Air Quality and Error Quantity: Pollution and Performance in a High-skilled, Quality-focused Occupation*

James Archsmith[†]
University of Maryland, College Park
UC Davis

Anthony Heyes[‡]
University of Ottawa
University of Sussex

Soodeh Saberian[§]
University of Ottawa

Draft Date: December 12, 2017

Please review and cite the published version of this manuscript available from
The Journal of the Association of Environmental and Resource Economists
doi:10.1086/698728

*We thank Stefan Ambec, Pierre Brochu, James Bushnell, Janet Currie, David Forrest, Richard Tol, Timo Goeschl, William Greene, Erich Muehlegger, Matthew Neidell, David Rapson, Roberton Williams III, two anonymous referees and numerous seminar participants at EAERE Venice 2015, University of Sussex and UC Davis for invaluable feedback. Errors are ours. The Online Appendix is available from econjim.com/WP1601a.

†Department of Agricultural and Resource Economics, University of Maryland, College Park, MD. University of California, Davis, One Shields Ave., Davis, CA 95616 Email: archsmith@econjim.com, URL: https://econjim.com. He is grateful for financial support from the UC Davis Office of Graduate Studies, College of Letters and Science: Division of Social Sciences, and the UC Davis Department of Economics.

‡Department of Economics, University of Ottawa, 120 University Private, Ottawa, Canada. K1N 6N5. He is also a part-time professor of economics at the University of Sussex. Heyes is Tier 1 Canada Research Chair (CRC) in environmental economics and the financial support of the CRC Program is acknowledged.

§Department of Economics, University of Ottawa, 120 University Private, Ottawa, Canada. K1N 6N5.

Abstract

We provide the first evidence that short-term exposure to air pollution affects the work performance of a group of highly-skilled, quality-focused employees. We repeatedly observe the decision-making of individual professional baseball umpires, quasi-randomly assigned to varying air quality across time and space. Unique characteristics of this setting combined with high-frequency data disentangle effects of multiple pollutants and identify previously under-explored acute effects. We find a 1 ppm increase in 3-hour CO causes an 11.5% increase in the propensity of umpires to make incorrect calls and a 10 $\mu g/m^3$ increase in 12-hour $PM_{2.5}$ causes a 2.6% increase. We control carefully for a variety of potential confounders and results are supported by robustness and falsification checks. Our estimates imply a 3% reduction in productive output is associated with a change in CO concentrations equivalent to moving from the 25th to the 95th percentile of the CO-distribution in many of the largest US cities.

JEL: J24, Q52, Q53

Keywords: Air quality, Labor productivity, Cognition

1 Introduction

Policymakers attach high priority to protecting air quality. The costs of air quality policies are known to be substantial (see *e.g.*, Greenstone, List, and Syverson (2012)) but the benefits of cleaner air are not as well understood and may accrue in various forms. The focus of most research has been on health impacts, which provide the usual rationale for policy intervention in this area.[1] However recent and emerging evidence suggests that polluted air may impose a more direct economic cost by negatively impacting how well people perform at work. Insofar as such effects are substantial, measuring improvements in labor productivity is likely to be an important component of valuing the benefits of clean air.[2]

In their important work Graff Zivin and Neidell (2012) and Chang et al. (2016a) provide persuasive evidence that short-term exposure to ozone (O_3) and fine particulate matter ($PM_{2.5}$) significantly reduce the daily productivity of laborers engaged in physical work (fruit picking and packing).[3] While these are seminal contributions, their direct implications - particularly for developed countries and in urban settings (where air quality problems are likely to be most pronounced) - are limited by their focus on unskilled *physical* work without significant mental dimension. In an economy like the United States, the share of workers in physically-demanding occupations comparable to fruit-picking is only around 15%, and even lower for older age groups likely to be the most susceptible to the effects of pollution (Rho (2010)).[4]

Most work, and in particular almost all *high value* work, in a modern economy is based on high levels of mental dexterity, often with little or no physical dimension (*e.g.* lawyers, air traffic controllers, surgeons, train drivers, computer programmers). Even in manufacturing, modern work practices mean that employment is increasingly about "brain rather than brawn" - operating precision machinery, for example, or supervising computer-controlled production processes, in a way that requires concentration and finesse.

We provide what we believe to be the first evidence of a *causal* effect of short-term (daily and intra-day) variations in air pollution on the *quality* of work done by a group of highly-skilled professionals engaged in mentally-demanding employment, namely Major League Baseball (MLB) umpires. We exploit attributes

[1] A large epidemiological literature provides evidence of the effect of short-term variations in common air pollutants on various health outcomes including heart attack (Gold et al. (2000)), stroke (Oudin et al. (2010)) and asthma (Neidell (2009)).

[2] Both Chay and Greenstone (2005) and Bento, Freedman, and Lang (2015) use hedonic analyses of house prices in areas regulated under the CAAA to estimate households' willingness-to-pay (WTP) for reductions in particulate matter pollution. Such estimates capture those benefits that are capitalized into housing prices but are agnostic to the source of those benefits. Given how poorly productivity effects are currently understood it is at least plausible that household WTP would fail fully to consider such things, in which case this sort of study would understate the benefits of pollution regulations. In addition, if labor is complementary with other factors of production, we would expect employers to capture some portion of the labor productivity improvements.

[3] $PM_{2.5}$ is particulate matter smaller than 2.5 microns in size. These particles are small enough to penetrate deep into the lungs and enter the bloodstream. It also penetrates indoors quickly and almost completely.

[4] The shares quoted are those of the common job categories identified as 'very physical' (including janitors, building cleaners, grounds maintenance workers, material movers, construction laborers, etc.).

of their employment setting favorable to causal identification and the measure of performance quality that is collected by their employer for the purposes of performance management.[5]

We will be cautious about the degree to which our results, based as they are on MLB umpires, may be generalized to a wider set of high-skilled workers. Other researchers have utilized this same employment situation and associated highly granular data, to identify effects which would likely be obscured by unobservables in other settings. For example, Parsons et al. (2011) use ball and strike calls by MLB umpires to examine racial discrimination (umpires are more likely to make mistakes that favor a pitcher of their own race). Chen, Moskowitz, and Shue (2016) use the measure to test for autocorrelation in decision-making. Kim and King (2014) use the quasi-random assignment of umpires to games to provide evidence supporting the so-called Matthew Effect whereby prior professional status (in their case, the number of career All Star Game appearances) affects third party performance evaluation.

Just as Parsons et al. (2011) is neither written, nor should be read as being, 'about' racism in baseball - but rather the MLB setting is taken as a microcosm for things that might be happening more broadly in society - Chen, Moskowitz, and Shue (2016) and Kim and King (2014) are of interest only because each points to an evaluative bias that might be expected to repeat in firms and organizations far away from professional baseball diamonds. While the work tasks that umpires execute are particular, they require repeated cognitive and sensory attention over an extended period of time. Many jobs that are important to the economy rely on tasks which tax similar mental and sensory systems.

While MLB provides the unique laboratory within which we test our hypotheses, this is *not* a paper about baseball. Researchers face a number of challenges in seeking to disentangle air quality from other determinants of productivity. Our setting allows us to overcome the three most important.

First, there exists a clean, consistent measure of individual-level performance or productivity - namely the production of correct 'calls' on balls and strikes. Recent technological developments mean that since 2008 performance has been observable with a high degree of accuracy. Unlike most work settings our measure of productivity is not jointly produced and does not suffer from potentially unobservable variations in other inputs (*e.g.*, capital, technology, effort from employees).

Second, the assignment of umpires to games (and therefore pollution treatments) is quasi-random. The schedules of umpires are determined and published weeks before the start of the season. As such we

[5] There is a separate strand of research that investigates the impact of air pollution on the athletic performance of athletes such as as marathon runners (Rundell (2012), Marr and Ely (2010)), Bundesliga soccer players (Lichter, Pestel, and Sommer (2015)) and ATP tour tennis players (**salvo2015**). Although such activities may have something in common with physical labor in an agricultural or other setting - and it is well-established that exposure to air pollution may compromise human physical capabilities - our paper does *not* fit into this strand. While our subjects happen to be employed in the sports industry, umpires are not sports-people nor are they engaged in a primarily physical endeavor.

can ignore issues of self-selection of umpires into particular air quality conditions.[6]

Third, the data landscape is extremely favorable. There are 30 MLB teams each of which plays 162 games per season, so even after some attrition we have a lot of data to work with - our main specifications are estimated on over 620 000 data points. More importantly we were able to find high-quality measures for all of the controls that we wished to apply, and we believe that our design controls very well for a wide variety of potential confounders.

We argue that the analysis makes plausible a link between air quality and day-to-day variations in workplace productivity for a broader range of jobs than baseball umpiring. As such the results complement and extend the nascent literature on physical and non-physical but semi-skilled work tasks. Evidencing portability of the results to other lines of employment is an important ambition in future research.[7]

The central results of our analysis are that ambient carbon monoxide (CO) and fine particulate matter ($PM_{2.5}$), at levels well below the respective EPA acute exposure standards of 9 ppm and 35 $\mu g/m^3$, have a significant negative effect on how well this group of workers do their job at any particular time. The fact that we find acute effects highlights the advantages of our research setting. The fine-grained structure, volume, and rich variation allow us to discern robust and well-identified effects larger than previously believed. Our use of more tightly-defined rolling time blocks - with duration chosen with reference to physiological fundamentals - gives us better traction in identifying the role of CO and $PM_{2.5}$.[8] Results make clear that exposure to elevated CO and $PM_{2.5}$ levels for just a few hours have a substantial effect on work performance which erodes quite quickly once ambient levels fall (the half-life of CO in the human body is between 3 and 4 hours).

Our preferred estimates indicate that a 1 ppm increase in 3-hour CO causes an 11.5% increase in the propensity of umpires to make incorrect calls (an extra 2.0 incorrect calls per 100 decisions). Likewise a 10 $\mu g/m^3$ increase in 12-hour $PM_{2.5}$ causes a 2.6% increase in the propensity of umpires to make incorrect calls (an extra 0.4 incorrect calls per 100 decisions). We control carefully for a variety of potential confounders

[6]In some other work settings there is a more pronounced concern about the extensive margin - that is whether air quality might influence a worker's decision whether or not to go to work, or how many hours to work. In addition in some professions the worker may be able vary the location of work. None of these concerns apply here.

[7]Another aspect of the work of umpires is that it takes place (predominantly) outdoors, whereas many other professionals in cognitively-demanding roles work exclusively indoors. However, unlike some other pollutants both CO and $PM_{2.5}$ efficiently penetrate buildings through physical openings and mechanical ventilation systems. The correlation between concentrations within and immediately outside a building are typically 90 to 100%. (see *e.g.*, Thatcher and Layton (1995), Vette et al. (2001), and Ozkaynak et al. (1995).) An additional concern may be that the vision-intensive nature of the job means that the challenge level is influenced *directly* by variations in air quality through changes in visibility. However the pollutants that we study have no discernible impact on visibility over the short distances involved (the distance from pitcher's mound to home plate is sixty feet). Carbon monoxide, for example, is an invisible gas.

[8]Most previous work (e.g., Aragón, Miranda, and Oliva (2016) and Chang et al. (2016b)) employs daily average pollution data. Graff Zivin and Neidell (2012) compute a workday average pollution level using hourly data. Chang et al. (2016a) use a 6-day average. Hausman, Ostro, and Wise (1984) and Ostro (1983) use annual data.

and the results prove robust to a battery of robustness and falsification checks.

To provide a better feel for what these effect sizes might mean in practice, we interact these point estimates with what we know about the distribution of air quality levels in the twenty largest US Metropolitan Statistical Areas (MSAs). Moving from the 25th to the 95th percentile in terms of CO pollution in Phoenix, for example, causes a decrement in the probability of production of a *correct* call of 2.9%. In Los Angeles that number is 2.7%. Moving from the 25th to the 95th percentile in terms of $PM_{2.5}$ reduces the probability of production of a correct call by 1.3% in Los Angeles, 1.0% Philadelphia. These effects are separately-identified and additive.

