The Cognitive and Emotional Impacts of Air Pollution: Evidence from Twitter

Jared Dean Katz*

April 2, 2023

Abstract

I study whether heightened air pollution leads to cognitive and emotional responses at the infra-marginal level. To address this question, I employ geolocated, timestamped Twitter microdata. Using an original dataset of over 30 million unique Tweets, I observe linguistic responses to varying levels of pollution across the U.S.. I find that Tweets from higher-pollution backgrounds are more negative and aggressive than Tweets from observably similar backgrounds with less pollution. Additionally, I find evidence that higher-pollution Tweets score cognitively lower at highest levels of pollution, but otherwise may result in better cognitive scores. I find that lower cognitive-scoring users Tweet less as air quality worsens, but individual users Tweet at a lower level. I also find that more negative and cognitively low-scoring Tweets are more vulnerable to air pollution's negative effects than high-scoring Tweets.

*Jared Dean Katz: Vanderbilt University, 2301 Vanderbilt Place PMB 354264, Nashville, TN 37235, jared.dean.katz@vanderbilt.edu. Thanks to Andrew Dustan (Advisor), Ariell Zimran (Honors Program Head), and Michelle Marcus (Thesis Committee) for their guidance and patience throughout.

1 Introduction

Air pollution is an ever-growing part of industrial life, as it is a byproduct of many processes that increase consumption and welfare. At the same time, air pollution can directly affect welfare itself if people have a preference for cleaner air. To determine the utility-maximizing level of air pollution, it is important to know all the direct costs of air pollution. Two of these costs are people's cognitive and emotional responses to air pollution. If worse air quality causes people to be more negative or cognitively impaired, it could result in significant economic costs policymakers need to consider.

What are the impacts of air pollution on the cognitive and emotional outcomes of those subject to it? This paper exploits a unique dataset of 30 million geolocated Tweets over a three-month period to examine linguistic responses to air pollution. From each Tweet, I extract measures of sentiment and writing level from the text using natural language processing techniques. Then, I exploit the availability of location and time markers to create a two-way fixed effects identification. I compare the cognitive and emotional outcomes for Tweets within the same region on high and low pollution days in that region, controlling for the overall pollution of the day and the weather in the day-location. This paper addresses the question of how the emotional and cognitive qualities of language are affected by air pollution.

Heightened air pollution has been studied in epidemiological and toxicological settings, and is related to both short and long-term health outcomes (Pope et al., 2009; Archsmith et al., 2018; Dockery et al., 1993). In economics specifically, research has demonstrated air pollution exposure can affect physical health, cognitive ability, mental health, and crime among other outcomes (Schlenker and Walker, 2015; Hanna and Oliva, 2015; Chen et al., 2018; Burkhardt et al., 2019). Health effects are large at an individual level, but relatively infrequent (Schlenker and Walker, 2015). Cognitive effects also have mainly been studied in test-like situations, which involve cognitive tasks such as math problems that are uncommon in daily life, and they are unstudied in writing contexts (Chen et al., 2018).

This paper adds to the literature in three ways. First, I analyze the effect of contemporaneous air pollution on infra-marginal outcomes in a new context by focusing on writing-related tasks. Previous studies on cognitive and emotional effects of pollution use tests or surveys, where I analyze writing that users themselves choose to publish, independent of research. As a result, this study may better approximate the true effect of pollution on emotional and cognitive outcomes and identifies the effect in a new context. Second, I improve on the external validity of previous work, which often uses smaller samples or populations outside the U.S. which typically are exposed to more air pollution. Third, I make small methodological improvements to the use of Twitter data in economics. I construct measurements of heterogeneity for outcomes by analyzing the effects of air pollution at different percentiles of Tweets.

I find that worse air quality can lead to a statistically significant increase in the share of negative words[1] from Tweets. Similarly, I find that worse air quality leads to a significant increase in the propensity of a Tweet to contain a curse word. In contrast, I find that changes in AQI have mixed effects on cognitive measures of a Tweet, specifically the Flesch-Kincaid "grade level" of a Tweet. These results are robust to a variety of specifications and are not explained by weather or time-related phenomena such as day-of-week or holidays. By incorporating user fixed-effects, I demonstrate that changes in the sample composition of users caused by air quality shifts have little impact on the results shown for emotional outcomes, but significant impacts on grade level. I find evidence that when air quality worsens, individual users Tweet at a lower cognitive level, but many previously low-level Tweeters will not Tweet at all. Next, I focus on understanding heterogeneity in my results. By concentrating on different percentiles in the distribution of Tweets, I show that additional air pollution affects Tweets at lower grade levels more than Tweets at higher grade levels.

My results have important implications for future research and policy. Air pollution is

[1]According to the VADER Negative Sentiment Score outlined in (Hutto and Gilbert, 2014). This is precisely calculated by determining the total negative valence score (the sum of all negative words weighted by their negativity) and then normalizing by the total valence score to be between 0 and 1.

a global issue that most economists agree demands regulation, so estimating its true cost is necessary (WSJ, 2019). My results indicate there are significant, but understudied, effects of pollution beyond high-stakes health shocks or cognitive tests. As a result, an estimated social cost of pollution might be understated by the current scope of research. This cost is imperative to be accurate, as it is used in policymaking to set pollution caps and taxes as well as alternative energy subsidies.

2 Literature Review

This paper fits in both the large literature on the cumulative costs of air pollution and the small but rapidly developing literature on the use of social media microdata in economic analysis. Studies from a variety of disciplines have concluded that the costs of air pollution are large and span a wide range of outcomes (Manisalidis et al., 2020; Zivin and Neidell, 2012). First, epidemiological, toxicological, and economic research have shown large effects of air pollution on health (Pope et al., 2009; Schlenker and Walker, 2015). Many of these studies focus on the long-term effects of continuous exposure to air pollutants rather than the effect of short-term exposure to air pollution. The most relevant paper in the health literature, Schlenker and Walker (2015) studies the impacts of contemporaneous air pollution on hospitalization rates, heart attacks, and asthma attacks. However, these studied events are all rare and costly, such that an increase in air pollution will not affect most people on the margin. My paper differentiates by focusing on small, but general, shifts in behavior, rather than previously identified large health shocks. This identifies the infra-marginal effects of contemporaneous air pollution that impact entire populations, rather than large shocks that may only affect relatively small groups. Together with Schlenker and Walker (2015), my paper presents a more complete view of the large, rare as well as the small, common harms of contemporaneous air pollution.

Next, there is a significant number of studies on the relationship between air pollution

and mental health Chen et al. (2018); Szyszkowicz et al. (2009). Calderón-Garcidueñas et al. (2014) identifies differences in prefrontal cortex size from MRI scans as a result of pollution exposure, and linked this with cognitive and emotional dysfunction. Chen et al. (2018) shows a link between air pollution and mental illness across China. Burkhardt et al. (2019) shows a relationship between contemporaneous air pollution and criminal activity, an indication that air pollution may spawn aggressive behavior. In these contexts, this paper shows the infra-marginal effects that contemporaneous pollution might have on populations without shifting individuals from healthy to mentally ill or law-abiding to aggressive criminals. Since behavioral outcomes are historically difficult to measure, this paper is the first to my knowledge to explore the low-stakes impact of air pollution on behavioral outcomes, as opposed to the previously studied catastrophic outcomes. Similar to my paper's relationship with the physical health literature, this paper forms a more complete picture of the individual-level effects of air pollution by estimating the small, wide-reaching costs of air pollution on mental health.

Heightened air pollution for a sustained period has also been suggested to shift cognitive abilities (Chen et al., 2018; Zare Sakhvidi et al., 2022; Braithwaite et al., 2019). Most relevantly, Chen et al. (2018) argues that short periods of exposure to increased pollution have limited effects on verbal and math test scores, but longer periods of exposure can have significant effects. By using Twitter microdata rather than a formal cognitive exam, this paper tests whether those results are robust in everyday situations.

Finally, there is a small and growing literature on the use of social media or microblogging data to perform economic analysis to which this paper belongs. Natural language processing techniques such as sentiment analysis have a history of use in economics Gentzkow et al. (2019). Twitter data has also been used in various contexts, such as determining the impacts of public figures' statements on financial markets, identifying extremist rhetoric, and understanding economic uncertainty (Bianchi et al., 2021; Mitts, 2021; Chen and Chen, 2022). Most related is Baylis (2020), which links negative sentiment and aggression in Tweets to

extreme temperatures. This paper follows much of the same approach as Baylis, with additional methodological extensions. First, I take extra focus on cognitive outcome variables in addition to the emotional outcomes Baylis studies. Furthermore, I also exploit Tweet-level data to identify the heterogeneous impacts of air pollution.

3 Data

3.1 Twitter Data

From May 30th to August 19th, 2022, I continuously[2] streamed Tweets from the Twitter API. Each Tweet consists of the Tweet's text, a Core Based Statistical Area (CBSA) region the Tweet was sent from, and the time at which it was sent. I drop retweets to focus on original content production. I drop non-geotagged Tweets and Tweets geotagged outside of the United States for practical reasons. The result is over 30 million unique Tweets across 3168 unique CBSAs in the United States. I aggregate Tweets at a CBSA-day level for computational reasons, making one unit of observation a CBSA-day in the main results. I choose to not present results at a User-day level to allow for infrequent Tweeters to be included in my analysis, improving the external validity of the results. At times, I take advantage of my data's granularity by breaking down my data to the user-day level, and I make a note when doing so.

My dataset consists of 130,000 CBSA-days. A major concern is that this data is not representative of the U.S. population. One possibility is that this data may be dominated by few CBSAs, as small geographic areas such as Los Angeles or New York create a disproportionately large number of Tweets. If only a few geographies drive my results, then it would be difficult to apply my findings externally. Figure 1 presents the distribution of the number of Tweets on each CBSA-day. It is clear that the data is not dominated by

[2]Tweets were continuously streamed apart from a handful of brief periods of downtime for technical issues, generally ranging between a few hours and a few days. These periods were random as a result of technical errors, and have no relationship to any outcome variables or air pollution at the time.

a few CBSA-days, as most of the Tweets in the sample come from CBSA-days with a few hundred to 2500 Tweets. A still significant portion of the Tweets come from CBSA-days with multiple thousands of Tweets. There exists a small, but significant tail of a few locations that contribute a large number of Tweets, but far from the majority of Tweets in the sample. The aggregation of Tweets at a CBSA-day level adds to the power of the results, as the data includes not solely the 130,000 observations, but the thousands of Tweets each observation contains. Another possibility that would present a problem is if geolocated users in each CBSA were fundamentally different than non-geolocated users in the same CBSA. I defer to Baylis (2020), which collects Twitter data using the same process as me and demonstrates that the geolocated Tweets have the same distribution of sentiment as non-geolocated Tweets, use the same most common words, and are sent at the same time of day. Finally, while Twitter has slightly different demographics from the U.S. population at-large, evidence from sentiment measures indicates that the two groups are similar enough that Twitter users can be considered representative of the population (Baylis, 2020).

While raw Tweets themselves may be interesting for qualitative analysis, they lack the ability to be used in quantitative analysis without being translated into data via natural language processing techniques. I categorize the data from natural language processing techniques in this paper into cognitive and emotional measures. I first compute measures at the Tweet level, then take the mean score of all Tweets in each observation to aggregate up to the observation's measure.

