.7%*	*23.6%*
95% CI Lower	*-3.7%*	*-9.4%*	*5.9%*	*-19.8%*	*-15.3%*	*-1.6%*	*-0.2%*	*8.4%*	*4.2%*
Coefficient	**-1.5e-05*** **	-2.8e-05	**2.2e-04*** **	-7.2e-05	6.9e-06	**-2.1e-05*** **	1.9e-06	**5.5e-04** **	5.0e-04
	(4.1e-06)	(2.0e-04)	*(3.6e-05)*	(5.3e-05)	(3.4e-05)	(6.7e-06)	(1.9e-06)	(2.2e-04)	(3.3e-04)
% Wildfire	-8.3%	-0.6%	*26.2%*	-6.9%	3.7%	-10.4%	1.5%	12.9%	12.1%
95% CI Upper	*-3.9%*	*7.6%*	*34.7%*	*3.1%*	*39.3%*	*-4.0%*	*4.5%*	*23.1%*	*27.6%*
95% CI Lower	*-12.8%*	*-8.7%*	*17.8%*	*-17.0%*	*-32.0%*	*-16.8%*	*-1.6%*	*2.7%*	*-3.4%*
Coefficient	-3.4e-06	4.9e-05	**2.4e-04*** **	**-9.0e-05***	3.1e-05	-1.0e-05	**5.1e-06** **	**7.0e-04*** **	**7.2e-04** **
	(4.7e-06)	(2.2e-04)	*(3.9e-05)*	(5.3e-05)	(3.7e-05)	(7.8e-06)	(2.1e-06)	(2.4e-04)	(3.4e-04)
% Wildfire	-1.9%	1.0%	*28.2%*	-8.6%	16.8%	-5.0%	4.0%	16.5%	17.5%
95% CI Upper	*3.2%*	*10.0%*	*37.3%*	*1.4%*	*56.0%*	*2.4%*	*7.3%*	*27.4%*	*33.9%*
95% CI Lower	*-7.0%*	*-7.9%*	*19.1%*	*-18.7%*	*-22.4%*	*-12.5%*	*0.7%*	*5.5%*	*1.2%*
Mean Concentration	2.80E-03	7.20E-02	*1.30E-02*	1.50E-02	2.80E-03	3.20E-03	1.90E-03	6.40E-02	6.20E-02
N	7850	15574	*7896*	15081	15611	8378	7850	15592	10628

Table 19: Regression of PM2.5 Metal Subspecies on Wildfire PM2.5, Set IV

	Rubidium	**Samarium**	**Scandium**	**Selenium**	**Silicon**	**Silver**	**Sodium**	**Sodium Ion**	**Strontium**
Coefficient	**1.2e-06****	-1.1e-05	**-5.7e-05*****	**4.6e-06*****	**-6.7e-04****	**-1.5e-05*****	-7.2e-05	**6.4e-04*****	1.4e-07
	(5.7e-07)	(1.1e-05)	(9.5e-06)	(1.6e-06)	(2.7e-04)	(5.1e-06)	(1.5e-04)	(2.1e-04)	(2.3e-06)
% Wildfire	2.6%	-4.3%	-14.1%	6.4%	-8.2%	-3.7%	-1.4%	7.8%	0.2%
95% CI Upper	*4.9%*	*4.8%*	*-9.5%*	*10.7%*	*-1.8%*	*-1.3%*	*4.1%*	*12.8%*	*5.7%*
95% CI Lower	*0.2%*	*-13.5%*	*-18.7%*	*2.1%*	*-14.7%*	*-6.1%*	*-6.9%*	*2.7%*	*-5.3%*
Coefficient	**4.2e-06*****	-5.1e-06	**-1.3e-04*****	**1.1e-05*****	**-7.6e-04****	9.8e-06	1.4e-04	3.4e-04	**-6.7e-06***
	(7.3e-07)	(2.0e-05)	(1.5e-05)	(2.6e-06)	(3.1e-04)	(6.3e-06)	(2.7e-04)	(2.9e-04)	(3.6e-06)
% Wildfire	8.8%	-2.1%	-31.2%	14.8%	-9.3%	2.4%	2.6%	4.1%	-8.2%
95% CI Upper	*11.8%*	*14.2%*	*-23.9%*	*21.8%*	*-1.8%*	*5.4%*	*12.9%*	*11.1%*	*0.3%*
95% CI Lower	*5.8%*	*-18.3%*	*-38.5%*	*7.8%*	*-16.8%*	*-0.6%*	*-7.7%*	*-2.9%*	*-16.8%*
Coefficient	**4.7e-06*****	-9.0e-06	**-1.1e-04*****	**1.1e-05*****	**-6.8e-04****	9.3e-06	2.3e-04	4.3e-04	-5.2e-06
	(7.9e-07)	(2.2e-05)	(1.6e-05)	(2.7e-06)	(3.3e-04)	(6.8e-06)	(3.0e-04)	(3.1e-04)	(3.9e-06)
% Wildfire	9.8%	-3.7%	-26.3%	15.6%	-8.4%	2.2%	4.4%	5.3%	-6.4%
95% CI Upper	*13.1%*	*14.1%*	*-18.6%*	*23.1%*	*-0.4%*	*5.5%*	*15.6%*	*12.8%*	*2.9%*
95% CI Lower	*6.6%*	*-21.5%*	*-34.1%*	*8.0%*	*-16.3%*	*-1.0%*	*-6.8%*	*-2.2%*	*-15.7%*
Mean Concentration	7.10E-04	3.80E-03	6.30E-03	1.10E-03	1.20E-01	6.20E-03	7.80E-02	1.20E-01	1.20E-03
N	15561	7850	7863	15611	15561	10319	15112	10697	15561

Table 20: Regression of PM2.5 Metal Subspecies on Wildfire PM2.5, Set VI

	Tantalum	**Terbium**	**Tin**	**Titanium**	**Tungsten**	**Vanadium**	**Yttrium**	**Zinc**	**Zirconium**
Coefficient	**1.6e-05****	-8.7e-06	3.8e-06	**-2.3e-05****	**1.2e-05****	2.3e-06	**2.5e-06***	-1.6e-06	**-4.2e-06****
	(6.3e-06)	(1.8e-05)	(6.5e-06)	(9.1e-06)	(5.6e-06)	(2.6e-06)	(1.4e-06)	(3.2e-05)	(2.1e-06)
% Wildfire	4.7%	-2.2%	0.5%	-8.2%	4.5%	1.9%	2.8%	-0.2%	-3.9%
95% CI Upper	*8.5%*	*6.9%*	*2.3%*	*-1.8%*	*8.8%*	*6.4%*	*5.9%*	*9.0%*	*-0.2%*
95% CI Lower	*1.0%*	*-11.4%*	*-1.3%*	*-14.6%*	*0.3%*	*-2.5%*	*-0.2%*	*-9.5%*	*-7.6%*
Coefficient	8.9e-06	6.7e-06	-1.7e-05	-1.1e-05	-1.1e-07	3.0e-06	**4.6e-06*****	5.4e-05	**-8.6e-06****
	(7.5e-06)	(3.1e-05)	(1.2e-05)	(1.2e-05)	(7.1e-06)	(4.8e-06)	(1.7e-06)	(4.0e-05)	(3.4e-06)
% Wildfire	2.7%	1.7%	-2.4%	-3.8%	0.0%	2.6%	5.2%	8.1%	-7.9%
95% CI Upper	*7.2%*	*17.4%*	*1.1%*	*4.5%*	*5.3%*	*10.7%*	*9.0%*	*19.8%*	*-1.7%*
95% CI Lower	*-1.8%*	*-14.0%*	*-5.8%*	*-12.1%*	*-5.4%*	*-5.5%*	*1.3%*	*-3.7%*	*-14.0%*
Coefficient	**2.7e-05*****	4.4e-06	-1.4e-05	-5.7e-06	1.3e-05	8.6e-07	**6.5e-06*****	6.4e-05	**-6.6e-06***
	(8.8e-06)	(3.3e-05)	(1.3e-05)	(1.2e-05)	(7.7e-06)	(4.6e-06)	(2.0e-06)	(4.1e-05)	(3.5e-06)
% Wildfire	8.1%	1.1%	-1.9%	-2.0%	4.9%	0.7%	7.3%	9.5%	-6.0%
95% CI Upper	*13.3%*	*17.9%*	*1.7%*	*6.3%*	*10.8%*	*8.4%*	*11.6%*	*21.6%*	*0.3%*
95% CI Lower	*2.8%*	*-15.6%*	*-5.5%*	*-10.4%*	*-1.0%*	*-6.9%*	*2.9%*	*-2.5%*	*-12.3%*
Mean Concentration	5.10E-03	6.00E-03	1.10E-02	4.20E-03	4.00E-03	1.70E-03	1.40E-03	1.00E-02	1.60E-03
N	7850	7850	10809	15561	7850	15574	8470	15574	15161

A.1 Estimating Fire Burn Durations

Wildfires can last for a period of hours to hundreds of days (for large, remote complex fires). The best measure in the FPA database of a fire's start time is the discovery time by the reporting agency, which is almost always reported. The time of the fire's containment, which indicates a judgment by the fire managing agency that the fire perimeter is secured from spreading further, is reported with similar frequency. Only some of the FPA database sources also have reports of their fires' extinguishment dates. Substantial emissions may still occur during the period between containment and extinguishment, especially for large fires. For fires greater than 300 acres, approximately 43% of burn time is post-containment. To better calibrate the time profile of emissions from fires, I use these fire events to fit a model and predict the burn duration for all fires in the absence of an explicitly-reported extinguishment or "put-out" time.

I merge fire extinguishment dates from the DOI-USGS database of fire reports from six major federal agencies. Then, estimate a linear model of a fire's burn duration:

$$D_i = c_i\xi + s(i)\theta_s + m(i)\theta_m + y(i)\theta_y + r_i$$

A fire's burn duration is a function of its time to containment D_i; its final land-area size, measured by a categorical "size class" c_i; some unobserved seasonal-, year-, and state-specific factors; and idiosyncratic factors r_i. I estimate this relationship using all fires from 2000-2010 larger than 300 acres. The containment time is naturally a strong predictor, as it is the earliest a fire can be extinguished. Its coefficient is sharply estimated close to 1, suggesting that time-to-containment is at least conditionally unrelated to unobservable characteristics of the fire that affect its total duration.

Where both containment and put-out dates are unavailable, a fire is assigned its duration based on the same model, estimated without including containment date as a covariate. All predictions less than 1 day due to the linear fit of the model are assigned a value of 1 day of burning. All fires with reported and predicted durations exceeding 160 days are assigned 160 days of burning to lower computational overhead. This is based on an assumption that fires reported to burn in excess

of 160 days have reporting error in their records or are long-burning smoldering fires, which do not have comparable emissions to flaming fires. This truncation procedure removes approximately 10 percent of fire emission days, and less than 4 percent of emissions when weighted by the total land area of the fire.

The purpose of these duration estimates is to improve the predictive power of modeled concentrations. Errors in the prediction from reporting errors or misspecification of the model for fire burn duration will result in emissions profiles of incorrect length. These errors will not affect the validity of the modeled pollutant concentrations as instruments for observed pollutant concentrations, provided they are statistically unrelated to the determinants of the observed pollutant concentrations I do not include in my first stage estimation.

A.2 Wildfire Modeling Details

A.2.1 Modeling Workflow

The fire events from the FPA FOD database are each input as individual events into the BSF. The CONSUME module reads the coordinate data of the event and determines the likely fuel type using the FCCS fuel map. CONSUME then divides the fuel consumption into flaming, smouldering, and residual emission phase, each of which has a distinct contribution to emissions volume for the same fuel (as a model of fuel combustion efficiency). Combining the fuel consumption profile with empirically-derived emissions factors, FEPS then estimates the quantities of heat and the pollutant of interest released by the fire. Using an empirically-derived diurnal (i.e., daily recurring) time profile embedded in FEPS, I generate a 24-hour emissions pattern that repeats for each day a fire burns and terminates at the estimated date of the fire's extinguishment. The pattern distributes the total emissions calculated by CONSUME among hours of the day. This modeling step is designed to improve downwind concentration estimates by accounting for fire burn cycles that vary with meteorological parameters that systematically vary with time of day, with lower emissions during nighttime hours.

The FEPS Plume Rise module estimates the buoyancy of the emitted pollutant due to the heat calculated by CONSUME and assigns 20 heights into which fractions of the hourly emissions are injected. This step reflects that quantities of a pollutant will be lofted higher from a fire location the more heat the fire releases, and that larger fires will also tend to have higher plumes that will result in longer-range transport. The result is a set of hourly point-source emissions for each fire event, with 20 emissions quantities in each hour released at the FEPS-calculated altitudes.

The point-source emissions generated by the CONSUME and FEPS models from the fire event data are then inputted into HYSPLIT, which calculates the trajectory and dispersion of the emitted pollutants and outputs a spatial field of concentrations over time. To calculate concentrations, HYSPLIT requires continuous meteorological data spanning the time period of the fire event and its corresponding downwind impacts of interest. Meteorological reanalysis data sets or archived forecasts are typically used for retrospective applications. Here, I use the Eta Data Assimilation System 40km (EDAS40), an archived 3-hourly forecast spanning 2004 to the present with a spatial resolution of 40km. This forecast system was developed and maintained by National Weather Service's National Centers for Environmental Prediction.

HYSPLIT represents the distribution of pollutants from a source through the behavior of a large number of individual "particles" (which are computational representations of pollutant masses, not to be confused with particulate pollutants in themselves). These particles are released over the duration of an emission and HYSPLIT models their advective motion using three-dimensional velocity vectors from the meteorological data. In addition, the particle approach adds a random component to their advective motion that approximates a random walk process calibrated by local atmospheric turbulence. HYSPLIT particles are assigned a proportional fraction of pollutant mass at the time of emission and shed mass through atmospheric removal processes (dry and wet deposition). Concentrations for a grid cell are calculated through the sum of masses of particles within the grid cell divided by the size of the grid cell. All HYSPLIT calculation methods are described in detail in Draxler and Hess (1997). I describe deposition processes and my choice of calibration parameters in the next section.

Each HYSPLIT run uses a 5-day set of hourly burning emissions at 21 vertical levels for a single fire location. I set HYSPLIT to release 300,000 particles per emissions hour, which are evenly divided among the vertical emission levels. I allow HYSPLIT to calculate the travel of particles for 920 hours (approximately five and a half weeks) from the hour of the first emission. From these calculations, HYSPLIT creates an hourly concentration grid for the CONUS model domain roughly matching the resolution of the meteorological data, with each grid square encompassing approximately 1,600 km sq. for 2004-2010. I sample concentrations from each fire event's grid at 10 meters above ground level at pollution monitoring sites and census tract centroids, sum concentrations across all fire events, and average the resulting hourly concentrations to daily average concentrations by each sampling site.

While the raw output is constructed from emissions measures and conversions that would denominate it in μgm^{-3} if it were to be taken literally, I remain agnostic about the units of the output and allow first-stage regressions to implicitly rescale the wildfire PM2.5 measure. In Appendix Section A.3, I establish that the output has a strongly logarithmic fit to observed pollution data and take a logarithmic transformation of the raw concentrations shifted by a small constant. This will be the wildfire PM2.5 instrument used for the remainder of the paper.

A.2.2 Deposition Processes

HYSPLIT's modeling of deposition, or the removal of pollutants from the atmosphere by precipitation and settling or impaction upon terrain, plays an important role in generating independent variation among pollutants to allow the separate identification of their health effects. HYSPLIT dynamically accounts for the amount of air pollution lost to precipitation by modeling the interaction of traveling parcels of air pollution from origin to destination with temporally and spatially smooth representations of precipitation events. HYSPLIT models particle pollutant wet deposition (also referred to as "wet removal" and "wet scavenging") via two processes described as in-cloud removal ("washout") and below-cloud removal ("rainout").[6] For gaseous pollutants, it uses a cal-

[6]There is some inconsistent usage of the terms "washout" and "rainout" in across some papers, their meanings occasionally swapped.

culation method based on gas solubility. HYSPLIT has one common process for both particles and gases for modeling dry deposition which assumes a rate of removal driven by wind speed. One pollutant-specific constant calibrates the intensity of each process: the washout ratio, representing an average ratio of pollutant concentration in air to concentration in water at the ground; the rainout rate, or a fixed rate of pollutant removal while pollutant concentrations are in a meteorological layer with precipitation (s^{-1}); the Henry's Law Constant for wet removal of soluble gases ($mol\,atm^{-1}$); and the dry deposition velocity (ms^{-1}). The constants I choose for each pollutant type, along with corresponding citations, are reported in Appendix Table 14. For reference, I also report constants for related pollutants that I do not model.

Wet deposition of particulate pollutants is characterized by HYSPLIT through one process in which polluted air is ingested over time into proximal atmospheric moisture (washout), and another in which rain falls through polluted air (rainout). Wet deposition processes play a relatively larger role in mass removal of fine particulate pollutants than they do for gaseous pollutants, up to an order of magnitude higher, though this relation varies by species. While there is substantial heterogeneity in the efficiency with which PM2.5 pollutants are removed by rain because of the many component subspecies and variation in the particle size distribution, a washout ratio of 1×10^5 is broadly used as an estimate for the washout ratio of general PM2.5. In the absence of well-established parameters for rainout rates, I use HYSPLIT's suggested particle rainout rate of $5 \times 10^{-5} s^{-1}$ which has been used in other HYSPLIT particulate modeling applications (Chand et al. 2008; Wen et al. 2013). I expect that empirically-derived washout ratios will capture most deposition since they are often measured without HYSPLIT's deposition process distinction in mind, and at least one study finds that below-cloud deposition is insignificant for fine particles except in extreme precipitation events (Andronache 2003).

Instead of explicit washout and rainout parameters, gaseous pollutants' wet deposition is calibrated by the appropriate Henry's Law constant for the water-soluble gas. Henry's Law holds that at a constant temperature, the solubility of gas in a liquid is proportional to the pressure of the gas surrounding the liquid. An intuitive example of Henry's Law at work is a carbonated soda: while

sealed, a soda bottle contains liquid with dissolved CO2 and a space above the liquid with CO2 gas. The opening of the bottle lowers the resulting pressure above the liquid, and over time the CO2 escapes from the liquid and into the open air through the bottle opening. The reverse process occurs if there is liquid in the same bottle with no CO2, and CO2 is injected into the empty space of the sealed bottle: the higher the pressure of the resulting air space in the bottle (and the greater the concentration of CO2), the greater the equilibrium concentration of CO2 in the liquid will be. Henry's Law constants are chosen from an extensive collection of estimates from academic papers (Sander 1999). Estimates are typically calculated in one of three ways: by theoretical calculations, extrapolations from other measured constants, or by field measurements and experiments. For each gas, I choose the most recent estimate from a literature review where available. If a literature review-based estimate is not available, I choose the modal Henry's Law constant reported in Sander (1999).

Dry deposition is modeled through gravitational settling and impaction at ground level which intensifies with wind velocity. In the absence of precipitation and chemical reactions, dry deposition is the primary determinant of a pollutant's lifetime in the atmosphere following emissions. I conduct a literature search for dry deposition velocities, using the compound name and "deposition velocity" as search terms. For deposition velocities for gases drawn from field observations, urban-setting deposition velocities are preferred. Many gases, such as NO and HCHO, do not have significant dry deposition fluxes over land. I use a deposition velocity of $0 ms^{-1}$ for such gases with trivial land deposition rates, and also for any gases for which I am unable to find any direct reference to deposition fluxes or velocities. The deposition velocities I choose are reported in Appendix Table 14.

A.3 Nonlinear First Stage Transformation

The relationship between measured concentrations and modeled concentrations is extremely nonlinear, requiring a monotonic transformation to maximize the predictive power of the wildfire pollution instrument. Figure 6a shows the estimated coefficients of a piecewise linear regression

of daily station PM2.5 on wildfire PM2.5 interacted with vigintile (5-percentile-block) indicator, representing an approximation of the first derivative of the true dose response function between measured and raw modeled PM2.5 across the raw modeled PM2.5 distribution. This regression controls for year, month, and county fixed effects, with standard errors clustered at the state level. The pattern is highly nonlinear, scaling multiple orders of magnitude, with the estimated slope monotonically decreasing in concentration. A function of the form $f(x) = \frac{a}{x+c}$ (with $a > 0, c \geq 0$) follows a comparable pattern, suggesting that a linear approximation better predicts station PM2.5 using as a regressor the natural logarithm of modeled wildfire PM2.5 plus some constant. This nonlinear pattern implies that some combination of the emissions calculations and HYSPLIT is resulting in systematic overestimation of large concentrations and underestimation of small concentrations. The monotonically decreasing slope across the domain of concentrations implicates the dispersion calculation of HYSPLIT, which relies on calibration from atmospheric parameters to determine turbulent velocities and a Gaussian random component that determine the random-walk-like dispersive behavior of the particle. One explanation for the subsequent logarithmic fit is that the calibration of the Gaussian component's variance does not account for how the true variance is itself positively related to concentration level, resulting in systematic underestimation of dispersion for large concentrations and overestimation for small concentrations (causing overestimated and underestimated concentrations, respectively).

A logarithmic transformation of the wildfire pollution measure in the first stage accounts for the implicit overdispersion of concentrations along trajectories by compressing the distribution of magnitudes. To accomplish this transformation without discarding zero values, I take the natural logarithm of daily average wildfire PM2.5 plus a constant. The choice of constant by which to shift the raw concentration before taking the logarithm, the "shift parameter" has two important impacts: it determines the position of zeroes on the log function, and relatedly, it changes the relative curvature of the fit of logged concentrations to observed concentrations. Shift parameters that are too small will result in the log transformation overestimating the contrast between the effect of positive wildfire concentrations relative to zero wildfire concentrations, while shift parameters

that are too large will cause an underestimated contrast. Large shift parameters may also distort the marginal effects for larger values in the distribution. One sensible choice of shift parameter is a point at which positive concentrations could be considered effectively zero for the dependent variable of interest. HYSPLIT's concentration outputs near zero can be reasonably framed as a sensitivity problem: there is a computational threshold below which it will never give a positive value, and the distribution of values approaching zero is continuous until the trivial minimum value at $4.92 \times 10^{-34} \mu gm^{-3}$. I choose a value corresponding to the 10th percentile of positive values ($7.21 \times 10^{-14} \mu gm^{-3}$), add it to the raw concentration value, and take the logarithm. For ease of interpretation later, I also shift all transformed values by the minimum of the transformed values to make all values nonnegative. Figure 6b shows the same regression as in Figure 6b, but now with logged daily wildfire PM2.5 interacted with vigintile indicator. The slopes now fall within the same order of magnitude, slightly increasing in vigintile (implying a gradual shift to underestimation of marginal changes in concentrations relative to smaller vigintiles).

Additionally, there are several numerically large outliers which may affect the fit, but station data provides a way of trimming outlier values sensibly. I account for these right-tail outliers by assigning the instrument the station PM2.5 value if the raw modeled wildfire PM2.5 exceeds the station-observed PM2.5 value, and both values are greater than $65 \mu gm^{-3}$. Empirically, the latter condition implies the former in 100 percent of cases, which motivated the selection of this cutoff. Less than 0.1 percent of station-days have measured PM2.5 exceeding $65 \mu gm^{-3}$. All other wildfire PM2.5 values exceeding $65 \mu gm^{-3}$ (approximately 0.6 percent of all values) are set to $65 \mu gm^{-3}$. This adjustment compresses the right tail of the distribution, enhancing the performance of the logarithmic transformation I take to improve the fit of the instrument (in exchange for losing some variation in extreme values). Because the observed nonlinear relationship and outliers are ostensibly due to HYSPLIT's dispersion calculation methods, which are not unique to any pollutant, I assume that concentrations for other pollutant species follow a comparable relationship with their observed values (in the absence of daily station data to do a pollutant-specific adjustment). For all control species, I reduce all right-tail values for other pollutant species to their 98th percentile of

positive values before taking the logarithmic transformation, since that is the approximate point at which the PM2.5 values always exceed station values. Then, I take the logarithm of the outlier-adjusted modeled concentration outputs plus the 10th percentile of their positive values added.

A.4 Coefficient Estimates under Non-Classical Measurement Error

Consider a simplified cross-sectional setting, with health outcome y as a function of true exposure x^*,

$$y_i = x_i^* \beta + \varepsilon_i$$

Assume x_i^* is uncorrelated with ε_i, and that the researcher only observes an imperfect measure x of x^*such that $x = x^* + e$. Define $var(x^*) = \sigma_{x^*}^2$, $var(e) = \sigma_e^2$, and $cov(x^*, e) = \sigma_{x^*e}$. Then, the probability limit of the ordinary least squares estimator of y on x can be written as

$$plim\hat{\beta} = \beta \frac{(\sigma_{x^*}^2 + \sigma_{x^*e})}{\sigma_{x^*}^2 + \sigma_e^2 + 2\sigma_{x^*e}}$$

By the Cauchy-Schwarz inequality, the denominator is always positive: it represents the variance of the error-prone regressor x. $\sigma_{x^*e} = 0$ corresponds to the classical errors-in-variables assumption, which results in attenuation bias. The probability limit of the OLS estimate $\hat{\beta}$ is both attenuated and incorrectly-signed if $\sigma_{x^*e} < 0$ and $\sigma_{x^*}^2 < |\sigma_{x^*e}|$. Negative correlation between the true regressor value and the size of the measurement error is plausible in the pollution setting if population density increases both pollution levels and reduces exposure measurement error asymmetrically across polluted areas.

A.5 Change in Finite-Sample Bias of IV when Fixed Effects are Included

Another possibility for the increase in estimates across different fixed effects specifications is that the inclusion of fixed effects potentially changes the finite-sample bias of the 2SLS estimate, even

with equally "strong" instruments in the Staiger and Stock (1997) nomenclature. This is because both the strength of the first-stage relationship and the amount of correlation between the endogenous variable and structural equation error term both determine finite-sample bias of IV estimators. In this interpretation, the inclusion of fixed effects chooses variation in observed PM2.5 that is less correlated with unobserved determinants of health than the total variation in PM2.5, while simultaneously not weakening the relationship between wildfire PM2.5 and station PM2.5 sufficiently to counterbalance the change. The corresponding OLS estimates for PM2.5 are close to zero and relatively precise, implying that IV estimates would be biased toward zero. One might be inclined to believe that between-county variation in pollution is more strongly associated with unobserved determinants of health than within-county variation is, both based on the mechanisms proposed for either correlation (e.g., residential and industrial sorting versus micro-level changes in economic activity) and more pragmatically through the revealed preference of researchers for multiple time-series and panel studies over cross-sectional studies. Murray (2006) provides a simplified approximation of the finite-sample bias of 2SLS based on Hahn and Hausman (2001) (where the structural and first-stage equation error terms have varianced normalized to one) as follows:

$$E(\hat{\beta}_{,2SLS}) - \beta \approx \frac{l\rho(1-\tilde{R}^2)}{N\tilde{R}^2}.$$

Here, β is the effect of PM2.5 on mortality, $l = 1$ is the number of instruments, ρ is the correlation between the structural and first-stage equation error terms (a measure of the level of endogeneity), $\tilde{R}^2$ is the partial R-squared of the first stage regression, and N is the sample size. If the inclusion of fixed effects decreases ρ to ρ_{fe}, but decreases $\tilde{R}^2$ to $\tilde{R}^2_{fe}$, then the approximate bias decreases as long as $\frac{\rho}{\rho_{fe}} > \left(\frac{(1-\tilde{R}^2)}{\tilde{R}^2}\right)\left(\frac{(1-\tilde{R}^2_{fe})}{\tilde{R}^2_{fe}}\right)^{-1}$.

REFERENCES

REFERENCES

[1] Anderson, G., Sandberg, D., & Norheim, R. (2004). Fire Emission Production Simulator (FEPS) User's Guide. USDA Forest Service Pacific Northwest Research Station, Fire and Environmental Research Applications Team, Portland, OR.

[2] Anderson, T. W., & Rubin, H. (1949). Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations. The Annals of Mathematical Statistics, 20(1), 46–63. doi:10.1214/aoms/1177730090

[3] Andronache, C. (2003). Estimated variability of below-cloud aerosol removal by rainfall for observed aerosol size distributions. Atmos. Chem. Phys., 3(1), 131–143. doi:10.5194/acp-3-131-2003

[4] Arceo-Gomez, E. O., Hanna, R., & Oliva, P. (2012). Does the Effect of Pollution on Infant Mortality Differ Between Developing and Developed Countries? Evidence from Mexico City (Working Paper No. 18349). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w18349

[5] Banzhaf, H. S., & Chupp, B. A. (2012). Fiscal federalism and interjurisdictional externalities: New results and an application to US Air pollution. Journal of Public Economics, 96(5), 449–464.

[6] Barreca, A. I. (2012). Climate change, humidity, and mortality in the United States. Journal of Environmental Economics and Management, 63(1), 19–34. doi:10.1016/j.jeem.2011.07.004

[7] Behrman, J. R., & Rosenzweig, M. R. (2004). Returns to Birthweight. Review of Economics and Statistics, 86(2), 586–601. doi:10.1162/003465304323031139

[8] Bell, M. L., & HEI Health Review Committee. (2012). Assessment of the health impacts of particulate matter characteristics. Research Report (Health Effects Institute), (161), 5–38.

[9] Bell, M. L., McDermott, A., Zeger, S. L., Samet, J. M., & Dominici, F. (2004). Ozone and short-term mortality in 95 US urban communities, 1987-2000. JAMA, 292(19), 2372–2378. doi:10.1001/jama.292.19.2372

[10] Bound, J., Jaeger, D. A., & Baker, R. M. (1995). Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak. Journal of the American Statistical Association, 90(430), 443–450. doi:10.2307/2291055

[11] Breton, C., Park, C., & Wu, J. (2011). Effect of Prenatal Exposure to Wildfire-generated PM2.5 on Birth Weight. Epidemiology, 22, S66. doi:10.1097/01.ede.0000391864.79309.9c

[12] Calvert, J. G., Atkinson, R., Becker, K. H., Kamens, R. M., Seinfeld, J. H., Wallington, T. J., & Yarwood, G. (2002). The mechanisms of atmospheric oxidation of aromatic hydrocarbons. Oxford University Press New York.

[13] Chand, D., Jaffe, D., Prestbo, E., Swartzendruber, P. C., Hafner, W., Weiss-Penzias, P., ... Kajii, Y. (2008). Reactive and particulate mercury in the Asian marine boundary layer. Atmospheric Environment, 42(34), 7988–7996. doi:10.1016/j.atmosenv.2008.06.048

[14] Chay, K., Dobkin, C., & Greenstone, M. (2003). The Clean Air Act of 1970 and Adult Mortality. Journal of Risk and Uncertainty, 27(3), 279–300. doi:10.1023/A:1025897327639

[15] Chay, K. Y., & Greenstone, M. (2003). Air Quality, Infant Mortality, and the Clean Air Act of 1970 (Working Paper No. 10053). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w10053

[16] Chen, B., Stein, A. F., Maldonado, P. G., Sanchez de la Campa, A. M., Gonzalez-Castanedo, Y., Castell, N., & de la Rosa, J. D. (2013). Size distribution and concentrations of heavy metals in atmospheric aerosols originating from industrial emissions as predicted by the HYSPLIT model. Atmospheric Environment, 71, 234–244. doi:10.1016/j.atmosenv.2013.02.013

[17] Chen, L. H., Knutsen, S. F., Shavlik, D., Beeson, W. L., Petersen, F., Ghamsary, M., & Abbey, D. (2005). The Association between Fatal Coronary Heart Disease and Ambient Particulate Air Pollution: Are Females at Greater Risk? Environmental Health Perspectives, 113(12), 1723.

[18] Collins, J., Williams, A., Paxton, C., & Davis, R. (2009). Geographical, Meteorological, and Climatological Conditions Surrounding the 2008 Interstate-4 Disaster in Florida. Papers of the Applied Geography Conferences, 153–162.