The layout of the rest of the paper is as follows. In Section 2 we provide a review on previous literature about air quality and productivity. In Section 3 we outline the key elements of the employment setting that we study. Section 4 and Section 5 describe data and methods. Primary results are presented and discussed in Section 6. A variety of robustness checks and falsification exercises are presented in Section 7. Section 8 concludes.

2 Air quality, productivity and mental performance

It has recently been established that - in addition to adverse health effects - exposure to pollution can significantly reduce workplace productivity (Chang et al. (2016a)). There are a number of ways in which air pollution might influence labor productivity. One obvious path is through attendance at work and absenteeism (though this sort of effect will not drive the results in our paper). Another is that it might impact the functioning of the human body or brain in ways that affect a worker's cognition, ability to concentrate, decision-making, etc. (Heyes, Neidell, and Saberian (2016)). It might also hinder visual perception. While we are going to have little to say about the precise physiological mechanism(s) at play in our setting, here we review some existing evidence that may be pertinent.

Using repeated cross-sectional surveys Ostro (1983) finds a 1 $\mu g/m^3$ increase in total suspended particulates (TSP) is associated with a 0.00145 day increase in work days lost during each two week survey period. Employing a similar dataset, Hausman, Ostro, and Wise (1984) find a one standard deviation increase in TSP results in 10% increase in work days missed. More recently, Aragón, Miranda, and Oliva (2016) find a non-linear response of household labor supply to increased levels of fine particulates in Peru. Recent research has employed micro-level data on worker output which allows researchers to control for individual-level heterogeneity and examine changes in productivity on both the extensive (decision to work) and intensive (level of productivity conditional on working) margins. Graff Zivin and Neidell (2012) find a 10 ppb decrease in O_3 concentrations leads to a 4.2% increase in productivity of outdoor agricultural workers. However, higher O_3 levels are not associated with increased absenteeism or reduced total hours worked, so the effects of O_3 are limited to reduced productivity while working. Chang et al. (2016a) find higher outdoor $PM_{2.5}$ levels lead to lower productivity for indoor workers at a pear-packing plant. They find the expected result that outdoor O_3 has no effect at this indoor plant.[9]

In related work Chang et al. (2016b) show that indoor workers at travel agency call centers in two highly-polluted Chinese cities handle fewer calls on high AQI days. Their work complements the results that we present. While the workers in their setting are engaged in non-physical tasks that work remains low- to semi-skilled, likely to require a fraction of the mental challenge and sustained concentration facing the subjects in our study. In an insightful decomposition of their results, they show that the reduction in daily calls handled is driven by workers taking longer breaks on more polluted days, rather than handling calls less quickly, so the central result is more akin to an intra-day labor supply effect - less time spent available for work - than a 'pure' productivity effect. Their setting also does not allow for observation of quality of work.

Outside employment contexts - but still pertinent for us given our interest in cognitively-intensive settings - Lavy, Ebenstein, and Roth (2014) separately examine the association between ambient concentrations of a number of local criteria pollutants on the performance of Israeli students taking the *Bagrut*, a high-stakes high school exit exam. They find a one-unit increase in $PM_{2.5}$ leads to a 0.046 standard deviation decrease in test scores. Likewise they find a one-unit increase in CO AQI leads to a 0.085 standard deviation decrease in test scores. They also find evidence that the effects of these pollutants are non-linear, with the majority of the effect occurring at levels above an AQI of 100. Roth (2016) exploits panel methods to identify a link from indoor measured fine particulate matter ($PM_{2.5}$) to reduced exam scores of a set of students taking university-level exams in London, though he is unable to account for the role of other (likely correlated) pollutants. Heyes, Rivers, and Schaufele (2016) find that elevated $PM_{2.5}$ in Ottawa significantly reduces the quality of speech - which they claim as a mentally-taxing task - of a panel of Canadian MPs, with a threshold effect at 15 $\mu g/m^3$ but little effect at lower levels.

Due to its known toxicity, few controlled experiments assessing the impacts of CO on cognition and mental acuity have been done. Beard and Wertheim (1967) expose human male subjects to CO levels between 50 and 250 ppm then test their ability to discern the relative duration of machine-generated tones. They find an approximately linear deterioration in correct responses over the range of exposure, with correct responses

[9] O_3 is is highly reactive and breaks down quickly indoors. This contrasts with the pollutants for which we will report significant effects in this paper.

decreasing by approximately 0.2% for each additional ppm of ambient CO. However, multiple attempts to replicate this study have failed to reproduce this result (see for example: Raub and Benignus (2002)). Amitai et al. (1998) find diminished performance of university students on some components of the Comparison of Neuropsychological Screening Battery (CONSB) when exposed to ambient CO concentrations between 17 and 100 ppm. Subjects in this study are exposed to much higher doses (levels eight to one hundred times higher) than those experienced by the workers we examine, allowing those authors to discern statistically significant effects in a study involving just 45 students.

Evidence on the impacts of other pollution, particularly particulate matter, is even sparser. Physiologically, short-term exposure to $PM_{2.5}$ is associated with inflammation and oxidative stress in the brain (Kleinman (2014)), microglial activation, cerebro-vascular dysfunction, and alterations in the blood-brain barrier of the central nervous system (Genc et al. (2012)). These effects can lead to symptoms such as memory disturbance, fatigue, loss of concentration and judgment (Kampa and Castanas (2008)), any of which could plausibly be linked to reduced mental acuity and so decreased performance in work tasks that require mental acuity.

3 Employment setting: The work of MLB umpires

Umpiring baseball is a skilled job that requires sustained concentration and mental effort. We study professional umpires in their places of employment, officiating baseball games in MLB venues. MLB employs around 100 umpires in any given season. They are organized into teams ("crews") of four with each serving as the "home plate" umpire every fourth game. The composition of each crew - and their work schedule for the season - is announced several weeks before the start of the season to allow for travel planning. It is a well paid career, with an experienced umpire commanding a base salary of 350 000 USD per season, which can be supplemented by post-season assignments and additional speaking or writing engagements.

The most significant task that the home plate umpire faces in a working day is 'calling' the game - arbitrating which pitches are balls and which are strikes. The accuracy and consistency of this calling is fundamental to the game. In this study we use the success of an umpire in the production of correct calls as our measure of performance. Of course this is only one element of what an umpire does in the course of work, but it is plausibly the most important, and one to which the employer attaches high weight in employee evaluation.

A pitch should be called a strike if any portion of the ball passes through the strike zone (see

Figure 1).[10] In an average game, the home plate umpire is required to adjudicate about 140 pitches, a little under half of the pitches thrown in a game (in many cases the umpire is not called upon to make a call - for example if the pitch is hit by the batter). On each pitch there is an objectively correct call, which means that we have an unambiguous measure of how well the umpire has performed. Success in generating correct calls is the key performance measure faced by this group of employees. MLB operates a robust system of monitoring and incentives which is called the Supervisor Umpire Review and Evaluation (SURE) system. This system "...uses on-site supervisors, semi-annual evaluations, high-end technology and incentives like play-off money and suspensions to keep track of how umpires are doing" (Drellich (2012)). Two reports on umpire performance are filed by supervisors after each game. The central component of one of these is 'zone evaluation' which uses a high-precision pitch-tracking technology called PITCHf/x. Since 2008 this technology has been in operation at every MLB ballpark and - amongst other things - provides an objective measure of balls and strikes against which an umpire's decision-making can be compared. An error rate above a certain threshold triggers a performance review and, more generally, this metric is central to how MLB appraises this group of employees.[11]

PITCHf/x supplies the raw data underpinning the on-screen pitch maps provided in real-time during ballgames by many US broadcasters. For each game it generates a spatial scatter-plot of the true locations of pitches upon which the umpire is required to call. Figure 3 is a plot of the locations of pitches from a single game. Correct calls are shown as hollow shapes and incorrect calls as solid black. Umpires make Type 1 and Type 2 errors. A black triangle captures a pitch that passed outside the strike zone, but which the umpire judged to have been inside. Conversely, a black circle is a pitch that passed through the strike zone, but which the umpire called as a ball.[12]

During a regular MLB season, the typical umpire handles 142 games, serving as the home plate umpire in one quarter of those games. Games are played between 30 teams in 26 different cities in the United States plus Toronto. To minimize travel MLB uses an optimization algorithm to set timetables for crews subject to a variety of constraints (Trick, Yildiz, and Yunes (2011)). Umpiring assignments are

[10] Rule 2.00 of the baseball rules defines the strike zone to be "...that area over home plate the upper limit of which is a horizontal line at the midpoint between the top of the shoulders and the top of the uniform pants, and the lower level is a line at the hollow beneath the kneecap, determined from the batter's stance as the batter is prepared to swing at a pitched ball."

[11] According to the current agreement between MLB and the umpire's union, MLB uses PITCHf/x data to provide feedback and evaluate umpires' performance: "...(S)ubstandard performance can influence his promotion to crew chief, assignment to lucrative post-season games, or even retention in MLB." (Drellich (2012)).

[12] Umpires have idiosyncrasies in how they will call pitches in particular locations relative to the strike zone. One umpire may have a tendency to call too many low strikes, for example. We control for such idiosyncrasies in our regressions with umpire-specific, nonparametric pitch location dummies. More specifically for each umpire and batter handedness, we include dummies for pitches farther left than the left 20% of the strike zone, pitches farther right than the right 20% of the strike zone, and pitches in the middle of the strike zone. Our results prove undisturbed to dropping these controls.

approved by the MLB commissioner about two months before the season begins.[13]

Importantly the setting makes plausible our identifying assumption, namely that after controlling for time and location fixed effects the assignment of umpires to air quality conditions is as good as random.

4 Data

Our objective is to explore whether changes in air quality impact how well an umpire calls balls and strikes. We exploit data from a variety of sources. A description of these data follows and Table 1 presents key summary statistics.

4.1 Pitches and calls

As already noted we rely on detailed information on the decision-making of MLB umpires using the PITCHf/x pitch-tracking system. This is a data-collecting system installed at all 30 MLB venues using multiple tracking cameras to record every pitch's trajectory with an accuracy of one inch as it travels from the pitcher to the batter. PITCHf/x data are collected by Sportsvision and provided through the MLB's website.

We collect data on pitches thrown in games officiated by full-time MLB umpires played in the 2008 through 2015 seasons inclusive. We exclude Toronto which is outside the US and for which we do not have consistent air quality data. We also exclude a small number of games in which the equipment was not operational or appears miscalibrated, or which were called by a non-full-time umpire (though in a robustness check we confirm that reinserting these makes little difference to the main results).

Our focus is on pitches where the umpire is forced to make a decision between one of two ex post objectively variable states, calling a pitch in flight a "ball" or a "strike". However, caution about the likelihood of measurement error introduced by manual input to the operation of PITCHf/x means that we will not rely on all such pitches in our main estimations. The location of each pitch is measured with a high degree of accuracy by the PITCHf/x technology. This is then compared to a strike zone *estimated* by PITCHf/x. The uprights of the strike zone are invariant between pitches and games because they are fixed at the edges of the home plate. The top and bottom edges, on the other hand, are defined with reference to the knee and shoulder of each particular hitter and are calibrated/estimated manually, pitch to pitch, by an operator pointing a sight. To avoid concerns that miscalibration by the operator may be confounding

[13] Later we probe the possibility that umpire travel might have a direct impact on umpire productivity and so threaten identification. At that point we will return to describe umpire travel patterns in more detail.

our results we limit attention to pitches lying further than 20% of the strike zone height below the top edge, and the same distance above the bottom edge. This means that we are restricting attention to a set of pitches where we can confidently ignore measurement error introduced by actions of the camera operator. It also means that results should strictly be interpreted as applying to that subset of pitches (in other words how umpires are making judgements with respect to the vertical boundaries, not the horizontal ones). In a robustness check we re-run preferred specification but without this restriction, obtaining attenuated results.