To measure the cognitive value of a Tweet, I calculate the Flesch-Kincaid (F-K) grade level score of each Tweet (Kincaid et al., 1979). The F-K grade level of a passage is designed to measure approximately how complex a passage of English is. It is particularly useful in this case because, unlike other standard tests of readability, it does not require a large passage to be sampled from, making it feasible in microblogging contexts, and is the standard readability measure used in the sciences (Albright et al., 1996; Badarudeen and Sabharwal, 2010; Cooley et al., 1995). The grade level is computed using the number of words per

sentence and the number of syllables per word. Sentences that have many words in each sentence and many syllables in each word are considered complex and therefore score a higher grade level. I define a Tweet without punctuation or a line-break as one individual sentence.[3] One important note is that while the F-K grade level can at most values be considered close to a school reading level, they are not direct parallels, especially at extrema. F-K grade levels have no upper bound if sentences become infinitely long or words have on average infinitely many syllables. Since scored grade levels can often supersede the standard levels of American schooling, I bound scores to $[0, 16]$. Results remain robust when including the original scores or when entirely removing outlier Tweets. In an extension of the main results, I also run regressions on various percentiles of grade levels. In these cases, rather than averaging all of the grade levels of a CBSA-day together, I take the respective percentile from the distribution. This allows me to capture the changes in the distribution of grade level– whether pollution's impact comes through a channel of affecting higher or lower grade level Tweets.

I use two techniques to estimate the emotional value of a Tweet. First, I use a standard and well-documented sentiment analysis dictionary, the VADER lexicon (Hutto and Gilbert, 2014). VADER is specifically designed for microblogging contexts, such as Tweets, and is considered a standard sentiment analysis method for a corpus of many Tweets because of its speed and accuracy on many different datasets. It works by scoring individual words, phrases, and key grammatical structures (i.e. "not good" is scored as "bad"), and then averaging the scores across the passage. The VADER "compound score" is the standard Python implementation in the Natural Language Toolkit package and scores a passage between -1 and 1, with a score of -1 as the most negative, a score of 0 considered to be neutral, and 1 as the most positive (Bird et al., 2009). I scale these scores by 100 to aid in interpretation, which leaves results unchanged. I similarly scale the VADER "negative score" to between 0 and 100,

[3]This would present a concern if less pollution resulted in less punctuation, artificially making sentences longer and increasing F-K scores. Since intuitively air pollution would lower the use of proper punctuation, this is likely to bias my estimates towards 0.

interpreted as the share of the passage that is negative. The passages that only use the most negative words are assigned a score close to 100, while a passage that has no negative words would be assigned a score of 0. I also calculate percentiles in the distribution of Tweets on each CBSA-day for further extensions into the distributional impacts of air pollution. Additionally, I take advantage of the aggregation of many Tweets into one CBSA-day to measure the emotionality of a Tweet–the likelihood of a Tweet containing a curse word.[4] Tweets that have many curses are likely to contain heightened emotions, are stylistically emphasized, indicate pain, or are about adult topics (Janschewitz and Kristin, 2012). From a list of curse words, I calculate the percentage (the share $\times$ 100) of Tweets in a CBSA-day containing a curse word, which is resistant to outlier Tweets that may contain many curses.

Table 1 summarizes the variables used in the empirical analysis. Rows 4 through 6 of Table 1 restrict the sample data to CBSA-days with specific AQI levels and provide important motivation for the rest of the paper. When air pollution is higher, a CBSA has more negative, lower grade levels, and higher cursing Tweets. I aim to identify the causality of this relationship in Sections 4 and 5. One potential concern is that there is limited variation in the distributions of these newly defined output variables. Figure 2 shows the distribution of the outcome variables across the dataset in order to alleviate this concern. One important characteristic of the data is that there must be significant variation within locations in order to identify any causal link between air pollution and outcomes. The final two rows of Table 1 show that this variation does exist.

3.2 Air Quality and Weather Data

I obtain air quality data from the EPA's Air Quality Survey (AQS) PM2.5 API, which provides air quality using the official U.S. Air Quality Index (AQI) (US Environmental Protection Agency). The AQS measures the AQI based on the PM2.5 concentration in the air around the monitor location, such that a lower score is considered to be cleaner air

[4] I use the Better Profanity package to identify curses, which works particularly well on Tweets because of its recognition of misspelled or altered curses, like ones that have vowels replaced with "*" (pro, 2021).

(United States Environmental Protection Agency). An important point is that the AQS does not collect data by CBSA, but by monitor. I aggregate my data up from the monitor to the CBSA level by averaging the reports of all monitors located in the CBSA. One concern is that each monitor does not necessarily collect data every day, leaving holes in the data. I drop all CBSA-days where there is no pollution information. [5]

Figure 3 displays the distribution of AQIs for CBSA-days over my sample. The average AQI reading in the sample is 33.33 with a standard deviation of 15.11. The EPA defines air quality to only begin posing a "Moderate Threat" when the AQI is above 50 and does not define air pollution as "Unhealthy for Sensitive Groups" until AQI reaches above 100. My sample does not include any data with AQIs the EPA deems "Unhealthy", "Very Unhealthy", or "Hazardous", as these conditions rarely occur across entire CBSAs in the U.S. As a result, my data spans only air pollution ranges that the EPA considers being safe and are unlikely to be explicitly noticeable by people. This is a strength of the data, as it captures effects at generally considered healthy levels of pollution. It also hints that, at higher levels of pollution, there is potential for larger magnitudes of effects. The effects of air pollution that I find occur in air quality regularly experienced in the U.S. without explicitly noticeable discomfort.

Baylis (2020) shows that temperature, which is related to pollution, can be linked to emotional outcomes, so it is important to control for the weather to identify a causal relationship (National Weather Service). As a result, I collect temperature and precipitation data from the National Oceanic and Atmospheric Administration (NOAA) through the python package meteostat (Lamprecht, 2023). I perform a similar aggregation approach for weather data as for air quality data, by taking the average of NOAA monitors across the CBSA. For temperature data, I find my results are robust to controlling for the maximum, average, and minimum temperature across a CBSA, as well as amount of precipitation.

[5] Zivin and Neidell (2012) use interpolation to fill in missing pollution data, which is a slight majority of all CBSA-days. I elect instead to drop these entries and can consider interpolation at a later point.

4 Empirical Strategy

I seek to determine the effect of air pollution on cognitive and emotional responses within writing. To that end, I identify the causal effect of air quality on an array of outcome variables using a panel fixed effects model. First, I treat air quality as a continuous exogenous variable to capture a linear relationship. Then, using a flexible functional form for my main results, I treat AQI as a series of bins in order to capture nonlinearities in the dose-response function. The flexibility is justified because the EPA's categorization of different AQI scores suggests different bins of severity, and intuition suggests that the dose-response function of air pollution may be nonlinear. Specifically, there might be a breaking point where the costs of harmful air ramp up or a point where all of the damage of harmful air has already been done. By using a panel dataset, I am able to control for unobservable cross-sectional or temporal characteristics that may affect my outcome variables. Cities, for instance, have on-average higher cognitive-scoring Tweets, and Tweets on Saturdays are more positive than on Tuesdays. Standard regression would not account for these threats if cities or Saturdays tended to have more pollution as well. Using CBSA and time fixed-effects, I am able to account for these unobservable characteristics to identify a causal relationship. I estimate the following models:

$$\bar{O}_{cd} = \beta A_{cd} + T_{cd} + P_{cd} + \phi_c + \phi_{time} + \epsilon_{cd} \tag{1}$$

$$\bar{O}_{cd} = f(A_{cd}) + T_{cd} + P_{cd} + \phi_c + \phi_{time} + \epsilon_{cd} \tag{2}$$

where c and d index CBSA and day. $\bar{O}_{cd}$ is the CBSA-day average of the outcome variables as described in Section II. T_{cd} and P_{cd} are the daily maximum temperature on the CBSA-day and the amount of precipitation on the CBSA-day, respectively. A_{cd} is the AQI for the CBSA-day, such that the coefficient of interest in equation (1) is β. In equation (2), let $f(A_{cd}) = \sum_b^B \beta_b A_{cd}^b$, where A_{cd}^b is an indicator variable equal to one if A_{cd} falls in the

given bin b. Here, the β's are the coefficients of interest. ϕ_c represents CBSA fixed effects, and ϕ_{time} represents temporal controls of day-of-week and holiday fixed effects. Results are robust using day-of-sample fixed effects as well. ϵ_{cd} is the unclustered error term, but results are robust to clustering both by CBSA and CBSA-day.

To allow for CBSA-days with more Tweets to be weighted stronger, I use weighted regressions based on the average number of Tweets in the CBSA over the sample period. A potential problem with weighting by Tweet count is that a user's decision to send a Tweet may itself be a function of pollution. By using the average Tweet count across the CBSA, I avoid these concerns and can weight without endogeneity threats. Another benefit of weighting my results is that I am able to estimate the Tweet-level response, rather than the location level response, which more accurately portrays the true costs of emotional and cognitive damage from air pollution.

A^b_{cd} specifies AQI bins of 10 points running between 0 and 80, with an edge bin for all observations with an AQI greater than 80. Results are robust to various bin size choices. In each case, I choose the bin that contains 25 AQI as the omitted category, which does not alter the shape or significance of any of my outcome response results. My identifying assumption is that deviations in air quality are as good as random after accounting for variation by CBSA, day-of-week, presence of holidays, temperature, and precipitation. Conditional on the assumption above, the coefficients βs are the average change in the outcome variable resulting from replacing a day in the omitted bin with a day in air quality bin b.

One potential concern to identification is additional variables that correlate both with AQI and our outcome measurements. For instance, heightened traffic can spur worse air quality, but also might make people unhappy directly. This relationship may lead to βs being overestimated, but only if daily traffic variation cannot be explained by holiday or day-of-week fixed effects. Another possible concern is heterogeneous measurement error in air quality. Because of the method used to estimate air quality, I often rely upon few AQI monitor readings to infer the air quality for a larger CBSA. Some CBSAs have more monitors

of higher accuracy that report more frequently than others, leading to more accurate AQI measures in certain CBSAs. This can bias βs towards 0. Similarly, air quality can fluctuate throughout the day, introducing further measurement error when we assume that a Tweet was sent with a certain level of pollution, further biasing my estimators towards 0. Another concern is that avoidance behavior may cause Tweeters to move inside under certain levels of air pollution, or make some Tweeters less likely to send a Tweet resulting in compositional sorting (Heckman 1979). The first concern biases β towards 0. The second has a more ambiguous effect, which I examine in Section 6. Finally, it is important to note that the β's represent an intention to treat effect rather than an average treatment effect. This is due to the fact that not every Tweeter in a CBSA with AQI A will actually be exposed to the air when sending their Tweet (they may be inside, for instance). This further biases β's towards 0.

5 Results

Table 1 suggests a relationship between air quality and the cognitive and emotional values associated with a Tweet. In this section, I estimate equations (1) and (2) in order to test this relationship causally. I find statistically significant declines in emotional measures resulting from worse air quality, but insignificant results from cognitive measures. This suggests a straightforward, linear relationship between the emotional value of a Tweet and the air pollution it was sent at. At the same time, the writing quality of Tweets is generally unaffected by changes in AQI. In Section 6, I explore some of the questions motivated by these main results.