[19] Currie, J., & Neidell, M. (2005). Air Pollution and Infant Health: What Can We Learn from California's Recent Experience? The Quarterly Journal of Economics, 120(3), 1003–1030.

[20] Currie, J., Zivin, J. S. G., Mullins, J., & Neidell, M. J. (2013). What Do We Know About Short and Long Term Effects of Early Life Exposure to Pollution? (Working Paper No. 19571). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w19571

[21] Daniels, M. J., Dominici, F., Samet, J. M., & Zeger, S. L. (2000). Estimating Particulate Matter-Mortality Dose-Response Curves and Threshold Levels: An Analysis of Daily Time-Series for the 20 Largest US Cities. American Journal of Epidemiology, 152(5), 397–406. doi:10.1093/aje/152.5.397

[22] Dieterle, S., & Snell, A. (2013). Exploiting Nonlinearities in the First Stage Regressions of IV Procedures.

[23] Donovan, G. H. (2006). Determining the optimal mix of federal and contract fire crews: A case study from the Pacific Northwest. Ecological Modelling, 194(4), 372–378.

[24] Draxler, R., Arnold, D., Chino, M., Galmarini, S., Hort, M., Jones, A., ... Wotawa, G. (n.d.). World Meteorological Organization's model simulations of the radionuclide dispersion and deposition from the Fukushima Daiichi nuclear power plant accident. Journal of Environmental Radioactivity. doi:10.1016/j.jenvrad.2013.09.014

[25] Draxler, R. R., & Hess, G. (1997). Description of the HYSPLIT4 modeling system.

[26] Dye, J. A., Lehmann, J. R., McGee, J. K., Winsett, D. W., Ledbetter, A. D., Everitt, J. I., ... Costa, D. L. (2001). Acute pulmonary toxicity of particulate matter filter extracts in rats: coherence with epidemiologic studies in Utah Valley residents. Environmental Health Perspectives, 109(Suppl 3), 395–403.

[27] Escudero, M., Stein, A., Draxler, R. R., Querol, X., Alastuey, A., Castillo, S., & Avila, A. (2006). Determination of the contribution of northern Africa dust source areas to PM10 concentrations over the central Iberian Peninsula using the Hybrid Single-Particle Lagrangian Integrated Trajectory model (HYSPLIT) model. Journal of Geophysical Research: Atmospheres, 111(D6), D06210. doi:10.1029/2005JD006395

[28] Federal fire history reports by date and organization: 1980 - 2013 DOI (BIA, BLM, BOR, NPS), USFWS, and USFS. (2013). U.S. Department of Interior. Retrieved from http://wildfire.cr.usgs.gov/firehistory/data.html

[29] Finlay, K., & Magnusson, L. M. (2009). Implementing weak-instrument robust tests for a general class of instrumental-variables models. Stata Journal, 9(3), 398–421.

[30] Fire Executive Council. (2009). Guidance for implementation of federal wildland fire management policy.

[31] Franklin, M., Koutrakis, P., & Schwartz, J. (2008). The Role of Particle Composition on the Association Between PM2.5 and Mortality. Epidemiology (Cambridge, Mass.), 19(5), 680–689.

[32] Hahn, J., & Hausman, J. (2002). Notes on bias in estimators for simultaneous equation models. Economics Letters, 75(2), 237–241. doi:10.1016/S0165-1765(01)00602-4

[33] Heutel, G., & Ruhm, C. J. (2013). Air Pollution and Procyclical Mortality (Working Paper No. 18959). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w18959

[34] Huang, Y.-C. T., & Ghio, A. J. (2006). Vascular effects of ambient pollutant particles and metals. Current Vascular Pharmacology, 4(3), 199–203.

[35] Jayachandran, S. (2009). Air Quality and Early-Life Mortality Evidence from Indonesia's Wildfires. Journal of Human Resources, 44(4), 916–954.

[36] Keane, R. E., & Karau, E. (2010). Evaluating the ecological benefits of wildfire by integrating fire and ecosystem simulation models. Ecological Modelling, 221(8), 1162–1172.

[37] Knapp, E. E., Estes, B. L., & Skinner, C. N. (2009). Ecological Effects of Prescribed Fire Season: A Literature Review and Synthesis for Managers. Retrieved from http://www.firescience.gov/projects/07-S-08/project/07-S-08_psw_gtr224-1.pdf

[38] Knittel, C. R., Miller, D. L., & Sanders, N. J. (2011). Caution, Drivers! Children Present: Traffic, Pollution, and Infant Health (Working Paper No. 17222). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w17222

[39] Kristensen, L. J., & Taylor, M. P. (2012). Fields and Forests in Flames: Lead and Mercury Emissions from Wildfire Pyrogenic Activity. Environmental Health Perspectives, 120(2), a56–a57. doi:10.1289/ehp.1104672

[40] Künzli, N., Jerrett, M., Mack, W. J., Beckerman, B., LaBree, L., Gilliland, F., ... Hodis, H. N. (2005). Ambient air pollution and atherosclerosis in Los Angeles. Environmental Health Perspectives, 201–206.

[41] Larkin, N. K., O'Neill, S. M., Solomon, R., Raffuse, S., Strand, T., Sullivan, D. C., ... Ferguson, S. A. (2009). The BlueSky smoke modeling framework. Int. J. Wildland Fire, 18(8), 906–920.

[42] MacNee, W., & Donaldson, K. (2003). Mechanism of lung injury caused by PM10 and ultrafine particles with special reference to COPD. European Respiratory Journal, 21(40 suppl), 47s–51s. doi:10.1183/09031936.03.00403203

[43] Moretti, E., & Neidell, M. (2011). Pollution, Health, and Avoidance Behavior: Evidence from the Ports of Los Angeles. Journal of Human Resources, 46(1), 154–175.

[44] Murray, M. P. (2006). Avoiding Invalid Instruments and Coping with Weak Instruments. Journal of Economic Perspectives, 20(4), 111–132. doi:10.1257/jep.20.4.111

[45] O'Neill, S. M., Larkin, N. (Sim) K., Hoadley, J., Mills, G., Vaughan, J. K., Draxler, R. R., ... Ferguson, S. A. (n.d.). Regional real-time smoke prediction systems, 8, 499–534.

[46] Ottmar, R. D., Miranda, A. I., & Sandberg, D. V. (n.d.). Characterizing sources of emissions from wildland fires, 8, 61–78.

[47] Ottmar, R. D., Sandberg, D. V., Riccardi, C. L., & Prichard, S. J. (2007). An overview of the Fuel Characteristic Classification System — Quantifying, classifying, and creating fuelbeds for resource planning. Canadian Journal of Forest Research, 37(12), 2383–2393. doi:10.1139/X07-077

[48] Pope, C. A., Burnett, R. T., Thurston, G. D., Thun, M. J., Calle, E. E., Krewski, D., & Godleski, J. J. (2004). Cardiovascular Mortality and Long-Term Exposure to Particulate Air Pollution Epidemiological Evidence of General Pathophysiological Pathways of Disease. Circulation, 109(1), 71–77. doi:10.1161/01.CIR.0000108927.80044.7F

[49] Pope III, C. A., Rodermund, D. L., & Gee, M. M. (2007). Mortality effects of a copper smelter strike and reduced ambient sulfate particulate matter air pollution. Environmental Health Perspectives, 679–683.

[50] Pozzi, R., De Berardis, B., Paoletti, L., & Guastadisegni, C. (2003). Inflammatory mediators induced by coarse (PM2.5–10) and fine (PM2.5) urban air particles in RAW 264.7 cells. Toxicology, 183(1–3), 243–254. doi:10.1016/S0300-483X(02)00545-0

[51] Prichard, S., Ottmar, R., & Anderson, G. (2006). Consume 3.0 User's Guide. USDA Forest Service. Pacific Northwest Research Station.(Seattle, WA) Available at Http://www. Fs. Fed. Us/pnw/fera/research/smoke/consume/index. Shtml [Verified 6 January 2012].

[52] Prinn, R., Cunnold, D., Rasmussen, R., Simmonds, P., Alyea, F., Crawford, A., ... Rosen, R. (1987). Atmospheric trends in methylchloroform and the global average for the hydroxyl radical. Science (New York, N.Y.), 238(4829), 945–950. doi:10.1126/science.238.4829.945

[53] Rappold, A. G., Cascio, W. E., Kilaru, V. J., Stone, S. L., Neas, L. M., Devlin, R. B., & Diaz-Sanchez, D. (2012). Cardio-respiratory outcomes associated with exposure to wildfire smoke are modified by measures of community health. Environmental Health, 11(1), 71. doi:10.1186/1476-069X-11-71

[54] Rolph, G. D., Draxler, R. R., Stein, A. F., Taylor, A., Ruminski, M. G., Kondragunta, S., ... Davidson, P. M. (2009). Description and Verification of the NOAA Smoke Forecasting System: The 2007 Fire Season. Weather and Forecasting, 24(2), 361–378. doi:10.1175/2008WAF2222165.1

[55] Samet, J. M., Dominici, F., Curriero, F. C., Coursac, I., & Zeger, S. L. (2000). Fine Particulate Air Pollution and Mortality in 20 U.S. Cities, 1987–1994. New England Journal of Medicine, 343(24), 1742–1749. doi:10.1056/NEJM200012143432401

[56] Sanders, N. J., & Stoecker, C. F. (2011). Where Have All the Young Men Gone? Using Gender Ratios to Measure Fetal Death Rates (Working Paper No. 17434). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w17434

[57] Schlenker, W., & Walker, W. R. (2011). Airports, Air Pollution, and Contemporaneous Health. National Bureau of Economic Research Working Paper Series, No. 17684. Retrieved from http://www.nber.org/papers/w17684

[58] Schwartz, J., Laden, F., & Zanobetti, A. (2002). The concentration-response relation between PM (2.5) and daily deaths. Environmental Health Perspectives, 110(10), 1025.

[59] Short, K. C. (2013). A spatial database of wildfires in the United States, 1992–2011. Earth System Science Data Discussions, 6(2), 297–366. doi:10.5194/essdd-6-297-2013

[60] Slama, R., Darrow, L., Parker, J., Woodruff, T. J., Strickland, M., Nieuwenhuijsen, M., ... Ritz, B. (2008). Meeting Report: Atmospheric Pollution and Human Reproduction. Environmental Health Perspectives, 116(6), 791–798. doi:10.1289/ehp.11074

[61] Sorensen, M., Daneshvar, B., Hansen, M., Dragsted, L. O., Hertel, O., Knudsen, L., & Loft, S. (2003). Personal PM2.5 exposure and markers of oxidative stress in blood. Environmental Health Perspectives, 111(2), 161–166.

[62] Staiger, D., & Stock, J. H. (1994). Instrumental Variables Regression with Weak Instruments (Working Paper No. 151). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/t0151

[63] Stieb, D. M., Burnett, R. T., Smith-Doiron, M., Brion, O., Shin, H. H., & Economou, V. (2008). A new multipollutant, no-threshold air quality health index based on short-term associations observed in daily time-series analyses. Journal of the Air & Waste Management Association, 58(3), 435–450.

[64] TAN, W. C., QIU, D., LIAM, B. L., NG, T. P., LEE, S. H., van EEDEN, S. F., ... HOGG, J. C. (2000). The Human Bone Marrow Response to Acute Air Pollution Caused by Forest Fires. American Journal of Respiratory and Critical Care Medicine, 161(4), 1213–1217. doi:10.1164/ajrccm.161.4.9904084

[65] Wen, D., Lin, J. C., Zhang, L., Vet, R., & Moran, M. D. (2013). Modeling atmospheric ammonia and ammonium using a stochastic Lagrangian air quality model (STILT-Chem v0.7). Geosci. Model Dev., 6(2), 327–344. doi:10.5194/gmd-6-327-2013

[66] Wiedinmyer, C., & Friedli, H. (2007). Mercury Emission Estimates from Fires: An Initial Inventory for the United States. Environmental Science & Technology, 41(23), 8092–8098. doi:10.1021/es071289o

[67] Wooldridge, J. M. (2010). Econometric Analysis of Cross Section and Panel Data. MIT Press.

[68] Yoder, J. (2004). Playing with fire: endogenous risk in resource management. American Journal of Agricultural Economics, 86(4), 933–948.

[69] Zelikoff, J. T., Chen, L. C., Cohen, M. D., Fang, K., Gordon, T., Li, Y., ... Schlesinger, R. B. (2003). Effects of Inhaled Ambient Particulate Matter on Pulmonary Antimicrobial Immune Defense. Inhalation Toxicology, 15(2), 131–150. doi:10.1080/08958370304478

Chapter 2. Finite Sample Properties and Empirical Applicability of Two-Sample Two-Stage Least Squares

(With Wei Lin)

1 Introduction

Instrumental variables (IV) methods enable the consistent estimation of endogenous variables' causal effects but suffer from poor finite-sample properties and data availability constraints. Bound, Jaeger, and Baker (1995) establish that estimation with weak instruments can lead to large inconsistencies and finite sample bias. IV estimates also tend to have relatively large standard errors, often inhibiting the interpretability of differences between IV and non-IV point estimates. Lastly, the idiosyncratic nature of valid instrumental variables reduces their availability in data sets alongside outcome and other variables of interest. Beginning with Klevmarken (1982), some researchers have sought to address the problem of data availability by using two-sample IV methods (TSIVM), which combine parameter estimates from multiple data sets into a final IV estimate. Under a set of ideal conditions, a TSIVM produces an estimate with identical bias to the otherwise inaccessible traditional IV estimate. However, the finite-sample properties of TSIVM estimators are generally unknown, and prior literature lacks clear guidelines for how researchers should interpret them. The potential for researchers to introduce additional data and estimate models by TSIVM to produce estimates superior to available single-sample estimates has also not been explored.

We establish some insights into the finite-sample properties of the two-sample two-stage least squares (TS2SLS) estimator. Likely owing to its ease of implementation and interpretation, the TS2SLS estimator is the most commonly-used TSIVM estimator in empirical applications (e.g., Arellano and Meghir 1992; den Berg et al. 2015; Devereux and Hart 2014; Nicoletti and Ermisch 2014; Rothstein and Wozny 2014). We broaden the set of potential applications of TS2SLS by

demonstrating that even where a one-sample 2SLS estimate is available, a TS2SLS estimator may sometimes be preferred or worth reporting alongside the one-sample estimate because of greater precision and smaller bias. We propose approximations for the bias and variance of the TS2SLS estimator that are dependent on both the typical set of parameters in the "weak instruments" literature for one-sample IV estimators and on three parameters unique to two-sample estimators: the distinct sample sizes of the first-stage and second-stage samples and the proportion of observations that "overlap" between them (i.e., the fraction of real population units from the first-stage sample which are also in the second-stage sample). To test the approximations, we conduct a series of Monte Carlo simulations and compute the average bias and standard errors across simulations.

We develop a data framework in which the TS2SLS estimator is computed from a complete theoretical sample of population units (the "super-sample") of which the first-stage and second-stage samples used to compute the TS2SLS estimate of interest are subsets containing potentially overlapping units. This approach formally reconciles two-sample and split-sample 2SLS estimators by nesting them in a common framework; for example, it makes equivalent the no-overlap TS2SLS estimator studied in Inoue and Solon (2008) and split-sample 2SLS (SS2SLS) estimator analogous to Angrist and Krueger (1995)'s split-sample IV (SSIV). We find that the TS2SLS estimator can be written as a convex linear combination of the 2SLS estimator computed using the overlapping units between the two samples, and the SS2SLS estimator computed using the remaining non-overlapping units. The weight on the 2SLS component is a function of sampling variation in the first-stage estimates for each of the overlapping and non-overlapping subsamples. The weight converges asymptotically to the "overlap" parameter, representing the proportion of units in the second-stage sample which are also in the first-stage sample. This linear partitioning of the TS2SLS estimator frames it as a function of two estimators with known properties, simplifying the development of approximations of bias and variance.

We find that the TS2SLS estimator has all bias dependent on the degree of sampling error in the first-stage parameters (i.e., the coefficients on the instruments) relative to the strength of the endogeneity. As with 2SLS, the bias is only decreasing for the finite-sample bias from first-stage

sampling error. Biases from invalid instruments or from violations of the key TS2SLS assumption (that the first-stage coefficients on the instruments are identical for both primary and secondary samples) are invariant to sample size. The TS2SLS estimator's variance is decreasing in each of its corresponding sample sizes, with variance reduction diminishing more rapidly from increasing first-stage sample size.

We demonstrate a hypothetical empirical application of TS2SLS using data from Angrist and Evans (1998), using their actual sample as the "super-sample" and examining the estimates they could have recovered had they been forced to use TS2SLS with subsets of their sample instead of 2SLS with their entire sample. Using TS2SLS with half the data for the first stage and half for the second stage (effectively a SS2SLS estimate), the estimate is closer to zero than the super-sample 2SLS estimate and less precise. We find that a TS2SLS estimate using only half of their observations for the first-stage estimation but their entire sample for the second-stage almost exactly recovers the super-sample 2SLS estimate with equivalent precision, providing evidence of the strength of the instrument used. This exercise suggests one situation in practice in which TS2SLS is most likely to yield a high return: when the researcher can estimate the first stage precisely with one set of data, but also has access to more observations containing the outcome and the instrument but not the endogenous variable.

The econometrics literature formally concerning TSIVM has explored the computation and asymptotic properties of various two-sample IV estimators. Angrist and Krueger (1995) provide the finite-sample properties of the split-sample IV estimator, only alluding to its relationship to the two-sample IV estimator, which they use in another study (Angrist and Krueger 1993). Inoue and Solon (2008) computationally distinguish the two-sample IV estimator, which is calculated explicitly using the ratio of covariance matrices each estimated from different data sets, from the TS2SLS estimator, which is calculated using ordinary least squares of the outcome against cross-sample first-stage fitted values. Their main finding is that the TS2SLS approach is asymptotically more efficient than the TSIV approach because TS2SLS takes into account differences in the sampling distribution of the instrument between the primary and secondary samples, while TSIV does

not. Both Angrist and Krueger (1995) and Inoue and Solon (2008) only consider two-sample estimators which use fully independent samples. One paper in the epidemiology literature, Pierce and Burgess (2013), makes some commentary on the finite-sample properties of two-sample IV estimators via simulation, though the paper is primarily focused on the use of TSIV methods to promote efficient study design by reducing the number of first-stage observations needed. This paper contributes to the TSIVM literature by formalizing results around the finite-sample behavior of the TS2SLS estimator, generalizing the results to potentially non-independent first-stage and second-stage samples, and applying the results to common applications in which TS2SLS-style estimators are used.

2 Properties of the TS2SLS Estimator

2.1 Model

To examine the TS2SLS estimator, we establish a conventional linear simultaneous equations framework while also defining the relationships between two arbitrary samples. Suppose S1 $\{(y_{1i}, \mathbf{z}_{1i})\}_{i=1}^{N_1}$ and S2 $\left\{\left(x_{2j}, \mathbf{z}_{2j}\right)\right\}_{j=1}^{N_2}$ are i.i.d. random vectors from the same underlying population, where $\mathbf{z}_{1i}^{'}$ and $\mathbf{z}_{2j}^{'}$ are $K \times 1$ vectors and y_{1i} and x_{2j} are scalars. $S1$ corresponds to what we call the "second-stage sample," and $S2$ corresponds to the "first-stage sample." We assume the following on z and x:

1. $E(\mathbf{z}_{1i}'\mathbf{z}_{1i}) = E(\mathbf{z}_{2j}'\mathbf{z}_{2j}) = \Omega_{\mathbf{z}}$.
2. $E(\mathbf{z}_{1i}'x_{1i}) = E(\mathbf{z}_{2j}'x_{2j}) = \Omega_{x\mathbf{z}}$.
3. Rank $E\left(\mathbf{z}_{1i}^{'}\mathbf{z}_{1i}\right) = \text{Rank } E\left(\mathbf{z}_{2j}^{'}\mathbf{z}_{2j}\right) = K$
4. Rank $E\left(\mathbf{z}_{1i}^{'}x_{1i}\right) = \text{Rank } E\left(\mathbf{z}_{2j}^{'}x_{2j}\right) = 1$

We follow a general single-endogenous variable framework:

$$
\begin{aligned}
x_{1i} &= \mathbf{z}_{1i}\gamma_1 + v_{1i} && (8)\\
x_{2j} &= \mathbf{z}_{2j}\gamma_2 + v_{2j}. && (9)\\
y_{1i} &= x_{1i}\beta + \varepsilon_i && (10)\\
&= \mathbf{z}_{1i}\gamma\beta + \beta v_{1i} + \varepsilon_i && (11)\\
&= \mathbf{z}_{1i}\gamma\beta + u_{1i} && (12)
\end{aligned}
$$

Without loss of generality, predetermined variables w (including constants) are "partialled out" of source variables to create x, z, and y. The outcome y is a function of an endogenous variable x, and x is a function of an instrument z and error v. The error ε may in general be correlated with error v, giving rise to the endogeneity of x. The data sets available for use in estimation consist of two subsamples, $S1$ and $S2$ of a broader data set of N units. $S1$ and $S2$ generally may partially or entirely overlap in terms of the underlying units for which they have data. The first subscript identifies the sample from which each observations of x, y, or z is drawn. The second subscripts, i for sample 1 and j for sample 2, index individuals within each sample. β is the causal effect of x on y, and the parameter of interest estimated by instrumental variables. γ_1 and γ_2 are the first-stage linear projection coefficients of x on z in each sample, and could differ in practice; here, we assume that $\boldsymbol{\gamma_1 = \gamma_2}$. The assumptions for this model are as follows:

Assumptions:

1. Define a function $\Phi(i)$ which takes on value 1 if unit i in $S1$ is also a member of $S2$, and zero otherwise.

2. The number of units in N_2 also in N_1 is ρN_2, with $0 \leq \rho \leq 1$.

   (a) The total number of distinct units represented by subsamples S1 and S2 is $N = N_1 + (1-\rho)N_2$.

3. The data are ordered so that sequence of the first ρN_2 number of observations from S2 are identical to the sequence of the first ρN_2 number of observations in S1 $\left\{\left(x_{2j}, \mathbf{z}_{2j}\right)\right\}_{j=1}^{\rho N_2} = \left\{\left(x_{1i}, \mathbf{z}_{1i}\right)\right\}_{i=1}^{\rho N_2}$.

4. $\mathbf{z}_{1i}$ are valid and relevant excluded instruments for x; that is,

$$E(\varepsilon_i|\mathbf{z}_{1i}) = E(\varepsilon_i|\mathbf{z}_{2i}) = 0, E(v_{1i}|\mathbf{z}_{1i}) = 0, E(v_{2j}|\mathbf{z}_{2j}) = 0, \gamma \neq 0$$

5. The ratio of S1 and S2 converges to a fixed positive number in large samples, that is,

$$\underset{N_1, N_2 \to \infty}{plim} \frac{N_1}{N_2} = \alpha.$$

6. Equation (8) is the structural equation with structural error ε_i. Equation (12) is the "reduced-form" equation, from substituting (8) into (10) and defining composite error $u_{1i} = \beta v_{1i} + \varepsilon_i$. The error terms are homoscedastic (implicitly conditioned on the partialled out exogenous variables w) with covariance matrix,

$$var\begin{pmatrix} \varepsilon_i \\ v_{1i} \\ v_{2j} \end{pmatrix} = \begin{pmatrix} \sigma_\varepsilon^2 & \sigma_{\varepsilon v} & \rho\sigma_{\varepsilon v} \\ \sigma_{\varepsilon v} & \sigma_v^2 & \rho\sigma_v^2 \\ \rho\sigma_{\varepsilon v} & \rho\sigma_v^2 & \sigma_v^2 \end{pmatrix}.$$

   Note that ε_i and v_{2j} are independent when $\Phi(i) = 0$, but when $\Phi(i) = 1$, $v_{2j} = v_{1j}$. This covariance matrix implies

   (a) $var(u_{1i}) = \beta^2\sigma_v^2 + \sigma_\varepsilon^2 + 2\beta\sigma_{\varepsilon v}$. Define $\sigma_u^2 \equiv var(u_{1i})$.

   (b) $\varepsilon_i = \theta v_{1i} + r_{1i}$, where $\theta = \frac{\sigma_{\varepsilon v}}{\sigma_v^2}$, and r_{1i} is independent of v_{1i}.

   (c) $E\left(x_{2j}\varepsilon_i\right) = \rho\sigma_{\varepsilon v}$.

Thus far, we have been agnostic regarding any practical case of data combination - this model will generally apply to any combination of two samples to estimate β. The model specifies how two

samples are drawn from the same underlying population and may share some units, describing the resulting covariance structure of the errors. For simplicity of argument, we have assumed zero conditional mean (i.e., valid instruments z) and homoscedasticity of the errors (implicitly conditioned on predetermined variables w). The bias and variance approximations presented in this paper are intended for directional insight regarding the relationship between the first- and second-stage samples, rather than explicit estimation. Invalid instruments will change the finite sample bias as a function of the magnitude and direction of covariance between z and ε; we conjecture that invalid instruments would not affect the nature of finite sample bias arising strictly from first-stage sampling error. However, heteroscedasticity would likely increase the finite sample bias and variance of the TS2SLS estimator relative to a model under homoscedasticity with the same parameters ρ and $\sigma_{\varepsilon v}$. Heteroscedastic error in the first stage results in an OLS first stage estimates no longer being minimum-variance and first-stage variance drives both bias and variance in the final 2SLS estimate. Practitioners will still need to compute accurate standard errors on their estimates, and thus variance estimates should be made robust to arbitrary heteroscedasticity.

2.2 Definitions of Estimators

In practice, the TS2SLS estimator involves generating an estimate of the first stage parameter γ, $\hat{\gamma}_2$, using N_2 observations with nonmissing values of x and z, generating N_1 cross-sample fitted values $\hat{x}_{1i} = z_{1i}\hat{\gamma}_2$, and then regressing y_1 on $\hat{x}_1$ via OLS to estimate β. To facilitate a clearer understanding of the estimator, we express TS2SLS as a weighted combination of 2SLS and SS2SLS estimators on different subsets of the data, whose sizes depend on the degree of overlap between samples $S1$ and $S2$. Intuitively, the 2SLS component of the estimator is estimated using all units which are shared between S1 and S2; the SS2SLS component is estimated using all units which lie exclusively within S1 or S2. Accordingly, the TS2SLS estimator is equivalent to 2SLS for $\rho = 1$ and SS2SLS for $\rho = 0$ when $N_1 = N_2$. Note that these individual estimates may rely on data that is not observed in the practical setting in which we are considering estimators; the expression of a weighted average of estimators primarily serves the purpose of providing a more interpretable and

algebraically convenient starting point for deriving the estimator's properties.

Let $Y_1 \equiv \begin{pmatrix} y_{11} \\ \vdots \\ y_{1N_1} \end{pmatrix} \equiv \begin{pmatrix} Y_{11} \\ Y_{12} \end{pmatrix}$, $\widehat{X}_1 \equiv \begin{pmatrix} \widehat{x}_{11} \\ \vdots \\ \widehat{x}_{1N_1} \end{pmatrix} \equiv \begin{pmatrix} \widehat{z}_{11}\hat{\gamma}_2 \\ \vdots \\ \widehat{z}_{1N_1}\hat{\gamma}_2 \end{pmatrix} \equiv \begin{pmatrix} \widehat{X}_{11} \\ \widehat{X}_{12} \end{pmatrix}$, $\mathbf{Z}_1 \equiv \begin{pmatrix} \mathbf{z}_{11} \\ \vdots \\ \mathbf{z}_{1N_1} \end{pmatrix} \equiv \begin{pmatrix} \mathbf{Z}_{11} \\ \mathbf{Z}_{12} \end{pmatrix}$, $\Xi_1 \equiv \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{N_1} \end{pmatrix} \equiv \begin{pmatrix} \Xi_{11} \\ \Xi_{12} \end{pmatrix}$ and $V_1 \equiv \begin{pmatrix} v_{11} \\ \vdots \\ v_{1N_1} \end{pmatrix} \equiv \begin{pmatrix} V_{11} \\ V_{12} \end{pmatrix}$ be the vectors in S1, where vectors Y_{11}, $\widehat{X}_{11}$, $\mathbf{Z}_{11}$, Ξ_{11}, V_{11} are the first ρN_2 rows, and vectors Y_{12}, $\widehat{X}_{12}$, $\mathbf{Z}_{12}$, Ξ_{12}, V_{12} are the remaining $N_1 - \rho N_2$ rows. Let $X_1 \equiv \begin{pmatrix} x_{11} \\ \vdots \\ x_{1N_1} \end{pmatrix} \equiv \begin{pmatrix} X_{11} \\ X_{12} \end{pmatrix}$, where we observe X_{11} but not X_{12}. Similarly, let $X_2 \equiv \begin{pmatrix} x_{21} \\ \vdots \\ x_{2N_2} \end{pmatrix} \equiv \begin{pmatrix} X_{11} \\ X_{22} \end{pmatrix}$, $\mathbf{Z}_2 \equiv \begin{pmatrix} \mathbf{z}_{21} \\ \vdots \\ \mathbf{z}_{2N_2} \end{pmatrix} \equiv \begin{pmatrix} \mathbf{Z}_{11} \\ \mathbf{Z}_{22} \end{pmatrix}$ and $V_2 \equiv \begin{pmatrix} v_{21} \\ \vdots \\ v_{2N_2} \end{pmatrix} \equiv \begin{pmatrix} V_{11} \\ V_{22} \end{pmatrix}$ be the vectors in S2, where vectors X_{11}, $\mathbf{Z}_{11}$ are the same as the first ρN_2 rows in S1, and vectors X_{22} and $\mathbf{Z}_{22}$ are the remaining $(1-\rho)N_2$ rows in S2.

The TS2SLS estimator is defined by

$$\widehat{\beta}_O = \left(\widehat{X}_1'\widehat{X}_1\right)^{-1}\widehat{X}_1'Y_1.$$

Proposition 1. *By the proof in the appendix, $\widehat{\beta}_O$ can be rewritten as*

$$\begin{aligned} \widehat{\beta}_O &= \left(\widehat{X}_1'\widehat{X}_1\right)^{-1}\left(\widehat{X}_{11}'\widehat{X}_{11}\right)\left(\widehat{X}_{11}'\widehat{X}_{11}\right)^{-1}\left(\widehat{X}_{11}'Y_{11}\right) \\ &\quad + \left(\widehat{X}_1'\widehat{X}_1\right)^{-1}\left(\widehat{X}_{12}'\widehat{X}_{12}\right)\left(\widehat{X}_{12}'\widehat{X}_{12}\right)^{-1}\left(\widehat{X}_{12}'Y_{12}\right) \\ &= \widehat{W}\widehat{\beta}_{2SLS}^{(1)} + \left(1-\widehat{W}\right)\widehat{\beta}_{SS2SLS}^{(2)} \end{aligned} \tag{13}$$

where

$$\hat{W} \equiv \left(\widehat{X}_1'\widehat{X}_1\right)^{-1}\left(\widehat{X}_{11}'\widehat{X}_{11}\right),$$

$$\widehat{\beta}_{2SLS}^{(1)} = \left(\widehat{X}_{11}'\widehat{X}_{11}\right)^{-1}\left(\widehat{X}_{11}'Y_{11}\right),$$

and

$$\widehat{\beta}_{SS2SLS}^{(2)} = \left(\widehat{X}_{12}'\widehat{X}_{12}\right)^{-1}\left(\widehat{X}_{12}'Y_{12}\right).$$

Proposition 2. *The probability limit of $\hat{W}$ as both samples approach infinity is the ratio of the overlap parameter to the asymptotic ratio of sample sizes.*

$$\underset{N_1,N_2\to\infty}{plim}\ \widehat{W} \equiv W = \frac{\rho}{\alpha}.$$

Remark 3. The expected value of $\hat{W}$ can be approximated as follows (see appendix):

$$E\left(\widehat{W}\right) \approx \frac{\rho N_2 \gamma' \Omega_z \gamma + K \cdot \sigma_v^2}{N_1 \gamma' \Omega_z \gamma + K \cdot \sigma_v^2 + (N_1 - \rho N_2)/(1-\rho) N_2 \sigma_v^2}.$$

Proposition 1 formalizes the partition of $\widehat{\beta}_O$ into a weighted average of $\widehat{\beta}_{2SLS}^{(1)}$ and $\widehat{\beta}_{SS2SLS}^{(2)}$. The weights represent the sum of squares used by each estimator relative to the total sum of squares in the entire vector of data. This variation has two components: variation explicitly from instruments Z, and variation from first stage error vectors V which manifests through the estimated first stage coefficients. Asymptotically, the weights are a function of the degree of overlap between samples and the ratio of sample sizes. For $\alpha = 1$, the weights are the ratio of the overlapping sample size to the second-stage sample size.