We collect data from PITCHf/x on a variety of pitch characteristics other than location. In particular: hand with which pitch is thrown, hand with which the batter is currently batting, pitch break angle, pitch break length, vertical pitch break distance, initial velocity, categorical indicator for pitch type within MLB definitions (*e.g.*, fastball, change-up).

For each pitch PITCHf/x also provides an indicator for current inning number, inning part (top or bottom), ball/strike count at time of pitch, team-specific fixed effects for the run surplus or deficit faced by the batting team. We also collect continuous measures of game time elapsed, the cumulative number of pitches thrown in the game, cumulative number of pitches thrown by current pitcher, game attendance, and venue specific controls for the time of day at which the pitch was thrown.

4.2 Air quality

Our focus is on the effects of carbon monoxide (CO), fine particulates ($PM_{2.5}$) and ozone (O_3).[14] Each has been linked to some aspect of reduced mental function in existing research. We extract data on ambient levels from the Environmental Protection Agency's Air Quality System (AQS) which provides hourly and daily data for monitors across the United States.

We assign pollutant levels during a game by taking the reading from the closest station. We exclude a venue for which data is not available for each pollutant from a monitor located within 10 miles (this cut-off distance has been used by, for example, Currie et al. (2009) to exclude schools from their analysis of the effect of air quality on pupil absences). As in all studies of this type a trade-off exists between the desire to have accurate pollution measures, and the desire to maintain sample size.[15]

In light of our high-frequency pollution measures it is important to consider the effective exposure.

[14] In support of a robustness exercise we will also collect data on sulfur dioxide (SO_2), nitrogen oxides (NO_2) and particulate matter smaller than 10 (PM_{10}). For these the distance to monitor cutoff that we applied was 10 miles and we applied 12 hour rolling time blocks.

[15] The Appendix provides a list of venues on which our preferred specification is estimated. We rely on hourly pollution levels from AQS monitors in our preferred specification and daily measures as a test of robustness. A portion of monitors have a "minimum detectable level" (MDL) of 0.5 ppm for CO. We rely on monitors with lower MDLs where available and account for monitors with higher MDLs in our preferred regression specification.

Since different pollutants reside in the human body for different periods these will typically differ from instantaneous ambient levels. A worker exposed to elevated levels of a pollutant may continue to suffer ill effects from exposure for some time after moving to a clean environment. Our primary specification measures exposure to ambient pollution from the time of the umpire's decision back over the approximate half-life of the pollutant in the human bloodstream. In particular we construct rolling exposure blocks specific to each pitch. For CO we compute the average ambient level in the 3 hour time block immediately before the pitch. For other pollutants we compute the average ambient level in the 12 hour time block immediately before each pitch. The much shorter block for CO reflects that carbon monoxide is expelled much more rapidly from the human body. As a robustness exercise, and consistent with many other studies in the literature, we re-estimate the preferred specification using daily-average pollution levels.

In common with most other papers on the health and non-health impacts of pollution in the US and elsewhere, our analysis is hampered somewhat by the absence of good quality data on ambient levels of other pollutants in the vicinity of our venues. We additionally collected data on PM_{10}, NO_2 and SO_2 for those venues for which a monitor was available within our 10 mile tolerance, in each case aggregated in rolling 12 hour time blocks. The loss of sample size, and degree of correlation among some of the pollutants, means that we need to be cautious in interpreting results. We return to discussion of this issue later.

4.3 Weather

Temperature, relative humidity and other weather conditions may impact worker performance. In our setting weather is variable across venues, within venues over time, and even within a single game (recall that a typical game lasts between 3 and 4 hours).

We compile hourly observations of temperature and relative humidity from NOAA's Quality Controlled Local Climatological Data (NCDC (2015), hereafter QCLCD) for all stations within 15 miles of each venue. We impute hourly values of these weather variables as the inverse distance-weighted average of those stations for each venue and linearly interpolate values between the hourly observations.

4.4 Additional data

We obtain additional data on umpire attributes including date and place of birth and career MLB umpiring experience from Retrosheet to compile experience profiles for all umpires from the start of their careers, prior to deployment of the PITCHf/x system. Using these we additionally compute an umpire experience measure.

We compile additional details of each venue by determining the latitude, longitude, elevation, stadium type, and orientation of each venue using aerial photographs on Google Earth. These data allow us to impute pollution levels at each venue and control for the potentially confounding factor of the position of the sun in the sky relative to the umpire's field of view.[16] The inclusion or exclusion of these from our regressions has no discernible impact on results.

5 Methods

Our setting has a number of desirable features. Of particular importance are the following;

First, workers are quasi-randomly assigned to a series of games (work days) in different cities that are scattered across the country. One of the constraints of the scheduling algorithm used by MLB is that each umpiring crew should be scheduled for a minimum of one series at each baseball venue, so umpires cannot sort in such a way that they only work in specific regions of the country.

Second, we have a clean, constant and objective measure of individual-level performance - namely the production of correct 'calls' - for a large portion of pitches. This is not jointly produced and does not suffer from potentially unobservable variations in other inputs (capital, technology, effort from other employees).

Third, the data landscape is extremely favorable. We observe the same umpires working in a variety of locations across the country and over a long period of time which enable us to disentangle the effects of multiple local criteria pollutants and account for worker-specific idiosyncrasies.

Finally, using high-frequency data on both pollution and worker decisions allows us to capture effects of acute exposure to pollutants even if the observed effects dissipate quickly after exposure. These would be lost in day-level analysis.

Our interest is in the effect of air pollution on the frequency with which umpires make correct and incorrect calls. Given the quality rather than quantity focus of this (and many other) professions, we can think in terms of the production of correct decisions, or in terms of the error rate - the propensity to make mistakes.

The central analysis is conducted at the level of the individual pitch. For each pitch we observe the decision of the umpire (the ball or strike call) and the actual position of the ball from PITCHf/x. The

[16] Detailed descriptions of the calculations are available at http://aa.quae.nl/en/reken/zonpositie.html (accessed on 8/3/2015). Stata code to implement these calculations is available on the author's website. This process does not reveal the orientation of the playing surface for domed stadiums. However, the position of the sun should be irrelevant in these venues and absorbed by venue fixed effects.

question is whether the likelihood that an umpire makes a correct call on a pitch is influenced by prevailing air quality conditions. For pitch p in venue v with umpire u at time t, we estimate the following linear probability model (LPM):

$$p_{pvut} = \beta_0 + \mathbf{P'}_{pvut}\beta_1 + \mathbf{W'}_{pvut}\beta_2 + \mathbf{X'}_{pvut}\beta_3 + \Phi_u + \Psi_v + \theta_{vt} + \varepsilon_{pvut}$$

where p_{pvut} is a binary variable that takes the value one if the umpire's call for pitch p in venue v with umpire u at time t is correct and zero otherwise. The vector $\mathbf{P}_{pvut}$ contains pollution level variables assigned to pitch p in venue v with umpire u at time t and contains controls for CO, $PM_{2.5}$ and O_3. The unit of analysis is a single pitch. All environmental factors are assigned to each pitch based on the date and location of the game in which the pitch was thrown as such β_1 is our coefficient of interest.

The vector $\mathbf{W}_{pvut}$ contains weather variables. These include flexible controls for temperature and humidity (indicators for each 5-degree Fahrenheit temperature bin, each 10-percent relative humidity bin) as well as interactions of the temperature and relative humidity indicators, sky cover, precipitation, wind speed and atmospheric pressure. Since the influence of weather can be expected to be quite different in indoor versus outdoor settings, we estimate separate parameters for games played outdoors versus indoors or at venues with a closed retractable roof.

The vector $\mathbf{X}_{pvut}$ contains a rich set of game and pitch characteristic controls that might impact umpire decision-making. Following Parsons et al. (2011) and Kim and King (2014) for each pitch we control for the hand with which the pitch is thrown, hand with which the batter is currently batting, pitch break angle, pitch break length, vertical pitch break distance, initial velocity and indicators for pitch type following MLB categorizations. Game controls comprise indicators for the current inning number, inning part, ball-strike count at moment of pitch, run surplus or deficit faced by batting team at time of pitch, current pitching and batting team, game time elapsed, cumulative number of pitches thrown in the game, cumulative number of pitches thrown by current pitcher. Also included are game attendance and venue-specific linear controls for local time of pitch.

There are potential individual-specific factors that may confound our analysis as such the vector Φ_u contains umpire fixed effects and a linear trend for umpire experience (measured as the number of career games officiated). Finally, umpires exhibit idiosyncratic tendencies for mistakes based on the pitch location. To control for these tendencies, we include umpire-specific nonparametric controls which contain dummies for pitches located in the right-hand 20% of the strike zone, pitches in the left-hand 20% of the strike zone, and those in the middle.[17]

The vector Ψ_v contains venue fixed effects.

We adjust for temporal factors that may be correlated with the umpire's decision by including θ_{vt} which contains venue-month-year, venue-day-of-week and venue-hour-of-day fixed effects in addition to a control for time to sunset at the time of each pitch.

The error term ε_{pvut} is clustered at game level to allow for arbitrary correlation within games. Our main identifying assumption is that pollution is assigned as good as randomly to umpires after controlling for spatial and temporal fixed effects.[18]

As already noted in the description of dataset construction in Section 4, concern about measurement error on the top and bottom edges of the strike zone mean that our central specifications exclude pitches within 20% of those edges. This allows us to focus in on pitches where there is only a trivial possibility that our estimation is confounded by the errors and whims of the PITCHf/x camera operator. As a test of robustness, we later re-estimate our primary specification including all pitches (*i.e.* including those closer to the horizontal edges).

6 Results

Table 2 presents linear probability model (LPM) results.

Moving rightwards across the table we go from sparsest to richest specification. Column (1) contains only venue fixed effects. The coefficients on our CO and $PM_{2.5}$ are negative and both significant at better than 1%, while O_3 does not achieve significance at standard levels. In column (3) we add our time fixed effects. In columns (4), (5) and (6) we allow for umpire idiosyncrasy by introducing umpire fixed effects, trends, and the dummies that capture umpire-specificity of strike zones. Column (7) introduces the set of detailed pitch characteristics (other than location) that are provided from PITCHf/x. Estimation including the fullest suite of controls is summarized in column (8). It is reassuring that the CO and $PM_{2.5}$ coefficients prove stable across specifications, and while the coefficient on ozone comes into significance with the inclusion of basic fixed effects, these are lost again once pitch characteristics are properly controlled for.

[17] Kim and King (2014) have some discussion about umpire-specific traits and the desirability of such controls. As an appendix exercise they implement a version of what we are doing here but with nine vertical stripes rather than three (page 2638). While we retain them in our preferred specification we confirm in a robustness check that dropping them altogether does not disturb results. Indeed so doing serves to increase the estimated coefficients and strength of significance on both CO and $PM_{2.5}$.

[18] We test for random assignment of umpires to pollution treatments, in addition to whether batter decision for swing or pitch location vary with pollution levels. These tests are explained in the Appendix. Results confirm that we cannot reject the null hypothesis that (1) umpires are randomly assigned to air quality treatments (2) decisions of hitters to swing and location relative to strike zone of pitches thrown by pitchers are insensitive to air quality.