5.1 Linear Results

Figure 4 and Table 2 report the estimated effect of AQI on our outcome variables of interest. As reported in column (2), a one-standard-deviation increase in AQI exposure (worse air

quality) is estimated to cause a 0.025 standard deviation decrease in VADER Compound Sentiment. This appears to be small, but its significance is worth noting as the total cost of the effect is across an entire population. I also estimate an intent-to-treat effect, which constructs a lower bound on the average treatment-on-the-treated effect. In column (1), I find results for compound sentiment that mirror column (2), such that compound sentiment is lower (worse-off) when pollution increases. In column (4), I similarly find that an increased AQI causes a statistically significant increase in the likelihood of a Tweet containing a curse word, evidence that air pollution may cause more adult, emotional, or extreme behavior. These results showcase the infra-marginal effects of contemporaneous air pollution. While other parts of the literature identify that high air pollution may make us sick or suffer worse exam scores, I show that even when air pollution does not cause a large shock such as a mental health diagnosis, it can have significant effects on emotional well-being enough to cause noticeable shifts in language use (Schlenker and Walker, 2015; Chen et al., 2018).

On the other hand, I find that the β for grade level is not statistically significant, is small, and is positive (the opposite sign of what is considered intuitive). These results are further explored in my main results in Table 3 and indicate that additional air pollution at the values in my sample has no significant effect on writing ability. There are a couple of possible explanations for why this effect may be taking place. First, my empirical design introduces a lot of noise. Not all users may be Tweeting while exposed to air pollution, and the F-K grade level in measuring Tweets can be imprecise and introduce randomness. By contracting the sample, I may find offsetting significant effects–grade level of some users may increase while others may decrease. For example, low-grade users may be made worse off as a result of air pollution, while high-grade users now talk about more complex topics like politics in response to heightened pollution, artificially increasing their grade level. Here, these effects may offset to create the null result, when in reality there are multiple competing effects. Another possibility is that selection into the sample of Tweeters plays a significant role. As air pollution worsens, some low-grade users may become less likely to Tweet, which

can make it appear that a CBSA's grade level remains unchanged, even if users within the CBSA are made worse off. I explore these possibilities further in Section 6.

5.2 Nonlinear Effects of AQI on Cognitive and Emotional Outcomes

Table 3 and Figure 5 present my main results, which utilize the same TWFE identification, but allow for nonlinearity in the effect of air pollution. By allowing for flexibility, they estimate a dose-response function to air pollution in the form of increased AQI. The results for $A \in [a, b)$ can be interpreted as the effect of replacing a day of AQI $A \in [20, 30)$ with AQI $A \in [a, b)$ instead. I find results largely consistent with Table 2, but the additional flexibility shows that effects do have nonlinearities across all outcome variables.

Column (1) of Table 3 and Panel A of Figure 5 reflect results for VADER compound sentiment, and column (2) and Panel B shows results for VADER negative sentiment. Both results contain time and CBSA fixed effects to absorb variation based on location, day of week, or holidays. Qualitatively, it seems effects on compound sentiment, negative sentiment, and cursing percentage are linear up until an AQI of around 60, where the effects flatten out. For the compound sentiment, most values are statistically significant from 0 and indicate that an increase in air pollution translates to a decrease in overall sentiment. For negative sentiment, all values are statistically significant from 0.[6] Panel D and Column (4) present the fraction of Tweets containing curses on a CBSA-day. I find a statistically significant relationship at some levels between cursing and air quality such that worsened air quality increases the fraction of Tweets containing curse words. However, a drawback of the flexible form model is reduced power, which this suffers from. A joint F-test is close to significant at a 10% level, and the dose-response function indicates my linear results from Table 3 are likely true. The flattening out of the dose-response functions for these outcomes suggests a

[6] I find similar results across a range of specifications, including a true TWFE model with day-of-sample fixed effects rather than time fixed effects, models that remove outlier Tweets in terms of sentiment and grade level, and models that remove small Tweets.

saturation effect. Since I think about a Tweet experiencing pollution as an Intent to Treat (but not necessarily a treatment), a Tweet is only vulnerable to pollution exposure if it is actually experiencing the treatment (sent while outside, for example). Then, this saturation effect occurs when all of the Tweets that are both treated and vulnerable to treatment, have been made more negative from the increased AQI. Another possible explanation is that as AQI becomes sufficiently large, people notice it and begin to exhibit avoidance behavior, or some people exposed choose to no longer Tweet. I explore this possibility in Section 6. Of course, without studying a sample with a wider range of AQI, it's impossible to know whether this trend will hold for larger pollution scores, but it presents an interesting nonlinear phenomenon that could have important political implications.

The negative relationships between emotional outcomes and increased pollution reflect similar results on contemporaneous pollution and mental health (Chen et al., 2018) and pollution and aggressive behavior (Burkhardt et al., 2019). By identifying nonlinearities, I add nuance to previous results found on pollution and emotional outcomes, which show a relationship between crime and air pollution. I also contribute to the literature on emotional outcomes by examining a strictly low-cost outcome. Unlike seeking a mental health diagnosis or committing a crime, changing the language of a Tweet is costless and immediate. Similarly, its measurement spans a continuous spectrum rather than a binary classification. As a result, I am able to capture a much more general infra-marginal effect. This adds an additional cost of air pollution to the literature that was previously unobserved. While previous literature shows that air pollution may cause people on the border to seek a diagnosis or commit a crime, I show that air pollution also has significant impacts on the people we previously did not see any effect on.

Lastly, Panel C and column (3) present results for grade-level responses to AQI. I find results that offer more insight into the null result presented in column (3) of Table 2. At small AQI levels (0-50), I find that grade level is not responding to increased air pollution. This mirrors the effect found in Figure 4. However at high levels of pollution, I actually

find that gradae level starts to respond negatively to increased air quality. This indicates that grade level may be impacted by air pollution, but a certain threshold of pollution is required before these effects materialize. Overall, it appears as if grade-level effects in writing from contemporaneous pollution are near zero and unnoticeable for much of the levels of air pollution the U.S. experiences, but at large enough levels of pollution, an effect may begin to take place. In the next section, I examine how this result may be the product of heterogeneous effects and selection into the sample.

6 Extensions

This section expands upon the main results of the paper to give extra care to understand the underlying mechanism behind the main results. First, I examine the effect that selection into the sample has on the results, providing insight into whether the underlying mechanism is a compositional one. Then, I examine the distributional effects of air pollution to better understand the types of Tweets that are most affected by air pollution.

6.1 Selection into the Sample

Twitter users choose where, when, and whether they will Tweet, which may introduce a selection bias into the main results of Figure 4 and Figure 5 as described by Heckman (1979) and Baylis (2020). I will test for this concern by following Baylis' argument and exploiting user-fixed effects on individual Tweets. The argument operates under the assumption that there exist positive (high cognitive-scoring) and negative (low cognitive-scoring) users. Positive users only Tweet high-scoring messages, and negative users only Tweet low-scoring messages. A concern would be that AQI only shifts the active balance between positive and negative users, but does not affect a user's writing individually. My main econometric approach does not account for the concern that CBSAs may shift in their composition of users as AQI varies, rather than making individual users better/worse off. Figure 6 plots

the dose-response function for the number of Tweets sent on a CBSA-day as a function of AQI. It is clear that the number of Tweets sent on a CBSA-day is statistically independent of the pollution in that observation, which indicates that users are not choosing Tweet based on the air quality. However, the possibility still remains that when air pollution worsens, a group of positive users is swapped out for an equal-sized group of negative users. This would create the main results, but would only mean that air pollution shifts the Twitter landscape, not make individuals more negative.

To address this concern, what Baylis (2020) and I perform introduces user fixed-effects. This removes the concern of different users tweeting on days with higher/lower pollution by controlling for the user directly. I estimate the following model:

$$\bar{O}_{cd} = \beta A_{cd} + T_{cd} + P_{cd} + \phi_i + \phi_{time} + \epsilon_{cd} \tag{3}$$

This model replaces equation (1) with user fixed-effects rather than CBSA fixed-effects. In order to avoid computational concerns while using a disaggregated sample, I trim my model to include only users who Tweeted on 5 or more unique days from the same CBSA, but sent less than 10 Tweets per day throughout the sample. While this model is useful to examine the effect selection may play in my results, it is limited by the added rigidity of user fixed-effects. Specifically, the incoporation of user effects means that infrequent Tweeters are excluded from results, reducing external validity as the population that Tweets frequently may be different than the population of general users. Linear results are presented in Figure 7, with nonlinear results in Figure 8.

For emotional outcomes, my results generally show the same pattern as Figure 4 and are consistent with my linear estimates in both direction and magnitude. Since my user fixed-effect results mirror my main results, I can rule out the idea that air pollution promotes emotional negativity by increasing the proportion of negative users who are Tweeting. This adds to evidence that air pollution's effect occurs at the user level–air pollution is something that affects the person individually. This also corroborates the surrounding literature on air

pollution and emotional and cognitive outcomes. Since large shocks identified in the rest of the literature happen at the user level, it makes sense that small shifts would as well. I present nonlinear results for Compound Sentiment in Panel A of Figure 7. The results mirror my main results in Panel A of Figure 5, but lack power for statistical significance as a result of the reduced sample. However, the existence of user fixed-effect results with the same trends and magnitudes as my main results indicates that selection is not driving my results for emotional outcomes.

Examining grade level, on the other hand, shows that selection *does* appear to have a role in determining the grade level of Tweets. Using user fixed-effects, Figure 6 shows a statistically significant, negative relationship between air pollution and grade level. This indicates that when air quality decreases, some low-cognitive scoring users stop Tweeting, which inadvertently makes it seem like grade level does not change. In reality, grade level at an individual level is actually remaining unchanged or decreasing slightly. This is corroborated in Panel B of Figure 7, which presents nonlinear results. As a result of reducing my sample, this lacks power for statistical significance. On the other hand, this new negative effect can potentially be explained from the change in sample, and the fact that frequent Tweeters may be substantially different than less frequent Tweeters. The next subsection adds creedence to this idea by demonstrating that the group being studied can play a strong role in the results found. It appears that most groups seem unaffected cognitively by AQI, but some groups experience substantial harm. If rare Tweeters are less likely to allow air pollution to affect their writing level than frequent Tweeters, then that would explain the results of Figure 7 independent of any concerns about selection into the sample.

6.2 Heterogeneity

In order to understand which Tweets are most affected by air pollution, I exploit Tweet-level data to understand heterogeneity in the responses of different Tweets belonging to the same CBSA-day. Using the same sample as for my main results in Table 3, I define

new measures for compound sentiment and grade level that take advantage of the Tweet-level data. Within each CBSA-day, there exists a distribution of Tweets according to each outcome variable. Intuitively, as air pollution varies, this distribution shifts. In my main results, I consider the score for a CBSA-day to be the average score of all Tweets within that CBSA-day in order to capture the shift that occurs from the center of the distribution. Now, I explore the possibility that the distribution of Tweets in a CBSA-day may also stretch out asymmetrically as air pollution changes. This can be interpreted as different parts of the distribution of Tweets being effected heterogeneously by changes in AQI. To capture this effect, I take the outcomes of Tweets at specific percentiles in each CBSA-day. Thus, I can now interpret results that show the impact of AQI on the n^{th} percentile sentiment or grade level Tweet. This is crucial for studying the channel from which air pollution affects my outcomes– I can now answer questions about the heterogeneous impacts of air pollution.