2.3 First-order bias approximation

There is an expansive literature on the finite-sample bias of IV estimators, with papers such as Nagar (1959), Bekker (1994), Staiger and Stock (1997), and Bun and Windmeijer (2010) offering approximations of various forms. For simplicity, we consider first-order approximations of the bias of 2SLS and SS2SLS and develop a bias approximation for TS2SLS. Hahn and Hausman (2002) offer a simple approximation of the bias of 2SLS, using the product of the inverse expected value of the variance of fitted values to the expected value of the covariance between the fitted values and the outcome. In a scalar case, this is in effect using the ratio of expectations in place of the expected value of the ratio. This approximation corresponds to a first-order Taylor Series expansion of the 2SLS estimator about each of these "numerator" and "denominator" terms. For 2SLS, this approximation sufficiently characterizes the directional responses of bias to sample size, number of instruments, error covariance, and instrument strength. The same holds for SS2SLS, which has a comparable shape of response but has bias characterized entirely by first-stage sampling error.

Proposition 4. *The finite-sample bias for the 2SLS estimator on the overlapping portion of the sample is approximated by a first-order Taylor expansion:*

$$E(\widehat{\beta}^{(1)}_{2SLS} - \beta) \approx \frac{K \cdot \sigma_{\varepsilon v}}{\rho N_2 \gamma_1' \Omega_z \gamma_1 + K \cdot \sigma_v^2}. \tag{14}$$

The approximate finite-sample bias for the SS2SLS estimator on the non-overlap portion of the sample is given by

$$\begin{aligned} E(\widehat{\beta}^{(2)}_{SS2SLS} - \beta) &\approx -\frac{\sigma_v^2 \beta / (1-\rho) N_2}{\gamma' \Omega_{\mathbf{z}} \gamma + \sigma_v^2 / (1-\rho) N_2} \\ &= -\frac{\beta}{(1-\rho) N_2 \gamma' \Omega_{\mathbf{z}} \gamma / \sigma_v^2 + 1} \end{aligned} \tag{15}$$

Equation (14) and equation (15) show that the bias of both the 2SLS estimator and SS2SLS estimator approach zero as N_2 gets large: they are asymptotically unbiased in first stage sample

size. This conforms with the intuition that all finite-sample bias in 2SLS (under valid instruments) originates from first-stage sampling error. Both the direction and magnitude of bias in the 2SLS estimator depends on $\sigma_{\varepsilon v}$, the covariance of error terms representing the endogenous portion of x. SS2SLS has a attenuation bias that is inversely proportional to $(1-\rho)N_2\gamma'\Omega_{\mathbf{z}}\gamma/\sigma_v^2$, the first-stage "concentration parameter," intuitively similar to a bias from measurement error. Note that this attenuation, found similarly in Angrist and Krueger (1995), is dependent on the assumption of a linear conditional expectation function in the first stage. An application of Jensen's inequality to the SS2SLS estimator suggests that the attenuation bias may not generally hold (see appendix).

Proposition 5. *For $0<\rho<1$ and $N_2 \neq N_1$, the finite sample bias of $\widehat{\beta}_O$ is approximated by*

$$\begin{aligned} E\left(\widehat{\beta}_O-\beta\right) &= E\left[\widehat{W}\left(\widehat{\beta}_{2SLS}^{(1)}-\beta\right)+\left(1-\widehat{W}\right)\left(\widehat{\beta}_{SS2SLS}^{(2)}-\beta\right)\right] \\ &\approx \frac{K\cdot\sigma_{\varepsilon v}-\left(\left(N_1-\rho N_2\right)/\left(1-\rho\right)N_2\right)\sigma_v^2\beta}{N_1\gamma'\Omega_z\gamma+K\cdot\sigma_v^2+\left(\left(N_1-\rho N_2\right)/\left(1-\rho\right)N_2\right)\sigma_v^2}. \end{aligned} \tag{16}$$

$\widehat{\beta}_O$ is approximately unbiased when the overlap proportion ρ is set according to the following formula (noting $\theta=\frac{\sigma_{\varepsilon v}}{\sigma_v^2}$ and K is the number of instruments):

For $\frac{N_2}{N_1}<1$ and $\beta<K\theta\frac{N_2}{N_1}$,

$$\rho=\frac{K\theta-\beta\frac{N_1}{N_2}}{K\theta-\beta}\in(0,1),$$

For $\frac{N_2}{N_1}>1$ and $\beta>K\theta\frac{N_2}{N_1}$, set the overlap proportion to

$$\rho=\frac{\beta\frac{N_1}{N_2}-K\theta}{\beta-K\theta}\in(0,1)$$

Proposition 5 follows from a first-order Taylor series approximation of $E(\widehat{\beta}_O)$, derived in the appendix. The second part is shown by setting the numerator in equation (16) to zero and solving for ρ. The net bias of TS2SLS is dependent on the overlap parameter ρ and sampling variation in the first stage parameters estimated from the units used in TS2SLS estimation expressed through the ratio $\hat{W}$. When the combination of β and $\sigma_{\varepsilon v}$ makes the direction of bias for the two estimators

have different signs, the overlap parameterρ can be tuned to yield an approximately unbiased estimator $\widehat{\beta}_O$.

Because of the nature of the approximation used, the approximation performs poorly near $\rho = 1$ and is undefined at $\rho = 1$.With $N_1 \neq N_2$ and $\rho = 1$, the expression evaluates to $-\beta$ after multiplying the numerator and denominator by $(1-\rho)$, suggesting that the first-order approximation does not capture useful properties of this important edge case of TS2SLS. With $N_1 = N_2$ and $\rho = 1$, the expression is undefined but has a defined limit as ρ approaches one:

$$lim_{\rho \to 1, N_1 = N_2} \left(\frac{K \cdot \sigma_{\varepsilon v} - \left((N_1 - \rho N_2) / (1-\rho) N_2 \right) \sigma_v^2 \beta}{N_1 \gamma' \Omega_z \gamma + K \cdot \sigma_v^2 + \left((N_1 - \rho N_2) / (1-\rho) N_2 \right) \sigma_v^2} \right) = \frac{K \cdot \sigma_{\varepsilon v} - \sigma_v^2 \beta}{N_1 \gamma' \Omega_z \gamma + K \cdot \sigma_v^2 + \sigma_v^2}.$$

This edge case is a likely inaccurate approximation, arising as an artifact of the first-order Taylor series approximation. We can intuit that the total overlap ($\rho = 1$) case with $N_1 = N_2$ results in an estimate computationally identical to 2SLS; the two-step nature of 2SLS means that the ability to explicitly link the units used in the steps has no bearing on the estimate if the units used in the steps are indeed the same. This approximation contains all elements of the bias approximation for 2SLS presented in equation (14) but has "artifacts" from the mixture of 2SLS and SS2SLS estimators. Incidentally, with $N_1 = N_2$, we intuit that the no-overlap ($\rho = 0$) case generates a computationally identical estimate to the the "split-sample" estimator that is biased toward zero, a result that would be consistent with Angrist and Krueger (1995).

2.4 Asymptotic Variance of TS2SLS

Proposition 6. *$\widehat{\beta}_O$ is consistent and asymptotically normally distributed with asymptotic variance*

$$\sqrt{N_1 + (1-\rho) N_2} \left(\widehat{\beta}_O - \beta \right) \overset{a}{\sim}$$

$$N \left\{ 0, \frac{(1+\alpha-\rho)\rho}{\alpha^2} \sigma_\varepsilon^2 \left[\Omega_{x\mathbf{z}} \Omega_{\mathbf{z}}^{-1} \Omega_{x\mathbf{z}} \right]^{-1} + \frac{(1+\alpha-\rho)(\alpha-\rho)}{\alpha^2} \left[\Omega_{x\mathbf{z}}' \left[\left(\sigma_\varepsilon^2 + \frac{\alpha-\rho}{1-\rho} \beta' \sigma_v^2 \beta \right) \Omega_{\mathbf{z}} \right]^{-1} \Omega_{x\mathbf{z}} \right]^{-1} \right\}. \tag{17}$$

This result follows from the fact that the limiting distribution of $\widehat{\beta}_O$ is a linear combination of the two estimators $\widehat{\beta}^{(1)}_{2SLS}$ and $\widehat{\beta}^{(2)}_{SS2SLS}$ with weight $\frac{\rho}{\alpha}$ (see appendix for proof).

Then, a natural approximation for $var(\hat{\beta}_o)$ is as follows:

$$var(\hat{\beta}_o) \approx \tilde{var}(\hat{\beta}_o) = (N_1 + (1-\rho)N_2)^{-1} avar(\widehat{\beta}_O)$$

This approximation nests conventional 2SLS, which corresponds to the case where $\alpha = 1$ and $\rho = 1$ under which the asymptotic variance is the conventional 2SLS asymptotic variance (Wooldridge 2010). TS2SLS has an asymptotic variance that accounts for variation in the final estimate $\widehat{\beta}_O$ due to first-stage sampling error, but only due to the error originating in the SS2SLS component. The 2SLS' asymptotic variance treats the first stage parameter as known (Wooldridge 2010), and so any variability from the first stage of the 2SLS component of the TS2SLS estimator is ignored. The presence of $(1-\rho)$ in the stabilizing factor, $(N_1 + (1-\rho)N_2)^{-1}$, implies that increasing first stage observations N_2 has a discounted and diminishing effect on precision relative to increasing second stage observations N_1.

Inoue and Solon (2008) derive the asymptotic variance of TS2SLS, but do so only for the case of independent primary and supplemental samples (i.e., $\rho = 0$). Allowing the samples to generally overlap, we find that the asymptotic variance is decreasing in ρ. The basic intuition underlying this property is that sampling variation in the first stage parameter coming from a secondary, independent sample is additional noise originating with error term v_2 unrelated to the outcome y_1. First stage estimate sampling variation coming from the same units used in second stage estimation (as they would in the typical 2SLS computation implied by $\rho = 1$) have a component with explanatory power for outcome y through the error component of x_1, v_1.

2.5 TS2SLS Under Data Availability Constraints

The traditional motivation behind using TS2SLS is to achieve an 2SLS estimate where all necessary variables are not available in a single sample. A second potential reason to use TS2SLS is that

it may provide lower bias or variance than the best available 2SLS estimator. Define $\hat{\beta}_{2SLS}(N^{'})$ as a 2SLS estimator using $N^{'}$ observations. Define $\hat{\beta}_{TS2SLS,N_2,N_1}$ as the TS2SLS estimator using N_2 observations to estimate the first stage and N_1 observations to estimate the second stage. Suppose a researcher has access to a single-sample 2SLS estimator. Provided model assumptions are met, we propose that a researcher with access to larger supplemental samples for either the first stage or second stage can provide a TS2SLS estimator that outperforms the 2SLS estimator on bias, variance, or both. Using a large supplemental sample for the first-stage in a TS2SLS estimator improves both the bias and variance over using the single-sample 2SLS estimator; using a large supplemental sample for the second stage in a TS2SLS estimator improves the variance. This motivates two conjectures:

Conjecture 7. *There exists a value* $N^{''} > N^{'}$ *such that* $var(\hat{\beta}_{TS2SLS,N^{''},N^{'}}) < var(\hat{\beta}_{2SLS,N^{'}})$ *and* $bias(\hat{\beta}_{TS2SLS,N^{''},N^{'}}) < bias(\hat{\beta}_{2SLS,N^{'}})$.

and

Conjecture 8. *There exists a value* $N^{''} > N^{'}$ *such that* $var(\hat{\beta}_{TS2SLS,N^{'},N^{''}}) < var(\hat{\beta}_{2SLS,N^{'}})$.

Because of the lack of closed-form expressions for both the bias and variance of TS2SLS and 2SLS estimators, these are only conjectures supported by the first-order bias approximations and asymptotic variance presented in sections 2.3 and 2.4. Given the approximations provided, we can only justify two narrower propositions:

Proposition 9. *(Larger first-stage supplemental sample)*

There exists a value $N^{''} > N^{'}$ *such that*
a) $\tilde{var}(\hat{\beta}_{TS2SLS,N^{''},N^{'}}) < \tilde{var}(\hat{\beta}_{2SLS,N^{'}})$ *if* $\rho \neq 1$, *where* $\tilde{var}$ *is the asymptotic variance approximation presented in equation 17 and*
b) $|\tilde{bias}(\hat{\beta}_{TS2SLS,N^{''},N^{'}})| < |\tilde{bias}(\hat{\beta}_{2SLS,N^{'}})|$, *where* $\tilde{bias}$ *is the approximation presented in equation (16).*

and

Proposition 10. *(Larger second-stage/second-stage supplemental sample)*

There exists a value $N^{''} > N^{'}$ *such that* $avar(\hat{\beta}_{TS2SLS,N',N''}) < avar(\hat{\beta}_{2SLS,N'})$.

Proposition 9a. and 10 hold trivially: comparing the two asymptotic variances, we increase either the first-stage or second-stage samples until the variance "penalty" from using TS2SLS is overcome. The following values for N'' (i.e., first stage observations) satisfy 9a.:

$$N'' > \frac{N'}{(1-\rho)} \left[avar(\hat{\beta}_{2SLS}) \right]^{-1} \left(avar(\hat{\beta}_{TS2SLS}) - avar(\hat{\beta}_{2SLS}) \right).$$

At a minimum, N'' must be at least as large as N', noting that $avar(\hat{\beta}_{TS2SLS}) - avar(\hat{\beta}_{2SLS})$ is positive semi-definite. In the limiting case of $\rho = 1$, TS2SLS has no variance penalty relative to 2SLS, and thus TS2SLS would have the same variance provided that $N' = N''$ for either $\hat{\beta}_{TS2SLS,N',N''}$ or $\hat{\beta}_{TS2SLS,N'',N'}$. The TS2SLS variance penalty is maximized for $\rho = 0$, requiring the largest increase in first-stage sample size to achieve equal variance to the single-sample estimator. Similarly, the following values for N'' (i.e., second-stage observations in the context of $\hat{\beta}_{TS2SLS,N',N''}$) satisfy 10:

$$N'' > N' \left[avar(\hat{\beta}_{2SLS}) \right]^{-1} \left(avar(\hat{\beta}_{TS2SLS}) - (1-\rho) avar(\hat{\beta}_{2SLS}) \right).$$

When $\rho = 1$ and $\alpha = 1$, $avar(\hat{\beta}_{TS2SLS}) = avar(\hat{\beta}_{2SLS})$, and unsurprisingly we need only $N'' > N'$ for TS2SLS to achieve lower variance than 2SLS according to the approximation.

Finally, for proposition 9b., we set N'' to a quantity such that

$$\left| \frac{K \cdot \sigma_{\varepsilon v} - \left((N' - \rho N'') / (1-\rho) N'' \right) \sigma_v^2 \beta}{N'' \gamma' \Omega_z \gamma + K \cdot \sigma_v^2 + \left((N' - \rho N'') / (1-\rho) N'' \right) \sigma_v^2} \right| < \left| E(\widehat{\beta}_{2SLS,N'} - \beta) \right| \approx \left| \frac{K \cdot \sigma_{\varepsilon v}}{\rho N' \gamma_1' \Omega_z \gamma_1 + K \cdot \sigma_v^2} \right|.$$

The exact solution depends on the direction and relative magnitude of the bias in each estimator.

3 Simulation Evidence

We conduct a series of Monte Carlo simulations of the model characterized in Section 2.1, setting the following parameters:

$$\beta = 1$$

$$\begin{pmatrix} \varepsilon \\ v \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} .5 & .4 \\ .4 & .5 \end{pmatrix} \right.$$

$$z = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}, K = 2$$

$$z_1 \sim N(0, 2.5), z_2 \sim N(0, 2.5)$$

$$\gamma = \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} = \begin{pmatrix} 0.0316 \\ 0.0316 \end{pmatrix}$$

Figure 7 plots the average parameter estimate across simulation repetitions for different first stage sample sizes N_2, approximating the expected values of 2SLS and TS2SLS estimators. The 2SLS estimator uses $N = N_2$ observations, while the TS2SLS estimator uses two non-overlapping data sets ($\rho = 0$) of $N_1 = 200$ and N_2 observations. In this setup, the OLS estimate is biased upward by about 80%, and the 2SLS estimate is biased toward the OLS estimate, while the TS2SLS estimate is biased towards zero (because it is equivalent to SS2SLS and the simulation uses a linear CEF). For both estimators, as N_2 increases the bias decreases. This simulation illustrates one of the potential bias advantages of using a second, larger sample for the first stage even when a single-sample estimate is available. A TS2SLS estimate using $N_1 = 200$ and $N_2 \geq 400$ is approximately unbiased while the 2SLS estimate with $N = 200$ is biased by around 6%, suggesting the TS2SLS

estimate is preferable (notwithstanding precision).

Figure 8 plots the average standard deviation across simulation repetitions for first stage sample sizes N_2, revealing that the variance of TS2SLS is decreasing in the number of first stage observations but eventually flattening as the first stage becomes precisely estimated while the 2SLS estimator continues to grow in precision (because of additional data to estimate the second stage). Finally, Figure 9 shows the same relationship but allowing N_1 to vary, showing that the TS2SLS standard error decreases with more N_1 at a rate comparable to OLS.

The bias consequences of overlap dovetail with the description of TS2SLS as a linear combination of 2SLS and SS2SLS estimators. Figure 10 plots the mean simulated point estimates for the TS2SLS estimator as a function of the percentage of overlap between samples for $N_1 = N_2 = 200$. Increasing overlap between samples weighs the estimate towards 2SLS, which is biased towards the probability limit of OLS. In this calibration, the sample-varying portion of $\hat{W}$ is small enough such that the weight is nearly the probability limit of $\hat{W}$, $\frac{\rho}{\alpha} = \rho$. The point of zero bias occurs approximately at 57 percent overlap. On the other hand, the effect on estimate variance of overlap is predicted by the asymptotic variance expression in proposition 6. Figure 11 shows a plot of the simulated standard error against the percentage overlap parameter, demonstrating the anticipated downward-sloping relationship. In this setting, 2SLS is always more efficient than SS2SLS using an equivalent number of observations per stage. Overlap results in a mixture of the two estimators, with more observations used in (and thus greater weight placed on) the 2SLS component when the degree of overlap increases.

4 Application

4.1 TS2SLS in Practice: Synthetic Example from Angrist and Evans (1998)

Angrist and Evans (1998) estimate the effect of fertility on parents' labor supply decisions. They use gender composition indicators of a family's first two children as instruments (z) for whether the family has more than two children (x). In one set of estimates, they use Census microdata

to estimate IV regressions of various labor market outcomes (e.g., whether the mother worked, how much she worked, and her labor income) on an indicator for more than two children, using indicators for the first two children being boys and first two children being girls as instruments.[7] We create hypothetical situations in which some data are missing and examine whether the study's findings could have been recreated using TS2SLS estimates under these conditions. The hypothetical TS2SLS estimates mimic what the authors might have needed to run if their data were only available in two samples, as they might have been if the Census had decided to release their first-stage variables in one anonymized microdata set and their second-stage variables in another.

We attempt this exercise for Angrist and Evans' estimates for married women ages 21-35 with 2 or more children. We develop hypothetical scenarios in which the authors receive their data in two data subsets of their true data on $\{y,x,z\}$. We simulate multiple draws of data sets for each estimator in order to isolate the expected value of each estimator across possible data draws the hypothetical data-constrained Angrist and Evans might have faced. For example, we randomly draw two mutually exclusive and exhaustive 25- and 75-percent subsets of the 254,654 observations and compute the SS2SLS estimate with each sample, repeating the procedure a total of 1000 times. In reality, practitioners would face a single draw of data, and sampling variation could result in a pattern of results that does not conform with theoretical predictions.

Table 21 shows a series of estimates for various hypothetical situations in which the authors have restricted access to some part of their data. First, we exactly replicate the estimates from Table 7, columns 4 and 6 (p. 465) in the original paper. As a baseline, we show what the 2SLS estimates would have been had the authors only had access to a 50% subsample of their data. Column 5 shows what the TS2SLS (effectively a split-sample 2SLS estimate) using 50% of their sample for the first stage and 50% for the second stage would have been, had they only had access to x and z in one set and y and z in another (with no overlapping observations). Column 6 shows the estimate if they had access to 100% of their original sample for the first stage, but only 50% for the second

[7]There is some controversy as to the validity of the family gender-composition instruments they use (e.g., Rosenzweig and Wolpin 2000). This exercise takes no position in this debate and only focuses on the comparison of the bias and variance mechanics of the estimators used as they relate to one-sample vs. two-sample 2SLS.

stage. Finally, column 6 shows the TS2SLS estimate if they had 50% of their original sample for the first stage, but 100% of their sample for the second stage (y and z).

The 50% 2SLS estimate counterintuitively moves 0.02 towards zero instead of OLS, but this is plausible given the large standard error (0.039) of the estimate. The 50-50 SS2SLS estimate also moves towards zero with a comparable, slightly larger standard error (0.041), consistent with the variance increase from using non-overlapping units in each stage. The bootstrapped estimate with only 50% subsample for the second stage in column 5 has a standard error roughly equal to the 2SLS and SS2SLS estimates in columns 3 and 4, suggesting that the variance improvement from additional first-stage observations is close to zero. As might be conversely expected, the estimate in the final column, instead with a 50% subsample for the first stage, has a comparable standard error to the full-sample 2SLS estimate in column 2. Recall that the finite sample bias of TS2SLS is the result of competing forces of bias towards OLS (from the overlapping part of the sample) and bias towards zero (from the non-overlapping part of the sample). In themselves, the positions of both the 50% 2SLS and 50% SS2SLS estimates suggest that finite sample bias is not a significant issue; we were able to discard 50% of observations altogether and still not alter the estimate substantially. Once we discard 50% of first-stage observations but use all of our second-stage observations, we find an estimate (-0.114) that is almost exactly the same as the 100% 2SLS estimate (-0.113) with the same standard error (0.028). This example strongly illustrates the potential variance improvements in practice one can get by using TS2SLS. It is more difficult to show any bias improvements in practice, particularly in this example: 2SLS and OLS point estimates are marginally indistinguishable to begin with. Decreasing the first stage sample size to increase bias is somewhat masked by the large standard error, which only increases as more observations are discarded.

One issue worth discussion is the assumption of a linear conditional expectation function (CEF) in the first stage required for the SS2SLS "attenuation bias" result to hold. Because Angrist and Evans' endogenous explanator is a binary variable, whether a mother has more than two children, the linear CEF assumption can potentially be called into question. However, it is important to note

that their main specification consists entirely of binary instruments, binary controls for race, and only two linear terms for mother's age (at Census collection and at first birth). The CEF is linear in saturated models–which increase by one parameter per potential combination of predictors–and the use of almost exclusively binary predictors makes the main specification already close to saturation. Indeed, fewer than 0.02% of observations have a predicted probability that lies outside the [0,1] interval, consistent with (though not sufficient for) a linear CEF. As a secondary check, we re-run the first stage using a fully saturated set of interactions among all predictors (with age categorical variables fully converted to binary variables) and re-compute fitted values. We find that 90 percent of predicted probabilities in the saturated model are within 0.05 of the main specification model, compared to an overall probability of approximately 0.38 of a mother having more than two children. We also expect some differences in individual predictions due to random chance, as the saturated model adds thousands of additional parameters, many of which are estimated from sparsely populated cells. To the extent that the saturated and main specifications produce statistically indistinguishable results, the assumption of linear CEF for the main specification becomes more plausible.

4.2 Other Considerations for Applications

There are several practical applications of two-sample estimators. The typical application in empirical literature using two-sample estimators is when the researcher only has access to one data set with the outcome and instrument and one data set with the endogenous variable and instrument; in this case, the researcher's only option to generate an IV estimate is to use a TSIVM. The second application, strongly suggested by the finite-sample findings here, is when researchers have access to a single-sample IV estimator but have additional observations usable to estimate the first-stage (endogenous variable and instrument) or second-stage (outcome and instrument) relationships.

The simulation evidence and synthetic working example from Angrist and Evans (1998) suggest that reporting a TS2SLS estimate may be the most rewarding when practitioners have access to additional second-stage observations. The TS2SLS estimate using all available observations

could then be substantially more precise without a need for having equal numbers of observations for the first-stage and second-stage samples. In that case, the strength of the instrument is likely to be established, and practitioners need only justify that their additional observations come from a population with the same first-stage projection coefficients. Even when that justification is lacking, the two-sample estimates can simply be presented alongside the single-sample estimates as additional evidence for a causal hypothesis, allowing the reader to decide the weight of evidence to assign to the TS2SLS estimate. A secondary case of interest is when practitioners have a single sample 2SLS estimate with weak instruments (i.e., an imprecisely-estimated first stage) but have access to additional observations with which to estimate the first stage. In this case, the TS2SLS estimate can potentially "solve" a weak instruments problem, but then the justification of first-stage coefficient equivalence between the two samples is of much greater importance, since the primary causal inference in the paper comes from the TS2SLS estimate.

In practice, these situations occur because of real-world limitations on the way data can be collected. Data on the endogenous variable may be costly to collect and thus limited (e.g., environmental monitoring of pollutants, accurate personal income measures), but sufficient to estimate a strong, representative first stage relationship with an instrument. With weak instruments bias not a concern, a study could then be scaled to have adequate power only through the expansion of data collection on the second-stage variables. This insight is also reflected in the findings of Pierce and Burgess (2013), who focus in particular on "Mendelian Randomization" designs, which use genetic factors as instruments for biological exposures. They conclude that it is possible to collect only a small first-stage sample and achieve comparable results to a "complete-data" design with equal first-stage and second-stage observations.

Practical issues with data linkage also make TSIVM potentially useful, especially where the providers of data can perfectly manipulate what data is available to the public. The complete interchangeability of two-sample estimators with one-sample estimators when the data sampling process is exactly known is a useful property for confidentiality applications. Consider a situation in which the combination of information on y, x, and z for subjects in a randomized study could

allow anonymous subjects to be identified, but providing information on just x and z or just y and z reduce that risk significantly. Data releases could then offer two data sets with those elements separate and unlinkable, but researchers could still generate IV estimates using TSIVM at little cost to their inferences. In this setting, two of the greatest sources of uncertainty in the validity of TS2SLS estimates are obviated: the populations constituting each of the samples are known to be equivalent and the level of overlap between the primary and secondary samples is generated by the study designers and passed on to the practitioner.

Lastly, as a generalization of split-sample IV methods, the results presented in this paper could also be used as part of an "eyeball test" for weak instruments. Practitioners can subdivide their samples arbitrarily and run a series of TS2SLS estimates using the subsamples, examining the sensitivity of their coefficient estimates to smaller first stage sample sizes or varying levels of overlap. This methodology may be superseded by other weak-instruments tests or bias-robust IV estimation methods such as Jackknife IV (Angrist et al. 1999), but it is also easy to implement and intuitively present to an audience using plots of estimates across researcher-manipulated overlap or sample size parameters.

4.3 The Practical Impact of Sample Overlap ρ

The overlap parameter ρ may in general be unknown and may have an impact on the performance of a TS2SLS estimator. We might imagine having a small sample of complete cases with which to estimate a parameter via 2SLS, but wish to get an improved estimate using TS2SLS using an supplemental sample of additional first-stage or second-stage observations. For example, we may have complete data to estimate a treatment effect for participants in a longitudinal survey like NLSY79, but wish to supplement with a significantly larger sample from the CPS or Census microdata. If we use supplemental data for the first stage, then the overlap moves toward zero. If we use supplemental data for the second stage, then the overlap approaches one (but for large enough populations the influence of the overlap is diminished).

Recall that ρ represents the fraction of units in the first-stage sample which also appear in the

second-stage sample. Suppose we have a sample of size N_1 with which to estimate the second stage, and can achieve a larger sample of size N_2 with which to estimate the first stage; then, as N_2 approaches the size of the population, ρ approaches its theoretical minimum. For infinitely large populations, this minimum is zero. In terms of the partitioning of the TS2SLS estimator into 2SLS and SS2SLS estimators on overlapping and non-overlapping units (per Proposition 1), this implies that growing N_2 with fixed N_1 results in the 2SLS component of the estimator decreasing towards zero, moving the TS2SLS estimator toward a SS2SLS estimator. However, it is also important to note that growing N_2 also results in shrinking finite sample bias of the TS2SLS estimator, obviating the role of the overlap parameter in bias. Furthermore, as N_2 becomes large, the variance of the TS2SLS estimator approaches the variance of the OLS estimator implied by regressing y on $z\gamma$ with N_1 observations.

Alternatively, we might attain a larger sample of size N_1 for estimating the second stage, while holding our sample for estimating the first stage fixed at size N_2, resulting in a ρ approaching one as N_1 approaches the size of the population. In this case, the TS2SLS estimator is expressed as a weighted combination of a SS2SLS estimator on $N_1 - N_2$ non-overlapping units and a 2SLS estimator on N_2 overlapping units. The 2SLS component on N_2 units will have impact approaching zero as the ratio N_1/N_2 grows larger. Intuitively, if the population size is large relative to the first-stage sample being held fixed, then the implied overlap will have minimal impact on the estimates.

It is possible in some circumstances to estimate ρ or bound its potential impact on TS2SLS estimates. If the sampling schemes and populations of two samples are known, the probability a unit in the first-stage sample is also in the second-stage sample can be estimated. For example, in the case of simple random sampling from the same population, ρ is simply the product of probabilities of an individual's inclusion into each sample. We can approximately bound the impact of ρ on variance by computing the asymptotic variance with $\rho = 0$ and $\rho = 1$. Bias is more challenging to address, given that the true bias is a nonlinear function of the unobserved covariance between the endogenous explanator and the error and the overlap parameter ρ. However, given that the limiting case of $\rho = 1$ corresponds to 2SLS and $\rho = 0$ corresponds to SS2SLS, and one

is confident to have approximately identified the probability limit of OLS, one should expect the TS2SLS estimate to lie somewhere between zero and what the single-sample 2SLS estimate would have been (and correspondingly between zero and the OLS estimate).

4.4 Computation of Standard Errors

As with 2SLS, the classic "generated regressors" problem results in inaccurate inference with a tendency towards overrejection of null hypotheses on β . Standard errors on $\hat{\beta}_{TS2SLS}$ can be estimated in one of two ways: using the asymptotic variance-based approximation in Section 2.4 or via a bootstrap method. The asymptotic variance formula captures the decreasing relationship between variance and ρ. Thus, because of uncertainty about the value of ρ in a given application, a conservative prescription is to compute standard errors using the formula with ρ set to zero. In practice, many data combination scenarios will either have have relatively small ρ or the impact of ρ will be diminished by a strong first stage.

As noted in 2.4, the asymptotic variance does not account for sampling variation in the first-stage parameters generated by the overlapping units (the "2SLS component" of the partitioned TS2SLS expression from Proposition 1). An alternative approach to account for generated regressors is to use a bootstrap procedure, resampling both first-stage and second-stage samples and estimating a new β_{TS2SLS} in each bootstrap repetition. The resulting distribution of estimates has a standard deviation that approximates the standard error.