The specification in column (8) delivers our preferred estimates of the marginal effect of pollution on performance.[19] Our preferred estimates indicate that a 1 ppm increase in 3-hour CO causes an 11.5% increase in the propensity of umpires to make incorrect calls (an extra 2.0 incorrect calls per 100 decisions).[20] Likewise a 10 $\mu g/m^3$ increase in 12-hour $PM_{2.5}$ causes a 2.6% increase in the propensity of umpires to make incorrect calls (an extra 0.4 incorrect calls per 100 decisions). We control carefully for a variety of potential confounders and the results prove robust to a battery of robustness and falsification checks. The effect of ozone is a precisely-estimated zero.[21]

Air quality varies across the day and is subject to different patterns in different population centers. To help the reader interpret these effect sizes we combine these preferred estimates with the realized distribution of air quality levels in the twenty largest Metropolitan Statistical Areas (MSAs) between 2008 and 2015.[22] This provides an indication of the impact of being at various percentile points in the distribution in a particular city against a comparator of the 25th percentile in that city. The results of this exercise are presented in Table 3. For example, moving from the 25th to the 95th percentile in CO pollution conditions in Los Angeles reduces performance by 2.66%, for Phoenix it is 2.90%, and so on. Of course, within-MSA variation would be expected to generate bigger local effects - these simulations are based on MSA-wide averages.

Some evidence (for example Lavy, Ebenstein, and Roth (2014), Aragón, Miranda, and Oliva (2016), and Chang et al. (2016a)) points to nonlinear impacts of pollution. Though the pollution levels that we observe at most venues are typically low compared to EPA standards, the volume of data and credibly exogenous assignment of workers to pollution treatments makes this a good setting within which to probe for such non-linearities. We divide the zero to 99th percentile support of each of the CO and $PM_{2.5}$ pollution spaces into 7 bins each.[23] We then re-estimate the preferred specification replacing continuous measure of pollution with binned data, with the bin containing zero as the omitted category. Figure 4 shows our primary linear regression effects in gray and nonparametric estimates of the effects in black. The $PM_{2.5}$ results appear very close to linear, for CO the effects to increase substantially above 1.5 ppm.

[19] We similarly tested other pollutants, finding no significant effect for NO_x, SO_2, or PM_{10}. Furthermore the inclusion or exclusion of these as controls has no meaningful impact on the estimated coefficients of interest.

[20] We use a dummy variable for those occasions that CO levels are less that 'minimum detectable level' of 0.5 ppm. Therefore all our point estimates should be interpreted as impacts of CO in excess of 0.5 ppm.

[21] In addition to not being statistically significant the point estimate for ozone is miniscule. The implied marginal effect is an additional 0.0015 incorrect calls per 100 decisions per ppb ozone. Ozone levels in our sample typically vary between 2 and 73 ppb.

[22] To reiterate: We do not have city-specific estimates of effects - the preferred results are estimated on the panel of venues. We then interact those estimated effects with what we know about patterns of air quality in the various locales.

[23] Specifically for $PM_{2.5}$ the bins start at 0, 5, 10, 15, 20, 25, and 30 $\mu g/m^3$. For CO we use bins starting at 0, 0.5, 0.75, 1, 1.25, 1.75, and 2.25 ppm. There is low density in observations with high levels of CO so making sense that the terminal bins be larger.

7 Robustness

Table 2 provided evidence that the coefficients of interest are robust in sign and significance to a variety of specifications (inclusion or exclusion of various controls). Furthermore we believe that the richness of the data to which we have access allows us to control convincingly for a wide set of potentially important confounders.

In this section we further probe the credibility of the main results with a series of additional robustness checks and falsification exercises.

7.1 Weather

A priori we expect weather factors to be potentially important confounders. Factors like temperature and humidity can be expected to have physiological effects on umpires, perhaps causing loss of concentration and/or inducing fatigue. They may also impact directly the difficulty of the task at hand - litter or leaves blown across the outfield may, for example, cause visual distraction.

While we have taken great care to include an exhaustive set of controls, including binned measures for the interaction of temperature and relative humidity, as a further check we re-estimate the preferred specification with a full suite of controls but excluding all weather variables. Our rich time-varying, venue-specific controls absorb the bulk of variation in weather at the time of each game. If failure to control adequately for weather are seriously confounding our estimates then we would expect omitting the whole set of weather covariates to appreciably change our results.

The outcome of this exercise, summarized in Table 4 column (2) shows that our estimates are largely undisturbed, with sign and significance maintained.

7.2 Travel

Umpires travel extensively across continental North America - spending time on planes, in airports, and adjusting to changes in time zones. Umpiring crews generally arbitrate in between three and four games in one location before moving cities. Often these moves will be short - New York to Boston, for example - but can be much longer.

Travel poses a potential challenge to identification in two ways.

First, issues around fatigue and habituation. Like many employers who require their staff to travel on business, MLB makes extensive efforts to schedule travel, rest days and work assignments such that employees are fresh and ready for each day of work.[24] However, these efforts may be less-than-perfect such that the process of travel may *in itself* influence umpire performance. Travel and/or changes in time zone may be fatiguing, for example. Or it may take some time for an umpire to get used to local light conditions or to become acquainted with stadium sight-lines when first arriving at a new venue.

Second, umpires traveling between cities may "import" environmental conditions from the departing city. Insofar as the effects of exposure persist from one day to the next - and most evidence points to such persistent effects being small to non-existent (Gemperli (2008) and Welty et al. (2008)) - performance on date t might then reflect not just environmental conditions in the game-city on that date, but some other location on date $t-1$, threatening identification.

Our prior judgment is these considerations are unlikely to be important. However to assuage concerns that travel (and also rest time) might be confounding results we conduct two additional exercises.

Insofar as the effects of travel persist and might impact umpire performance, it is plausible to suppose that those impacts would be most pronounced when the umpire first arrives in a particular city. As such we re-estimate the preferred specification on that sub-sample of games where we know that the umpire did not change cities on the previous day. The result of that exercise is reported in Table 4, column (3). Again sign and significance of our two coefficients of interest are maintained, and the value of each coefficient is little changed.

To test the possibility that time off may 'refresh' the umpire, and perhaps change susceptibility to variations in air quality, column (4) of Table 4 takes a slightly different approach. We re-estimate our preferred specification but adding the log hours since the last game officiated (if it is less than 40 hours) as a linear control into our regression. The sign and significance of our coefficients of interest are maintained.

Taken together this set of results confirm our conjecture that despite the travel-intensity of the work of this set of employees, travel does not appear to have an important influence on the relation between air quality and performance.

[24] Trick, Yildiz, and Yunes (2011) provide a detailed institutional account of the process of the scheduling of games and the assignment of umpires to those games. MLB uses an algorithm to assign umpiring crews to games while meeting a range of constraints. For teams some of these are discretionary (for example the Boston Red Sox always play at home on Patriots' Day and the Toronto Blue Jays always play at home on Canada Day) but most are designed to ease the rigours of travel on both players and officials. The contract between MLB and the umpires' union specifies a number of constraints on the scheduling of umpiring crews. Hard constraints include: (a) No umpire should travel from West Coast to East Coast without an intermediate day-off; (b) No umpire should travel more than 300 miles on the day preceding a series whose first game starts before 4 pm; (c) No umpire should work more than 21 consecutive days; (d) All umpires should visit each MLB city at least once; (e) Each umpire should officiate a series involving each MLB team at home and away at least once in the season, but no more than four series in total; etc. Umpire union rules also mean umpires receive four week-long vacations during the baseball season, three as a crew and one individually. "The umpire scheduler, whose main goal is to minimize the miles that each crew travels, must adhere to many rules." (Trick, Yildiz, and Yunes (2011):p. 234). Furthermore union rules require that umpires have what they call 'balanced' schedules - they should travel a similar number of miles, handle approximately the same number of games, and have the same number of days off.

7.3 Alternative pollution measures

Different pollutants affect the body in different ways and - importantly for us - those effects wear off at very different speeds. Our preferred estimates are based on hourly readings from air quality monitors which are combined into rolling time blocks of lengths that we have argued to be appropriate based on physiological fundamentals, in particular the longevity of the various pollutants in an adult human.

An alternative, that mirrors a common approach in the literature, would be to use daily average measures of air quality in the vicinity of the venue at which a pitch was thrown. While the simplicity of such an approach is appealing, the cost in terms of measurement error of what is the biologically-pertinent measure of exposure is potentially severe. Pollution levels in cities vary substantially *within* the course of a day so daily average pollution levels measure actual exposure (at and around game time) with error. The effect of exposure can be short-lived, depend on human physiology which in most cases does not synchronize with the accounting practices embodied in EPA databases. In particular CO is largely expelled from the human body within a few hours of exposure. Further, much of what is picked up in a calendar day measure will reflect pollution levels *after the game in question has finished* - particularly true for games played early in the day - which are clearly irrelevant for what happens during a game.[25]

However, for purposes of completeness we report in column (5) the results of re-estimating the preferred specification but with each pitch assigned pollution levels equal to the daily average at that location on the date of the game in question. The point estimates on the two coefficients of interest remain negative and similar in size to those from the preferred specification. Significance is maintained for $PM_{2.5}$ (coefficient value attenuated somewhat) but lost for CO. This is not surprising given the discussion in the last paragraph, and reinforces our preference for using the rolling time block approach for studying the impacts of short-lived pollutants.

7.4 Player identity

One may be concerned that identity of players involved in a particular call could have a systematic effect on the umpire's decision. Kim and King (2014), for example, find that a more 'famous' pitcher - as measured by number of All-Star Game appearances - is more likely to have a call made in his favor than would a less-celebrated colleague. To address these possibilities we re-estimate our preferred specification with the addition of pitcher, batter and catcher fixed effects (the three players primarily involved when

[25] As an additional challenge to such an approach we can note that if daily peaks in pollution levels in the vicinity of venues correlates with timing of games (for example, early evening) then the measurement error introduced would be non-classical and could bias parameter estimates in either direction.

an umpire makes a ball/strike call). Column (6) reports that the results of the preferred specification are undisturbed.

7.5 Reinserting exclusions

As noted we excluded (a) a small number of games in which equipment appeared miscalibrated or the game was officiated by a temporary umpire and, (b) within each game pitches that did not travel through the band defined by a horizontal line 20% of the height of the strike zone below the upper edge of the PITCHf/x-estimated strike zone, and 20% above the bottom edge. The rationale for the former should be apparent. The latter, as already noted, allowed us to concentrate on a subset of pitches where we can ignore operator error in the calibration of the PITCHf/x equipment.

In column (7) of Table 4 we re-estimate the preferred specification but reinserting the games excluded in (a). Signs and significance on the three coefficients of interest are maintained, and coefficient values are similar in magnitude.

In column (8) we re-estimate the preferred specification including all pitches on which the umpire was required to make a call. Again the sign and significance of results are maintained. The magnitude of the estimated coefficient is in each case quite a bit smaller. This likely reflects two things. First, by considering many pitches close to the upper and lower edges of the strike zone we are introducing measurement error and hence attenuation bias. Second, we are at the same time excluding many pitches that travel well away from the strike zone that the umpire could be expected to call correctly almost all of the time irrespective of conditions.

7.6 Umpire specific strike zones

Our preferred specification employs a set of controls including umpire-specific strike zone dummies. These reflect that different umpires are idiosyncratic in how they call pitches that travel through different parts of the strike zone.

In column (9) of Table 4 we report the result of dropping these controls altogether. Compared to the preferred specification both coefficient values become a little larger in absolute value with this exclusion, with the effect size implied for CO around 15% larger.

7.7 Alternative estimation

Throughout the paper we have focused on LPM results. In columns (3) through (5) of Table 5 we contemplate the results of three alternative estimation strategies.

A coherent alternative to pitch-level analysis would have been to treat the game as a 'day of the work' for the umpire, and to develop game-level results in which the dependent variable is the percentage of correct calls by an umpire in a particular game. Such an approach has two primary disadvantages; (a) It prevents us from controlling meaningfully for the rich set of pitch-specific characteristics that determine the degree of difficulty facing the umpire in evaluating any particular pitch and which PITCHf/x provides. (b) It requires a measure of environmental conditions averaged across a game, whereas our preferred approach uses rolling time blocks to assign a measure of exposure to the time of each pitch. As we have already noted, such temporal averaging may hinder us in identifying acute effects of pollutants like CO, which is rapidly expelled from a human body. Nonetheless, for completeness we summarize in column (2) of Table 5 the results of conducting such a game-level analysis. This uses proportion of correct calls per game as the dependent variable and environmental measures averaged over the period of the game. The results prove very similar to those from our central specification.