Figure 9 shows the results for the linear effect of AQI on Compound Sentiment and Grade Level at various points in the distribution of Tweets. In Panel A, air pollution is shown to have the largest impact on the 10th percentile of Tweets. As the percentiles increase, the magnitude of the effect decreases, with the effects on the 75th and 90th percentiles as insignificant. Results are similar in Panel B, which shows the same set of results for grade level. The prior section suggested that grade level results are sensitive to the choice of sample, which Panel B confirms. I find that AQI has a limited or even slightly positive effect on grade level except at the left tail of the distribution (bottom 10%), where it has a significant negative effect. This helps to explain the puzzle regarding grade level I find in my results thus far. It appears that only a small fraction of users are vulnerable to the cognitive damages from air pollution, but for that subset of users the damages caused by pollution can be statistically significant. Figure 10 shows the nonlinear results for compound sentiment and grade level. In both Panels A and B of Figure 10, it is again confirmed that right-tail Tweets ("better" Tweets) are affected significantly less by AQI, while low-sentiment/grade-level Tweets bear the brunt of the effect. This captures the idea that the within-CBSA

distribution of Tweets stretches out as AQI increases. It also shows that AQI's primary channel into our output variables is through affecting the Tweets/Tweeters who are already the most negative or low-scoring.

The results shown in Figures 9 and 10 have important policy and equity implications. They indicate that increased air quality has the largest impact on those who are already the most vulnerable– air pollution does not affect us all equally even at the smallest levels. These results mirror results from the literature indicating the inequities of air pollution, but they are particularly interesting because they showcase the inequities even at non-catastrophic levels, where previous inequities were related to large health or test score shocks.

6.3 Contextualizing Costs

While my results show that air pollution does affect the emotional and (to a lesser extent) cognitive outcomes of users, it is important to quantify the magnitude of these effects in a way that would be useful to policymakers. The primary goal of measuring the costs of air pollution is to determine an optimal level of pollution, but this requires the economic cost to be measured in a unit that can be compared to other costs and benefits. Figure 11 presents the linear results for each of my outcome variables in comparison to the effects caused by other phenomena, such as the weather, day of the week, or whether the day was the 4th of July. When compared to other effects, I find relatively large costs associated with a standard deviation increase in AQI. Panels A and B suggest that a standard deviation increase in AQI in my sample is comparable in effect to a standard deviation increase in temperature (just over 7 degrees Celsius) in the same period. Panel B finds that a standard deviation increase in AQI increases negative sentiment by same magnitude as the 4th of July decreases negative sentiment. Panel D finds that a standard deviation increase in AQI increases cursing at around the same magnitude as switching from a Saturday to a Tuesday. Panel C is presented for completeness, but since the effect of AQI on grade level in my linear results is near-zero and insignificant, the effects are relatively small compared to other

drivers of writing ability.

A policymaker may be interested in a rough estimate of the monetary costs of these changes in outcome. To predict this, I perform a back-of-the-envelope calculation using the work from Baylis (2020), which estimates the per-SD daily value of sentiment, taken from the same data source using the same sentiment measures, at 196.77 By multiplying this value with my standardized regression coefficient from Column (2) of Table 2, I estimate loosely that a standard deviation increase in AQI for one day can be valued as a loss of $4.82 per person. Figure 12 presents a complete set of nonlinear results, which can be interpreted literally as the daily cost associated with swapping a day with AQI $[20, 30]$ with a day of AQI $[a, b]$.

7 Conclusion

Understanding the infra-marginal effects that air pollution may broadly have on human behavior is essential for finding a complete cost of air pollution, and thus crafting efficient environmental policies. Often, a lack of data on cognitive and emotional responses that can be coupled with pollution makes it difficult to identify a causal effect of pollution on these outcomes. This paper solves the measurement issue by exploiting the granularity of geotagged Twitter data as a broad survey tool. I identify cognitive and emotional responses to short-term variations in levels of air pollution, then estimate response functions for sentiment, profanity use, and grade level outcomes as a result of an AQI flexible form model. I find significant intent to treat effects of pollution exposure on emotional outcomes in Tweets: increased profanity use, increased share of negative sentiment, and decreased compound sentiment. I find smaller and less significant effects of pollution on the grade level of the Tweet, but these effects are sensitive to specification and significant for certain groups of users. These results show that there are statistically significant infra-marginal effects of AQI on emotional measures that are separate from large shocks previously studied, giving an idea

that population-wide effects may contribute more significantly than previously considered to the total cost of air pollution. They also show that there are potential infra-marginal effects of AQI on writing ability that may emerge at higher levels of pollution or among certain populations. This identifies an additional lens in which we should evaluate our approach to determining air quality regulation, depending on the value a policymaker places on broad cognition and writing abilities.

This paper adds to a literature primarily studying rare and costly output variables, presenting a more complete picture on the effects of air pollution. My measure of increased profanity, interpreted as increased aggressiveness, offers the smaller-scale version of the idea that crime increases with air pollution (Burkhardt et al., 2019). My measures of decreased sentiment similarly corroborates infra-marginally the relationship between air pollution and mental illness (Chen et al., 2018). Lastly, my puzzling, but statistically significant results on grade level may help explain discrepancies in pollution and cognition work, where papers have both shown no and large effects of contemporaneous pollution. I offer data that suggests these mixed impacts are a result of the dosage size of the treatment.

Furthermore, I have expanded and am continuing to expand the scope of the research by examining what types of Tweets, and therefore individuals, are most susceptible to the cognitive shocks we find in the first section of the paper. By taking full advantage of the microdata, I can identify heterogeneity amongst high cognitive scoring and low cognitive scoring Tweeters, and also understand compositional changes in the types of Tweets being sent. A final small thing I want to potentially explore given time constraints include interpolation of missing air pollution data outlined in Zivin and Neidell. Interpolation will allow my main results to have additional power and almost triple my number of main-result observations, given measurement errors do not become too large.

This paper also identifies some important future work to be done on the infra-marginal effects of contemporaneous air pollution. First, expanding the work of this paper in different countries is important in order to better understand how different cultures, infrastructures,

and economies heterogeneously react to air pollution. An additional strength of studying pollution in alternative contexts is that it expands the range of AQIs that there exists data, which is necessary to examine what nonlinearities in the dose-response function look like at high levels of pollution. Another important contribution to be made is structurally identifying a true Average Treatment on the Treated effect. For policy, it is not enough to know that air pollution has broad effects across populations emotionally and cognitively, but these effects need to be estimated precisely and translated into true economic costs to determine an approximate value to use in policy calculations. Finally, it may be interesting to exploit alternative cognitive measures to determine if the null results shown on cognitive outcomes are robust. I am particularly excited about the possibility of more advanced artificial intelligence to determine the vocabulary level, amount of grammar mistakes, or tone of different passages in microblogging contexts. Utilizing the most up-to-date natural language models can allow for deeper insights into how pollution exposure affects language generation.

Lastly, some of these forthcoming methods, such as examining heterogeneity, used to estimate the consequences of air pollution are novel to this paper and could be useful in other contexts. A benefit of Twitter data is that it acts as real-time, location-specific survey data, as long as you can tease out outcomes of interest using natural language processing tools on microblogging posts. I exploit recent advancements in NLP research to identify my outcome variables. If an outcome of interest is able to be inferred from Tweets from a specific CBSA-day, it can be measured in real-time and with previously unattainable levels of granularity.

Tables

Table 1: Summary statistics for Tweet Outcomes

	Compound	Negative	Grade	Curse	AQI
Count	139808	139808	139808	139808	139808
Mean	9.551	5.395	7.337	6.583	33.3379
Std. Dev	14.961	5.232	3.284	9.902	15.1075
AQI $\leq$ 20	**9.814**	**5.274**	**7.385**	**6.200**	
20 $\leq$ AQI $<$ 40	**9.587**	**5.356**	**7.348**	**6.494**	
40 $\leq$ AQI $<$ 60	**9.289**	**5.543**	**7.282**	**7.011**	
SD within CBSAs	13.459	4.576	2.981	8.346	12.2875
SD between CBSAs	6.345	1.983	1.552	3.995	8.7315

Note: One unit of observation refers to a CBSA-day, where the variables are all averages taken over all Tweets in a unique CBSA-day. Rows AQI $<$ # refer to the mean of the variable when restricting the sample to AQI's within certain ranges.

Table 2: Linear Effects of AQI on Cognitive and Emotional Outcomes

	Compound (1)	Negative (2)	Grade Level (3)	Curse % (4)
AQI	$-.0070^{***}$ (0.0016)	0.0039^{***} (0.0006)	0.0005 (0.0004)	$.0041^{***}$ (0.0011)
Max Temperature (C)	-0.0297^{***} (0.0047)	0.0062^{***} (0.0017)	-0.0020 (0.0013)	0.0085^{***} (0.0032)
Precipitation (in)	≈ 0 (0.0001)	≈ 0 (≈ 0)	≈ 0 (≈ 0)q	≈ 0 (≈ 0)
CBSA fixed effects	Yes	Yes	Yes	Yes
Time fixed effects	Yes	Yes	Yes	Yes
Standardized Regression Coefficient	-0.0131	0.0245	0.0115	0.0125
CBSA-days	111242	111242	111242	111242
Adjusted R^2	0.0043	0.0019	0.0021	0.0006

Note: One unit of observation refers to a CBSA-day, where the variables are all averages taken over all Tweets in a unique CBSA-day. Compound is the average VADER compound sentiment, Negative is the average VADER negative sentiment, Grade Level is the Flesch-Kincaid Grade Level. Cursing is the percentage of Tweets containing a curse word. Estimates equation (1) using TWFE approach. Standardized Regression Coefficients are βs multiplied by the AQI standard deviation divided by the outcome standard deviation. Standard errors are unclustered.

Table 3: Nonlinear Responses to AQI

	Compound (1)	Negative (2)	Grade Level (3)	Curse % (4)
Daily AQI Estimate A				
$A \in [0, 10)$	0.2719	-0.0949	0.0148	-0.2159
	(0.2494)	(0.0769)	(0.0)	(0.1479)
$A \in [10, 20)$	0.1745^{*}	-0.801^{**}	0.0201	-0.0897
	(0.0905)	(0.0353)	(0.0297)	(0.0592)
$A \in [30, 40)$	-0.0177	0.0369^{*}	0.0045	0.0401
	(0.5014)	(0.0218)	(0.0207)	(0.0409)
$A \in [40, 50)$	-0.1675^{***}	0.0522^{**}	0.0204	0.0174
	(0.0571)	(0.0223)	(0.0261)	(0.0486)
$A \in [50, 60)$	-0.0816	0.0750^{*}	0.0826^{*}	0.0982
	(0.1006)	(0.0412)	(0.0474)	(0.0877)
$A \in [60, 70)$	-0.3328^{**}	0.2451^{***}	-0.1924^{*}	0.20^{*}
	(0.1513)	(0.0539)	(0.0436)	(0.1082)
$A \in [70, 80)$	-0.3325	0.2442^{***}	0.0247	0.2751^{*}
	(0.2369)	(0.0866)	(0.0965)	(0.1647)
$A \geq 80$	-0.1834	0.1365	-0.1392^{**}	0.1217
	(0.2286)	(0.0935)	(0.0965)	(0.1404)
Weather Controls	Yes	Yes	Yes	Yes
CBSA fixed effects	Yes	Yes	Yes	Yes
Time fixed effects	Yes	Yes	Yes	Yes
CBSA-days	111242	111242	111242	111242
Adjusted R^2	0.0046	0.0021	0.0028	0.0006
F-Test on Bins	23.2933^{***}	37.3674^{***}	14.3051^{*}	11.3558

Note: One unit of observation refers to a CBSA-day, where the variables are all averages taken over all Tweets in a unique CBSA-day. Compound is the average VADER compound sentiment, Negative is the average VADER negative sentiment, Grade Level is the Flesch-Kincaid Grade Level. Cursing is the percentage of Tweets containing a curse word. Estimates equation (2) using TWFE approach. Omitted bin is $A \in [20, 30)$.