5 Conclusion

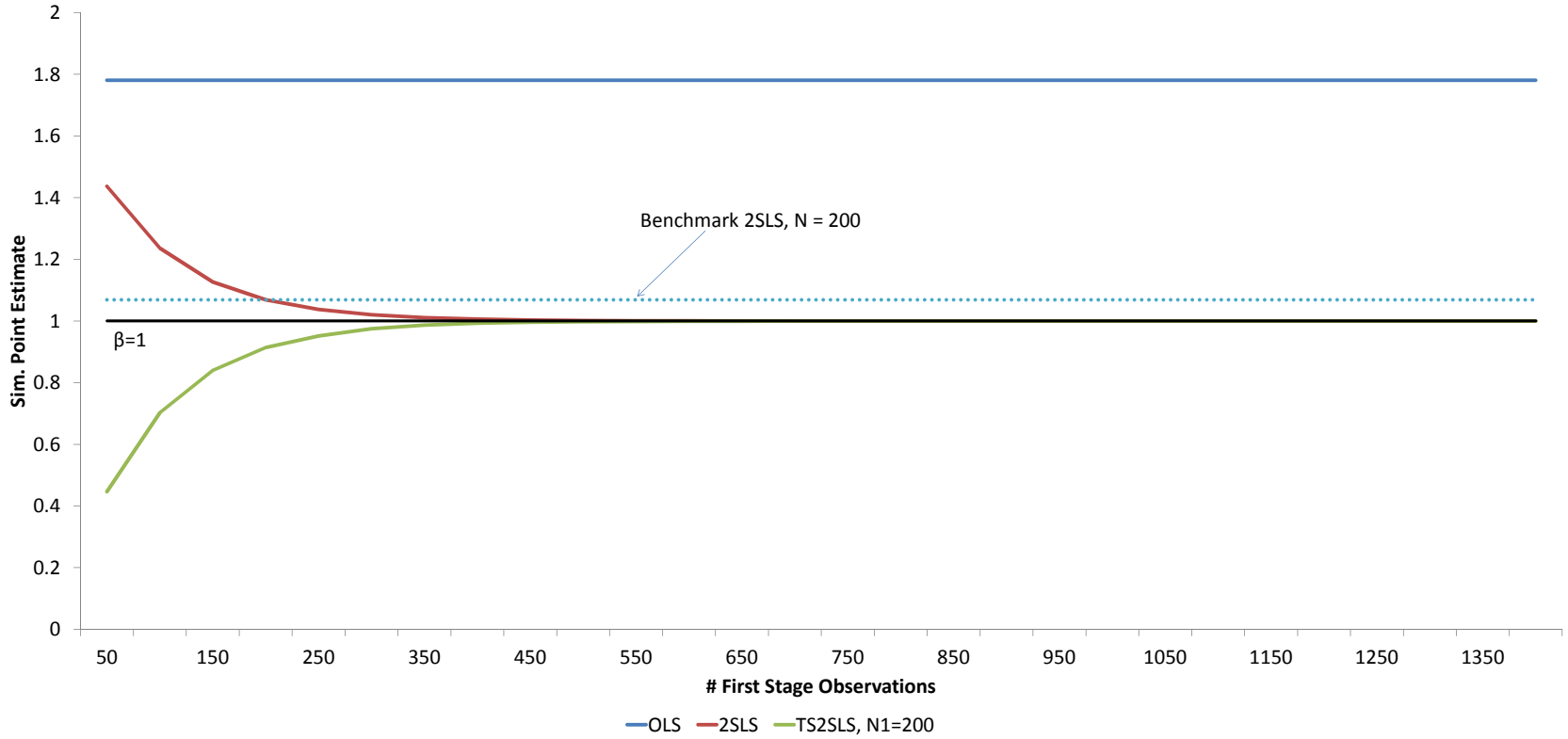
This paper introduces new considerations for applied researchers using TS2SLS. First, we offer a new way of expressing the TS2SLS estimator as a weighted average of 2SLS and SS2SLS estimators using the underlying sample units. The weights structurally depend on the degree to which the researcher's supplemental sample overlap with those in the primary sample, overlapping units used in the 2SLS component and only non-overlapping units used in the SS2SLS component. A

first-order approximation characterizes the behavior of the TS2SLS estimator in finite samples in conformity with the typical intuition from the weak instruments literature for 2SLS: the magnitude of bias is always related to the degree of sampling error in the first stage parameter estimate. However, the bias is pulled in two competing directions: toward zero to the extent that the two samples are non-overlapping, and toward the probability limit of OLS to the extent that the two samples are overlapping. We also show that the variance of the TS2SLS estimator is decreasing in both first-stage and second-stage sample sizes as well as the degree of overlap. We replicate our theoretical findings using both simulation methods and an empirical example from Angrist and Evans (1998), noting that TS2SLS can perform as well as 2SLS with equivalent sample size when instruments are strong.

These results show the potential for TS2SLS to be useful in empirical studies beyond its traditional use as a solution to missing data, while also suggesting that TS2SLS estimates should be interpreted with caution when samples have an unknown level of overlap and instruments are not strong. Researchers may have access to many different data sets with multiple options for how to combine them into IV estimates, also potentially having access to both single-sample and two-sample estimates. Any uses of TS2SLS, whether to present alone or alongside single-sample estimates, must carefully consider the populations represented by each sample.

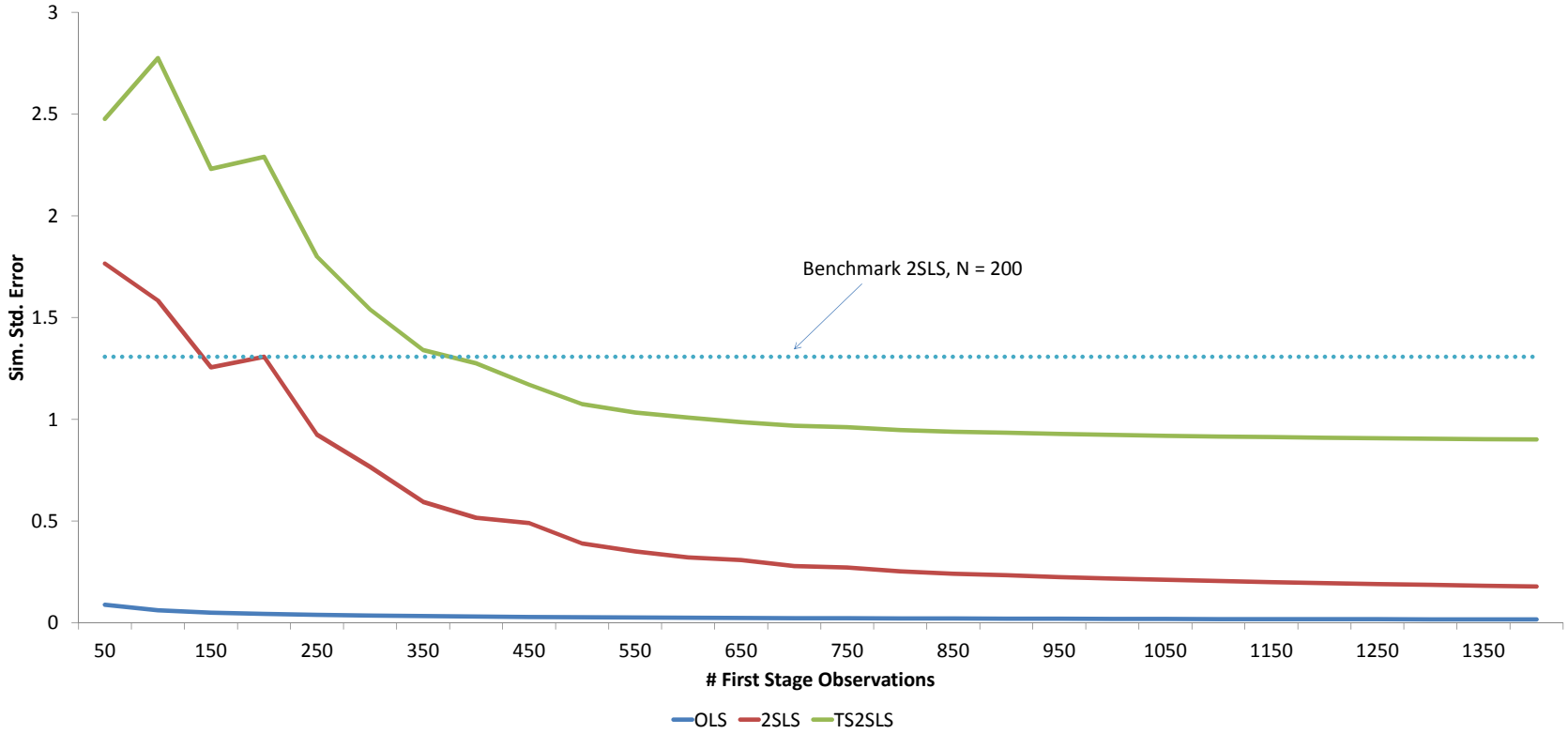
APPENDIX

Figure 7: Mean Simulated TS2SLS Point Estimate by First Stage Sample Size N_2



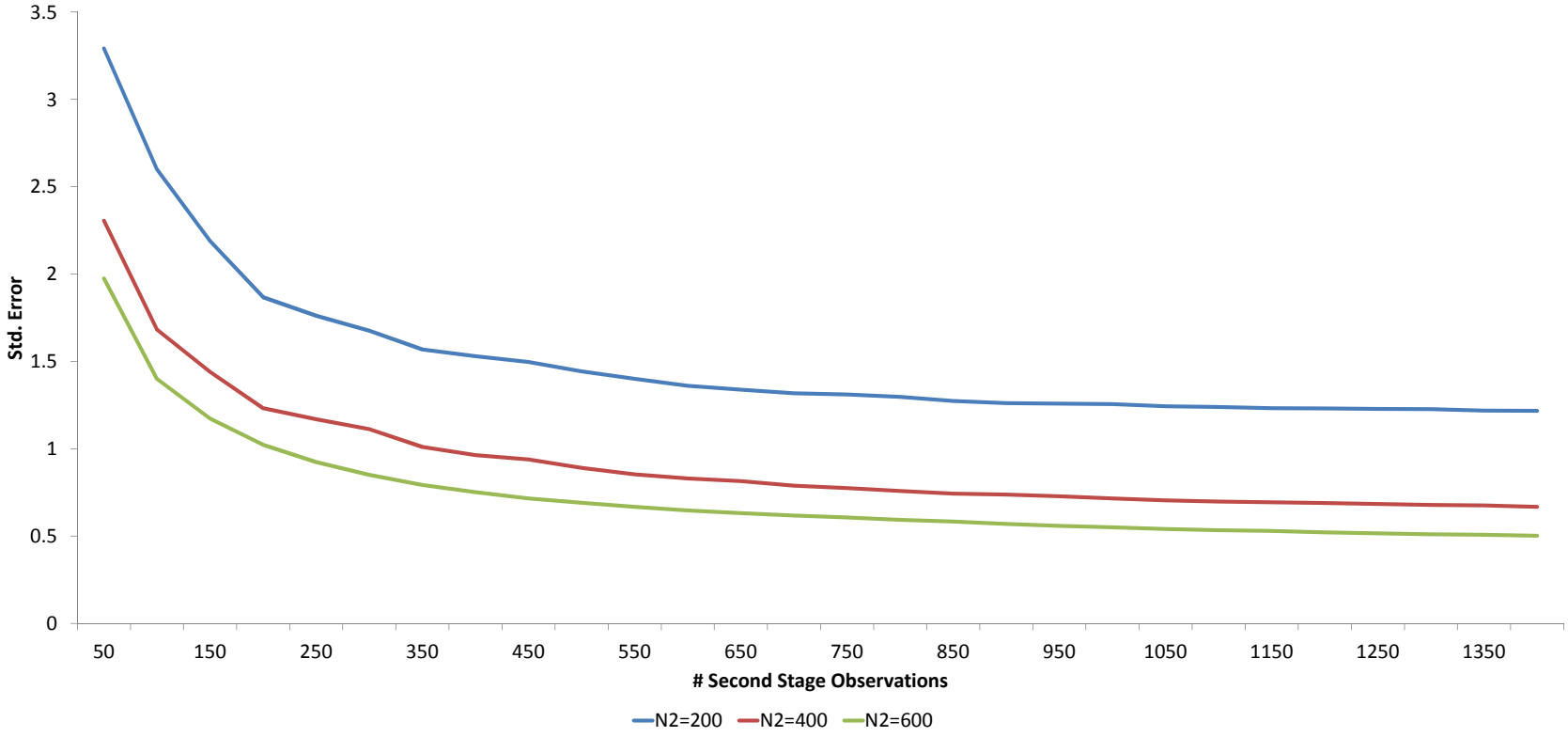
This is a plot of mean simulated coefficients of three estimators: OLS, 2SLS, and TS2SLS. OLS and 2SLS have their sample sizes grow according to the horizontal axis. TS2SLS holds the second-stage sample at N1 = 200, but the first stage sample size grows according to the horizonal axis. The degree of overlap is held fixed at $\rho = 0$.

Figure 8: Simulated TS2SLS Standard Error by First Stage Sample Size N_2



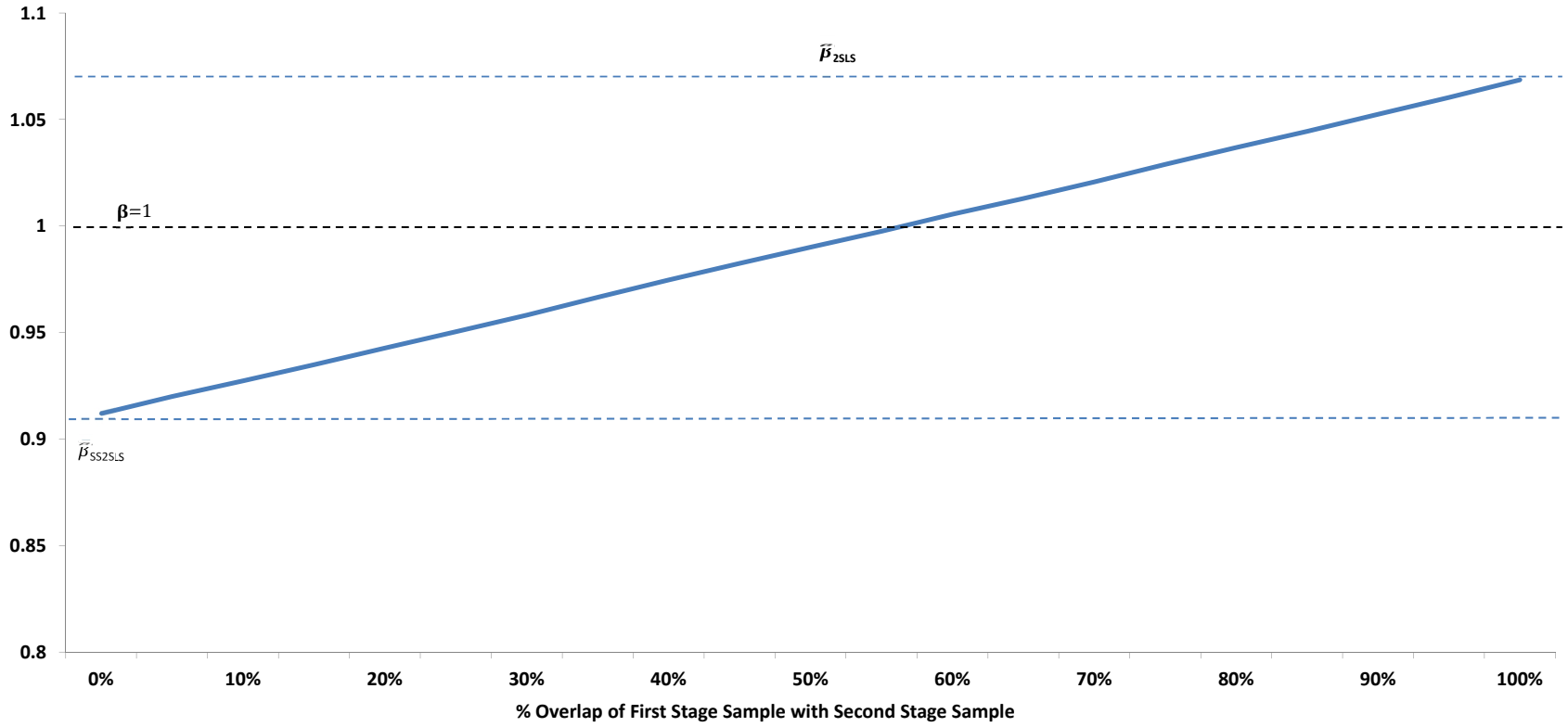
This is a plot of mean simulated standard errors of three estimators: OLS, 2SLS, and TS2SLS. OLS and 2SLS have their sample sizes grow according to the horizontal axis. TS2SLS holds the second-stage sample at N1 = 200, but the first stage sample size grows according to the horizonal axis. The degree of overlap is held fixed at $\rho = 0$.

Figure 9: Simulated TS2SLS Standard Error by second-stage Sample Size N_1



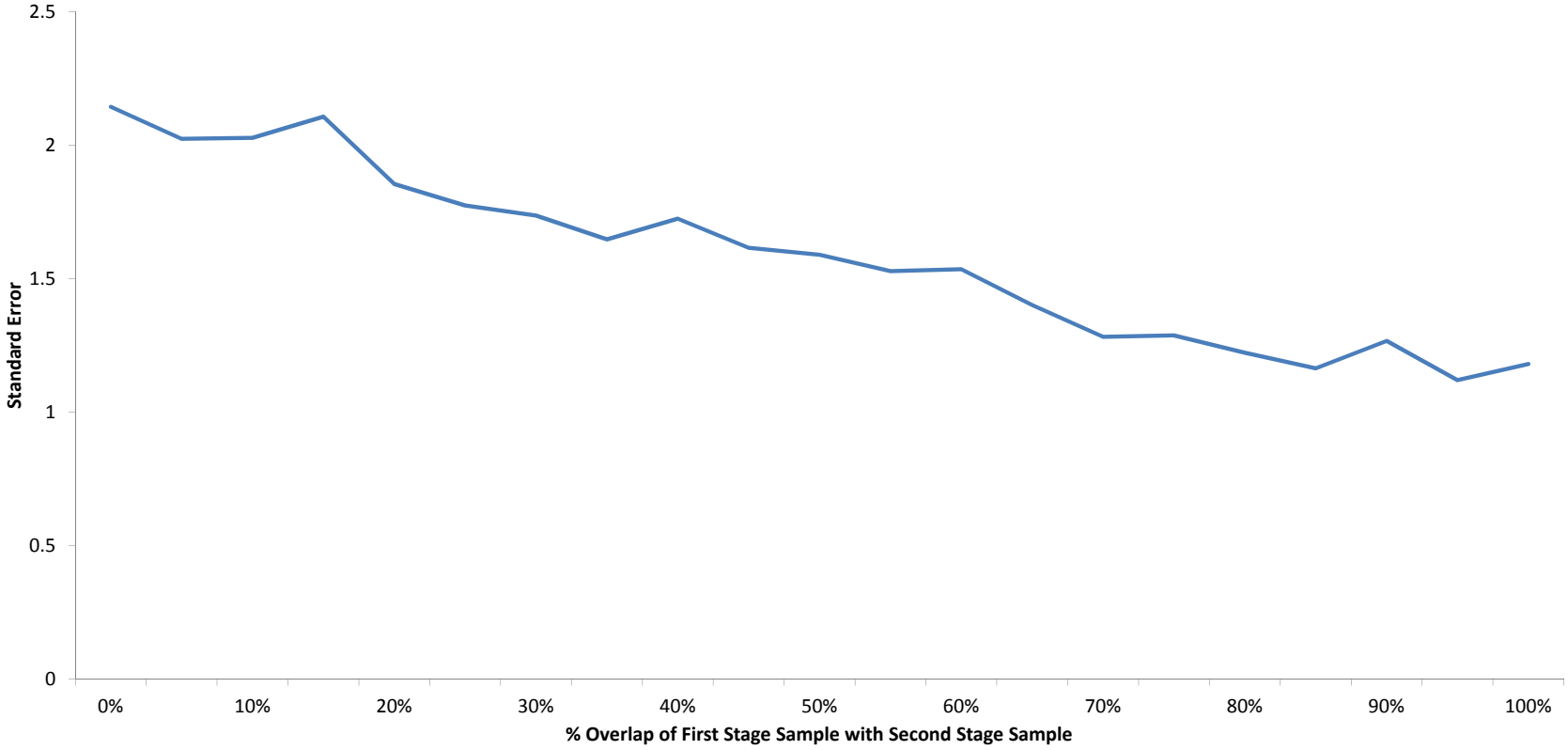
This is a plot of mean simulated standard errors of three estimators: OLS, 2SLS, and TS2SLS. OLS and 2SLS have their sample sizes grow according to the horizontal axis. Three different TS2SLS estimates are plotted holding the first stage sample sample sizes at N2 = 200, 400, and 600 with the second-stage sample size N1 growing according to the horizonal axis. The degree of overlap is held fixed at $\rho = 0$.

Figure 10: Mean Simulated TS2SLS Point Estimate by Proportion of Overlap Between Samples



This graph plots the mean simulated coefficient estimates for the TS2SLS estimator holding sample sizes fixed but varying the degree of overlap between the two samples (the percentage of units contained in first stage sample of size N2 that are also contained in the second sample). Note that the lower and upper extremes correspond to estimates computationally identical to SS2SLS and 2SLS estimators, respectively.

Figure 11: TS2SLS Simulated Standard Error by Proportion of Overlap Between Samples



This graph plots the mean simulated standard error for the TS2SLS estimator holding sample sizes fixed but varying the degree of overlap between the two samples (the percentage of units contained in first stage sample of size N2 that are also contained in the second sample).

Table 21: Hypothetical Two-Sample Estimates for Angrist and Evans (1998), Effects of 2 or More Children on Labor Supply for Married Women, 21-35

	Original Estimates		Split-Sample Estimates (Bootstrapped)			
	OLS	**2SLS**	**50% Sample 2SLS**	**50-50 SS/TS2SLS**	**50% Sample + 100% First Stage**	**50% Sample + 100% Reduced-Form**
Worked?	-0.167	-0.113	-0.086	-0.092	-0.091	-0.114
	(0.002)	(0.028)	(0.039)	(0.041)	(0.041)	(0.028)
# Weeks Worked	-8.043	-5.164	-4.461	-5.037	-4.981	-5.242
	(0.086)	(1.156)	(1.694)	(1.716)	(1.694)	(1.158)
# Hours/Week	-6.021	-4.613	-3.678	-2.972	-2.938	-4.682
	(0.075)	(1.023)	(1.412)	(1.521)	(1.5)	(1.038)
Labor Income	-3165.4	-1321.2	-836.5	-567.6	-569.7	-1344.3
	(39.41)	(550)	(793.53)	(819.14)	(813.72)	(550.03)
N1	254,654	254,654	127,327	127,327	127,327	254,654
N2	--	--	--	127,327	254,654	127,327

Huber-white robust standard errors in parenthesis. This table presents the replicated estimates from Angrist and Evans (1998) for the effect of 2 or more children on labor supply for married women aged 21-35 alongside a series of hypothetical downsampled 2SLS and TS2SLS estimates.

Proof for proposition 1

Under the regularity condition that the data matrix $\widehat{X}_1'\widehat{X}_1$ is nonsingular w.p.a. 1,

$$\begin{aligned}
\widehat{\beta}_O &= \left(\widehat{X}_1'\widehat{X}_1\right)^{-1}\widehat{X}_1'Y_1 \\
&= \left(\widehat{X}_1'\widehat{X}_1\right)^{-1}\left[\left(\widehat{X}_{11}',\widehat{X}_{12}'\right)\begin{pmatrix}Y_{11}\\ Y_{12}\end{pmatrix}\right] \\
&= \left(\widehat{X}_1'\widehat{X}_1\right)^{-1}\left(\widehat{X}_{11}'Y_{11}+\widehat{X}_{12}'Y_{12}\right) \\
&= \left(\widehat{X}_1'\widehat{X}_1\right)^{-1}\left(\widehat{X}_{11}'\widehat{X}_{11}\right)\left(\widehat{X}_{11}'\widehat{X}_{11}\right)^{-1}\left(\widehat{X}_{11}'Y_{11}\right) \\
&\quad +\left(\widehat{X}_1'\widehat{X}_1\right)^{-1}\left(\widehat{X}_{12}'\widehat{X}_{12}\right)\left(\widehat{X}_{12}'\widehat{X}_{12}\right)^{-1}\left(\widehat{X}_{12}'Y_{12}\right)
\end{aligned}$$

Denote $W \equiv \left(\widehat{X}_1'\widehat{X}_1\right)^{-1}\left(\widehat{X}_{11}'\widehat{X}_{11}\right)$, $1-W \equiv \left(\widehat{X}_1'\widehat{X}_1\right)^{-1}\left(\widehat{X}_{12}'\widehat{X}_{12}\right)$, and that

$$\begin{aligned}
\widehat{\beta}_{2SLS}^{(1)} &= \left(\widehat{X}_{11}'\widehat{X}_{11}\right)^{-1}\left(\widehat{X}_{11}'Y_{11}\right), \\
\widehat{\beta}_{SS2SLS}^{(2)} &= \left(\widehat{X}_{12}'\widehat{X}_{12}\right)^{-1}\left(\widehat{X}_{12}'Y_{12}\right),
\end{aligned}$$

from which $\widehat{\beta}_O = \widehat{W}\widehat{\beta}_{2SLS}^{(1)} + \left(1-\widehat{W}\right)\widehat{\beta}_{SS2SLS}^{(2)}$ follows.

Proof for proposition 2

Decompose $\widehat{W}$ into the following block vectors,

$$\begin{aligned}
\widehat{W} &= \left(\widehat{X}_1'\widehat{X}_1\right)^{-1}\left(\widehat{X}_{11}'\widehat{X}_{11}\right) \\
&= \left(\widehat{\gamma}_1'\mathbf{Z}_{11}'\mathbf{Z}_{11}\widehat{\gamma}_1+\widehat{\gamma}_2'\mathbf{Z}_{12}'\mathbf{Z}_{12}\widehat{\gamma}_2\right)^{-1}\left(\widehat{\gamma}_1'\mathbf{Z}_{11}'\mathbf{Z}_{11}\widehat{\gamma}_1\right) \\
&= \left[\rho N_2\widehat{\gamma}_1'\frac{\mathbf{Z}_{11}'\mathbf{Z}_{11}}{\rho N_2}\widehat{\gamma}_1+(N_1-\rho N_2)\widehat{\gamma}_2'\frac{\mathbf{Z}_{12}'\mathbf{Z}_{12}}{N_1-\rho N_2}\widehat{\gamma}_2\right]^{-1}\left(\rho N_2\widehat{\gamma}_1'\frac{\mathbf{Z}_{11}'\mathbf{Z}_{11}}{\rho N_2}\widehat{\gamma}_1\right)
\end{aligned}$$

By assumption 1.(a) $\text{plim}\frac{\mathbf{Z}_{11}'\mathbf{Z}_{11}}{\rho N_2} = E(\mathbf{z}_{1i}'\mathbf{z}_{1i}) = \text{plim}\ \frac{\mathbf{Z}_{12}'\mathbf{Z}_{12}}{N_1-\rho N_2} = E(\mathbf{z}_{2i}'\mathbf{z}_{2i}) = \Omega_{\mathbf{z}}$, and assumption

1.(b), $\text{plim}\,\widehat{\gamma}_1 = \text{plim}\,\widehat{\gamma}_2 = \gamma$. Therefore, by Slutsky's theorem,

$$\begin{aligned}
\text{plim}\widehat{W} &= \left[\rho N_2 \gamma' \Omega_{\mathbf{z}} \gamma_1 + (N_1 - \rho N_2)\gamma' \Omega_{\mathbf{z}} \gamma_1\right]^{-1} \left(\rho N_2 \gamma' \Omega_{\mathbf{z}} \gamma_1\right) \\
&= \left[\rho N_2 + (N_1 - \rho N_2)\right]^{-1} \rho N_2 \left(\gamma' \Omega_{\mathbf{z}} \gamma_1\right)^{-1} \left(\gamma' \Omega_{\mathbf{z}} \gamma_1\right) \\
&= \frac{\rho}{\alpha}.
\end{aligned}$$

Derivation of Remark 3

Use the first order Taylor expansion of $\widehat{W}$ at the point of $\left[E\left(\widehat{X}_{11}'\widehat{X}_{11}\right), E\left(\widehat{X}_{12}'\widehat{X}_{12}\right)\right]$

$$\begin{aligned}
E\left(\widehat{W}\right) &= E\left[\left(\widehat{X}_{11}'\widehat{X}_{11} + \widehat{X}_{12}'\widehat{X}_{12}\right)^{-1}\left(\widehat{X}_{11}'\widehat{X}_{11}\right)\right] \\
&\equiv E\left[g\left(\widehat{X}_{11}'\widehat{X}_{11}, \widehat{X}_{12}'\widehat{X}_{12}\right)\right] \\
&\approx E\left\{\left[E\left(\widehat{X}_{11}'\widehat{X}_{11}\right) + E\left(\widehat{X}_{12}'\widehat{X}_{12}\right)\right]^{-1} E\left(\widehat{X}_{11}'\widehat{X}_{11}\right) + \frac{E\left(\widehat{X}_{12}'\widehat{X}_{12}\right)\left[\widehat{X}_{11}'\widehat{X}_{11} - E\left(\widehat{X}_{11}'\widehat{X}_{11}\right)\right]}{E\left(\widehat{X}_{11}'\widehat{X}_{11}\right) + E\left(\widehat{X}_{12}'\widehat{X}_{12}\right)}\right. \\
&\quad \left. - \frac{E\left(\widehat{X}_{11}'\widehat{X}_{11}\right)\left[\widehat{X}_{12}'\widehat{X}_{12} - E\left(\widehat{X}_{12}'\widehat{X}_{12}\right)\right]}{\left[E\left(\widehat{X}_{11}'\widehat{X}_{11}\right) + E\left(\widehat{X}_{12}'\widehat{X}_{12}\right)\right]^2}\right\} \\
&= \left[E\left(\widehat{X}_{11}'\widehat{X}_{11}\right) + E\left(\widehat{X}_{12}'\widehat{X}_{12}\right)\right]^{-1} E\left(\widehat{X}_{11}'\widehat{X}_{11}\right)
\end{aligned}$$

The numerator is equal to

$$\begin{aligned}
E\left(\widehat{X}_{11}'\widehat{X}_{11}\right) &= E\left(X_{11}' P_{\mathbf{Z}_{11}} X_{11}\right) \\
&= E\left[\left(\gamma' \mathbf{Z}_{11}' + V_{11}'\right) P_{\mathbf{Z}_{11}} \left(\mathbf{Z}_{11}\gamma + V_{11}\right)\right] \\
&= E\left(\gamma' \mathbf{Z}_{11}' \mathbf{Z}_{11} \gamma\right) + E\left(V_{11}' P_{\mathbf{Z}_{11}} V_{11}'\right) \\
&= \rho N_2 \gamma' \Omega_z \gamma + K \cdot \sigma_v^2.
\end{aligned}$$

where $P_{\mathbf{Z}_{11}} \equiv \mathbf{Z}_{11}\left(\mathbf{Z}_{11}'\mathbf{Z}_{11}\right)^{-1}\mathbf{Z}_{11}'$ is the projection matrix. By assumption 1, $E(\mathbf{z}_{1i}'\mathbf{z}_{1i}) = \Omega_{\mathbf{z}}$. Since $rank\left(P_{\mathbf{Z}_{11}}\right) = K$, w.p.a 1, we have $\frac{V_{11}'}{\sigma_v} P_{\mathbf{Z}_{11}} \frac{V_{11}}{\sigma_v} \sim \chi_K^2$, and $E\left(\chi_K^2\right) = K$ for the last equality.