Returning to pitch-level estimation, alternatives to the LPM would have been to estimate the same model using Logit or Probit nonlinear estimators. These approaches require additional parametric assumptions on the error structure and are more efficient if they represent the true underlying model. However that potential efficiency gain comes at a cost. First, under both Logit and Probit, misspecifying the model can bias parameter estimates. Probit poses the additional difficulty that inconsistent incidental parameters (such as the venue-by-month fixed effects) poison the consistency of all parameters. Columns (3) and (4) of Table 5 report the results of reestimating the preferred model using Logit and Probit models respectively. The coefficients can be interpreted as marginal effects, in each case evaluated at the mean of all covariates. It can be seen that the marginal effects for both CO and $PM_{2.5}$ are essentially identical to those derived from the LPM.

7.8 Other pollutants

We need to be cautious in attributing effects to particular pollutants because of the existence of other correlated pollutants.

Our attempts to account for other pollutants are hampered by data limitations. However, adding PM_{10}, NO_2 and SO_2 to our preferred specification - either in separate exercises or jointly - for those 15 venues for which data is available, delivers a close-to-zero coefficient on that additional pollutant and leaves the results on CO and $PM_{2.5}$ qualitatively undisturbed.[26] The results of these exercises are reported in columns (2) through (5) of Table 6.

Indeed it is noteworthy that when jointly included (column (5)), the additional pollutants cause the coefficient estimates on CO and $PM_{2.5}$ to be substantially bigger. We opted against column (5) as a preferred specification primarily because the less favorable availability of monitors means that it is estimated on a much smaller sample (15 venues instead of 29). However the evidence is strongly suggestive that if data availability were to allow us to control for these additional criteria pollutants at a wider set of venues our estimated effect sizes would be larger.

The problem of isolating the role of individual pollutants out of the cocktail of pollution to which people are exposed on 'bad air' days is a challenge throughout the literature. In general researchers study a single or subset of pollutants, with that subset often determined by data availability. For example, in their excellent recent study on health outcomes, Schlenker and Walker (2016) deploy only data on CO, NO_2 and Ozone.[27] However, they are explicit in "... acknowledging that we may be picking up the health effects of other pollutants" (page 787). Later they insert the three pollutants in the same regression with qualitative loss of results. The omission of a measure for particulate matter, with clear links to a number of the health outcomes that they study, is clearly a challenge for the interpretation of their results. As such they note that: "We believe that some amount of caution is warranted in interpreting CO as the unique pollution-related causal channel leading to adverse health outcomes; there may in fact be other unobserved sources of air pollution that covary with CO that may also effect health" (page 800). We are similarly circumspect in interpretation of our results, though the evidence of Table 6 is helpful in pointing to CO and $PM_{2.5}$ as the pollutants of interest.

7.9 Placebos

Recent debates regarding causal inference in the social sciences have led to a growing desire for "tests of design". In the design-based inference literature such tests serve to address concerns that the research design may *itself* be tending to generate apparent causal effect.

In Table 7 we present tests of our primary result using placebo treatments. In each test, we estimate our preferred regression specification on a set of alternate pollution data where, if our hypothesis is true, one

[26] For each pollutant we again applied a 10 mile cut-off in distance from venue to monitor.

[27] Indeed their central results (in fact all but one table in the paper) are derived from exercises in which each of these pollutants is used as explanatory variable separately, absent controls for the other two.

would expect to find no statistically significant result. These tests lend evidence that the primary result is not driven by some underlying systematic trend in the data or shortcoming in study design.

Column (1) in the table repeats the preferred specification.

Column (2) shifts the pollution data along the temporal dimension, substituting each imputed pollution level with the level imputed at the same venue precisely one year earlier. Given the inclusion of venue-year-month fixed effects identification is driven by variations around the monthly, venue-specific means. As such, in a well-designed model we would not expect calls on a particular date to be significantly affected by air quality conditions one year earlier. The estimates in column (2) are consistent with this prior, point estimates are near zero and do not achieve statistical significance.

Column (3) instead shifts pollution data along the spatial dimension and replaces pollution values at the game venue with those prevailing at the venue belonging to the away (visiting team) at game time. We exclude from this exercise games between teams located within the same US Census Bureau commuting zone.[28]

Column (4) uses as placebo conditions taken from the EPA pollution monitor in the Continental US that is *farthest* (in great circle distance) from the venue in question.[29]

A limitation of the approach in column (4) is that while it provides a placebo series of conditions at a location far from the venue at which any particular game is being played, we end up drawing very frequently from just two locations (Seattle and Miami). To provide more variation in the source of the placebo, while still ensuring pollution conditions are taken from far enough away that they cannot reasonably be expected to influence outcomes at the game of interest, in column (5) the placebo series for each venue is taken from the *closest* EPA monitor that is *more than* 1000 miles distant from the game location.

Consistent with the hypothesis of a null effect from a placebo treatment, the estimated coefficients on CO and $PM_{2.5}$ in each of columns (2) through (5) are smaller (typically much smaller) in absolute value than those from the preferred specification, mixed in sign and in no case come close to achieving statistical significance at conventional levels.

[28]Despite the exclusion this is perhaps the least appealing of the placebo exercises reported because of the number of games played between teams not sharing a US Census Bureau commuting zone but still located comparatively close to each other (for example the Milwaukee Brewers playing the Chicago Cubs).

[29]The number of venues that can be included here grows. While not all MLB venues have close enough pollution monitors to be included in the main results, all venues have a most distant monitor. We have also verified that the placebo 'works' if we restrict attention only to the 29 venues included in the estimation of the preferred specification.

8 Conclusions

Recent evidence points to the effect that air pollution may have on how well people do their work. If detrimental impacts are significant in size and sufficiently widespread, then the economic burden associated with such effects could rival the direct health effects.

We contribute to the emerging but important literature in this area. While existing research has looked at low wage workers engaged in manual, or non-manual but low-skilled work, our focus is on a group of highly-skilled professional engaged in 'mental output'.

As with many professions work performance in our setting is defined by quality, not quantity, and that is what we - as well as the employer - focus on. The central results of our analysis are that ambient carbon monoxide (CO) and fine particulate matter ($PM_{2.5}$), at levels well below the respective EPA acute exposure standards of 9 ppm and 35 $\mu g/m^3$, have a significant negative effect on the performance of this group of workers. Our preferred estimates indicate that a 1 ppm increase in 3-hour CO causes an 11.5% increase in the propensity of umpires to make incorrect calls (an extra 2.0 incorrect calls per 100 decisions). Likewise a 10 $\mu g/m^3$ increase in 12-hour $PM_{2.5}$ causes a 2.6% increase in the propensity of umpires to make incorrect calls (an extra 0.4 incorrect calls per 100 decisions). We control carefully for a variety of potential confounders and the results prove robust to a battery of robustness and falsification checks.The effect of ozone is a precisely-estimated zero.

The effect sizes are robust to alternative specifications, and occur well below NAAQS acute exposure standards. As with other contributions to this literature we need to be cautious in attributing effects to particular pollutants because of the existence of other correlated pollutants. Our attempts to account for other pollutants are hampered by data availability issues. However, adding PM_{10}, NO_2 and SO_2 to our preferred specification - either in separate exercises or jointly - delivers a close-to-zero coefficient on that additional pollutant and leaves the results on CO and $PM_{2.5}$ qualitatively undisturbed. Indeed joint inclusion makes the effects sizes on CO and $PM_{2.5}$ meaningfully larger, though estimated on a smaller sample.

It is useful to reflect briefly on how our results complement recent and emerging evidence on the productivity effects of air pollution. The seminal work of Graff Zivin and Neidell (2012) and Chang et al. (2016a) related to physically-oriented workers in an agricultural setting. The work of the call center employees studied by Chang et al. (2016b) was not physical, but remains low-skilled (indicative of this is that the average annual pay of a call center worker in China is around 2 000 USD, less than half the average pay in that country).[30] Our analysis extends this line of inquiry to highly-skilled, highly-trained, highly-

[30] Interestingly the reductions in call processing per day identified by Chang et al. (2016b) are driven by workers spending more time logged-off on more polluted days, rather than handling calls less quickly. As such the result is something more akin

remunerated specialists engaged in a work setting that requires sustained mental acuity. While the task they execute is particular, other jobs that are important in a modern economy make similar demands on mental and sensory systems. Furthermore, while Chang et al. (2016b) find only effects at the extensive margin (supply of labor), we find effects on quality of work. This and other recent papers cited should motivate further work to understand more generally the sorts of jobs and work tasks where effects arise.

Our analysis does not allow us to speak to mechanism. There is established research linking exposure to the pollutants that we study to reduced mental acuity, but it remains unclear whether this works through loss of oxygen to the brain, fatigue, or through other channels either singly or in combination. Given the idiosyncratic nature of the work task studied we cannot rule out that the effect works through a limited channel - such as decreased attention due to respiratory irritation - rather than mental function more generally. Understanding mechanism(s) should be a priority for future work. Such understanding might inform design of mitigative interventions.

The results also provide some evidence consistent with a previously perplexing result from Greenstone, List, and Syverson (2012). They found that while more binding particulates and ozone regulations led to a 2.6% decrease in TFP, CO regulation is associated with a statistically significant 2.2% *increase* in the level of TFP. Our results on CO are comparatively stronger than might have been expected from a reading of the extant literature, and we believe that our careful treatment of short term exposure using the rolling 3 hour time blocks allowed us to tease out previously under-explored acute effects which could readily translate into increased TFP associated with improvement in air quality.

Looking to future research, while air quality clearly impacts MLB umpires, as a group umpires may differ substantially from the general population. These are individuals who are, during the period of our sample, all males of working age. Further, the highly-selective process through which individuals advance to the ranks of MLB umpire may eliminate candidates who are particularly sensitive to air quality, so the effect on a more general pollution may be more pronounced. Future research could identify portions of the population most at-risk of having their work performance impacted by pollution and, as already suggested, probe further the types of work task that are likely to be most impacted.

to an intra-day labor supply effect than the 'pure' effect on execution of tasks that we uncover.

References

Amitai, Y, Z Zlotogorski, V Golan-Katzav, A Wexler, and D Gross (1998). "Neuropsychological impairment from acute low-level exposure to carbon monoxide." In: *Archives of neurology* 55(6), pp. 845–848. ISSN: 00039942. DOI: 10.1001/archneur.55.6.845.

Aragón, Fernando M., Juan Jose Miranda, and Paulina Oliva (2016). "Particulate matter and labor supply: Evidence from Peru." In: (February). URL: http://www.sfu.ca/econ-research/RePEc/sfu/sfudps/dp16-01.pdf.

Beard, Rodney R and George A Wertheim (1967). "Behavioral Impairment Associated With Small Doses of Carbon Monoxide." In: *American Journal of Public Health and the Nations Health* 57(11), pp. 2012–2022. ISSN: 0002-9572. DOI: 10.2105/AJPH.57.11.2012.

Bento, Antonio, Matthew Freedman, and Corey Lang (2015). "Who benefits from environmental regulation? Evidence from the Clean Air Act Amendments." In: *Review of Economics and Statistics* 97(3), pp. 610–622. ISSN: 15309142. DOI: 10.1162/REST_a_00493.

Chang, Tom, Joshua S Graff Zivin, Tal Gross, and Matthew Neidell (2016a). "Particluate Pollution and the Productivity of Pear Packers." In: *American Economic Journal: Economic Policy* 8(3), pp. 141–169. DOI: http://dx.doi.org/10.1257/pol.20150085.