Figures

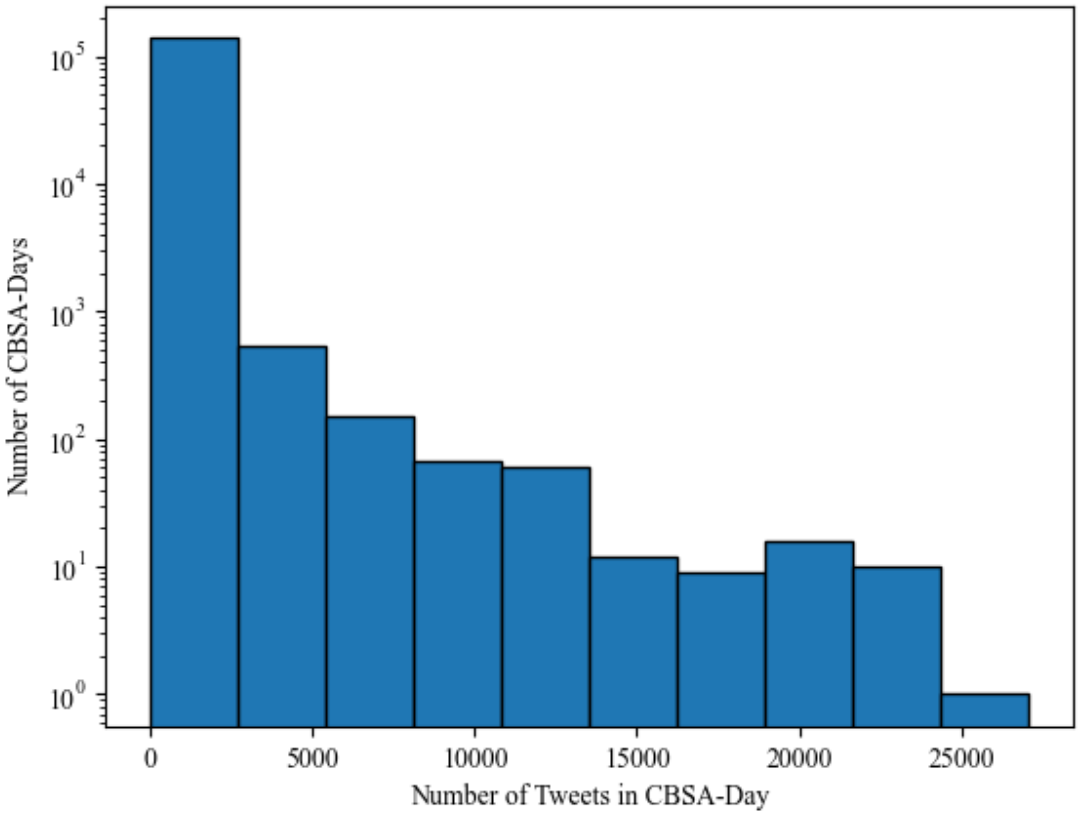

Figure 1: Number of Tweets across Units of Observation

Note: This figure plots a histogram of the Number of Tweets for each CBSA-Day (unit of observation). The y-axis represents the count of observations which are contained in the bins of number of Tweets on the x-axis.

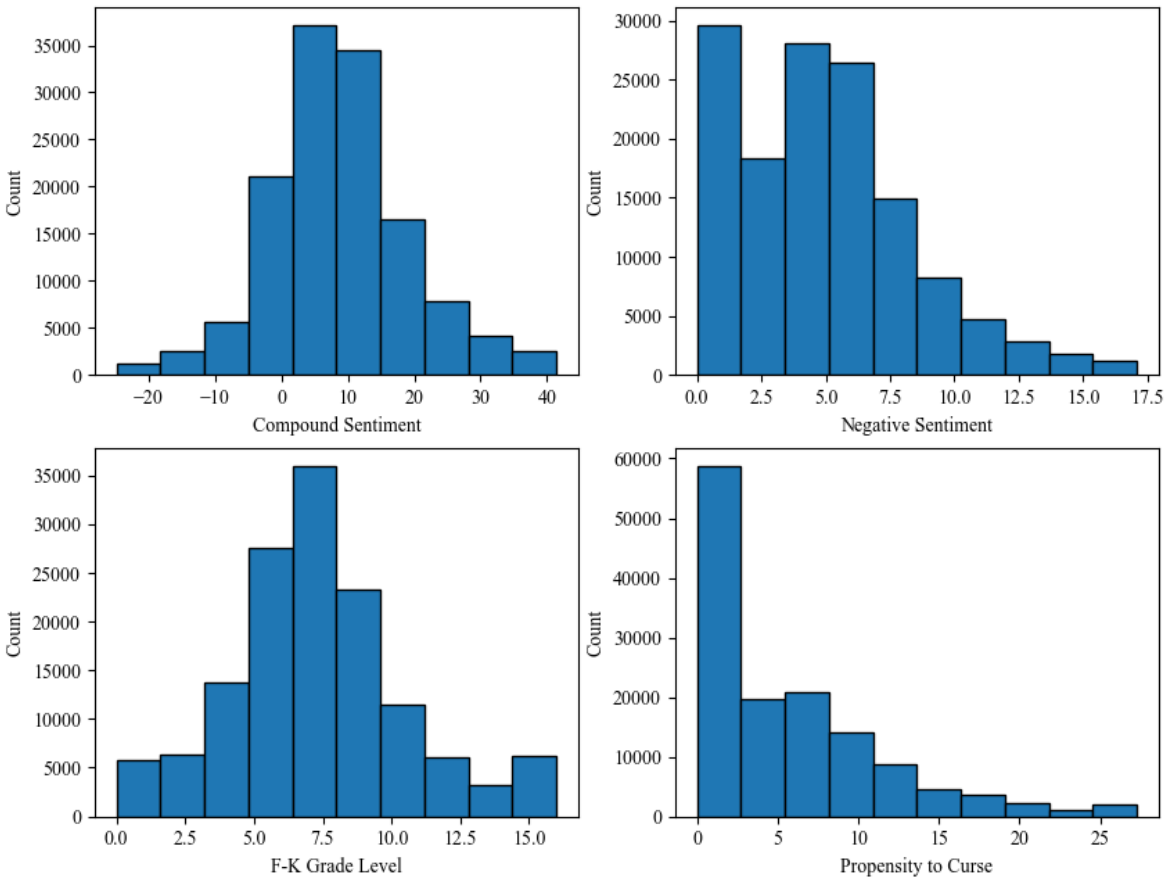

Figure 2: Distribution of Outcome Variables across Units of Observation

Note: This figure plots the distributions for our outcome variables of interest. Outliers are removed from the histograms for visibility. The y-axis represents the number of observations whose variable belongs to the bin it is placed in.

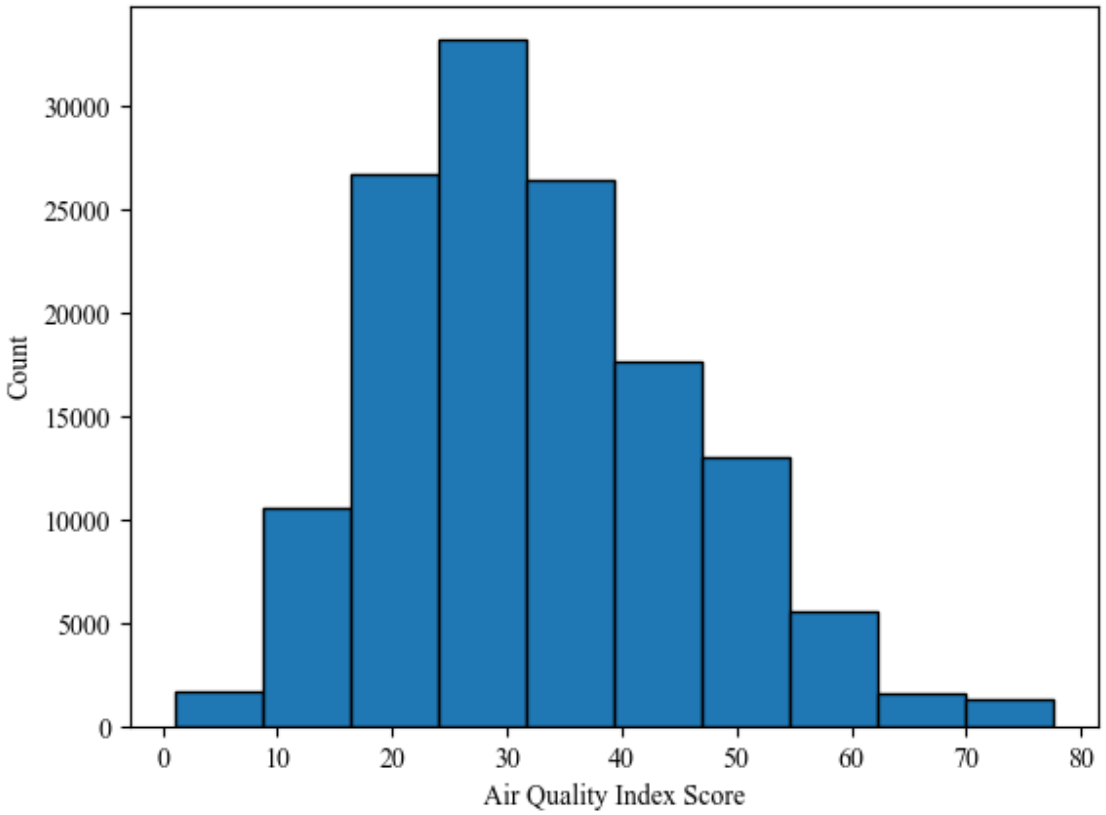

Figure 3: Distribution of AQI across Units of Observation

Note: This figure plots the distributions for the AQI score. Outliers are removed from the histograms for visibility. The y-axis represents the number of observations whose AQI belongs to the bin it is placed in.