Another component of the denominator is

$$\begin{aligned}
E\left(\widehat{X}_{12}'\widehat{X}_{12}\right) &= E\left\{\left[\mathbf{Z}_{12}\gamma+\mathbf{Z}_{12}\left(\mathbf{Z}_{22}'\mathbf{Z}_{22}\right)^{-1}\mathbf{Z}_{22}'V_{22}\right]'\left[\mathbf{Z}_{12}\gamma+\mathbf{Z}_{12}\left(\mathbf{Z}_{22}'\mathbf{Z}_{22}\right)^{-1}\mathbf{Z}_{22}'V_{22}\right]\right\} \\
&= E\left(\gamma'\mathbf{Z}_{12}'\mathbf{Z}_{12}\gamma\right)+E\left(\gamma'\mathbf{Z}_{12}'\mathbf{Z}_{12}\left(\mathbf{Z}_{22}'\mathbf{Z}_{22}\right)^{-1}\mathbf{Z}_{22}'V_{22}\right) \\
&\quad +E\left(V_{22}'\mathbf{Z}_{22}\left(\mathbf{Z}_{22}'\mathbf{Z}_{22}\right)^{-1}\mathbf{Z}_{12}'\mathbf{Z}_{12}\gamma\right) \\
&\quad +E\left(V_{22}'\mathbf{Z}_{22}\left(\mathbf{Z}_{22}'\mathbf{Z}_{22}\right)^{-1}\mathbf{Z}_{12}'\mathbf{Z}_{12}\left(\mathbf{Z}_{22}'\mathbf{Z}_{22}\right)^{-1}\mathbf{Z}_{22}'V_{22}\right) \\
&= (N_1-\rho N_2)\gamma'\Omega_{\mathbf{z}}\gamma \\
&\quad +E\left(V_{22}'\mathbf{Z}_{22}\left(\mathbf{Z}_{22}'\mathbf{Z}_{22}\right)^{-1}\mathbf{Z}_{12}'\mathbf{Z}_{12}\left(\mathbf{Z}_{22}'\mathbf{Z}_{22}\right)^{-1}\mathbf{Z}_{22}'V_{22}\right)
\end{aligned}$$

The third equality follows because $\mathbf{Z}_{12}$ and $\mathbf{Z}_{22}$ are independent and by Assumption 5,

$$E\left(V_{22}|\mathbf{Z}_{22}\right)=0.$$

Because $E\left(V_{22}'\mathbf{Z}_{22}\left(\mathbf{Z}_{22}'\mathbf{Z}_{22}\right)^{-1}\mathbf{Z}_{12}'\mathbf{Z}_{12}\left(\mathbf{Z}_{22}'\mathbf{Z}_{22}\right)^{-1}\mathbf{Z}_{22}'V_{22}\right)$ is a scalar,

$$\begin{aligned}
& E\left(V_{22}'\mathbf{Z}_{22}\left(\mathbf{Z}_{22}'\mathbf{Z}_{22}\right)^{-1}\mathbf{Z}_{12}'\mathbf{Z}_{12}\left(\mathbf{Z}_{22}'\mathbf{Z}_{22}\right)^{-1}\mathbf{Z}_{22}'V_{22}\right) \\
&= E\left[tr\left(V_{22}'\mathbf{Z}_{22}\left(\mathbf{Z}_{22}'\mathbf{Z}_{22}\right)^{-1}\mathbf{Z}_{12}'\mathbf{Z}_{12}\left(\mathbf{Z}_{22}'\mathbf{Z}_{22}\right)^{-1}\mathbf{Z}_{22}'V_{22}\right)\right] \\
&= E\left[tr\left(\mathbf{Z}_{22}'V_{22}V_{22}'\mathbf{Z}_{22}\left(\mathbf{Z}_{22}'\mathbf{Z}_{22}\right)^{-1}\mathbf{Z}_{12}'\mathbf{Z}_{12}\left(\mathbf{Z}_{22}'\mathbf{Z}_{22}\right)^{-1}\right)\right] \\
&= E\left\{tr\left[\mathbf{Z}_{22}'E\left(V_{22}V_{22}'|\mathbf{Z}_{22}\right)\mathbf{Z}_{22}\left(\mathbf{Z}_{22}'\mathbf{Z}_{22}\right)^{-1}\mathbf{Z}_{12}'\mathbf{Z}_{12}\left(\mathbf{Z}_{22}'\mathbf{Z}_{22}\right)^{-1}\right]\right\} \\
&= E\left\{tr\left[\mathbf{Z}_{12}'\mathbf{Z}_{12}\left(\mathbf{Z}_{22}'\mathbf{Z}_{22}\right)^{-1}\right]\right\}\sigma_v^2 \\
&= tr\left\{E\left[\mathbf{Z}_{12}'\mathbf{Z}_{12}\left(\mathbf{Z}_{22}'\mathbf{Z}_{22}\right)^{-1}\right]\right\}\sigma_v^2 \\
&\approx \frac{(N_1-\rho N_2)}{(1-\rho)N_2}\sigma_v^2
\end{aligned}$$

The third equality follows by the law of iterated expectations. The independence between $\mathbf{Z}_{12}$ and $\mathbf{Z}_{22}$ and Assumption 1.(a) and 5 enables us to pull $E\left(V_{22}V_{22}'|\mathbf{Z}_{22}\right)=E\left(V_{22}V_{22}'\right)=\sigma_v^2 I_n$ out of the expectation. The last equality follows because $\mathbf{Z}_{12}'\mathbf{Z}_{12}$ is independent of $\mathbf{Z}_{22}'\mathbf{Z}_{22}$ and $E\left(\mathbf{Z}_{12}'\mathbf{Z}_{12}\right)=(N_1-\rho N_2)\Omega_{\mathbf{z}}$, $E\left(\mathbf{Z}_{22}'\mathbf{Z}_{22}\right)=(1-\rho)N_2\Omega_{\mathbf{z}}$, and the first order Taylor expansion

gives $tr\left\{E\left[\mathbf{Z}_{12}'\mathbf{Z}_{12}\left(\mathbf{Z}_{22}'\mathbf{Z}_{22}\right)^{-1}\right]\right\} \approx \frac{(N_1-\rho N_2)}{(1-\rho)N_2}$.

Therefore,

$$E\left(\widehat{W}\right) \approx \frac{\rho N_2 \gamma' \Omega_z \gamma + K \cdot \sigma_v^2}{N_1 \gamma' \Omega_z \gamma + K \cdot \sigma_v^2 + (N_1 - \rho N_2)/(1-\rho) N_2 \sigma_v^2}$$

Proof for proposition 4

$$\widehat{\beta}_{2SLS}^{(1)} - \beta = \left(\widehat{X}_{11}'\widehat{X}_{11}\right)^{-1}\left(\widehat{X}_{11}'\Xi_{11}\right)$$

We have derived the expectation of the denominator in remark 3.

Similarly, the expectation of the numerator is

$$\begin{aligned} E\left(\widehat{X}_{11}'\Xi_{11}\right) &= E\left(X_{11}'P_{\mathbf{Z}_{11}}E_{11}\right) \\ &= E\left[\left(\gamma'\mathbf{Z}_{11}' + V_{11}'\right)P_{\mathbf{Z}_{11}}\left(\theta V_{11} + R_{11}\right)\right] \\ &= \theta E\left(V_{11}'P_{\mathbf{Z}_{11}}V_{11}\right) + E\left(V_{11}'P_{\mathbf{Z}_{11}}R_{11}\right) \\ &= \theta\sigma_v^2 \cdot E\left(\chi_1^2\right) \\ &= K \cdot \sigma_{\varepsilon v}, \end{aligned}$$

The second equality follows from assumption 4, that is,

$$E\left(\gamma'\mathbf{Z}_{11}'P_{\mathbf{Z}_{11}}\Xi_{11}\right) = E\left(\gamma'\mathbf{Z}_{11}'\Xi_{11}\right) = 0$$

Also, due to assumption 3.(c), write $\Xi_{11} = \theta V_{11} + R_{11}$, and R_{11} is independent of V_{11}.

So $E\left(V_{11}'P_{\mathbf{Z}_{11}}R_{11}\right) = 0$ for the fourth equality to hold.

The first-order Taylor expansion of

$$g\left(X_{11}'P_{\mathbf{Z}_{11}}X_{11}, X_{11}'P_{\mathbf{Z}_{11}}\Xi_{11}\right)$$

at point

$$\left[E\left(X_{11}'P_{\mathbf{Z}_{11}}X_{11}\right),E\left(X_{11}'P_{\mathbf{Z}_{11}}E_{11}\right)\right]$$

is

$$\begin{aligned} E\left(\widehat{\beta}_{2SLS}^{(1)}-\beta\right) &\approx E\left(\widehat{X}_{11}'\widehat{X}_{11}\right)^{-1}E\left(\widehat{X}_{11}'\Xi_{11}\right) \\ &= \frac{K\cdot\sigma_{\varepsilon v}}{\rho N_2\gamma_1'\Omega_z\gamma_1+K\cdot\sigma_v^2}. \end{aligned}$$

Following Angrist and Kruger (1995),

$$\begin{aligned} E\left(\widehat{\beta}_{SS2SLS}^{(2)}-\beta\right) &= E\left[\left(\widehat{X}_{12}'\widehat{X}_{12}\right)^{-1}\left(\widehat{X}_{12}'Y_{12}\right)\right]-\beta \\ &= E\left\{\left[\widehat{X}_{12}'\widehat{X}_{12}\right]^{-1}\left[\widehat{X}_{12}'\left(X_{12}\beta+V_{12}\right)\right]'\right\}-\beta \\ &= E\left[\left(\widehat{X}_{12}'\widehat{X}_{12}\right)^{-1}\left(\widehat{X}_{12}'X_{12}\right)\right]\beta-\beta \end{aligned}$$

The second equality follows from Assumption 3 ($E(v_{2j}|\mathbf{z}_{2j})=0$) and the independence between V_{12} and $(X_{22},\mathbf{Z}_{12})$,

$$\begin{aligned} & E\left[\left(\widehat{X}_{12}'\widehat{X}_{12}\right)^{-1}\left(\widehat{X}_{12}'\Xi_{12}\right)\right] \\ = \; & E\left\{\left\{\left[\mathbf{Z}_{12}\left(\mathbf{Z}_{22}'\mathbf{Z}_{22}\right)^{-1}\mathbf{Z}_{22}'X_{22}\right]'\left[\mathbf{Z}_{12}\left(\mathbf{Z}_{22}'\mathbf{Z}_{22}\right)^{-1}\mathbf{Z}_{22}'X_{22}\right]\right\}^{-1}X_{22}'\mathbf{Z}_{22}\left(\mathbf{Z}_{22}'\mathbf{Z}_{22}\right)^{-1}\mathbf{Z}_{12}'V_{12}\right\} \\ = \; & 0 \end{aligned}$$

The following shows that $E(x_{1i}|\widehat{x}_{1i}) = \widehat{x}_{1i}\left\{E\left[\widehat{x}'_{1i}\widehat{x}_{1i}\right]^{-1}E\left[\widehat{x}'_{1i}x_{1i}\right]\right\}$ for $i = 1,...,\rho N_2$:

$$\begin{aligned}
E(x_{1i}|\widehat{x}_{1i}) &= E\left(z_{1i}\gamma + v_{1i}|z_{1i}\widehat{\gamma}_2\right) \\
&= E\left(z_{1i}\gamma|z_{1i}\widehat{\gamma}_2\right) \\
&= z_{1i}\widehat{\gamma}_2\frac{E\left(\widehat{\gamma}'_2 z'_{1i}z_{1i}\gamma\right)}{E\left(\widehat{\gamma}'_2 z'_{1i}z_{1i}\widehat{\gamma}_2\right)} \\
&= z_{1i}\widehat{\gamma}_2\frac{E\left(\widehat{\gamma}'_2 z'_{1i}z_{1i}\gamma\right) + E\left(\widehat{\gamma}'_2 z'_{1i}v_{1i}\right)}{E\left(\widehat{\gamma}'_2 z'_{1i}z_{1i}\widehat{\gamma}_2\right)} \\
&= \widehat{x}_{1i}\frac{E\left(\widehat{x}'_{1i}x_{1i}\right)}{E\left(\widehat{x}'_{1i}\widehat{x}_{1i}\right)},
\end{aligned}$$

Because $\widehat{\gamma}_2 = \left(z'_{2j}z_{2j}\right)^{-1}\left(z'_{2j}x_{2j}\right)$ and the fact that v_{1i} is independent of z_{1i}, z_{2j}, x_{2j}, the second equality follows from $E\left(v_{1i}|z_{1i}\widehat{\gamma}_2\right) = 0$. The third equality is clear since $\widehat{\gamma}_2$ can be seen as a constant so that $z_{1i}\gamma$ is linear in $z_{1i}\widehat{\gamma}_2$.

By stacking the number of the observations, it follows that $E(X_{12}|\widehat{X}_{12})$ is linear as well and that $E(X_{12}|\widehat{X}_{12}) = \widehat{X}_{12}\left\{E\left[\widehat{X}'_{12}\widehat{X}_{12}\right]^{-1}E\left[\widehat{X}'_{12}X_{12}\right]\right\}$. Once more, by the law of iterated expectations,

$$E\left[\left(\widehat{X}'_{12}\widehat{X}_{12}\right)^{-1}\left(\widehat{X}'_{12}X_{12}\right)\right] = E\left(\widehat{X}'_{12}\widehat{X}_{12}\right)^{-1}E\left(\widehat{X}'_{12}X_{12}\right)$$

The numerator simplifies to

$$\begin{aligned}
E\left(\widehat{X}'_{12}X_{12}\right) &= E\left\{\left[\mathbf{Z}_{12}\gamma + \mathbf{Z}_{12}\left(\mathbf{Z}'_{22}\mathbf{Z}_{22}\right)^{-1}\mathbf{Z}'_{22}V_{22}\right]'\left(\mathbf{Z}_{12}\gamma + V_{12}\right)\right\} \\
&= E\left(\gamma'\mathbf{Z}'_{12}\mathbf{Z}_{12}\gamma\right) + E(V'_{22}\mathbf{Z}_{22}\left(\mathbf{Z}'_{22}\mathbf{Z}_{22}\right)^{-1}\mathbf{Z}'_{12}\mathbf{Z}_{12}\gamma) + E\left(\gamma'\mathbf{Z}'_{12}V_{12}\right) \\
&\quad + E\left(V'_{22}\mathbf{Z}_{22}\left(\mathbf{Z}'_{22}\mathbf{Z}_{22}\right)^{-1}\mathbf{Z}'_{12}V_{12}\right) \\
&= E\left(\gamma'\mathbf{Z}'_{12}\mathbf{Z}_{12}\gamma\right) \\
&= (N_1 - \rho N_2)\gamma'\Omega_{\mathbf{z}}\gamma
\end{aligned}$$

The third equality follows from assumption 5 and V_{22} and V_{12} coming from independent sam-

ples. Also, recall $E(\mathbf{z}_{2i}'\mathbf{z}_{2i}) = \Omega_{\mathbf{z}}$.

We already derived in remark 3

$$E\left(\widehat{X}_{12}'\widehat{X}_{12}\right) = (N_1 - \rho N_2)\gamma'\Omega_{\mathbf{z}}\gamma + ((N_1 - \rho N_2)/(1-\rho)N_2)\sigma_v^2,$$

so the approximate bias of SS2SLS follows as,

$$E\left(\widehat{\beta}_{SS2SLS}^{(2)} - \beta\right) = -\frac{\sigma_v^2\beta/((1-\rho)N_2)}{\gamma'\Omega_{\mathbf{z}}\gamma + \sigma_v^2/((1-\rho)N_2)}.$$

Proof for Proposition 5

$$\begin{aligned}
& E\left[\widehat{W}\left(\widehat{\beta}_{2SLS}^{(1)} - \beta\right) + \left(1-\widehat{W}\right)\left(\widehat{\beta}_{SS2SLS}^{(2)} - \beta\right)\right] \\
= \; & E\left[\widehat{W}\left(\widehat{\beta}_{2SLS}^{(1)} - \beta\right)\right] + E\left[\left(1-\widehat{W}\right)\widehat{\beta}_{SS2SLS}^{(2)}\right] - E\left(1-\widehat{W}\right)\beta \\
= \; & E\left(\frac{\widehat{X}_{11}'\Xi_{11}}{\widehat{X}_{11}'\widehat{X}_{11} + \widehat{X}_{12}'\widehat{X}_{12}}\right) + E\left(\frac{\widehat{X}_{12}'X_{12}}{\widehat{X}_{11}'\widehat{X}_{11} + \widehat{X}_{12}'\widehat{X}_{12}}\right)\beta - E\left(\frac{\widehat{X}_{12}'\widehat{X}_{12}}{\widehat{X}_{11}'\widehat{X}_{11} + \widehat{X}_{12}'\widehat{X}_{12}}\right)\beta \\
\approx \; & \frac{E\left(\widehat{X}_{11}'\Xi_{11}\right) + E\left(\widehat{X}_{12}'X_{12}\right)\beta - E\left(\widehat{X}_{12}'\widehat{X}_{12}\right)\beta}{E\left(\widehat{X}_{11}'\widehat{X}_{11} + \widehat{X}_{12}'\widehat{X}_{12}\right)} \\
= \; & \frac{K\cdot\sigma_{\varepsilon v} - ((N_1 - \rho N_2)/(1-\rho)N_2)\sigma_v^2\beta}{N_1\gamma'\Omega_z\gamma + K\cdot\sigma_v^2 + ((N_1 - \rho N_2)/(1-\rho)N_2)\sigma_v^2}.
\end{aligned}$$

The third approximation uses a first-order Taylor expansion. All the moments in this approximation can be found in the proofs of proposition 1 and 2.

Proof for Proposition 6

The asymptotic variance of 2SLS estimator is

$$\sqrt{\rho N_2}\left(\widehat{\beta}_{2SLS}^{(1)} - \beta\right) \overset{a}{\sim} N\left[0, \sigma_\varepsilon^2 E\left(x_{1i}^{*\prime}x_{1i}^*\right)^{-1}\right] = N\left\{0, \sigma_\varepsilon^2\left[\Omega_{x\mathbf{z}}\Omega_{\mathbf{z}}^{-1}\Omega_{x\mathbf{z}}\right]^{-1}\right\}$$

where $x_{1i}^* = \mathbf{z}_{1i}\gamma = \mathbf{z}_{1i}E\left(\mathbf{z}_{1i}'\mathbf{z}_{1i}\right)^{-1}E\left(\mathbf{z}_{1i}'x_{1i}\right)$ (Wooldridge 2010).

The asymptotic variance of the SS2SLS estimator is adapted from Inoue and Solon (2010) in this context as a special case of TS2SLS:

$$\sqrt{N_1-\rho N_2}\left(\widehat{\beta}^{(2)}_{SS2SLS}-\beta\right) \overset{a}{\sim} N\left\{0,\left[\Omega'_{x\mathbf{z}}\left[\left(\sigma_u^2+\frac{\alpha-\rho}{1-\rho}\beta'\sigma_v^2\beta\right)\Omega_{\mathbf{z}}\right]^{-1}\Omega_{x\mathbf{z}}\right]^{-1}\right\}.$$

By the asymptotic equivalence theorem,

$$\begin{aligned}
&\sqrt{N_1+(1-\rho)N_2}\left(\widehat{\beta}_O-\beta\right)\\
=\;&\sqrt{\frac{N_1+(1-\rho)N_2}{\rho N_2}}\sqrt{\rho N_2}\widehat{W}\left(\widehat{\beta}^{(1)}_{2SLS}-\beta\right)\\
&+\sqrt{\frac{N_1+(1-\rho)N_2}{\alpha-\rho N_2}}\sqrt{N_1-\rho N_2}\left(1-\widehat{W}\right)\left(\widehat{\beta}^{(2)}_{SS2SLS}-\beta\right)\\
&\xrightarrow{p}\frac{\sqrt{(1+\alpha-\rho)\rho}}{\alpha}\sqrt{\rho N_2}\left(\widehat{\beta}^{(1)}_{2SLS}-\beta\right)+\frac{\sqrt{(1+\alpha-\rho)(\alpha-\rho)}}{\alpha}\sqrt{N_1-\rho N_2}\left(\widehat{\beta}^{(2)}_{SS2SLS}-\beta\right)\\
&\overset{a}{\sim} N\left[0,\frac{(1+\alpha-\rho)\rho}{\alpha^2}\sigma_\varepsilon^2\left[\Omega_{x\mathbf{z}}\Omega_{\mathbf{z}}^{-1}\Omega_{x\mathbf{z}}\right]^{-1}\right.\\
&\left.+\frac{(1+\alpha-\rho)(\alpha-\rho)}{\alpha^2}\left[\Omega'_{x\mathbf{z}}\left[\left(\sigma_u^2+\frac{\alpha-\rho}{1-\rho}\beta'\sigma_v^2\beta\right)\Omega_{\mathbf{z}}\right]^{-1}\Omega_{x\mathbf{z}}\right]^{-1}\right].
\end{aligned}$$

REFERENCES

REFERENCES

[1] Anderson, M. L., & Matsa, D. A. (2011). Are restaurants really supersizing America? American Economic Journal: Applied Economics, 152–188.

[2] Angrist, J. D., & Evans, W. N. (1998). Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size. The American Economic Review, 88(3), 450–477.

[3] Angrist, J. D., Imbens, G. W., & Krueger, A. B. (1999). Jackknife instrumental variables estimation. Journal of Applied Econometrics, 14(1), 57–67. http://doi.org/10.1002/(SICI)1099-1255(199901/02)14:1<57::AID-JAE501>3.0.CO;2-G

[4] Angrist, J. D., & Krueger, A. B. (1992). The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples. Journal of the American Statistical Association, 87(418), 328–336. http://doi.org/10.2307/2290263

[5] Angrist, J. D., & Krueger, A. B. (1995). Split-Sample Instrumental Variables Estimates of the Return to Schooling. Journal of Business & Economic Statistics, 13(2), 225–235. http://doi.org/10.2307/1392377

[6] Arellano, M., & Meghir, C. (1992). Female Labour Supply and On-the-Job Search: An Empirical Model Estimated Using Complementary Data Sets. The Review of Economic Studies, 59(3), 537–559. http://doi.org/10.2307/2297863

[7] Bekker, P. A. (1994). Alternative Approximations to the Distributions of Instrumental Variable Estimators. Econometrica, 62(3), 657–681. http://doi.org/10.2307/2951662

[8] Bound, J., Jaeger, D. A., & Baker, R. M. (1995). Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak. Journal of the American Statistical Association, 90(430), 443–450. http://doi.org/10.2307/2291055

[9] Bun, M. J. G., & Windmeijer, F. (2011). A comparison of bias approximations for the two-stage least squares (2SLS) estimator. Economics Letters, 113(1), 76–79. http://doi.org/10.1016/j.econlet.2011.05.047

[10] den Berg, G. J., Pinger, P. R., & Schoch, J. (2015). Instrumental variable estimation of the causal effect of hunger early in life on health later in life. The Economic Journal. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/ecoj.12250/abstract

[11] Devereux, P. J., & Hart, R. A. (2010). Forced to be Rich? Returns to Compulsory Schooling in Britain*. The Economic Journal, 120(549), 1345–1364.

[12] Gong, H., Leigh, A., & Meng, X. (2012). Intergenerational income mobility in urban China. Review of Income and Wealth, 58(3), 481–503.

[13] Hahn, J., & Hausman, J. (2002). Notes on bias in estimators for simultaneous equation models. Economics Letters, 75(2), 237–241. http://doi.org/10.1016/S0165-1765(01)00602-4

[14] Hahn, J., Hausman, J., & Kuersteiner, G. (2004). Estimation with weak instruments: Accuracy of higher-order bias and MSE approximations. The Econometrics Journal, 7(1), 272–306.

[15] Hausman, J. A., Newey, W. K., Woutersen, T., Chao, J. C., & Swanson, N. R. (2012). Instrumental variable estimation with heteroskedasticity and many instruments. Quantitative Economics, 3(2), 211–255. http://doi.org/10.3982/QE89

[16] Inoue, A., & Solon, G. (2010). Two-Sample Instrumental Variables Estimators. Review of Economics and Statistics, 92(3), 557–561. http://doi.org/10.1162/REST_a_00011

[17] Klevmarken, N. A. (1982). On the Stability of Age-Earnings Profiles. The Scandinavian Journal of Economics, 84(4), 531–554. http://doi.org/10.2307/3439516

[18] Nagar, A. L. (1959). The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations. Econometrica, 27(4), 575–595. http://doi.org/10.2307/1909352

[19] Nicoletti, C., & Ermisch, J. F. (2008). Intergenerational earnings mobility: changes across cohorts in Britain. The BE Journal of Economic Analysis & Policy, 7(2). Retrieved from http://www.degruyter.com/view/j/bejeap.2007.7.2/bejeap.2007.7.2.1755/bejeap.2007.7.2.1755.xml

[20] Olivetti, C., & Paserman, M. D. (2014). In the Name of the Son (and the Daughter): Intergenerational Mobility in the United States, 1850-1940. Retrieved from http://people.bu.edu/olivetti/papers/Olivetti-Paserman_NameOfTheSon_July2014.pdf

[21] Pierce, B. L., & Burgess, S. (2013). Efficient Design for Mendelian Randomization Studies: Subsample and 2-Sample Instrumental Variable Estimators. American Journal of Epidemiology, 178(7), 1177–1184. http://doi.org/10.1093/aje/kwt084

[22] Rosenzweig, M. R., & Wolpin, K. I. (2000). Natural "natural experiments" in economics. Journal of Economic Literature, 827–874.

[23] Rothstein, J., & Wozny, N. (2013). Permanent income and the black-white test score gap. Journal of Human Resources, 48(3), 510–544.

[24] Wooldridge, J. M. (2010). Econometric analysis of cross section and panel data. MIT press.

Chapter 3. Estimating and Validating Nonlinear and Heterogeneous Classroom Peer Effects

1 Introduction

Studies of educational peer effects suffer from the classic tensions between the validity, power, and cost-effectiveness of research designs. Natural experiments generally provide credible sources of identification of limited peer-effects models, but frequently lack the necessary statistical power to fully explore nonlinear and heterogeneous effects. Observational data provides power in excess, but without the prima facie validity of estimates conferred by randomized assignment. Large-scale randomized trials are costly and may fail to yield effects with clear policy implications, as there exists both empirical and theoretical evidence that peer effects in the presence of purely random assignment differ from those with endogenous assignment (Weinberg 2007; Duflo, Dupas, and Kremer 2011). Observational approaches may provide useful advantages over quasi-experimental or experimental methods if inferences from observational studies can be made credibly robust to potential biases. Combined with the growth in availability of administrative educational data sets, robust observational methods can supply a low-cost, scalable method for developing peer effect estimates directly relevant to local policy.

This paper uses an observational approach to estimate the nonlinear shape of peer effects, examines whether effects vary depending on a student's relative ability in the classroom, and checks the plausibility of estimated patterns using a placebo testing approach. Using administrative data for students in North Carolina high schools from 2006-2013, I estimate a model of ability peer effects with linear-in-shares and linear-in-means components for standardized tests in Algebra II, Geometry, Biology, Physical Science, U.S. History, Civics, and English I. I control for several potential confounding factors, including student past test scores, teacher quality, and school quality. In contrast with prior studies using a single test score or an unweighted average of test scores,

I measure subject-specific peer ability using a regression-calibrated nonlinear function of prior test scores which better measures the underlying ability construct associated with a particular test. Estimates reveal some evidence of peer effects operating through mean ability in a classroom, but I also find evidence that they are biased upward by sorting or non-classroom peer effects. I find robust evidence of peer effects nonlinear in ability, with effects monotonically increasing in peer ability in most cases. I also find that peer effects are decreasing in relative ability of a student—higher-achieving students within a classroom tend to receive smaller test score increases than their lower-achieving peers from improving peer ability in any part of the distribution.

To assess the extent to which sorting or non-classroom peer effects may be driving the observed linear and nonlinear associations, I estimate a series of placebo regressions. The regressions test whether a student's test score in a specific core subject is predicted by the ability composition of a student's classrooms for other core subjects (the "placebo") conditional on the outcome classroom ("treatment") ability composition. While there are almost always both significant estimated linear and nonlinear coefficients for the classrooms corresponding to the treatment classrooms, placebo classrooms almost always return significant linear coefficients but largely insignificant nonlinear coefficients. I interpret this pattern as evidence that the estimated coefficients for peer mean ability are driven by sorting or peer effects external to the classroom, but that the model's estimates of peer ability's nonlinear effects are valid. I provide a more formal econometric interpretation for placebo tests, showing how the placebo estimand in large samples is directly proportional to the magnitude of omitted variables bias in the main estimates.

This study contributes new knowledge about the nonlinear shape of peer effects for high school students and how effects are heterogeneous depending on a student's own ability. It also applies a unique placebo methodology to examine the validity of each estimate, providing new evidence of bias in linear-in-means estimates from observational data. Previous studies have provided some evidence of nonlinear effects, and this study expands the body of evidence on nonlinear effects to include richer divisions of students into ability group and subject-specific nonlinear effects for seven core high school subjects. Either nonlinear or heterogeneous peer effects are necessary con-

ditions for there to be any improvement in total education production by sorting into classrooms based on peer groups (Carrell, Sacerdote, and West 2013). However, it is important to distinguish between nonlinear response to the ability of individual students in contrast with nonlinearity in the aggregate composition of a peer group. Linear-in-shares specifications capture the degree to which one student can have a nonlinear response to the ability of another student. However, there may be further emergent effects from classroom composition: two high ability students may have an effect on one low ability student that is greater than the sum of individual effects of high ability on low ability, which linear-in-shares models do not directly account for. For both linear-in-means and linear-in-shares models (with fixed class sizes), improvements from reassignment of a peer to another classroom are offset by losses in another. Nonetheless, the methods used here to validate the coefficients can be applied to broader nonlinear or heterogeneous-effect specifications that capture the nonlinear compositional effects required for there to be gains from changing peer ability grouping. In Section 2.4, I estimate a model with heterogeneous effects by students' own absolute ability and use the estimates to demonstrate an example of optimal ability grouping for two Algebra II classrooms. I find that the underlying estimated heterogeneous effects are robust to unobservables according to the same placebo tests as the main results. I calculate that the mean achievement gains of optimal ability tracking (based only on the modeled heterogeneity) over random assignment are at least 0.03 SD. Further extensions of the model to accommodate nonlinearity in composition and heterogeneous effects can be used to identify ability sorting strategies that result in significant additional test score gains.

1.1 The Peer Effects Literature

The educational peer effects literature aims to measure the interdependence of outcomes among students. The existence of significant academic and disciplinary spillovers between students has policy implications for the optimal grouping of students between or within schools. The quintessential problem in the measurement of peer effects is endogeneity of assignment to a peer group: low-performing students may tend to be assigned to low-performing schools and classrooms and

have low-performing friends. One strand of literature attempts to isolate idiosyncratic variation in peer group composition over time, essentially controlling for (or exploiting) cross-group variation within schools through adjacent differencing techniques (Hoxby 2000; Lavy, Paserman, and Schlosser 2011; Bifulco et al. 2011). Burke and Sass (2013) estimate peer effects observationally as well, using estimated student fixed effects from past achievement as measures of peer ability. Some papers have used explicitly identifiable, plausibly exogenous natural sources of variation in peer groups, such as quasi-random college dormitory assignment (Sacerdote 2001) and Air Force academy squadron assignments (Carrell, Fullerton, and West 2009).

Several studies test for whether peer effects have any nonlinear shape. Lavy, Paserman, and Schlosser (2012) find that the proportion of grade-repeating students in a classroom has a negative impact on classwide achievement in Israeli middle schools. Using survey data, they find that higher proportions of lower-achieving students alter teachers' pedagogical practice, increase the frequency of violent or disruptive behavior, and harm student-teacher relationships. Using within-student regressions, Lavy, Silva, and Weinhardt (2012) find that the fraction of students in the bottom 5 percent of the ability distribution (based on prior test scores) is negatively associated with test scores.