Chang, Tom, Joshua S Graff Zivin, Tal Gross, and Matthew Neidell (2016b). "The Effect of Pollution on Worker Productivity: Evidence from Call-Center Workers in China." In: *NBER Working Paper* #22328. URL: http://www.nber.org/papers/w22328.

Chay, Kenneth Y. and Michael Greenstone (2005). "Does air quality matter? Evidence from the housing market." In: *Journal of Political Economy* 113(2), pp. 376–424. ISSN: 0022-3808. DOI: 10.1086/427462. URL: http://www.nber.org/papers/w6826.

Chen, Daniel, Tobias J Moskowitz, and Kelly Shue (2016). "Decision-Making under the Gambler's Fallacy : Evidence from Asylum Judges, Loan Officers, and Baseball Umpires." In: *The Quarterly Journal of Economics* 131(3), pp. 1181–1242. DOI: 10.1093/qje/qjw017.

Currie, Janet, Eric a Hanushek, E. Megan Kahn, Matthew Neidell, and Steven G Rivkin (2009). "Does Pollution Increase School Absences?" In: *Review of Economics and Statistics* 91(4), pp. 682–694. ISSN: 0034-6535. DOI: 10.1162/rest.91.4.682.

Drellich, Evan (Aug. 2012). *Complex system in place to evaluate umpires.* URL: http://mlb.mlb.com/news/print.jsp?ymd=20120828%7B%5C&%7Dcontent%7B%5C_%7Did=37468304%7B%5C&%7Dvkey=news%7B%5C_%7Dmlb%7B%5C&%7Dc%7B%5C_%7Did=mlb.

Gemperli, A (2008). "The time-lagged effect of exposure to air pollution on heart rate variability." In: *Epidemiology* 19(6), S151. DOI: 10.1097/01.ede.0000339969.09289.ff.

Genc, Sermin, Zeynep Zadeoglulari, Stefan H. Fuss, and Kursad Genc (2012). "The adverse effects of air pollution on the nervous system." In: *Journal of Toxicology* 2012. ISSN: 16878191. DOI: 10.1155/2012/782462.

Gold, Diane R et al. (2000). "Ambient pollution and heart rate variability." In: *Circulation* 101(11), pp. 1267–1273.

Graff Zivin, Joshua S and Matthew Neidell (2012). "The Impact of Pollution on Worker Productivity." In: *American Economic Review* 102(7), pp. 3652–3673. DOI: 10.1257/aer.102.7.3652.

Greenstone, Michael, John A. List, and Chad Syverson (2012). "The Effects of Environmental Regulation on the Competitiveness of U.S. Manufacturing." In: *NBER Working Paper* #18392. DOI: 10.3386/w18392. URL: http://www.nber.org/papers/w18392.

Hausman, Jerry A, Bart D Ostro, and David A Wise (1984). "Air pollution and lost work." In: *NBER Working Paper*. NBER Working Paper #1263.

Heyes, Anthony, Matthew Neidell, and Soodeh Saberian (2016). "The Effect of Air Pollution on Investor Behavior: Evidence from the S&P 500." In: *NBER Working Paper* 22753.

Heyes, Anthony, Nicholas Rivers, and Brandon Schaufele (2016). "Politicians, Pollution and Productivity." In: *unpublished manuscript, Ivey School of Business, Western University*.

Kampa, Marilena and Elias Castanas (2008). "Human health effects of air pollution." In: *Environmental Pollution* 151(2), pp. 362–367. ISSN: 02697491. DOI: 10.1016/j.envpol.2007.06.012.

Kim, Jerry W. and Brayden G. King (2014). "Seeing Stars: Matthew Effects and Status Bias in Major League Baseball Umpiring." In: *Management Science* 60(11), pp. 2619–2644. ISSN: 15265501. DOI: 10.1287/mnsc.2014.1967. URL: http://dx.doi.org/10.1287/mnsc.2014.1967.

Kleinman, Michael (2014). *Central Nervous System Effects of Ambient Particulate Matter : The role of Oxidative Stress and Inflammation*. Tech. rep. April 08.

Lavy, Victor, Avraham Ebenstein, and Sefi Roth (2014). "The Impact of Short Term Exposure to Ambient Air Pollution on Cognitive Performance and Human Capital Formation." In: *NBER Working Paper* #20648. URL: http://www.nber.org/papers/w20648.

Lichter, Andreas, Nico Pestel, and Eric Sommer (2015). "Productivity Effects of Air Pollution: Evidence from Professional Soccer." In: *IZA Discussion Papers* 8964. URL: http://hdl.handle.net/10419/110687.

Marr, Linsey C. and Matthew R. Ely (2010). "Effect of air pollution on marathon running performance." In: *Medicine and Science in Sports and Exercise* 42(3), pp. 585–591. ISSN: 01959131. DOI: 10.1249/MSS.0b013e3181b84a85.

NCDC (2015). *Quality Controlled Local Climatological Data.* DOI: gov.noaa.ncdc:C00679.

Neidell, Matthew (2009). "Information, Avoidance behavior, and health the effect of ozone on asthma hospitalizations." In: *Journal of Human Resources* 44(2), pp. 450–478.

Ostro, Bart D (1983). "The effects of air pollution on work loss and morbidity." In: *Journal of Environmental Economics and Management* 10(4), pp. 371–382. ISSN: 00950696. DOI: 10.1016/0095-0696(83)90006-2.

Oudin, Anna, Ulf Strömberg, Kristina Jakobsson, Emilie Stroh, and Jonas Björk (2010). "Estimation of short-term effects of air pollution on stroke hospital admissions in southern Sweden." In: *Neuroepidemiology* 34(3), pp. 131–142.

Ozkaynak, H et al. (1995). "Personal exposure to airborne particles and metals: results from the Particle TEAM study in Riverside, California." In: *Journal of Exposure Analysis and Environmental Epidemiology* 6(1), pp. 57–78.

Parsons, Christopher A, Johan Sulaeman, Michael C Yates, and Daniel S. Hamermesh (2011). "Strike Three : Discrimination, Incentives, and Evaluation." In: *American Economic Review* 101(4), pp. 1410–1435. DOI: 10.1257/aer.101.4.1410. URL: http://www.jstor.org/stable/23045903.

Raub, J. a. and V. a. Benignus (2002). "Carbon monoxide and the nervous system." In: *Neuroscience and Biobehavioral Reviews* 26(8), pp. 925–940. ISSN: 01497634. DOI: 10.1016/S0149-7634(03)00002-2.

Rho, Hye Jin (2010). *Hard Work? {P}atterns in Physically Demanding Labor Among Older Workers.* Tech. rep. Center for Economic and Policy Research.

Roth, Sefi (2016). "The Contemporaneous Effect of Indoor Air Pollution on Cognitive Performance: Evidence from the UK." In:

Rundell, Kenneth William (2012). "Effect of air pollution on athlete health and performance." In: *British journal of sports medicine* 46(6), pp. 407–12. ISSN: 1473-0480. DOI: 10.1136/bjsports-2011-090823. URL: http://bjsm.bmj.com/content/46/6/407.

Schlenker, Wolfram and W. Reed Walker (2016). "Airports, air pollution, and contemporaneous health." In: *Review of Economic Studies* 83(2), pp. 768–809. ISSN: 1467937X. DOI: 10.1093/restud/rdv043.

Thatcher, Tracy L and David W Layton (1995). "Deposition, resuspension, and penetration of particles within a residence." In: *Atmospheric Environment* 29(13), pp. 1487–1497. URL: thatcher:Layton:1995.

Trick, Michael a., Hakan Yildiz, and Tallys Yunes (2011). "Scheduling Major League Baseball Umpires and the Traveling Umpire Problem." In: *Interfaces* 42(3), pp. 232–244. ISSN: 0092-2102. DOI: 10.1287/inte.1100.0514. URL: http://interfaces.journal.informs.org/cgi/doi/10.1287/inte.1100.0514.

Vette, Alan F et al. (2001). "Characterization of indoor-outdoor aerosol concentration relationships during the Fresno PM exposure studies." In: *Aerosol Science \& Technology* 34(1), pp. 118–126.

Welty, L J, R D Peng, S L Zeger, and F Dominici (2008). "Bayesian distributed lag models: estimating effects of particulate matter air pollution on daily mortality." In: *Biometrics* 65(1), pp. 282–291. ISSN: ISSN 0006-341X EISSN 1541-0420. DOI: 10.1111/j.1541-0420.2007.01039.x. URL: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve%7B%5C&%7Ddb=PubMed%7B%5C&%7Ddopt=Citation%7B%5C&%7Dlist%7B%5C_%7Duids=18422792%7B%5C%%7D5Cnhttp://dx.doi.org/10.1111/j.1541-0420.2007.01039.x.

Figure 1: Definition of strike zone

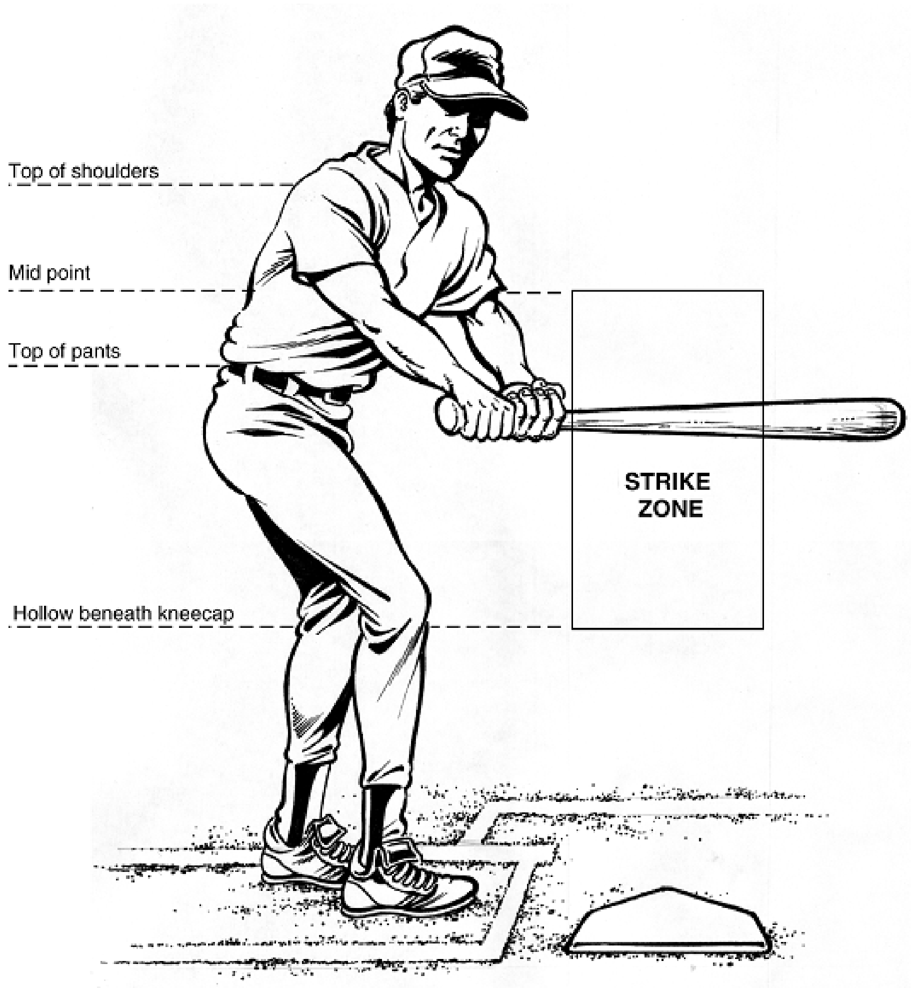
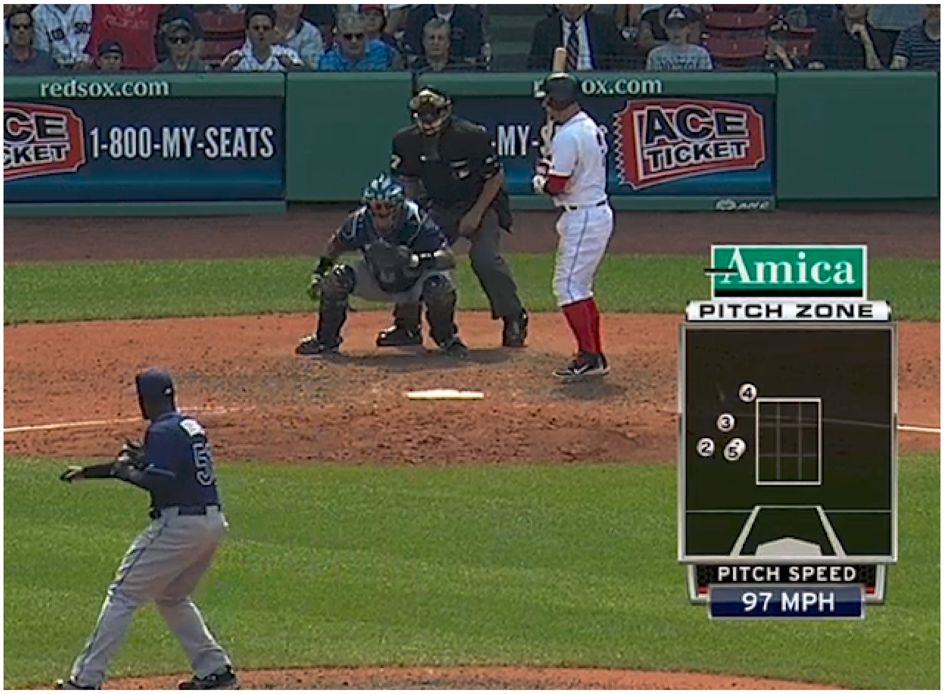