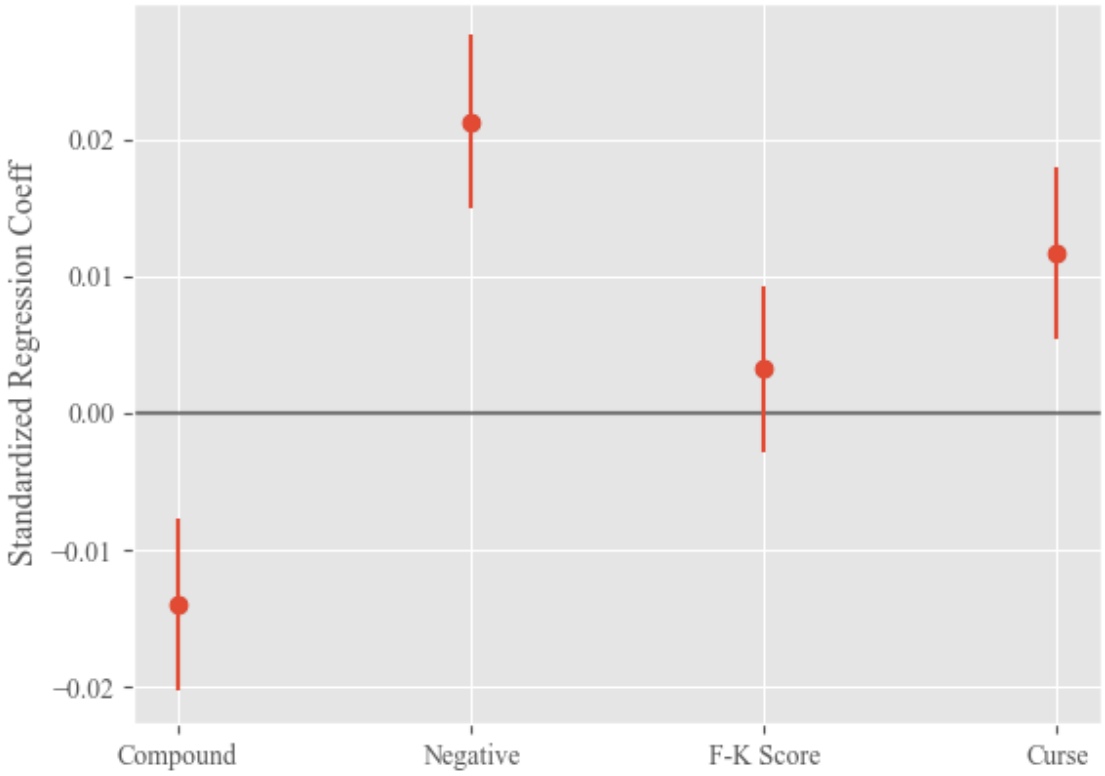

Figure 4: Linear Responses to AQI Increase

Note: Each column represents a standardized TWFE coefficient for AQI on one of the outcome variables of interest. Compound and Negative are the VADER Compound and Negative Sentiment scores, respectively. F-K Score shows the effect on Flesch-Kinkaid Grade Level. Curse shows the effect on the fraction of Tweets in a CBSA-day that contain curse words. Each regression controls for temperature and precipitation, and includes fixed effects for CBSA, day of week, holiday. Error bars are 95% confidence intervals.

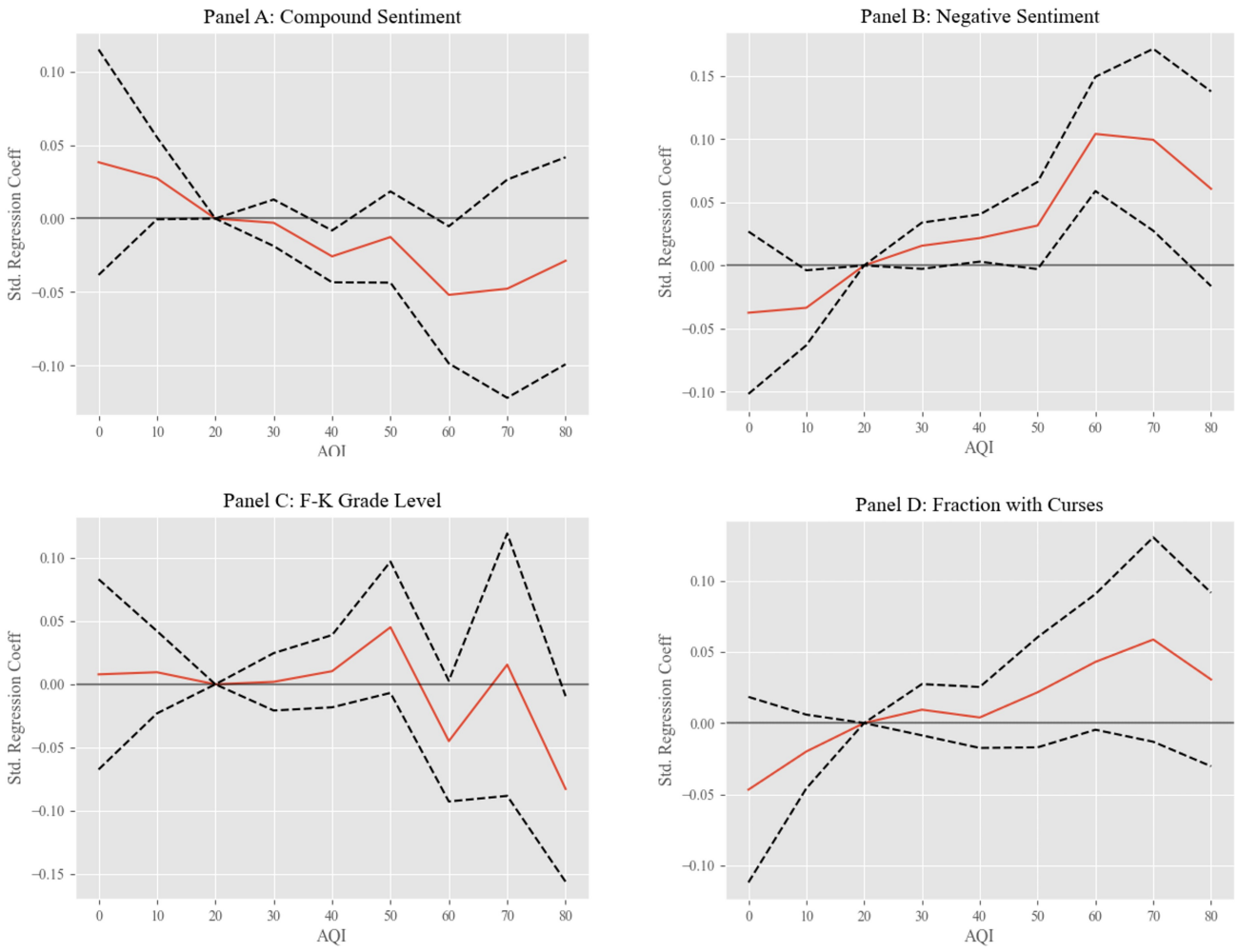

Figure 5: Nonlinear Responses to AQI

Note: Panel A, B show standardized TWFE coefficients for AQI on VADER Compound Sentiment, Negative Sentiment, respectively. Panel C shows standardized TWFE coefficients for AQI on F-K Grade Level. Panel D shows standardized TWFE coefficients for AQI on the fraction of Tweets that contain curse words. Solid lines represent the regression coefficients on AQI and represent the difference (measured in standard deviations) in CBSA-date sentiment for the AQI bin $A \in [a, b)$ relative to $[20, 30]$, controlling for temperature and precipitation and with fixed effects for CBSA, day of week, holiday. Dotted lines are 95% confidence intervals. Standard errors are unclustered.

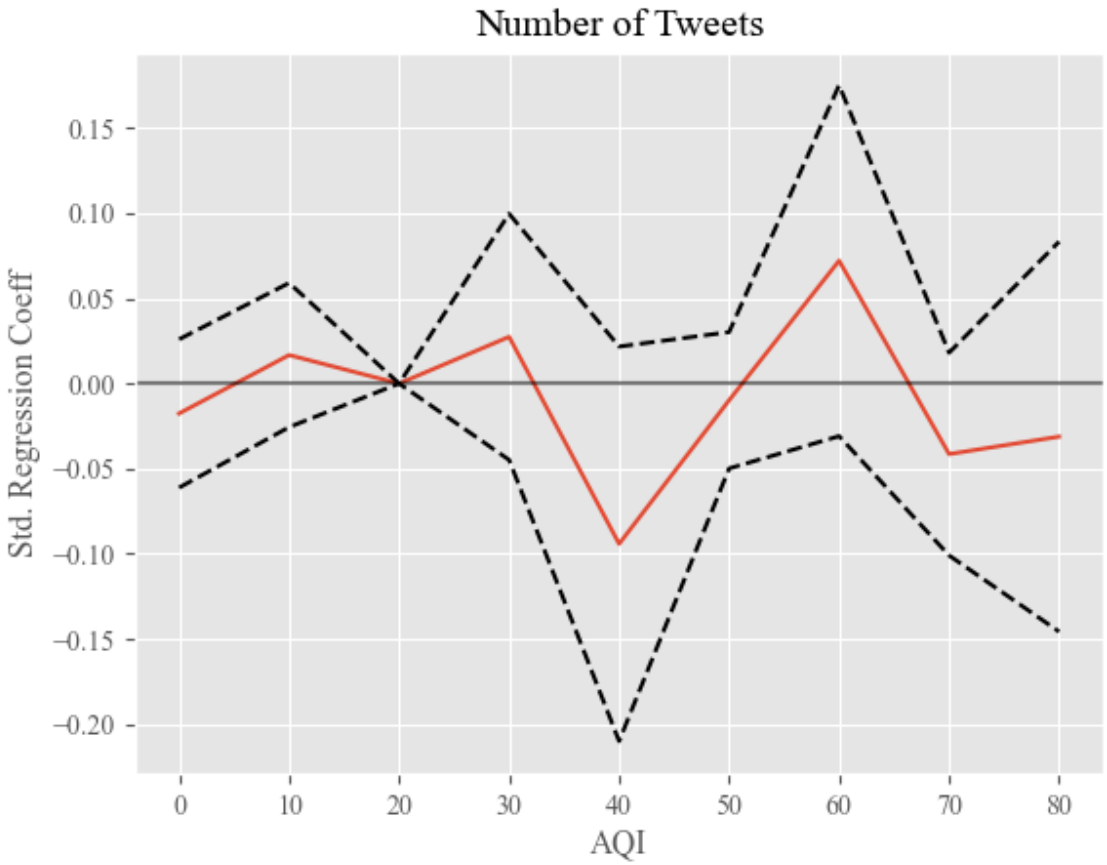

Figure 6: Selection into Sample–Number of Tweets

Note: Nonlinear effect of AQI on number of Tweets in CBSA-day. Error bars are 95% confidence bounds. Coefficients are estimated using TWFE model with controls for temperature, precipitation, day-of-week, holidays, and User.

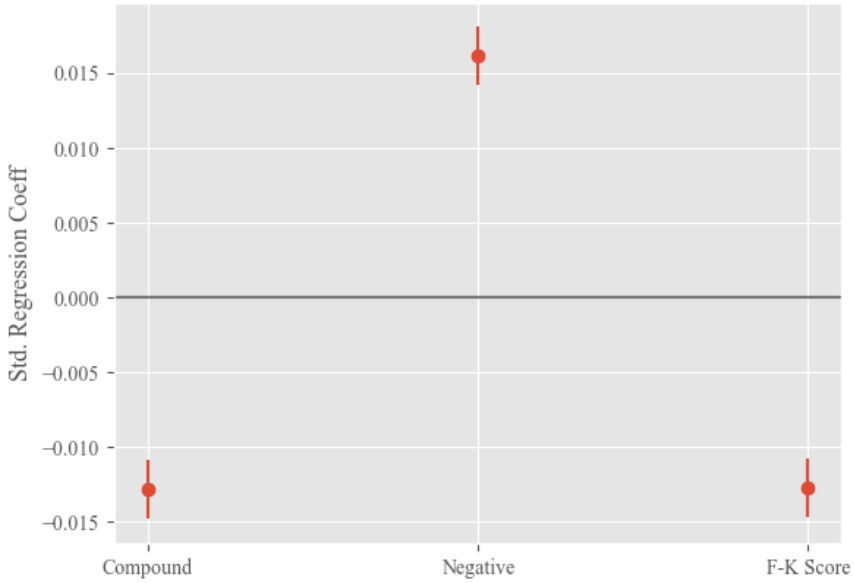

Figure 7: Selection into Sample–User FE

Note: Linear effect of AQI on Compound Sentiment, Negative Sentiment, and Flesch-Kincaid Grade Level. Mean is from the main results and is the mean outcome of all Tweets in the CBSA-day. Error bars are 95% confidence bounds. Coefficients are estimated using TWFE model with controls for temperature, precipitation, day-of-week, holidays, and User. Users who Tweet across 10 or more days, but have less than 100 total Tweets across the data are included for computational practicality.