1.2 Peer Effects Model

I employ a typical education production function using a combination of "linear-in-means" and "linear-in-shares" peer effects:

$$y_{rigst} = \overline{X}_{(-i)gst}\lambda + X^{d}_{(-i)gst}\beta_1 + A^{d}_{igst} \star X^{d}_{(-i)gst}\beta_2 + X_{igst}\eta + G_{gst}\gamma + \alpha_s + T_{gs} + \varepsilon_{igst} \tag{18}$$

This model relates a standardized test score y in any subject r for student i in classroom g in school s to individual, classroom, school, and peer characteristics at time t. Students are only observed for each subject in one time period, resulting in a repeat cross-section of students nested in classrooms and schools. $\overline{X}_{(-i)gst}$, the "linear component" of the peer effect, is part of the classic

linear baseline specification for peer effects studies (e.g., Carrell, Fullerton, and West 2009), measuring the mean ability of student i's classroom peers. $X^d_{(-i)gst}$ is a vector of shares of a student's classroom peers in each decile of ability in the given subject, representing a nonlinear component of the peer effect. A^d is a vector of dummy indicators for the decile of ability within the classroom into which student i falls. $A^d_{igst} \star X^d_{(-i)gst}$ is the interaction between the two factors, its vector of coefficients β_2 measuring heterogeneous nonlinear response to peer ability levels with respect to a student's relative ability ranking. X_{igs} and G_{gs} represent observable individual- and group-level characteristics, while α_s and T_{gs} correspond to school- and teacher-unobserved heterogeneity (which are controlled for in estimation by fixed effects). The coefficients λ and β can be understood as Manski's (1993) "exogenous effects," measuring the effects of pre-existing peer background characteristics rather than the effects of contemporaneous peer performance. In practice, estimates for these coefficients may also reflect the effects of endogenous decisions made by students that are predicted by peer composition variables but are not accounted for by controls.

While each coefficient on peer characteristics reflects a meaningful aggregate of peer effects, the exact mechanisms driving the effects are difficult to discern. Low-achieving peers can impact classrooms by altering teachers' pedagogical approaches or by diverting teacher effort from other students. Students engaging in disruptive behavior can also divert teacher effort to disciplinary action instead of instruction, affect other students' ability to engage in learning, or cause further disruptive behavior among other students. Furthermore, both low achievement and poor classroom discipline are highly correlated. Measures of both are characterized by measurement error. Thus, estimates of the impacts of ability may in part reflect the effects of disciplinary peer effects.

Generally, high-achieving peers are likely to have positive effects on classroom learning. In group activities such as science labs or in-class open study they may facilitate additional learning and even be explicitly tasked by the teacher to assist with instruction. Students of higher ability tend to have significantly fewer disciplinary infractions and so will reflect positive effects on classroom achievement to the extent that they displace more disruptive students. Social ties plausibly span achievement levels and may result in top students assisting other students with studying or

homework assignments. On the other hand, some models of peer effects such as the "invidious comparison" model suggest that additional high-achieving students may create a climate discouraging to lower-achieving peers by diminishing the perceived returns to additional academic effort or affecting pedagogical practices (e.g., encouraging the teacher to cover more advanced material). If teachers aim to maximize the share of their students passing, pedagogical practices are unlikely to be affected by high-achieving students. Akerlof (1997) and Lantis (2014) put forward models that imply that if students are engaged in a "tournament" within their peer group for various rewards, the ability levels of competing students affect their effort choices. The most commonplace instance of the tournament occurs due to grade curving, where relative performance within the classroom dictates the reward of high grades.

1.3 Biases from Sorting and Identifying Nonlinear vs. Linear Effects

In section 2.2, I conclude that estimates of the linear component of peer effects are likely to be biased upward by sorting or non-classroom peer effects even given a rich set of controls, but the nonlinear components are approximately unbiased from these sources. There are several meaningful reasons why the estimates for the linear components are not credible, while the nonlinear components are. The bulk of the correlation between own performance and peers' ability is likely to be accounted for by a student's own ability measure based on past test scores, school quality, teacher quality, and course level. However, there may remain sorting between multiple classrooms at the same level with the same teacher, and there also may be time-varying changes within groups which explain part of the correlation between own academic performance and peers' ability. Mean classroom peer ability thus captures at least part of the sorting mechanisms and shocks not accounted for by the existing set of controls. This amounts to changing the empirical role of mean peer ability from a variable of interest to a control variable, allowing more robust inference for the nonlinear effects of peer ability.

Consider a simple one school, two-classroom setting with a repeat cross-section structure. One classroom is classified as "high ability" (e.g., honors) and the other "low ability" (regular), and

students are probabilistically sorted on an ability cutoff into each, but the class type is not directly observed by the researcher. In this stylized setting, mean peer ability would perfectly distinguish (in expectation) between the high- and low-ability classrooms and mute any corresponding bias from sorting, leaving only variation in the shares of peer ability orthogonal to the course type identifying their effects. The high ability classroom has a higher average share of 90th-percentile ability students than the low ability classroom, but only deviations from the expected share of 90th-percentile ability students for each class type are used to identify the nonlinear effect of 90th percentile students. In that manner, mean peer ability plays a comparable role to a fixed effect that could be included if class type were explicitly observed.

In this setting, I control for two key characteristics of courses that are strongly related to sorting into them: their formal level and their teachers. In addition, I control for a students' own underlying ability, which is correlated with academic performance and hence likely correlated with classroom sorting. However, there may be ability differentiation between classrooms which persists even conditional on these controls because of transitory shocks in time or because of the structure of a school's classrooms in a subject. School- or cohort-specific annual shocks could shift both a student's performance and the absolute ability composition of his peers through sample attrition (i.e., dropouts and retention). Alternatively, a teacher may have multiple classrooms in the same formal course level that are achievement-differentiated, resulting in correlated sorting not accounted for by the controls. As in the stylized model, mean peer ability in these classrooms captures at least part of the residual unobserved characteristics and time-varying shocks.

Variation in the distribution of peer ability in a classroom is likely to be coming from a mixture of cross-cohort variation in ability composition, students' course-picking decisions (based on preferences for teachers or friends) and constraints (e.g., the timing of their other courses), and any administrative rules for or direct intervention in student schedules. Sorting into subjects is moderated by the number of classrooms available into which students can sort: if there is only one classroom available for a subject, then all variation in peer ability composition in the class is only driven by cross-cohort variation. The included fixed effects narrow this to the number of

classrooms in a subject per teacher and per course level. A large number of classrooms potentially implies greater unobserved differentiation between classrooms.

1.4 Data Description

1.4.1 North Carolina Administrative Data

The North Carolina Education Research Data Center (NCERDC) is a centralized database of administrative data for North Carolina public schools. It contains student-level data for all students in grades 3-12 spanning from as early as the 1997-1998 school year to the 2013-2014 school year. Students can be tracked across years using the database's anonymized master identifier. Based on the overlapping availability of multiple variables used in the analysis, I use all available historical data for students who were in grades 9-12 from 2006-2014.

The student-level master build files form the core data set for the analysis, providing a record identifying each student by school membership, grade level, and school year. The course membership files allow students to be matched to their classroom peers, and classrooms to be matched to the corresponding subject tests. The attendance and demographic file provides students' month and year of birth, sex, ethnicity, school membership and number of days therein for each year, and annual attendance. The master suspension files document suspensions and other disciplinary actions taken by the school against the student, including information on the offense and the extent of the punishment. Finally, the measures of secondary school achievement I use come from North Carolina's state-mandated End of Course (EOC) tests for Algebra I, Algebra II, Physical Science, and English. Each testing mandate covers a different period, with some tests no longer being administered. For example, Algebra II EOC tests were no longer administered after the 2011 school year. For each test subject, I standardize students' test scores within school year across all students with available test scores in the database to account for annual statewide variation in score distributions.

1.4.2 Determining Course Membership

During 2006-2013, North Carolina high schools typically operated in a 4x4 semester block schedule. Under 4x4 block scheduling, students take 4 courses in Fall and 4 courses in Spring with 90-minute daily instructional periods. Most core courses last one semester, though some schools offer year-long versions of courses. The NCERDC data provides records on students' membership in courses collected from databases one time on each of the first days of Fall and Spring semesters. This includes the course title, semester, teacher ID (when matched by NCERDC), and section (i.e., class period) of the course. Students' classrooms contain a course title that is used to assign them into subject groups. Depending on keywords in the course titles, courses are classified as math, science, social studies, or English. For example, classes with "Algebra" or "Alg" in the title are classified as math, "Biology" or "Bio" science, "English" or "Eng" English, and "Civics" or "Civ" social studies. The keyword lists I use are largely comprehensive—I manually review the remainder of course titles to ensure that only elective courses remain and no core courses are omitted.

I am able to track students' changing classroom memberships between Fall and Spring semesters. Students' course choices are generally stable, but in some cases students may transfer or change courses within a semester. To account for classroom changes from Fall to Spring (either due to teacher changes or class period changes with the same teacher), a student with records in multiple courses of the same type (e.g., Algebra I) is counted as being included in the course discovered in the Spring data collection.

There are multiple types of courses for which the same EOC tests are administered; they have varying pedagogical designs and target different populations of students. Students may endogenously track into courses of different levels. To help account for variation in course levels, I use the hierarchical nature of North Carolina's state course codes to differentiate courses into level groups. For example, both "Algebra I" and remedially-themed "Foundations of Algebra" courses are associated with students who take the Algebra I EOC test in the same semester, but are appropriately assigned different 4-digit prefixes in their course codes. Each 4-digit prefix in a subject is

represented by a dummy indicator contained in G_{gst}.

1.4.3 Test Scores and Ability

Previous studies have attempted to measure peer ability using past test scores (Lantis 2014; Lavy, Silva, and Weinhardt 2012; Hoxby and Weingarth 2006) or other traits correlated with low achievement, such as grade repetition (Lavy, Paserman, and Schlosser 2012). I introduce a more flexible way of approximating the underlying ability construct for a specific outcome of interest using a regression of the outcome on a nonlinear function of past test scores. A typical approach is to use a single test itself or an unweighted average of tests. In contrast, the regression-based method allows a broader set of information on students' characteristics to be incorporated into ability measures and empirically calibrates the relative weights of each characteristic instead of imposing them externally. Among other potential benefits, the regression approach flexibly incorporates multiple prior test score measures (and characteristics) nonlinearly and adjusts for the partial correlation between past scores and current scores generated by fadeout, subject differences, and normal test score variation, and to incorporate multiple prior test score measures.

The procedure is to regress the test score in a subject on the chosen nonlinear function of past test scores and generate the fitted values, which form the ability score composite. Students can then be ranked on this score and the score used to generate peer ability averages. This can result a different linear-in-means (LIM) peer effect coefficient estimate even using only a linear function of a single past test score, since the ability composite regression coefficient "rescales" the estimated LIM peer effect to account for the partial correlation between past and present tests.

If multiple test scores are available, the regression-based ability score results in a different ability ranking than if students were ranked on a single test score. The fitted values reflect an empirically-calibrated weighted average of the two past test scores, resulting in a higher-quality measure of the underlying ability construct most relevant to a given high school subject than to just use a single test. For example, 8th-grade reading scores are more highly predictive of high school science and English test scores than are 8th-grade math scores. Using a single score or unweighted

average of the two scores would fail to take this into account, resulting in additional measurement error in the peer ability measure.

For this setting, I use a piecewise function of North Carolina's 8th-grade End-of-Grade math and reading tests equivalent in fit to a 20-piece regression spline with cut points at vigintiles of each test score. Using OLS, I estimate the following model of student i's test score in subject s:

$$y_{ri} = M_i^v\psi_1 + M_i\psi_2 + M_i * M_i^v\psi_3 + R_i^v\psi_4 + R_i\psi_5 + R_i * R_i^v\psi_6 + v_{ri}$$

M_i and R_i are continuous test scores (standardized within years) for math and reading. M_i^v and R_i^v are vectors of indicator variables identifying which vigintile of the score distribution each subject test score falls into. I generate an estimated ability composite using the fitted values from this regression:

$$A_i^r = \hat{y}_{ri} = M_i^v\hat{\psi}_1 + M_i\hat{\psi}_2 + M_i * M_i^v\hat{\psi}_3 + R_i^v\hat{\psi}_4 + R_i\hat{\psi}_5 + R_i * R_i^v\hat{\psi}_6$$

This ability score is then used to calculate own-ability decile indicators A_{igst}^d and peer ability variables in equation 18. The splines of math and reading scores used to generate this ability score are included for individuals directly in X_{igs} for additional flexibility.

Using the regression-based ability composite may induce minor econometric problems that likely do not significantly impact the results in this context. Because the realization of the outcome of interest is used in estimation, a particular student's past test scores and the dependent variable (current test score) enters into the computation of the coefficient that serves as the weight on prior test scores. Thus, the main regressions of the dependent variable on the ability measure partially use variation from the dependent variable itself through the estimated ability weights. The problem is similar in concept to the finite sample bias issue in instrumental variables, and can be understood in the same way: as the sample size grows, bias from this phenomenon shrinks to zero.[8] A second econometric issue is that the ability measures are a form of generated regressor, and using them

[8] A jackknife-type procedure can also be potentially used to exclude a student's own score from the estimation of the weights and sidestep this problem directly.

without adjusting for their variability results in underestimated standard errors. While this is of greatest concern in using a continuous ability measure, the primary use of the ability measure is for estimation of decile cut points for ability classification, and any variability is only expressed through reclassification of marginal students from one decile to the other. Moreover, the only direct use of continuous ability in the regression corresponding to equation 18 is through the mean peer ability measure, which to some extent dilutes the impact of sampling variability by averaging over several students.

2 Results

2.1 Linear vs. Non-linear Peer Effects Estimates

Tables 22 and 23 present the estimated coefficients for only peer mean ability, percentage male, and percentage peers who were held back a grade at least once in grades 3-8 for each subject's EOC standardized test across four specifications of the complete model (including nonlinear interactions). All specifications control for 8th grade math and reading scores (in vigintile dummies), gender, the semester of test administration, and year, grade, and school fixed effects. Column (2) for each subject is the same specification as Column (1), but with controls for the course level of the course taken in the same semester as the test. Column (3) adds teacher-by-school fixed effects, which are roughly comparable to teacher and school fixed effects independently (but additionally accounting for teacher-school-specific effects for teachers who switch schools). Column (4) is the specification of column (2) plus school-by-year fixed effects. Standard errors are clustered at the school level.

As is typical of observational estimates of peer effects, there is a strong positive relationship between classroom mean ability and own achievement for all subjects across all specifications, with the exception of column (3) for Biology. Controlling for course level and teacher fixed effects tends to weakly decrease the magnitude of all peer ability estimates for all subjects, with the largest decreases occurring for Algebra II and Biology. As the most stringent specification, column (3) is

preferred. The mean ability coefficients are interpreted as the effect on students in the lowest decile of ability in the class of raising mean ability in the classroom but holding the shares of students in each ability decile fixed. For example, a 1 S.D. increase in mean peer ability in a student's Algebra II classroom corresponds to a baseline increase of 0.13 standard deviations in the student's test score. In practice, increasing mean peer ability also means increasing the share of students in higher ability deciles, and I find that peer effects are monotonically increasing in peer ability.

Figure 12 and the top plots of figures 7-12 plot the 100 estimates contained in model coefficient vectors β_1 and β_2 for Algebra II, Geometry, Physical Science, English I, U.S. History, Civics, and Biology, respectively. The coefficient estimates correspond to the interaction terms between shares of peers in each absolute ability decile and individual student *i*'s own relative ability within the classroom. The omitted category is for a student in the bottom decile of relative ability with a classroom composed of 100% students in the bottom decile of absolute ability. The baseline coefficients on the peer ability decile shares measure the effect of increasing the share of peers in an ability decile by one unit (100%) for a student in the first decile of relative ability. Each dot in the plot shows the sum of the baseline coefficient and the interaction coefficient for each relative-ability-decile-by-peer ability-decile pair. Horizontal lines within each division represent the average effect for each peer ability decile.

Across all subjects, there is a general upward trend in effect size by ability decile. For example, in Algebra II, the average effect on test scores of a 10 percentage point increase in students in the 40th-50th percentile (replacing students in the bottom decile) is approximately 0.015 standard deviations; the average effect of a comparable increase of students in the for the 70th-80th percentile, 0.03 standard deviations; the 90th-100th percentile, 0.05 standard deviations. Figure 13 shows the same estimates for Algebra II as Figure 12, but now inverted to have own relative ability in the top axis defining each group and peer ability decile shares within each group. The sawtooth pattern of Figure 13 demonstrates a strongly nonlinear peer effect that is increasing in ability. For a student in the 40th-50th percentile, the effect of increasing the share of peers in an ability decile ranges from 0.01 SD for the 2nd decile to 0.05 SD for the top decile.

The apparent nonlinear profile of peer effects is similar across all subjects except Physical Science, which is flat or even negative at higher peer ability shares. This somewhat surprising result may be an effect of varying student populations or classroom structures across subjects. Physical Science precedes specialized science courses (Biology, Chemistry, and Physics) in the curriculum and is not a required part of the course track, but is designed as an additional option for students to meet their science requirements (North Carolina Department of Public Instruction 2004). Both the mean and standard deviation of 8th-grade math and reading test scores among students taking the physical science exam are the smallest among the subjects presented here, suggesting both lower-achieving students in general and smaller dispersion in ability. Mean 8th-grade scores in both math and science for physical science students are 0.2 SD below the mean. Peer effects may be heterogeneous over a student's own absolute ability (rather than relative ability within a classroom). Alternatively, there may be unobserved heterogeneous effects arising from differences in ability peer effects across grades.

There are also generally flat or downward-sloping estimates across individuals' own relative ability in the classroom. Lower-achieving students may be more sensitive to peer influences for obvious reasons. Instructional time may be relatively more valuable for their achievement than self-study, and so disruptions to it are more harmful. To the extent that these ability effects are biased by disciplinary spillovers, we might also expect that low-achieving students are more likely to commit infractions that result in missed school time or decreases in teacher investment in their success. In accordance with the tournament concept, in which students compete with each other for rewards including grades or academic opportunities, higher-achieving students may also reduce their effort choices in the presence of greater academic competition.

Endogenous ability sorting into classrooms is most likely source of bias in peer effects estimates generated from observational data. Worse-performing students may be more likely to sort into classrooms with higher proportions of low-ability students, suggesting negative bias. On the other hand, if schools effectively sort students into classrooms which maximize their potential test scores by adapting pedagogical practices, a higher proportion of low-performing students may sig-

nal a remedial or other curriculum-adapted class which can increase a low-performing student's test score. Conditioning on own 8th-grade test scores accounts for some of this sorting, but is limited by measurement error in the scores and the potential for down-trending performance from 8th grade to high school. The inclusion of course level and teacher fixed effects also account for some sorting but some unobserved differentiation of classrooms within course levels or teachers may persist. For example, a teacher may teach courses for high ability and for low ability students separately which have the same course code prefix (and hence the same course level), such as Algebra I and Algebra I honors. Furthermore, there may be peer effects operating in a context broader than the classroom. Peer composition in a classroom may be correlated with peer composition in other courses or in other aspects of school life (such as the lunch period or after school activities), and meaningful academic or disciplinary spillovers can occur in these other contexts. Because these potential correlated unobservables are likely to manifest themselves across all of a student's core classrooms, I test for their existence through a series of classroom-based placebo tests in the next section.

2.2 Placebo Tests – Alternate Classrooms

I conduct several placebo tests to address the threat of bias in peer effects estimates from individuals' unobservable characteristics that are correlated with peer ability group composition. Test scores are highly correlated across high school subjects, suggesting that the underlying correlations of peer abilities are also large. Thus, sorting into classrooms on ability is likely to be a related process across subjects. In practice, its most distinctive manifestation is the tracking of high ability students into multiple "honors" courses and low ability students into multiple remedial courses.

The primary placebo test is to regress student test scores in subject A on peer ability composition in the corresponding classrooms for subject A and subject B. A significant positive estimate for peer ability composition in subject B is evidence of correlated unobservables which may bias the estimates for subject A's peer composition. Specifically, for some unobservable affecting test scores, the compositions for both A and B classrooms are both correlated with the unobservable.

The parts of the compositions that are correlated with each other are not reflected in the coefficient estimates, but each subject's classroom peer composition may have a part that is uncorrelated with the other subject's but correlated with the unobservable.[9] Statistically significant placebo estimates in the same direction as the treatment estimates are evidence that the primary regression results are at least partially driven by either correlated peer effects occurring at a level above the classroom or ability sorting into classrooms unaccounted for by controls. Section 2.3 provides a formal econometric interpretation of the placebo test for a single treatment, showing that the placebo coefficient is proportional to the amount of omitted variables bias in effect estimates from the primary specification.

Table 24 shows estimates for the linear effects of treatment and placebo classroom mean ability, gender composition, and proportion retained for Algebra II and Biology, using the classrooms of three other core subjects as the placebo classrooms. Most of the placebo coefficient estimates are significant and in the same direction as the treatment effect, indicating that unobservable characteristics or broader peer effects are driving the effect of peer mean ability for the treatment classrooms. The same reasoning applies to estimates for both classroom gender composition and percentage of students retained.

Figure 14 shows the peer ability-relative ability coefficients for the English class placebo for Algebra II, and Figure 13 is the inverted version of the same plot. Both of the patterns observed in the treatment effect plots in Figures 12 and 13 are not upheld in the English placebo, with most estimates noisy and narrowly distributed around zero or opposite-signed to the treatment effect. I repeat this exercise using science classes in Figure 16 and social studies classes in Figure 17. In contrast with the mean ability case, this placebo is evidence in favor of the estimated nonlinear relationship being approximately unbiased by unobservable mechanisms or characteristics that would be related across classrooms, such as sorting and non-classroom peer effects.

The bottom plots of figures 7-12 are the placebo tests for the corresponding subjects, using

[9] An even stronger placebo test would be to repeat the above regression excluding the true treatment; a null finding would strongly indicate that correlated unobservables are not driving the main result. However, any correlation between A and B's compositions would cause part of the true treatment effect of classroom A to appear in estimates for B.

English classroom composition for all subjects for the placebo test except English (which uses social studies classroom composition). In most cases, any nonlinear pattern in peer ability is not upheld in the placebo case. Where nonlinear patterns are replicated in the placebo, there is cause to distrust the estimated magnitude of that specific portion of the nonlinear peer effect. For example, for Biology, the effects of the shares of students in the 2nd and 3rd ability deciles in the placebo classroom are slightly larger than those in the treatment classroom, suggesting that those specific coefficients are identified using variation in shares associated with some unobserved sorting process. This would arise if there were several classes in the data targeted for and successful in raising the test scores of struggling students. The remainder of the Biology deciles show a monotonically increasing nonlinear pattern, while the placebo estimates are small in magnitude and tightly clustered near zero, underscoring the credibility of the estimates for those peer ability deciles.

In some cases, the pattern over relative ability is mirrored by the placebo. For biology, civics, and U.S. History, though the average magnitude of the effects is smaller, the placebo estimates show a downward-sloping pattern in relative ability in several categories. A possible explanation is a peer effect occurring outside of the classroom with a similar heterogeneous response over own ability as the classroom peer effect, such as disciplinary spillovers or some other cultural effect. Another explanation is sorting, which would require both foresight and intentional manipulation of a student's position in the relative ability distribution in the classroom, but is less plausible in light of these requirements combined with the typically student-driven classroom selection in high schools.

There is a useful interpretation to the nonlinear effect coefficients alone. Multiple distributions of peer abilities correspond to the same mean ability, but there are potentially well-defined rankings of test score production over those distributions of peers. For a function of form $y = x\lambda + f(x)\beta$, the marginal effect of increasing x is $dy/dx = \lambda + f'(x)\beta$. Suppose that x represents the linear (mean ability) peer component of the educational production function and $f(x)$ a continuous, positive nonlinear component (e.g., x^2). Assuming $\lambda \geq 0$ (i.e., test scores are weakly increasing in peer ability for at least some part of the domain of ability), then $f'(x)\beta < \lambda + f'(x)\beta$; the non-

linear marginal effect bounds the total effect of a change in x from below. This generalizes to non-continuous functions $f(x)$ as long as they are also positive and increasing in x. In practice, this means that if we have an estimate of λ that is not trustworthy but an estimate of β that is trustworthy, we can determine the minimum increase in test scores due to changing the peer ability distribution within a classroom. We only need to rely on the assumption that $\lambda \geq 0$. However, this result generalizes further, irrespective of the value of λ. For a fixed pool of students and group of classrooms to assign to, the score changes from reassignment that operate through λ will result in a net zero change in mean achievement. For some functions $f(x)$, such as the linear-in-shares specification used thus far, the changes also net to zero. When $f(x)$ reflects compositional changes in the classroom, mean overall achievement will be responsive to classroom sorting and the gains will be captured by the term $f(x)\beta$ regardless of the value of λ. In Section 2.4, I pursue an exercise that demonstrates a gain in mean overall achievement by sorting based on a heterogeneous effect of peer ability.

2.3 Note on the Interpretation of the Placebo Test

Using a simplified setup with one treatment classroom and one placebo classroom, I show how the placebo estimand in large samples is directly proportional to the magnitude of omitted variables bias in the estimated peer effect. I also show some conditions under which the placebo estimate produces either a false negative or false positive result. I begin by writing outcome y_{i1} as a function of classroom peer ability in classroom 1, T_{i1}, student characteristics, and peer ability in a student's other classroom, T_{i2}. All observable student characteristics are "partialled out", leading to the following specification:

$$y^*_{i1} = T^*_{i1}\beta + T^*_{i2}\beta^p + \varepsilon^*_{i1} \tag{19}$$

$$\varepsilon^*_{i1} = T^*_{i1}\rho + \varepsilon^*_{i1} \tag{20}$$

$$T_{i2}^{*} = T_{i1}^{*}\gamma + v_{i2}^{*} \tag{21}$$

ρ represents the projection of the residual error on the treatment of interest, T_{i1}^{*}, capturing the bias in the OLS estimator of y_{i1}^{*} on T_{i1}^{*}. We can imagine any linear projection coefficient being partitioned in two parts, a coefficient that represents correlations from causal relationships and one that represents non-causal correlations. For the "placebo coefficient",$\beta^{p} = \beta^{p,c} + \beta^{p,b}$. The placebo test's critical assumption is that $\beta^{p,c} = 0$.

$$\begin{aligned} plim_{N\to\infty}\hat{\beta}^{p} &= \left(var(\tilde{T}_{i2})\right)^{-1} cov(\tilde{T}_{i2}, y_{i1}) \\ &= \left(var(\tilde{T}_{i2})\right)^{-1} cov(T_{i2}^{*} - \gamma T_{i1}^{*}, \varepsilon_{i1}^{*}) \\ &= \left(var(\tilde{T}_{i2})\right)^{-1} cov(v_{i2}^{*}, \varepsilon_{i1}^{*}) \end{aligned}$$

Then, substituting in 20,

$$= \left(var(\tilde{T}_{i2})\right)^{-1} \quad cov(v_{i2}^{*}, T_{i1}^{*}\rho + \varepsilon_{i1}^{*}) \tag{22}$$

Now suppose students are sorted into classrooms for each subject on the basis of a general ability level, A_i, and other idiosyncratic factors specific to each subject a_{ic}, which includes subject-specific ability (e.g., ability in math that does not predict sorting into English classes). A linear reduced form representation of the ability level of one's peers in their classroom for subject c chosen through this sorting mechanism is

$$T_{ic}^{*} = A_i^{*}\Gamma_c + a_{ic}^{*} \tag{23}$$

Then, beginning from the relationship between classroom ability levels across subject in 21, substitute for peer ability level as a function of own ability described by the sorting mechanism:

$$T_{i2}^* = T_{i1}^*\gamma + v_{i2}^*$$

$$\Rightarrow A_i\Gamma_2 + a_{i2}^* = (A_i\Gamma_1 + a_{i1}^*)\gamma + v_{i2}^*$$

Rearranging gives

$$v_{i2}^* = A_i(\Gamma_2 - \Gamma_1\gamma) + (a_{i2}^* - a_{i1}^*\gamma) \tag{24}$$

Note that

$$\begin{aligned}\gamma &= var(T_{i1}^*)^{-1} cov(T_{i2}^*, T_{i1}^*) \\ &= var(T_{i1}^*)^{-1}\sigma_A^2\Gamma_1\Gamma_2\end{aligned}$$

Finally, substituting 24 and 23 into 22 yields

$$\begin{aligned}plim_{N\to\infty}\hat{\beta}^p &= \left(var(\tilde{T}_{i2})\right)^{-1} cov\left(A_i(\Gamma_2 - \Gamma_1\gamma) + (a_{i2}^* - a_{i1}^*\gamma), \rho(A_i\Gamma_1 + a_{i1}^*) + \varepsilon_{i1}^*\right) \\ &= \left(var(\tilde{T}_{i2})\right)^{-1}\left[\sigma_A^2\Gamma_2\Gamma_1\left(1 - \frac{\sigma_A^2}{\sigma_{T_1}^2}\Gamma_1^2 - \frac{\sigma_{a_1}^2}{\sigma_{T_1}^2}\right) + cov(a_{i2}^*, a_{i1}^*)\right]\rho \end{aligned} \tag{25}$$

where $\sigma_{T_1}^2 = var(T_{i1}^*)$, $var(A_i) = \sigma_A^2$, $var(a_{i1}) = \sigma_{a_1}^2$.

As intuitively proposed in all papers using placebo tests, a zero placebo coefficient estimate is evidence of the unbiasedness of the main estimates. Indeed, the placebo coefficient in this setting is directly proportional to ρ. If $\rho = 0$, implying no bias in the estimate of β from OLS of y_{i1} on T_{i1} conditional on observables, then the placebo coefficient is correspondingly zero (as it would be if T_{i1}^* were randomly assigned, for example).

Reflecting the inherent weaknesses of placebo tests, the placebo coefficient may also be zero because the placebo test is irrelevant. The first term is zero if either $\Gamma_1 = 0$ or $\Gamma_2 = 0$, meaning that there is no sorting on "overall ability" into one of the two classrooms conditional on observables. The second term is zero if $cov(a^*_{i2}, a^*_{i1}) = 0$, which would occur if there were no unobserved common factors that drive sorting into classrooms besides overall ability. These conditions will potentially mask nonzero ρ (i.e., selection on unobservables), suggesting a notable limitation of placebo tests: it is possible to create poor placebo tests of little relevance that will always fail to reject the null of $\rho = 0$. Thus, the credibility of the placebo estimates shown hinge on our belief that $cov(a^*_{i2}, a^*_{i1}) = 0$ and ($\Gamma_1 = 0$ or $\Gamma_2 = 0$) are only true when $\rho = 0$; otherwise, it is possible that the placebo test will give a misleading null result (a "false negative"). There is also some risk of false positives: the placebo coefficient may be nonzero despite $\rho = 0$ if our assumption that the true causal effect is zero ($\beta^{p,c} = 0$) is false.

2.4 Policy Implications: Optimal Ability Tracking

The estimates presented thus far have only shown "ability-nonlinear" peer effects; that is, the estimates only represent a nonlinearity in the ability level of each individual peer. It has also only shown heterogeneity in effects by a student's relative ability ranking in the classroom. Models with nonlinearities in composition of the classroom and broader dimensions of heterogeneity are required to derive policy implications for optimal classroom assignment of students. "Composition-nonlinear" effects are those which are nonlinear in the effect of some aggregate peer statistic, such as the proportion of students who are high ability.

The central policy implication of composition-nonlinear or heterogeneous peer effects is that some arrangements of classrooms result in higher total educational achievement than others. Because linear-in-composition model components always result in net zero estimated changes in total achievement from reassignment of students between classrooms, only credible estimates of the nonlinear components of any peer effects model are needed in order to determine the total-achievement-maximizing arrangement of students across multiple classrooms. As discussed in

Section 2.2, various placebo estimates suggest that some nonlinear effect estimates are valid while most linear components are not.