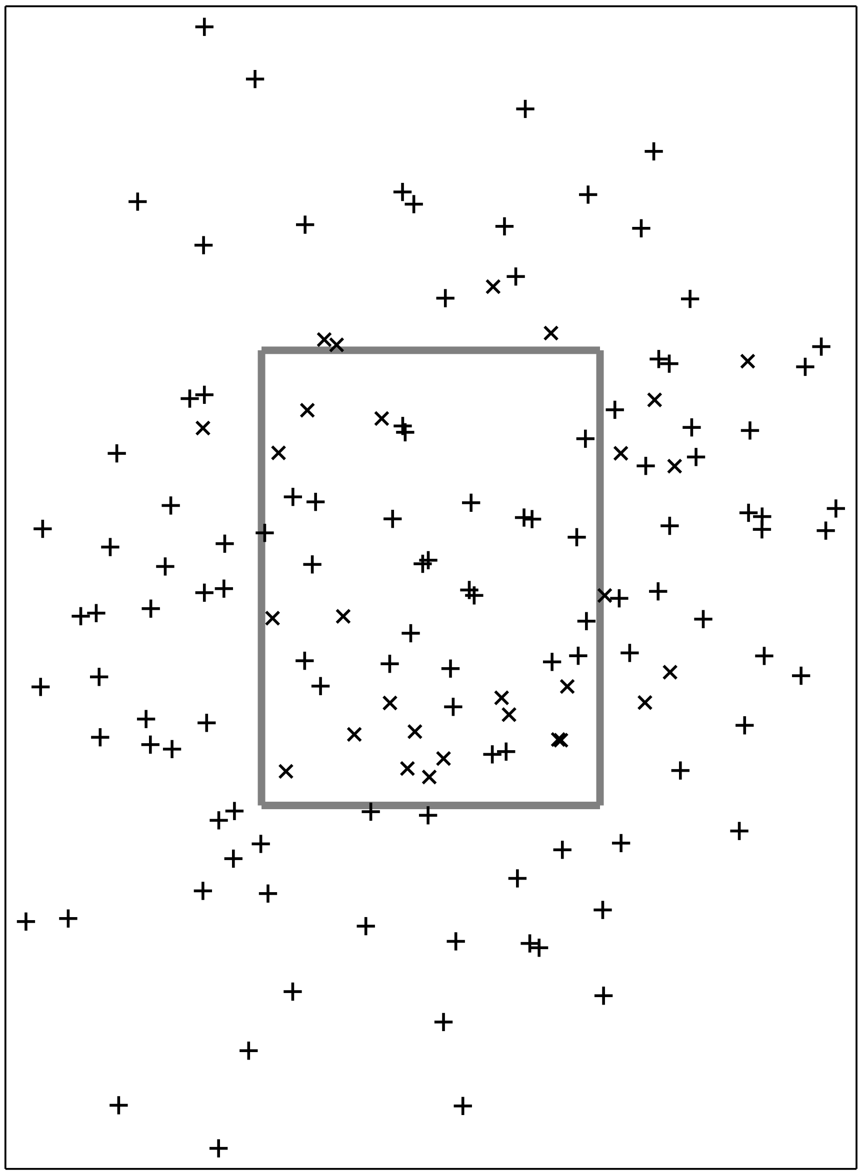
Diagram of the MLB strike zone by rule during the sample period (2008 to 2015) Source: http://mlb.mlb.com/mlb/downloads/y2008/official_rules/02_definition_of_terms.pdf (Accessed 14 June 2016).

Figure 2: Real-time PITCHf/x data during a television broadcast



Screen capture of an MLB game showing, from left to right, pitcher, catcher, umpire, and batter. The graphic in the lower right corner uses the same PITCHf/x data as the analyses presented here to show the locations of all pitches thrown during this at-bat relative to the strike zone. Source: http://www.sportsvision.com

Figure 3: Location of pitches for a single game

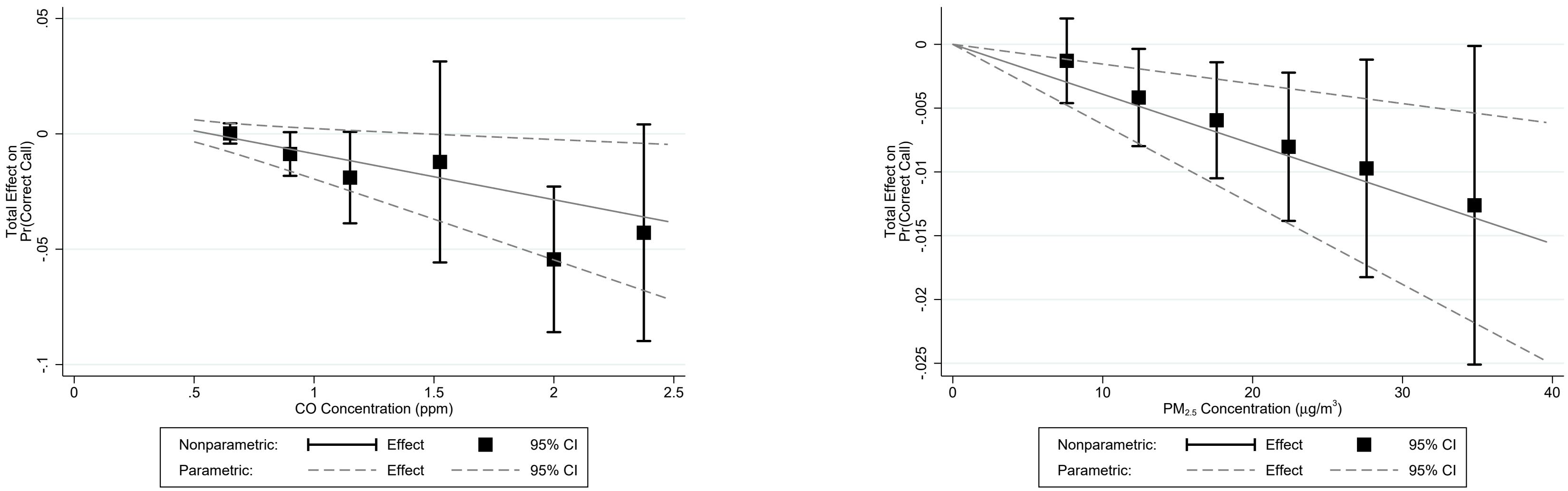


The location of all pitches, from the perspective of the pitcher, for which the umpire made a ball/strike decision in the single game between the Philadelphia Phillies and the New York Mets on 9 April 2008. The strike zone, standardized on the vertical dimension for each batter is the gray rectangle. Circles represent "ball" calls and triangles represent "strike" calls. Hollow shapes are correct calls and solid shapes are incorrect calls.

Figure 4: Linear and nonparametric estimates of effects of air pollution on work performance

(a) CO

(b) $PM_{2.5}$



These figures plot the comparison of linear and nonparametric estimated total effects of CO and $PM_{2.5}$ on the probability of a correct decision. Dash lines show the 95% confidence intervals clustered at the game level. Marginal effects for the parametric model are constant in pollution level, so confidence intervals are increasing in pollution levels for total effects. Nonparametric effects estimated using seven bins over the zero to 99th percentile of support in the observed pollution values. The omitted category in nonparametric estimates is the bin containing zero pollution. Regressions include venue, time fixed effects and controls for weather, pitch characteristics, game situation and umpire. See notes in Table 2 for a full description of controls.

Table 1: Summary statistics

	Mean	Std. Dev.
Correct call	0.827	0.378
Pitch in strike zone	0.541	0.498
Game indoors	0.135	0.342
Attendance	30,977	10,689
Pitch speed (mph)	87.82	6.00
Outdoor temperature (F)	72.31	11.84
Relative humidity (%)	59.07	18.55
Wind speed (mph)	7.567	5.125
Outdoor air pressure (inHg)	29.52	0.74
CO (ppm)	0.295	0.139
$PM_{2.5}$ (10 μg/m^3)	1.09	0.58
Ozone (ppm)	0.034	0.015
Observations	623,573	
Number of games	12,543	
Number of venues	29	
Number of umpires	86	

This table represents statistics for the primary estimation sample.

Table 2: LPM estimates of effect of air pollution on work performance

	Venue FEs (1)	Weather controls (2)	Time FEs (3)	Umpire FEs (4)	Umpire trends (5)	Ump-specific strike zone (6)	Pitch controls (7)	Game situation (Preferred) (8)
CO (>0.50 ppm)	-0.021	-0.021	-0.024	-0.024	-0.026	-0.022	-0.022	-0.020
	(0.008)***	(0.008)***	(0.010)**	(0.010)**	(0.010)**	(0.008)***	(0.008)***	(0.008)**
$PM_{2.5}$ (10 $\mu g/m^3$)	-0.005	-0.010	-0.005	-0.004	-0.005	-0.005	-0.005	-0.004
	(0.001)***	(0.001)***	(0.001)***	(0.001)***	(0.001)***	(0.001)***	(0.001)***	(0.001)***
Ozone (ppm)	-0.019	0.074	0.138	0.111	0.118	0.068	0.087	0.029
	(0.041)	(0.051)	(0.055)**	(0.054)**	(0.053)**	(0.047)	(0.047)*	(0.048)
N obs	624,358	624,354	624,354	624,354	623,573	623,573	623,573	623,573
N clusters	12,560	12,560	12,560	12,560	12,543	12,543	12,543	12,543
N venues	29	29	29	29	29	29	29	29
Venue	Y	Y	Y	Y	Y	Y	Y	Y
Time	N	N	Y	Y	Y	Y	Y	Y
Weather	N	Y	Y	Y	Y	Y	Y	Y
Pitch characteristics	N	N	N	N	N	N	Y	Y
Game situation	N	N	N	N	N	N	N	Y
Umpire	N	N	N	FE	FE+Trend	SZ+Trend	SZ+Trend	SZ+Trend

Linear probability model. Dependent variable is a binary indicator for a correct call. Standard errors clustered at the game level shown in parentheses. *** p<0.01, ** p<0.05, * p<0.1. All specifications include *venue* fixed effects. *Time* dummies include venue-month-year, venue-day-of-week, venue-hour-of-day fixed effects and a control for time to sunset at the time of each pitch. *Weather* controls include temperature (indicators for each 5-degree Fahrenheit bin), humidity (indicators for each 10-percent relative humidity bin), interactions of temperature and humidity indicators, sky cover, precipitation, wind speed and atmospheric pressure. We estimate separate weather controls for outdoor venues and domed venues with closed roof. *Pitch characteristics* include controls for the pitcher hand, batter hand, pitch break angle, pitch break length, vertical pitch break distance, initial velocity and indicators for pitch type. *Game situation* for the current pitch controls for the current inning number, inning part, ball and strike count, run surplus or deficit faced by the batting team, current pitching and batting teams, game time elapsed, cumulative number of pitches thrown in the game, cumulative number of pitches thrown by current pitcher, game attendance and venue-specific linear trend in the local time of day. *Umpire controls*: FE denote umpire fixed effects; trend denotes linear umpire-specific experience trends; SZ denotes umpire-specific nonparametric strike zone controls.