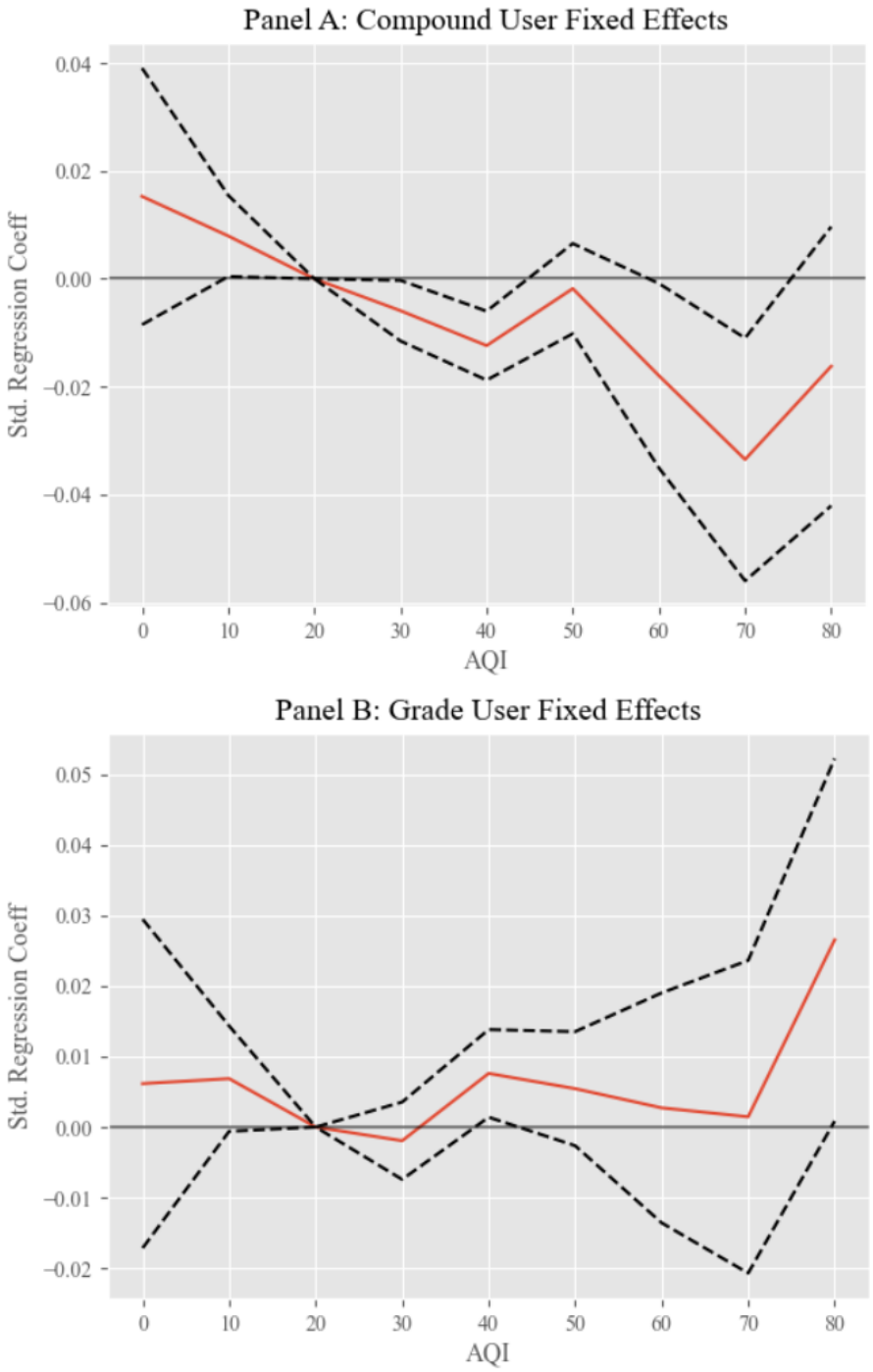

Figure 8: Selection into Sample–User FE Nonlinear

Note: Linear effect of AQI on Compound Sentiment and Flesch-Kincaid Grade Level. Mean is from the main results and is the mean outcome of all Tweets in the CBSA-day. Error bars are 95% confidence bounds. Coefficients are estimated using TWFE model with controls for temperature, precipitation, day-of-week, holidays, and User. Users who Tweet across 10 or more days, but have less than 100 total Tweets across the data are included for computational practicality.

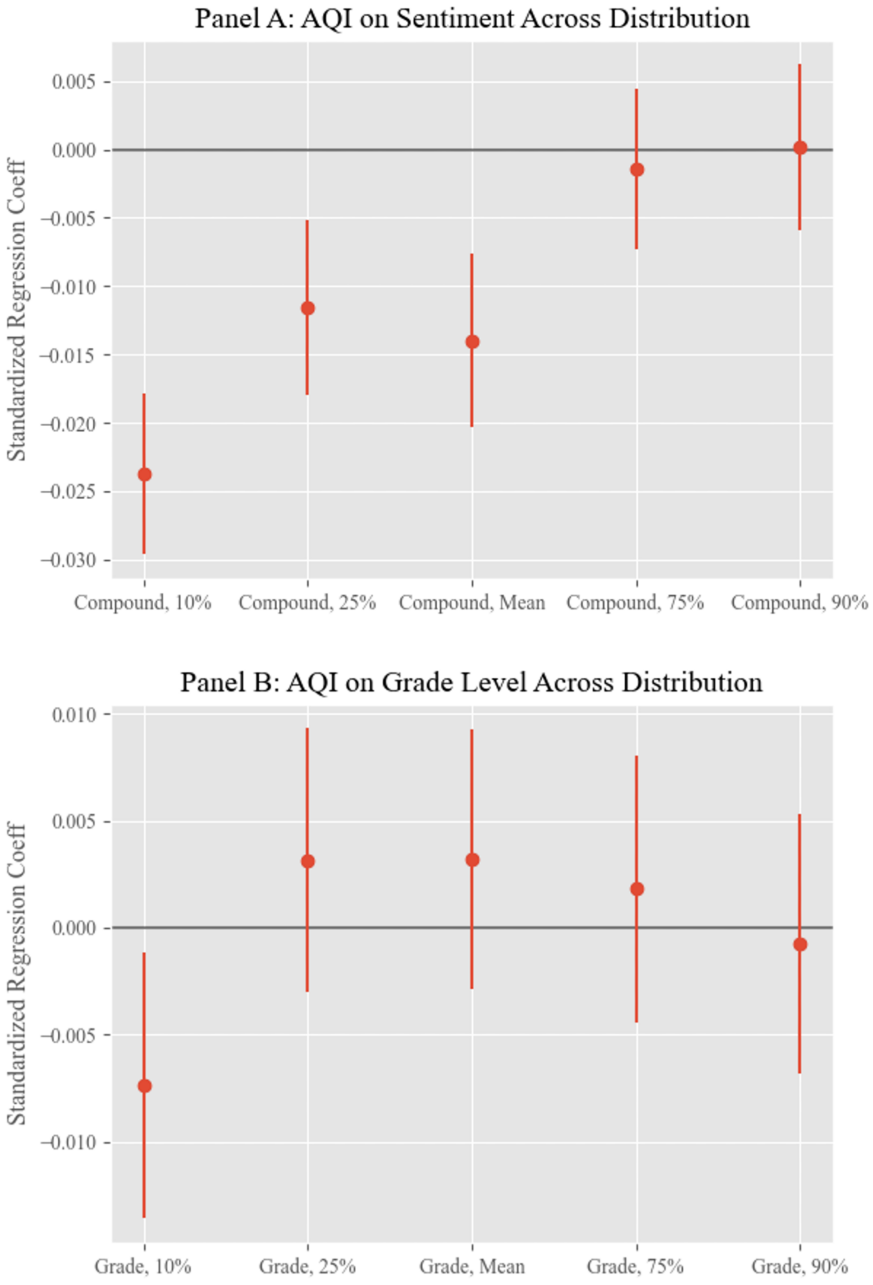

Figure 9: Heterogenous Effects

Note: Effect of AQI on Compound Sentiment and Flesch-Kincaid Grade Level. Each column represents a different measure of the outcome variable. Mean is from the main results and is the mean outcome of all Tweets in the CBSA-day, while other columns are various percentiles. Error bars are 95% confidence bounds. Coefficients are estimated using TWFE model with controls for temperature, precipitation, day-of-week, holidays, and CBSA.

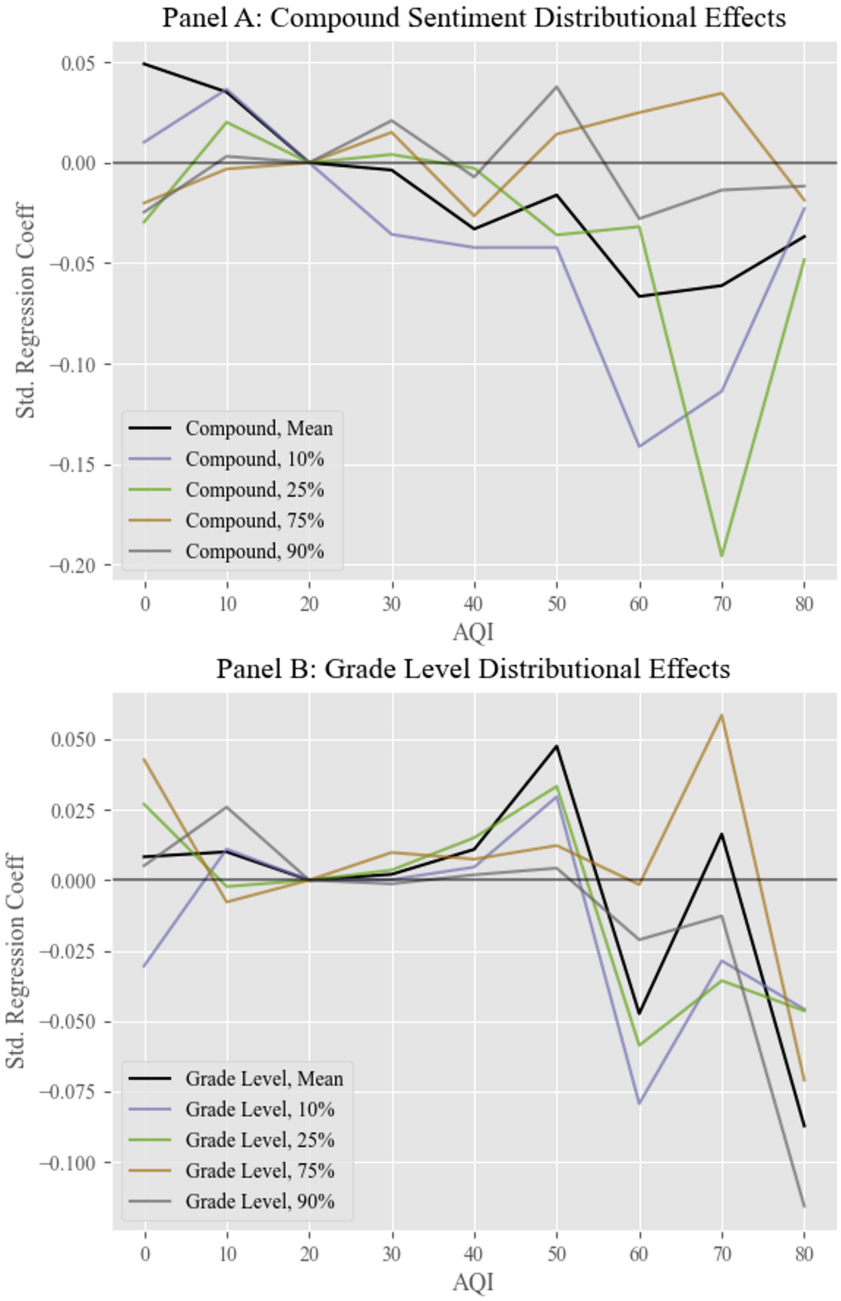

Figure 10: Heterogeneous Effects–Nonlinear

Note: Effect of AQI on Compound Sentiment (Panel A) and Flesch-Kincaid Grade Level (Panel B). Black line is result from main results and is the mean of all Tweets in the CBSA-day, while other lines are various percentiles of all Tweets in the CBSA-day, and capture the idea of understanding how the within-CBSA distribution of Tweets changes with AQI.