To put this idea into practice, I estimate a similar peer effects model as used in previous estimates, but instead use a student's decile of absolute ability rather than relative ability in a classroom. The ability-nonlinear component interacted with student's own ability results in average test scores that vary by classroom assignment regime, opening the possibility of optimal sorting into classrooms. I consider a hypothetical case of 40 students with 4 students per absolute ability decile who must be assigned to two Algebra II classrooms of 20 students. Using the nonlinear Algebra II peer effect estimates just presented, I simulate all assignment possibilities of students to classrooms and identify the allocation of students between classrooms that maximizes average Algebra II achievement. The score impact from heterogeneous peer effects for each classroom is the sum of products of the classroom's peer ability shares interacted with student ability decile indicators and their corresponding estimated average effects. The net gain for one classroom assignment regime over another is the difference between their heterogeneous-composition score gains. By construction, the average effect of a sorting regime from this simulation generalizes to a real-world, multiple classroom case (holding all inputs such as teachers and class size fixed).

The results of the simulation show that the nonlinear-heterogeneous profile of effects for Algebra II imply some meaningful benefits of ability tracking. Choosing a regime optimized for the heterogeneous effects identified results in a 0.03 SD average gain in test scores over a random assignment regime. Figure 13 shows the shares of students in each ability decile for each classroom for the maximum-achievement assignment regime. The effects imply that the optimal sorting regime is consistent with a traditional tracking model, separating students into high- and low-ability classrooms. The practical strength of this result is tempered by the limited criterion used to choose the optimal sorting—merely average test score—and ignores the effects of grouping on other outcomes of interest, such as test score distribution.

Whether these potential gains can be easily achieved in practice is a complicated administrative matter. The policy implications of ability peer effects in secondary schools are not directly in

parallel to the obvious "ability tracking" implications commonly explored in peer effects studies for primary school students. The secondary school setting has a comparatively large degree of student autonomy in classroom selection and it is less feasible than in primary schools for secondary schools' administrators (or district policies) to intervene directly in course selection. Instead, they would need to manipulate course offerings, academic requirements, or other incentives, which may be fraught with their own problems. Generally, knowledge of ability peer effects in secondary school classrooms may serve the role of guiding optimal curricular design and classroom offering structure instead of providing information for school administrators to directly implement (as they might in elementary schools).

3 Conclusion

This paper is another of several studies estimating peer effects using observational techniques while attempting to account for endogenous variation in assignment to peer groups. I focus on the classroom as one of the most salient peer groups with respect to educational outcomes. Classrooms are the core of educational production, yet have several avenues of social interaction that lead to causal dependency between students' outcomes. With a linear-in-shares model, I find evidence of highly nonlinear response to peer ability for high school students in nearly all tested subjects in North Carolina. I also find that a student's relative ability ranking in a classroom moderates the estimated effect, with lower-ability students benefitting the most from the nonlinearity. Drawing on complete data on students' classroom assignments to core subjects, I develop a set of placebo tests based on the composition of peers in other core subjects. The findings of the placebo tests strongly suggest that estimates of the effect of mean ability in a high school classroom are biased upward, even when accounting for school quality, teacher quality, and formal course level. However, they also suggest that most of the estimated nonlinear effects are approximately unbiased conditional on mean ability.

Peer effects that heterogeneous or nonlinear in classroom composition in general are needed to determine ability groupings of students into classrooms that result in higher overall achievement.

The combined linear-in-means and linear-in-shares model used to estimate the main results only uncovers students' nonlinear responses to individual peers' ability rather than nonlinear responses to the peer ability composition of the classroom. However, the placebo testing framework applied to the main results can be applied flexibly to any nonlinear or heterogeneous-effects specification. Using a specification allowing heterogeneous effects by own absolute (rather than relative) ability, I find that optimal sorting only on heterogeneous effects by own absolute ability produces a 0.03 SD gain in average test scores. The implied optimal sort is nearly a classic ability tracking structure, with a clearly delineated low- and high-ability classrooms, though both high- and low-ability students appear in both classrooms (likely to tap into the relatively large benefits for low-ability peers of high-ability peers).

There are several limitations to the estimates presented that future work can easily address with richer, more general specifications of peer effects. First, these estimates potentially reflect a composite of both the effect of those background characteristics and changes in endogenous student decisions (e.g., effort) that are correlated with peer ability, and understanding the mechanisms at play can result in more robust policy recommendations. Second, future models should more generally account for nonlinearities in composition and heterogeneous effects in order to develop classroom assignment algorithms that maximize potential score gains. Models such as the one used in the simulation in Section 2.4 only capture part of the potential gains of intentional ability grouping. On that point, there are further dimensions of heterogeneity that also must be explored to make optimal sorting predictions more robust. For example, there is evidence that peer effects differ by gender and race (Hoxby 2000; Hoxby and Weingarth 2006; Lantis 2014b) both in terms of the peer composition and the characteristics of the individual student. While generalization to these many nuances and complications makes for a multiplicity of additional parameters to estimate, the growth in availability of large, rich administrative data sets like North Carolina's may allow the statistical analysis of peer effects to attain greater relevance and scope.

APPENDIX

Table 22: Linear Peer Effect Component Estimates (Math and Science)

	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
	Algebra II				**Geometry**			
Mean Ability (SD)	0.22	0.22	0.13	0.2	0.17	0.17	0.18	0.17
	(0.05)	(0.05)	(0.04)	(0.04)	(0.04)	(0.04)	(0.03)	(0.04)
% Male	-0.11	-0.11	-0.08	-0.11	-0.08	-0.08	-0.07	-0.09
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
% Retained Gr. 3-8	-0.15	-0.14	-0.15	-0.14	-0.17	-0.17	-0.14	-0.11
	(0.07)	(0.07)	(0.06)	(0.07)	(0.06)	(0.06)	(0.05)	(0.06)
	Physical Science				**Biology**			
Mean Ability (SD)	0.14	0.17	0.11	0.19	0.06	0.04	0.02	0.05
	(0.03)	(0.04)	(0.04)	(0.04)	(0.02)	(0.02)	(0.02)	(0.02)
% Male	-0.04	-0.06	-0.05	-0.03	-0.07	-0.06	-0.06	-0.07
	(0.03)	(0.03)	(0.02)	(0.03)	(0.01)	(0.01)	(0.01)	(0.01)
% Retained Gr. 3-8	-0.07	-0.06	-0.05	-0.02	-0.13	-0.12	-0.1	-0.08
	(0.06)	(0.06)	(0.05)	(0.05)	(0.03)	(0.03)	(0.03)	(0.03)
School FE	X	X	X	X	X	X	X	X
Course Level FE		X	X	X		X	X	X
Teacher X School FE			X				X	
School X Year FE				X				X

Table presents estimates corresponding to the coefficient of the mean ability component of the model (λ) by specification and by class subject. Standard errors are in parentheses.

Table 23: Linear Peer Effect Component Estimates (Social Studies and English)

	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
	Civics				**English I**			
Mean Ability (SD)	0.07	0.06	0.07	0.07	0.08	0.08	0.08	0.09
	(0.03)	(0.03)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
% Male	-0.07	-0.07	-0.07	-0.07	-0.08	-0.08	-0.07	-0.07
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
% Retained Gr. 3-8	-0.17	-0.16	-0.12	-0.13	-0.09	-0.09	-0.1	-0.06
	(0.03)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)
	U.S. History							
Mean Ability (SD)	0.27	0.25	0.24	0.22				
	(0.03)	(0.03)	(0.03)	(0.03)				
% Male	-0.05	-0.04	-0.05	-0.04				
	(0.02)	(0.02)	(0.02)	(0.02)				
% Retained Gr. 3-8	-0.11	-0.1	-0.1	-0.12				
	(0.05)	(0.05)	(0.04)	(0.05)				
School FE	X	X	X	X	X	X	X	X
Course Level FE		X	X	X		X	X	X
Teacher X School FE			X				X	
School X Year FE				X				X

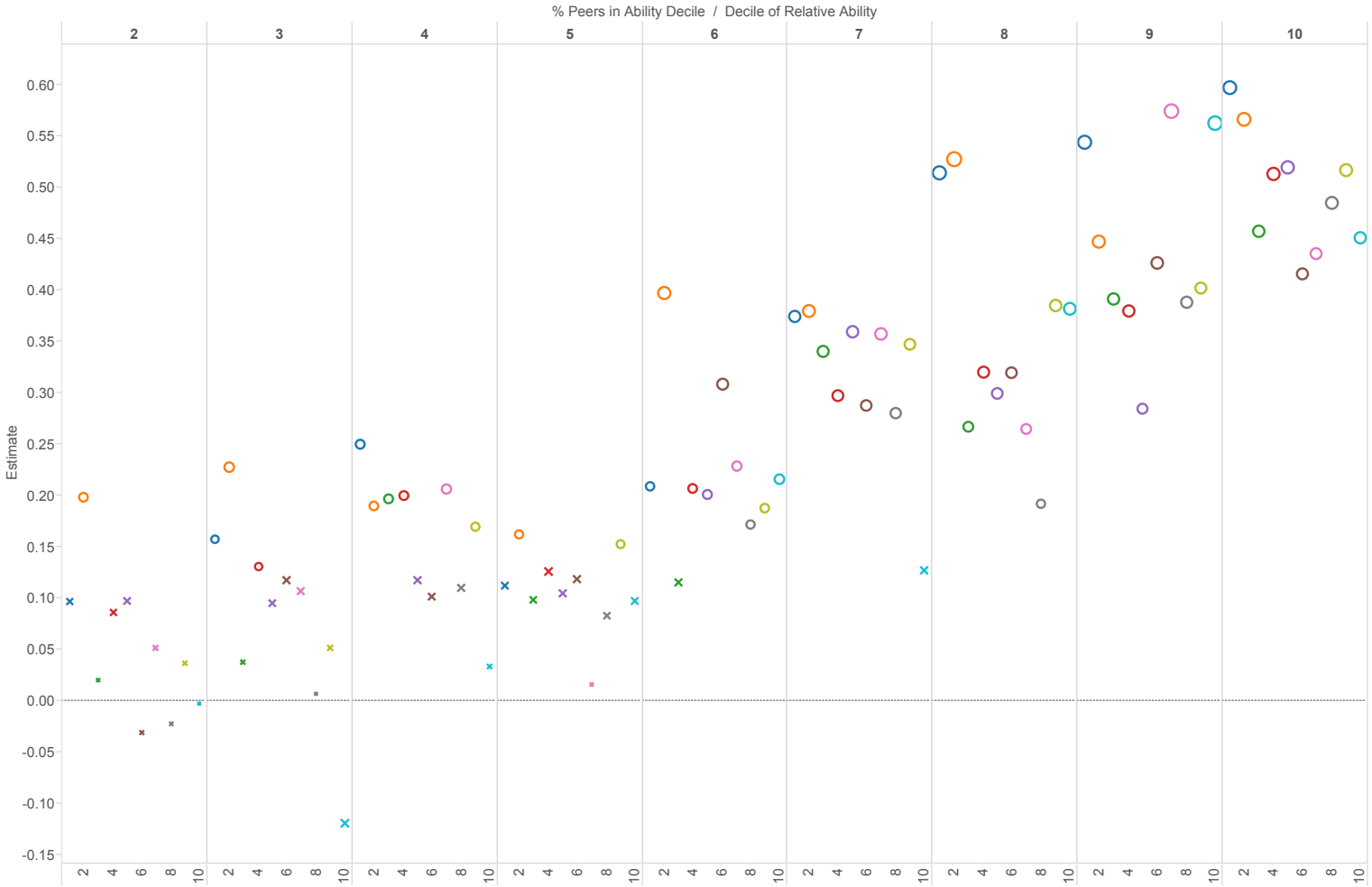
Table presents estimates corresponding to the coefficient of the mean ability component of the model (λ) by specification and by class subject. Standard errors are in parentheses.

Table 24: Linear Peer Effect Component Placebo Test (Algebra II and Biology)

Subject	**Math Placebo Sample**		**English Placebo Sample**		**Science Placebo Sample**		**Social Studies Placebo Sample**	
Algebra II	Main Class	Placebo	Main Class	Placebo	Main Class	Placebo	Main Class	Placebo
Mean Ability (SD)	-	-	0.11**	0.08***	0.08	0.10***	0.06	0.10***
	-	-	(0.06)	(0.03)	(0.05)	(0.02)	(0.06)	(0.03)
% Male	-	-	-0.08***	-0.07***	-0.06***	-0.02	-0.07***	-0.02
	-	-	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)
% Retained Gr. 3-8	-	-	-0.13*	-0.15***	-0.14*	0.01	-0.19**	-0.20***
	-	-	(0.08)	(0.05)	(0.08)	(0.04)	(0.09)	(0.05)
	Math Placebo Sample		**English Placebo Sample**		**Science Placebo Sample**		**Social Studies Placebo Sample**	
Biology	Main Class	Placebo	Main Class	Placebo	Main Class	Placebo	Main Class	Placebo
Mean Ability (SD)	-0.00	0.07***	-0.00	0.10***	-	-	-0.05*	0.11***
	(0.02)	(0.02)	(0.03)	(0.02)	-	-	(0.03)	(0.02)
% Male	-0.03**	-0.03**	-0.04**	-0.04***	-	-	-0.06***	-0.04***
	(0.01)	(0.01)	(0.02)	(0.01)	-	-	(0.01)	(0.01)
% Retained Gr. 3-8	-0.06**	-0.09***	-0.10***	-0.16***	-	-	-0.08**	-0.07**
	(0.03)	(0.03)	(0.04)	(0.03)	-	-	(0.03)	(0.03)

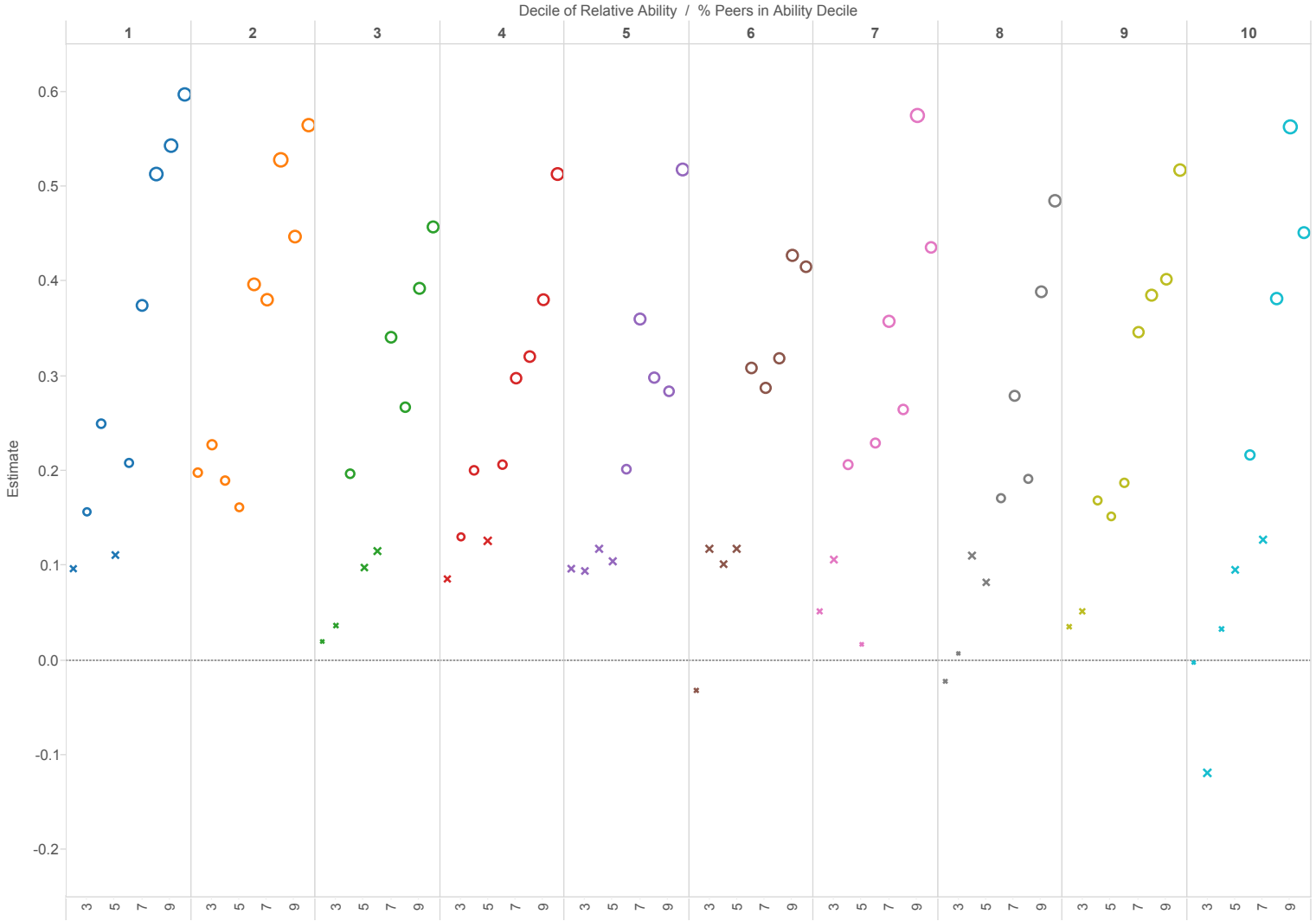
*** corresponds to p < .01, ** p < .05, * p < 0.1. Table presents estimates corresponding to the coefficient of the mean ability component of the model (λ) by specification for Algebra II and Biology alongside the corresponding placebo coefficient using mean ability from other class subjects. Standard errors are in parentheses.

Figure 12: Algebra II: Effects of Peer Ability Shares by Own Relative Ability in Classroom



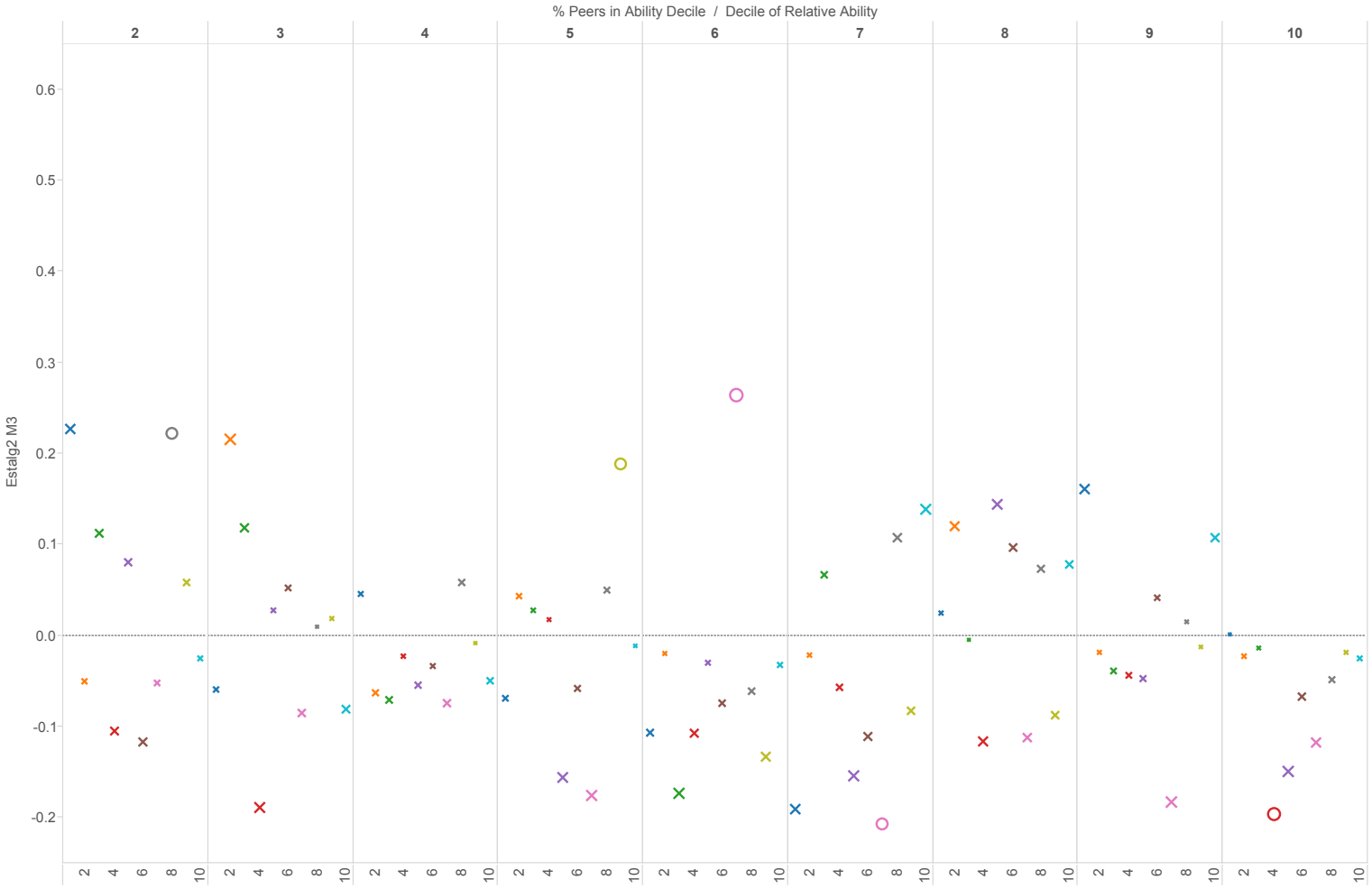
Each point represents the effect of increasing the proportion of students in a classroom with an ability score in a given decile (top axis) on a student with an ability score in a given within-classroom ability decile (bottom horizontal axis, within bins denoted by top axis). X's denote estimates statistically insignificantly different from zero, O's otherwise. The sizes of points are scaled to the absolute value of the t-statistic.

Figure 13: Algebra II: Inverted Plot of Effects of Peer Ability Shares by Own Relative Ability in Classroom



Each point represents the effect of increasing the proportion of students in a classroom with an ability score in a given decile (bottom horizontal axis, within bins denoted by top axis) on a student with an ability score in a given within-classroom ability decile (top axis). X's denote estimates statistically insignificantly different from zero, O's otherwise. The sizes of points are scaled to the absolute value of the t-statistic.

Figure 14: Algebra II: English Class Composition Placebo Test



Each point represents the estimated effect of increasing the proportion of students in the *placebo* classroom with an ability score in a given decile (top axis) on a student with an ability score in a given within-classroom ability decile (bottom horizontal axis, within bins denoted by top axis). X's denote estimates statistically insignificantly different from zero, O's otherwise. The sizes of points are scaled to the absolute value of the t-statistic.

Figure 15: Algebra II: English Class Composition Placebo Test (Inverted)