Table 3: Estimated effect of air pollution on work performance at selected percentiles compared to 25th percentile by MSA

Metropolitan Statistical Area	CO Effect (%) 75th Pctile	CO Effect (%) 90th Pctile	CO Effect (%) 95th Pctile	$PM_{2.5}$ Effect (%) 75th Pctile	$PM_{2.5}$ Effect (%) 90th Pctile	$PM_{2.5}$ Effect (%) 95th Pctile
Atlanta-Sandy Springs-Roswell, GA	0.00	-0.32	-0.80	-0.42	-0.68	-0.87
Boston-Cambridge-Newton, MA-NH	0.00	0.00	-0.08	-0.31	-0.58	-0.77
Chicago-Naperville-Elgin, IL-IN-WI	-0.32	-1.05	-1.53	-0.44	-0.73	-0.94
Dallas-Fort Worth-Arlington, TX	0.00	-0.08	-0.32	-0.33	-0.55	-0.70
Denver-Aurora-Lakewood, CO	-0.32	-0.81	-1.53	-0.31	-0.61	-0.87
Detroit-Warren-Dearborn, MI	-0.73	-0.97	-1.45	-0.38	-0.66	-0.90
Houston-The Woodlands-Sugar Land, TX	0.00	-0.20	-0.81	-0.34	-0.59	-0.77
Los Angeles-Long Beach-Anaheim, CA	-0.97	-1.93	-2.66	-0.57	-0.95	-1.28
Miami-Fort Lauderdale-West Palm Beach, FL	-0.73	-1.21	-1.69	-0.24	-0.43	-0.60
Minneapolis-St. Paul-Bloomington, MN-WI	-0.32	-1.05	-1.53	-0.38	-0.71	-0.95
New York-Newark-Jersey City, NY-NJ-PA	-0.32	-0.81	-1.29	-0.40	-0.69	-0.91
Philadelphia-Camden-Wilmington, PA-NJ-DE-MD	-0.08	-0.56	-1.05	-0.43	-0.75	-0.99
Phoenix-Mesa-Scottsdale, AZ	-1.21	-2.18	-2.90	-0.28	-0.56	-0.80
Riverside-San Bernardino-Ontario, CA*	-0.73	-1.45	-2.18	-0.57	-0.95	-1.28
San Diego-Carlsbad, CA	-0.97	-1.93	-2.66	-0.41	-0.66	-0.85
San Francisco-Oakland-Hayward, CA	-0.48	-1.21	-1.69	-0.33	-0.62	-0.90
Seattle-Tacoma-Bellevue, WA	0.00	-0.32	-0.64	-0.26	-0.50	-0.73
St. Louis, MO-IL	0.00	-0.08	-0.56	-0.42	-0.70	-0.90
Tampa-St. Petersburg-Clearwater, FL	0.00	-0.32	-0.81	-0.27	-0.45	-0.58
Washington-Arlington-Alexandria, DC-VA-MD-WV	-0.56	-1.29	-1.77	-0.38	-0.66	-0.85

Estimated percentage change in work performance for the twenty largest Metropolitan Statistical Areas (MSAs) by population in the United States. Columns represent the effect of 75th, 90th, and 95th percentile pollution levels relative to the 25th percentile level for each MSA from 2008 to 2015. With the exception of Riverside-San Bernardino-Ontario (denoted with *), each MSA was home to at least one MLB team from 2008 to 2015. Effects computed using the preferred specification.

Table 4: Robustness

	Preferred	Weather exclusion	Travel days exclusion	Log hours inclusion	Daily avg. pollution	Player FEs	No game exclusion	No sample restrictions	No ump-spec sz
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
CO (>0.50 ppm)	-0.020	-0.020	-0.014	-0.018	-0.005	-0.022	-0.019	-0.009	-0.023
	(0.008)**	(0.008)**	(0.008)*	(0.008)**	(0.006)	(0.008)***	(0.008)**	(0.005)**	(0.010)**
$PM_{2.5}$ (10 $\mu g/m^3$)	-0.004	-0.003	-0.004	-0.005	-0.003	-0.004	-0.003	-0.002	-0.004
	(0.001)***	(0.001)***	(0.001)***	(0.001)***	(0.001)***	(0.001)***	(0.001)***	(0.001)**	(0.001)***
Ozone (ppm)	0.029	0.010	0.058	0.050	0.059	0.045	0.032	-0.015	0.047
	(0.048)	(0.042)	(0.053)	(0.052)	(0.074)	(0.047)	(0.045)	(0.031)	(0.054)
N obs	623,573	623,577	510,262	521,499	637,087	623,186	688,687	1,510,332	623,573
N clusters	12,543	12,543	10,253	10,499	12,805	12,543	13,908	12,543	12,543
N venues	29	29	29	29	29	29	29	29	29
Base controls	Preferred	Preferred	Preferred	Preferred	Preferred	Preferred	Preferred	Preferred	Preferred
Weather controls	Y	N	Y	Y	Y	Y	Y	Y	Y
Exclude travel days	N	N	Y	N	N	N	N	N	N
Log hours control	N	N	N	Y	N	N	N	N	N
Player FEs	N	N	N	N	N	Y	N	N	N
Temporary ump	N	N	N	N	N	N	Y	Y	N
Miscalibrated eq	N	N	N	N	N	N	Y	Y	N
Boundaries res	Y	Y	Y	Y	Y	Y	Y	N	Y
Ump-specific sz	Y	Y	Y	Y	Y	Y	Y	Y	N

Linear probability model. Dependent variable is a binary indicator for a correct call. Standard errors clustered at the game level shown in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$. All regressions include venue, time fixed effects and controls for weather, pitch characteristics, game situation and umpire. See notes in Table 2 for a full description of controls. Column (2) excludes weather variables. Column (3) limits the sample to only game days where the umpire did not travel the previous day. Column (4) adds the log hours since the last game officiated as a linear control. Column (5) replaces pitch level interpolated air quality levels with daily mean levels. Column (6) adds fixed effects identifying the batter and pitcher for each observation. Column (7) includes all games with miscalibrated equipment or officiated by temporary umpires. Column (8) includes all pitches. Column (9) excludes umpire-specific nonparametric strike zone controls.

Table 5: Alternative estimation approaches

	Preferred	Game level analysis	Logit marginal effects	Probit marginal effects
	(1)	(2)	(3)	(4)
CO (>0.50 ppm)	-0.020	-0.021	-0.020	-0.019
	(0.008)**	(0.011)*	(0.008)***	(0.007)**
$PM_{2.5}$ (10 μg/m^3)	-0.004	-0.005	-0.004	-0.004
	(0.001)***	(0.002)***	(0.001)***	(0.00)***
Ozone (ppm)	0.029	0.064	0.028	0.016
	(0.048)	(0.066)	(0.047)	(0.046)
N obs	623,573	12,513	624,622	624,622
N clusters	12,543	12,513	12,543	12,543
N venues	29	29	29	29
Base controls	Preferred	Game level	Preferred	Preferred

Columns (1) is a linear probability model, Column (2) is a linear model, and Columns (3) and (4) are marginal effects from the specified binary outcome model. Dependent variable in Columns (1), (3), and (4) is a binary indicator for a correct call, and in Column (2) is portion of correct calls in a game. Standard errors clustered at the game level shown in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$. All regressions include venue, time fixed effects and controls for weather, pitch characteristics, game situation and umpire. See notes in Table 2 for a full description of controls. Column (2) aggregates the unit of observation to each game and uses the mean rate of correct calls as the outcome, excluding all controls which vary within game. Column (3) and (4) report marginal effects using nonlinear Logit and Probit models evaluated at the mean of all covariates and standard errors computed using the delta method.

Table 6: Other pollutants

	Preferred (1)	Add PM_{10} (2)	Add NO_2 (3)	Add SO_2 (4)	Add all (5)
CO (>0.5 ppm)	-0.020	-0.040	-0.023	-0.020	-0.043
	(0.008)**	(0.014)***	(0.009)***	(0.008)**	(0.014)***
$PM_{2.5}$ (10 $\mu g/m^3$)	-0.004	-0.004	-0.004	-0.004	-0.005
	(0.001)***	(0.002)**	(0.001)***	(0.001)***	(0.002)***
Ozone (ppm)	0.029	0.006	0.029	0.033	0.072
	(0.048)	(0.077)	(0.050)	(0.049)	(0.083)
PM_{10} (10 $\mu g/m^3$)		-0.000			-0.000
		(0.000)			(0.000)
NO_2 (ppm)			0.000		0.000
			(0.000)		(0.000)*
SO_2 (ppm)				0.000	-0.000
				(0.000)	(0.000)
N obs	623,573	222,947	598,890	591,984	217,945
N clusters	12,543	4,584	12,060	1,1907	4,482
N venues	29	15	29	27	15
Base controls	Preferred	Preferred	Preferred	Preferred	Preferred

Linear probability model. Dependent variable is a binary indicator for a correct call. Standard errors clustered at the game level shown in parentheses. *** p<0.01, ** p<0.05, * p<0.1. All regressions include venue, time fixed effects and controls for weather, pitch characteristics, game situation and umpire. See notes in Table 2 for a full description of controls.

Table 7: Placebos

	Preferred Spec (1)	Home Venue Lagged 1 Yr (2)	Away Venue (3)	Farthest Monitor In Lower 48 (4)	Closest Monitor More than 1000 mi (5)
CO (>0.5 ppm)	-0.020	0.004	-0.004	0.016	-0.000
	(0.008)**	(0.006)	(0.011)	(0.074)	(0.012)
$PM_{2.5}$ (10 $\mu g/m^3$)	-0.004	-0.001	-0.001	-0.001	0.001
	(0.001)***	(0.001)	(0.001)	(0.001)	(0.001)
Ozone (ppm)	0.029	0.038	-0.002	0.077	-0.031
	(0.048)	(0.044)	(0.044)	(0.057)	(0.040)
N Obs	623573	529773	578369	802561	745800
N Clusters	12543	10686	11619	16217	14869
N Venues	29	26	29	32	32
Air Quality Placebo:					
AQ Time	Contemp.	Lag 1 Yr	Actual	Contemp	Contemp
AQ Location	Actual	Actual	Away Venue	Farthest	Over 1000 mi

Linear probability model. Dependent variable is a binary indicator for a correct call. Standard errors clustered at the game level shown in parentheses. *** p<0.01, ** p<0.05, * p<0.1. All regressions include venue, time fixed effects and controls for weather, pitch characteristics, game situation and umpire. See notes in Table 2 for a full description of controls. Column (2) replaces contemporaneous pollution levels with levels lagged by precisely one year. Column (3) uses pollution values from venue of visiting team. Games between two teams based in the same US Census commuting zone are excluded. Column (4) uses pollution levels from the pollution monitor in the continental United States most distant from the venue. Column (5) applies pollution levels from the closest monitor at least 1000 miles from the venue.