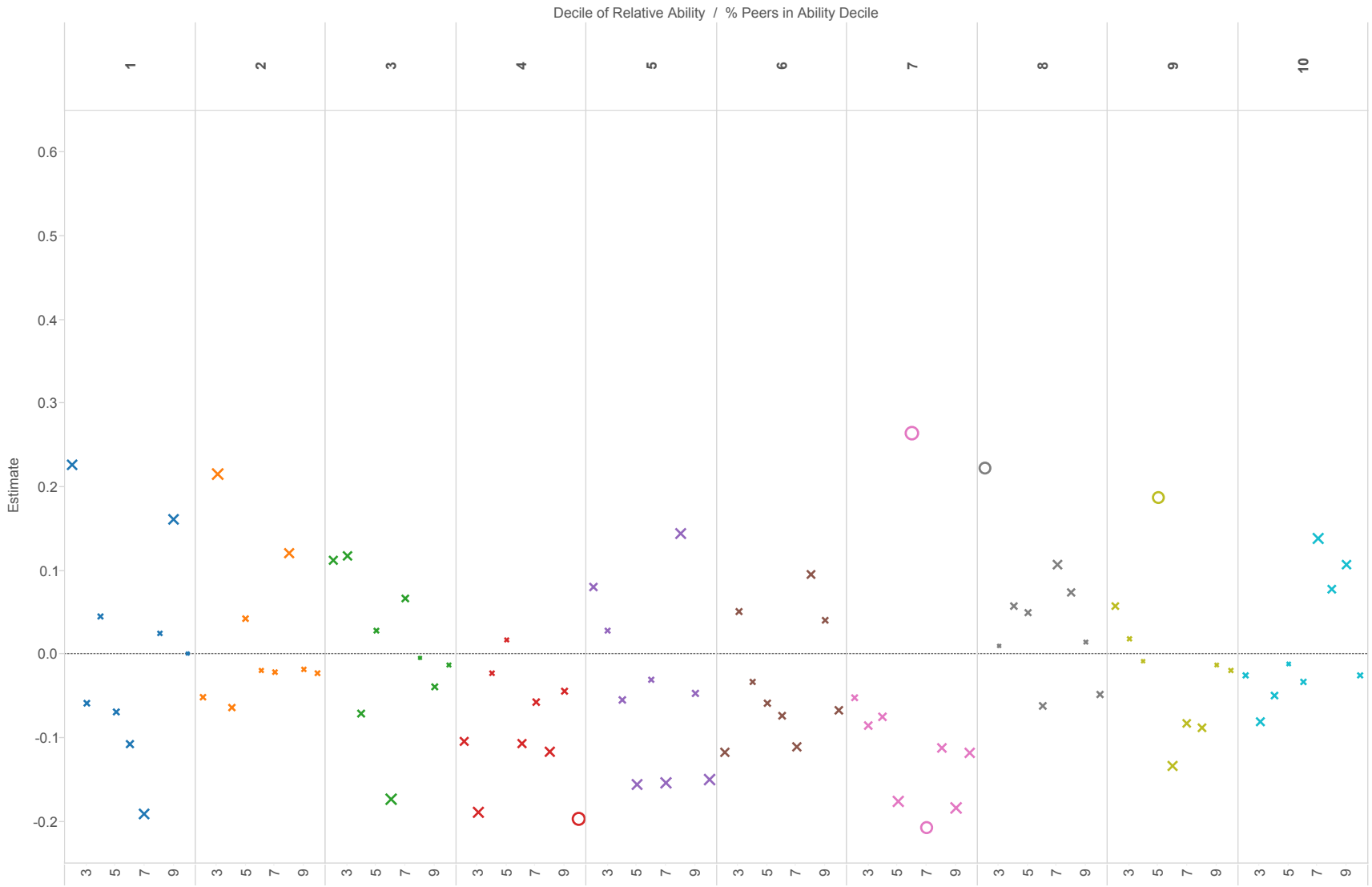


Figure 16: Algebra II: Science Class Composition Placebo Test

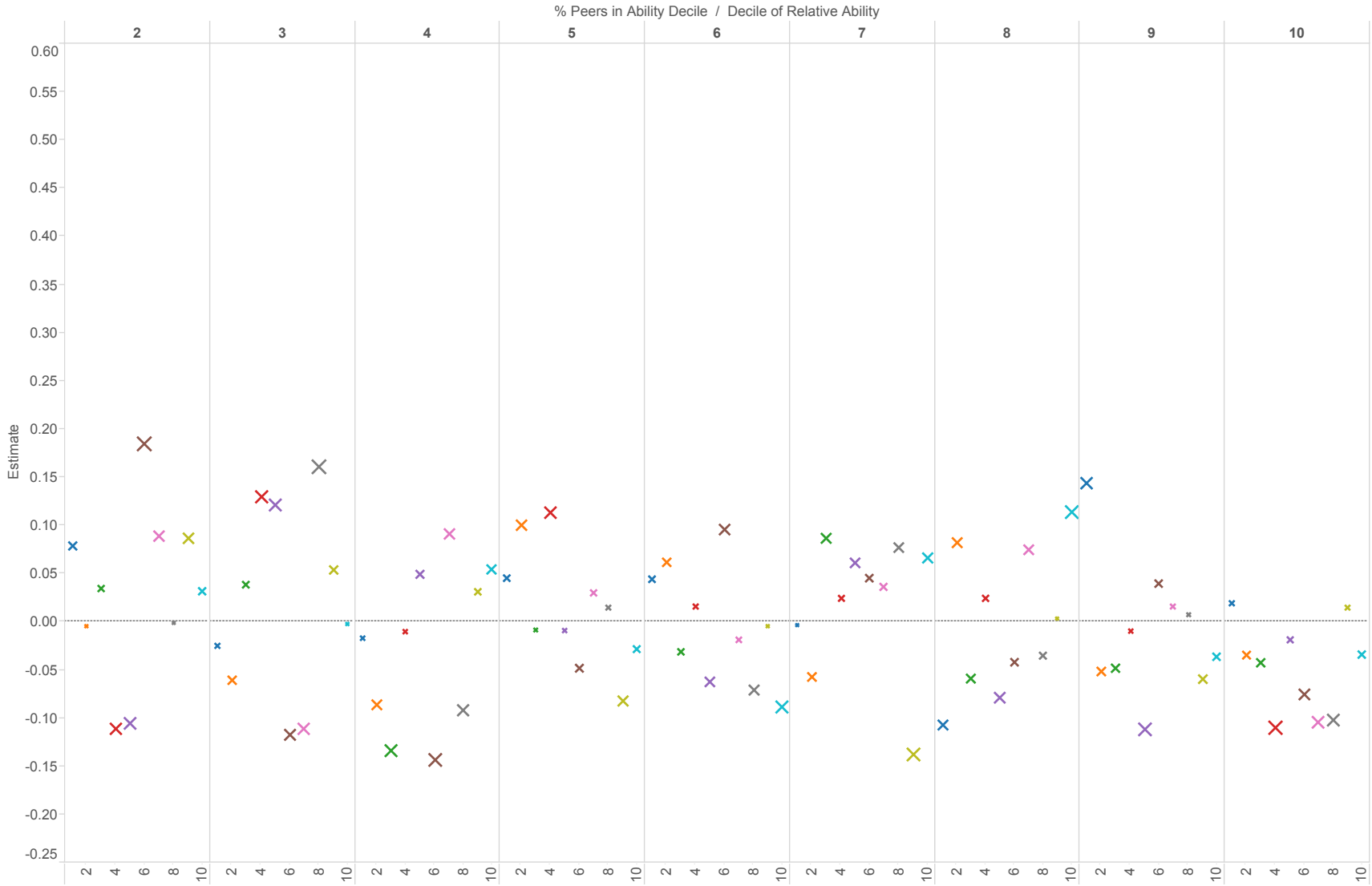


Figure 17: Algebra II: Social Studies Class Composition Placebo Test

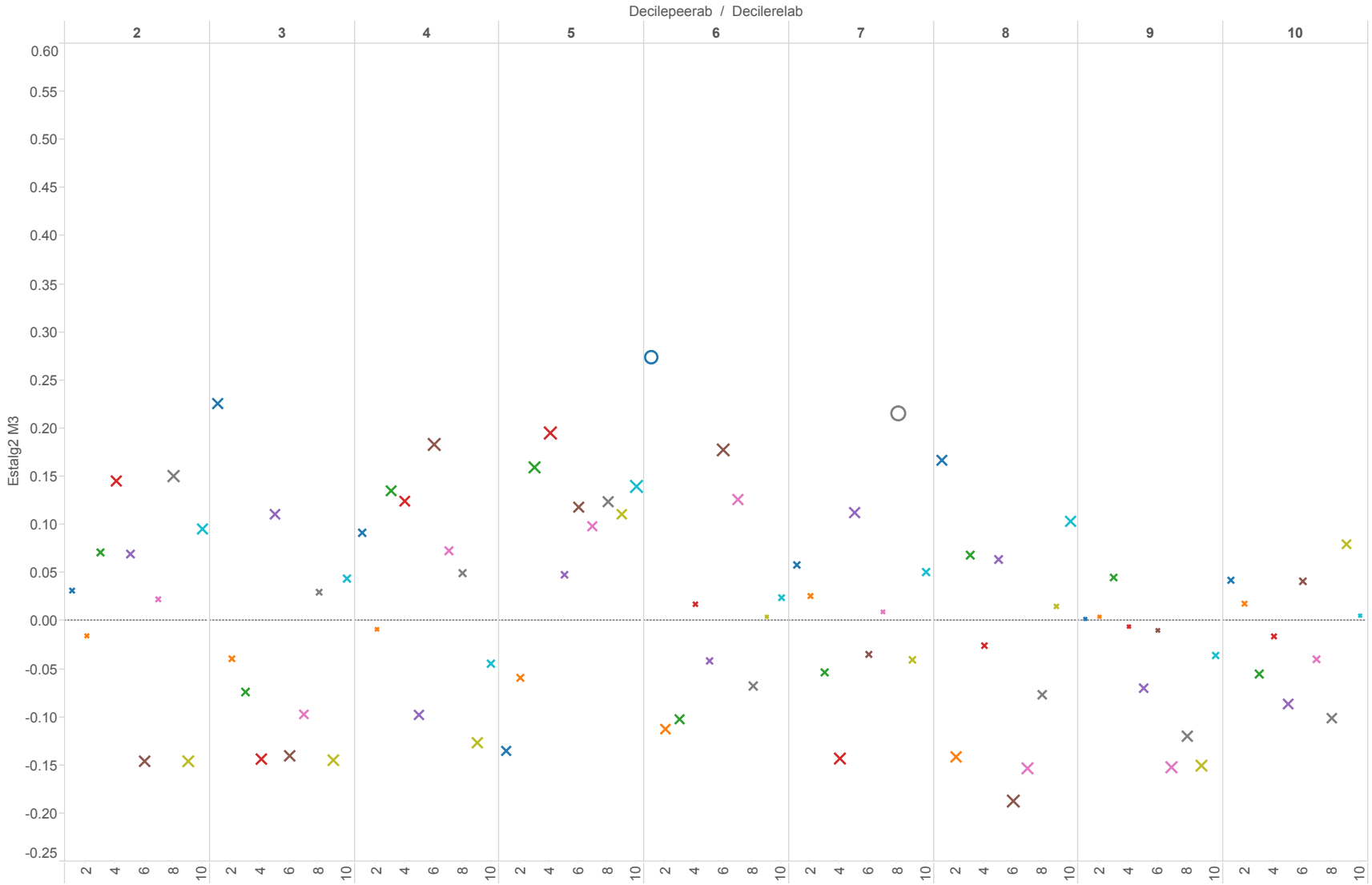


Figure 18: Peer Effects: Geometry with English Class Placebo

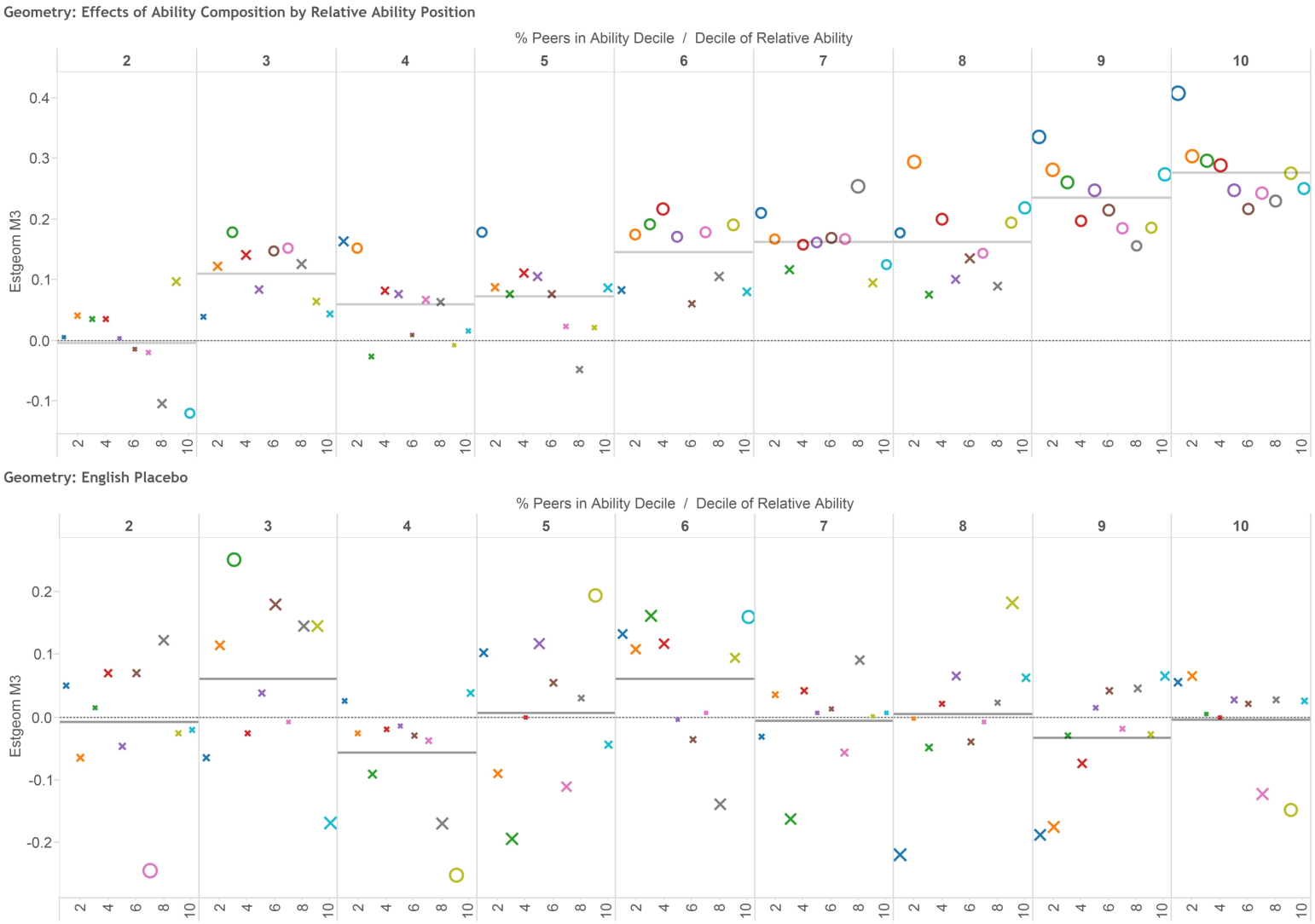


Figure 19: Peer Effects: Science with English Class Placebo

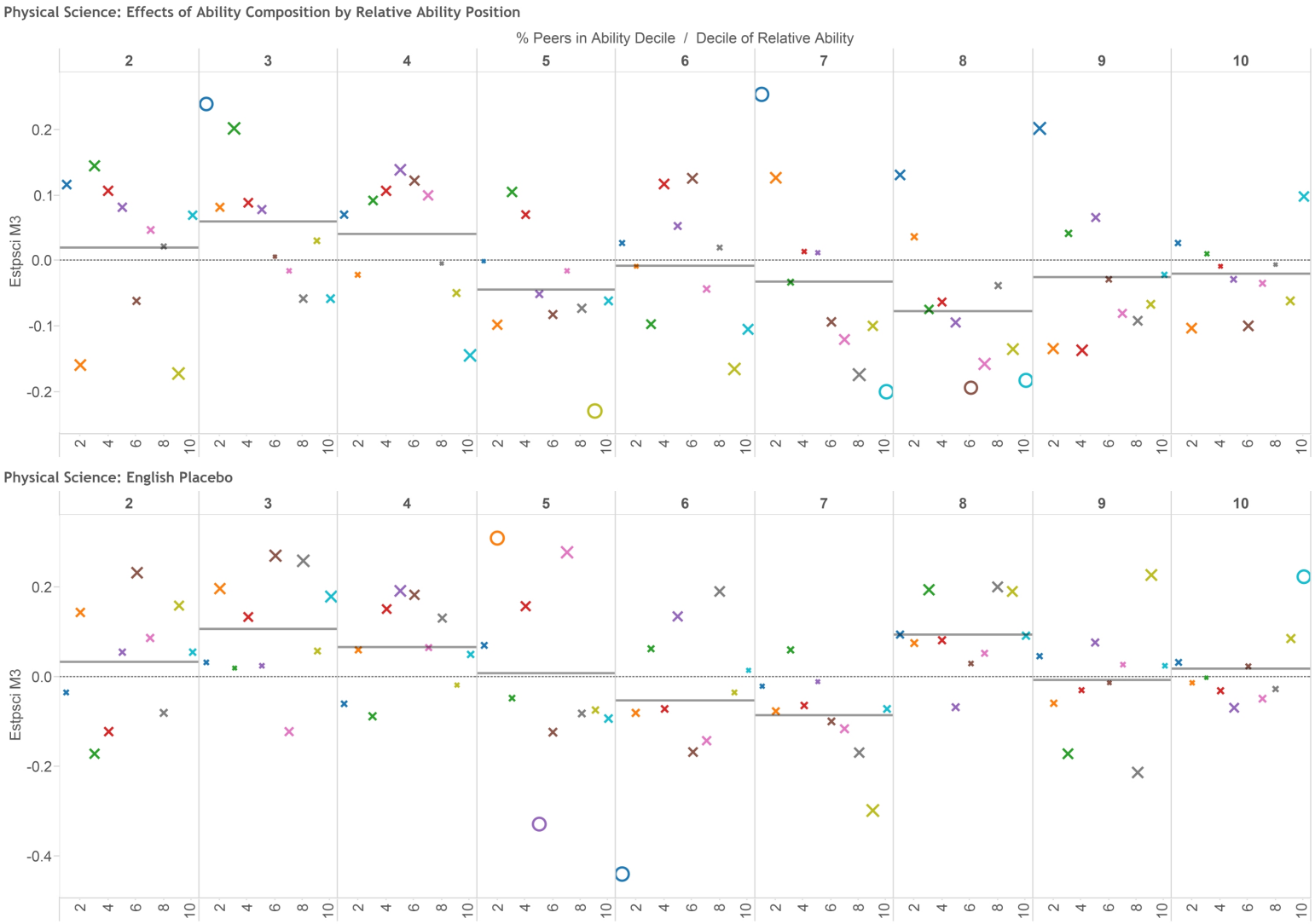


Figure 20: Peer Effects: English with Social Studies Class Placebo

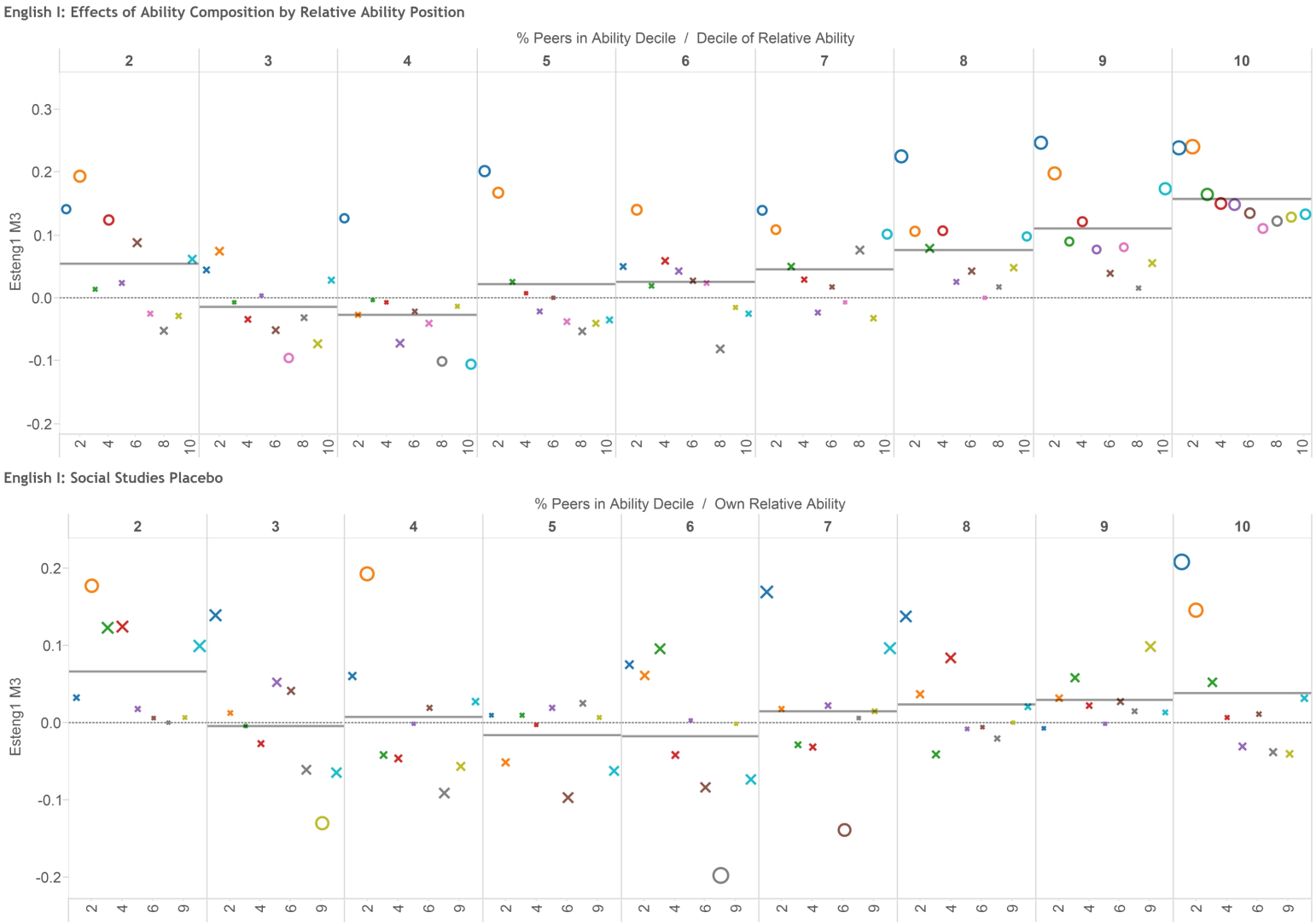


Figure 21: Peer Effects: U.S. History with English Class Placebo

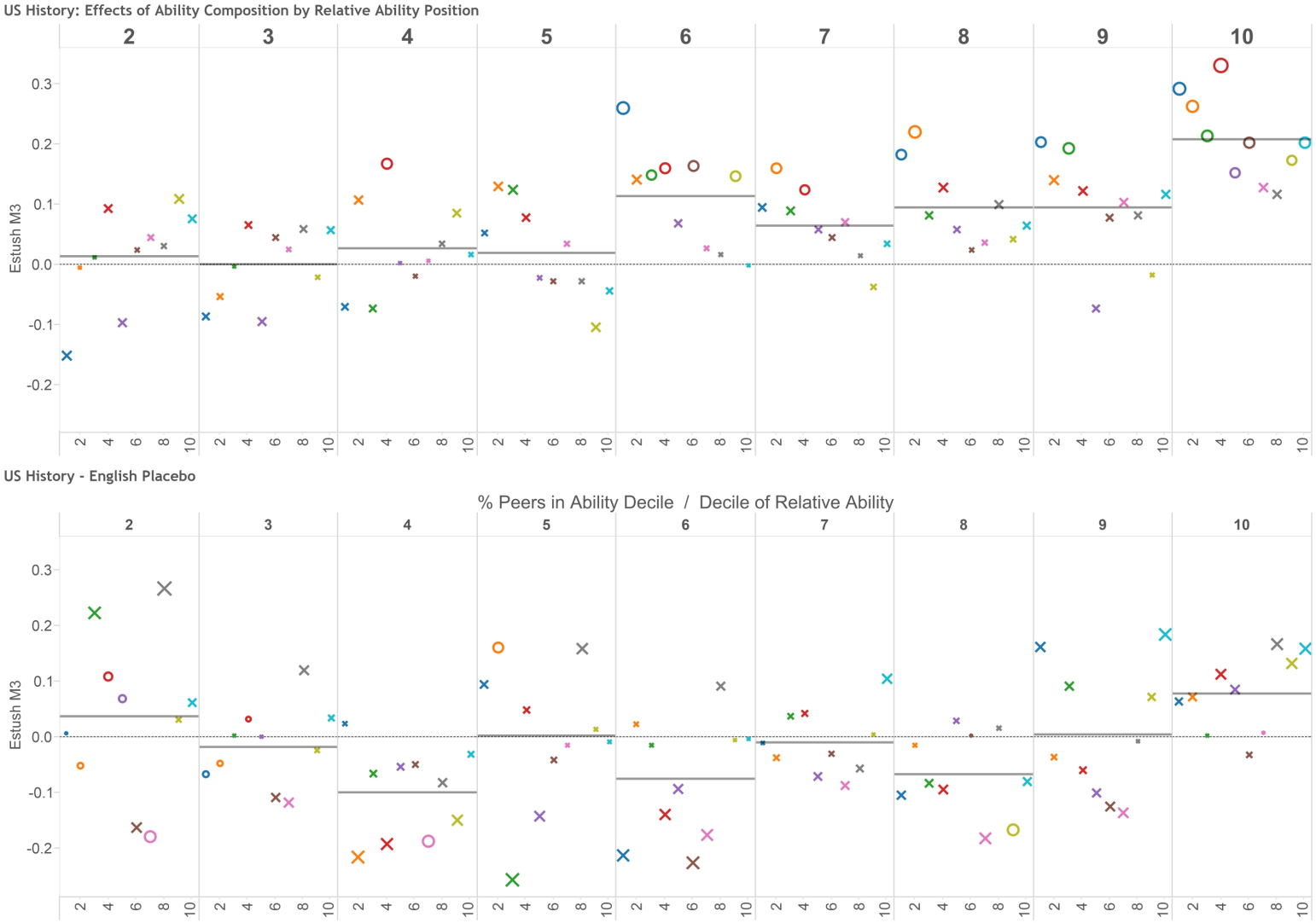


Figure 22: Peer Effects: Civics with English Class Placebo

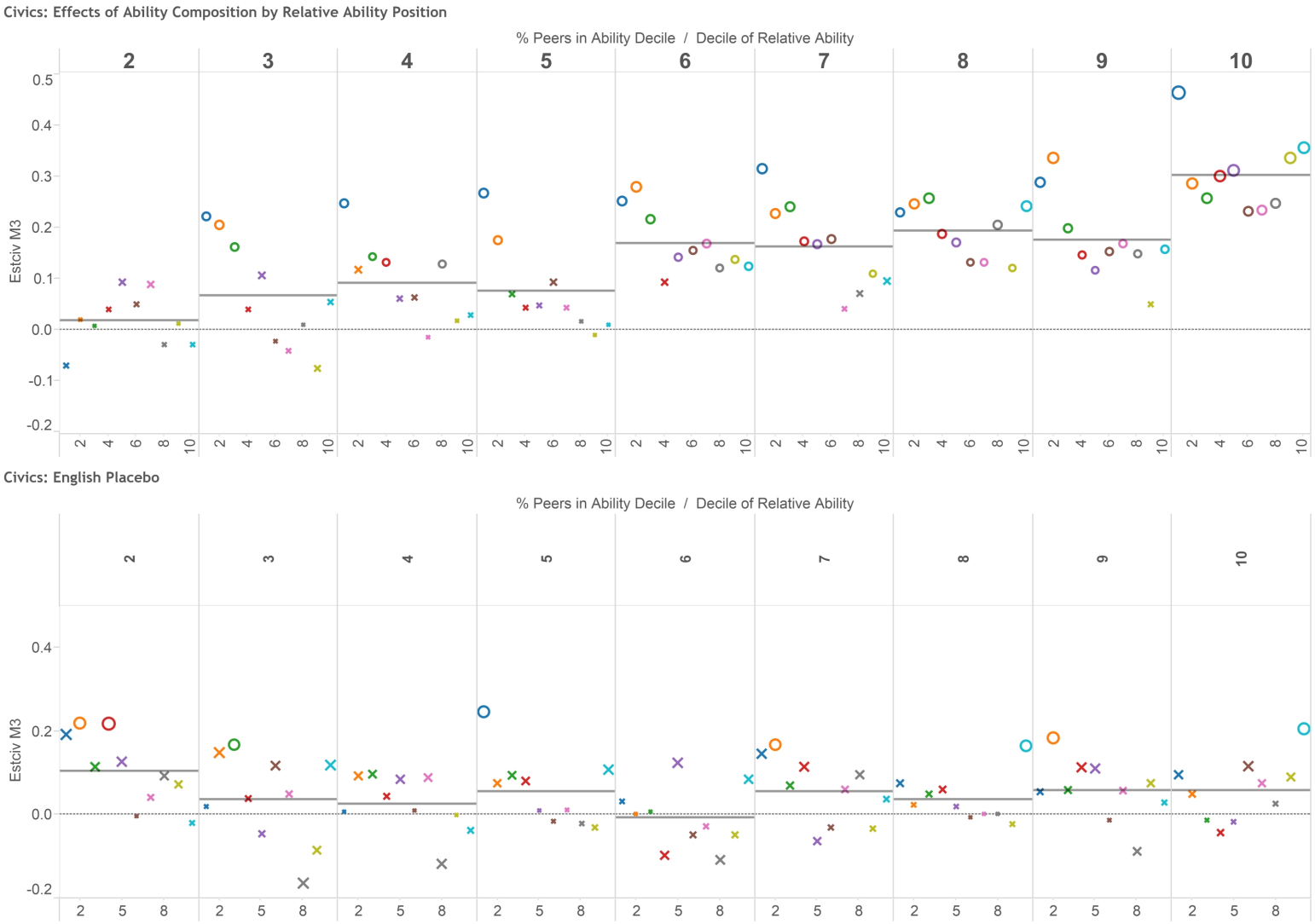


Figure 23: Peer Effects: Biology with English Class Placebo

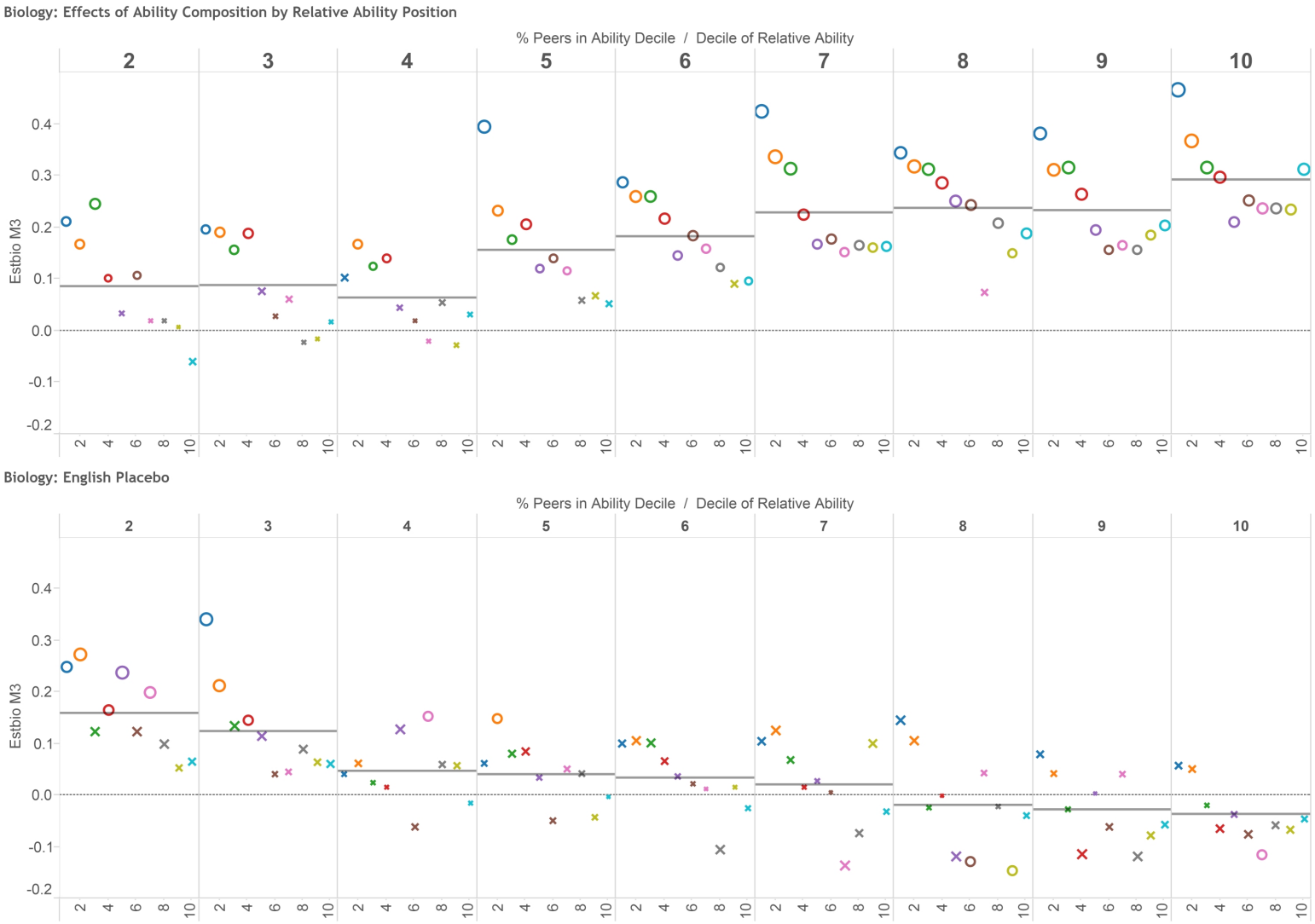
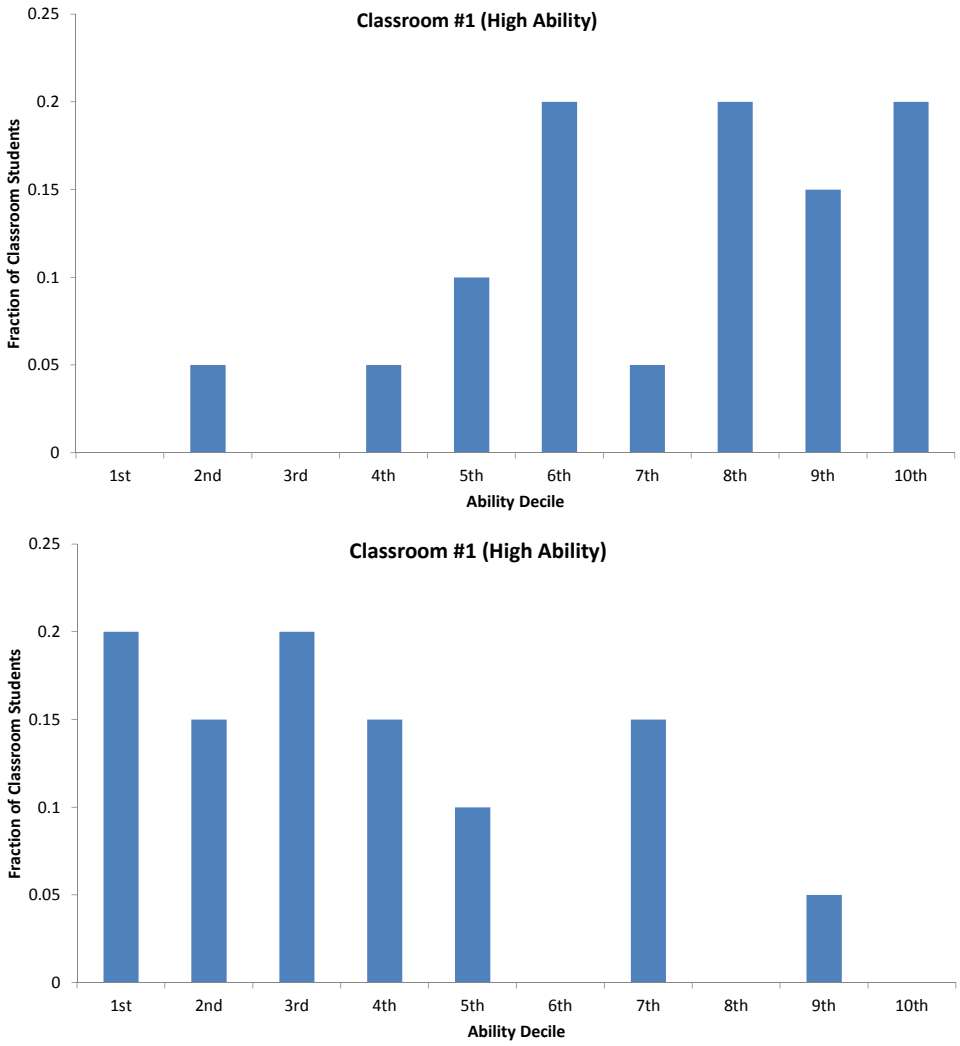


Figure 13: Optimal Classroom Sorting for Peer Effect Model with Heterogeneous Peer Effects by Own Absolute Ability



This chart shows the distribution of absolute ability in two simulated Algebra II classrooms that optimizes total mean test scores in accordance with the peer effect estimates in Section 2.4. The top plot is the first classroom, corresponding to a "high ability" mixture of students, and the second plot is for the second, "low-ability" classroom.

REFERENCES

REFERENCES

[1] Akerlof, G. A. (1997). Social distance and social decisions. Econometrica: Journal of the Econometric Society, 1005–1027.

[2] Bifulco, R., Fletcher, J. M., & Ross, S. L. (2011). The Effect of Classmate Characteristics on Post-Secondary Outcomes: Evidence from the Add Health. American Economic Journal: Economic Policy, 3(1), 25–53.

[3] Burke, M. A., & Sass, T. R. (2013). Classroom Peer Effects and Student Achievement. Journal of Labor Economics, 31(1), 51–82. http://doi.org/10.1086/666653

[4] Carrell, S. E., Fullerton, R. L., & West, J. E. (2009). Does Your Cohort Matter? Measuring Peer Effects in College Achievement. Journal of Labor Economics, 27(3), 439–464. http://doi.org/10.1086/600143

[5] Carrell, S. E., Sacerdote, B. I., & West, J. E. (2013). From natural variation to optimal policy? The importance of endogenous peer group formation. Econometrica, 81(3), 855–882.

[6] Duflo, E., Dupas, P., & Kremer, M. (2011). Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya. American Economic Review, 101(5), 1739–74. http://doi.org/10.1257/aer.101.5.1739

[7] Hoxby, C. (2000). Peer Effects in the Classroom: Learning from Gender and Race Variation (Working Paper No. 7867). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w7867

[8] Lantis, R. (2014a). Academic Performance, Effort Choice, and the Role of Peers. Retrieved from http://web.ics.purdue.edu/~rlantis/Peer%20EffectAug28.pdf

[9] Lantis, R. (2014b). Birds of a Feather Flock Together, but Does it Matter?: Inter and Intra Race Effects of Peer Ability. Retrieved from http://web.ics.purdue.edu/~rlantis/RaceEffectsEdits.pdf

[10] Lavy, V., Paserman, M. D., & Schlosser, A. (2012). Inside the Black Box of Ability Peer Effects: Evidence from Variation in the Proportion of Low Achievers in the Classroom*. The Economic Journal, 122(559), 208–237. http://doi.org/10.1111/j.1468-0297.2011.02463.x

[11] Lavy, V., Silva, O., & Weinhardt, F. (2012). The Good, the Bad, and the Average: Evidence on Ability Peer Effects in Schools. Journal of Labor Economics, 30(2), 367 – 414.

[12] Manski, C. F. (1993). Identification of Endogenous Social Effects: The Reflection Problem. The Review of Economic Studies, 60(3), 531–542. http://doi.org/10.2307/2298123

[13] Sacerdote, B. (2001). Peer Effects With Random Assignment: Results For Dartmouth Roommates. The Quarterly Journal of Economics, 116(2), 681–704.

[14] Weinberg, B. A. (2007). Social Interactions with Endogenous Associations (Working Paper No. 13038). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w1303