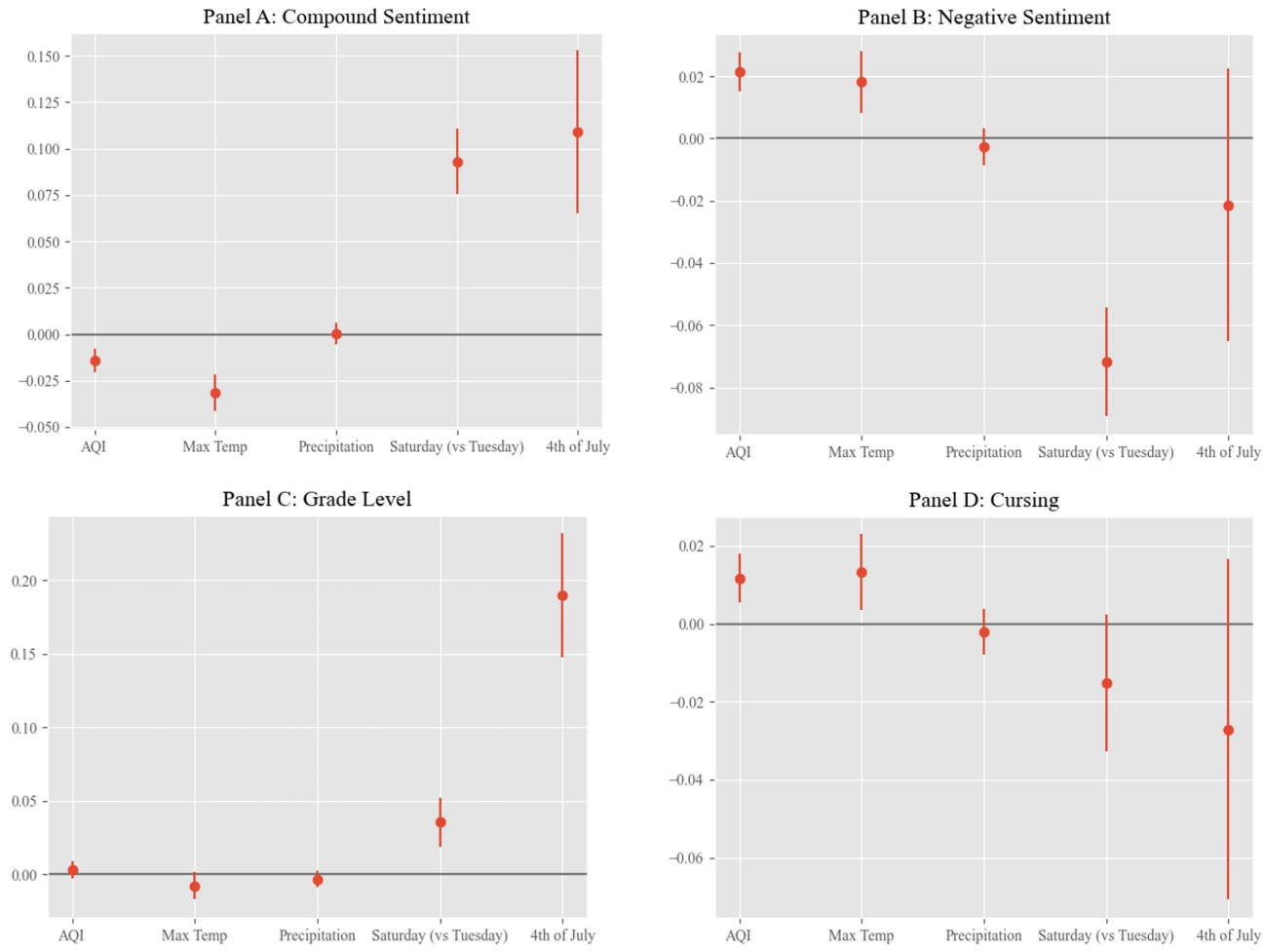

Figure 11: Comparison of Effects

Note: Effect of AQI on Compound Sentiment (Panel A), Negative Sentiment (Panel B), Flesch-Kincaid Grade Level (Panel C), Cursing Percentage (Panel D). The first column in each panel represents the standardized regression coefficient–the effect of a standard deviation increase of AQI in terms of standard deviations of the outcome variable. This represents the same estimates found in Figure 4. Max Temp and Precipitation measure the same standardized regression coefficient for the daily maximum temperature and the total precipitation. The following columns represents the effect of switching the day-of-week from a Tuesday to a Saturday and switching to the 4th of July holiday respectively, both effects measured in terms of standard deviations of the outcome variable.

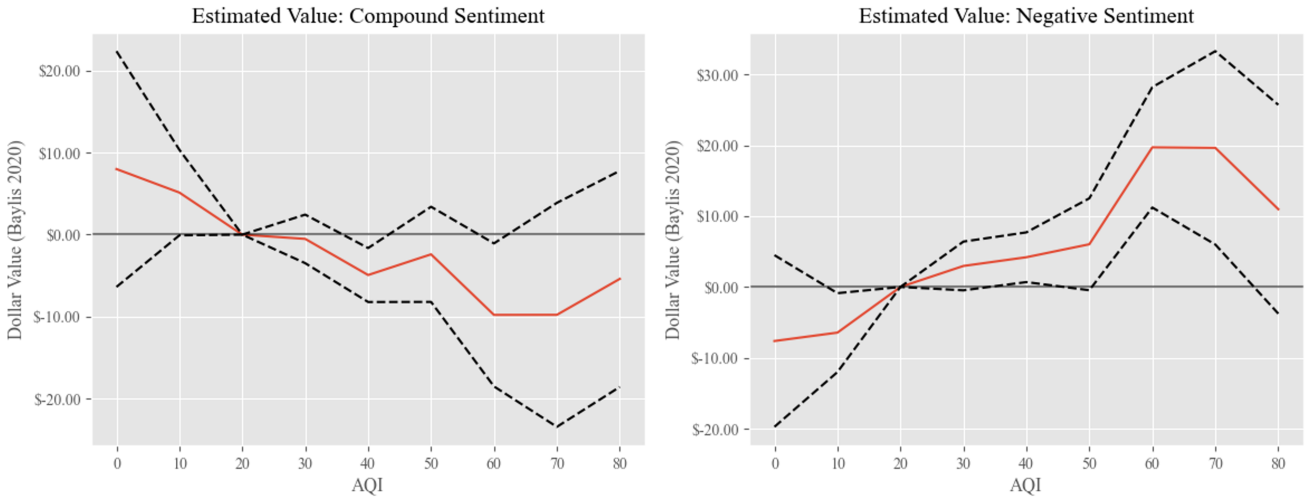

Figure 12: Estimated Dollar Costs

Note: Back-of-envelope estimation of the daily cost of AQI through Compound Sentiment (Panel A) and Negative Sentiment (Panel B). Orange lines represent how much a person is expected to pay in order to swap a single day with AQI $[20, 30]$ with a day of AQI $[a, b]$, black lines represent 95% confidence intervals. Value of a SD increase in sentiment for estimation is taken from Baylis (2020).

References

Opinion — economists' statement on carbon dividends, Jan 2019. URL https://www.wsj.com/articles/economists-statement-on-carbon-dividends-11547682910.

Better profanity. https://github.com/snguyenthanh/better_profanity, 2021. [Online; accessed Jan-2023].

"textstat", Mar 2022.

J Albright, C de Guzman, P Acebo, D Paiva, M Faulkner, and J Swanson. Readability of patient education materials: implications for clinical practice. *Appl. Nurs. Res.*, 9(3): 139–143, August 1996.

James Archsmith, Anthony Heyes, and Soodeh Saberian. Air quality and error quantity: Pollution and performance in a high-skilled, quality-focused occupation, Oct 2018. URL http://dx.doi.org/10.1086/698728.

Sameer Badarudeen and Sanjeev Sabharwal. Assessing readability of patient education materials: current role in orthopaedics. *Clin. Orthop. Relat. Res.*, 468(10):2572–2580, October 2010.

Patrick Baylis. Temperature and temperament: Evidence from twitter, Apr 2020. URL http://dx.doi.org/10.1016/j.jpubeco.2020.104161.

Francesco Bianchi, Roberto Gomez Cram, and Howard Kung. Using social media to identify the effects of congressional partisanship on asset prices, 2021. URL http://dx.doi.org/10.2139/ssrn.3823756.

Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with python*. O'Reilly, 2009.

Isobel Braithwaite, Shuo Zhang, James B. Kirkbride, David P. J. Osborn, and Joseph F. Hayes. Air pollution (particulate matter) exposure and associations with depression, anxiety, bipolar, psychosis and suicide risk: A systematic review and meta-analysis, Dec 2019. URL http://dx.doi.org/10.1289/EHP4595.

Savannah Brown. what guys look for in girls, Jan 2015. URL https://www.youtube.com/watch?v=N-ezbk0AOCI.

Jesse Burkhardt, Jude Bayham, Ander Wilson, Ellison Carter, Jesse D. Berman, Katelyn O'Dell, Bonne Ford, Emily V. Fischer, and Jeffrey R. Pierce. The effect of pollution on crime: Evidence from data on particulate matter and ozone, Nov 2019. URL http://dx.doi.org/10.1016/j.jeem.2019.102267.

Lilian Calderón-Garcidueñas, Ana Calderón-Garcidueñas, Ricardo Torres-Jardón, José Avila-Ramírez, Randy J. Kulesza, and Amedeo D. Angiulli. Air pollution and your brain: what do you need to know right now, Sep 2014. URL http://dx.doi.org/10.1017/S146342361400036X.

Shuai Chen, Paulina Oliva, and Peng Zhang. Air Pollution and Mental Health: Evidence from China. Nber working papers, National Bureau of Economic Research, Inc, June 2018. URL https://ideas.repec.org/p/nbr/nberwo/24686.html.

Wen-Yi Chen and Mei-Ping Chen. Twitter's daily happiness sentiment, economic policy uncertainty, and stock index fluctuations. *The North American Journal of Economics and Finance*, 62(C), 2022. doi: 10.1016/j.najef.2022.1017. URL https://ideas.repec.org/a/eee/ecofin/v62y2022ics1062940822001243.html.

M E Cooley, H Moriarty, M S Berger, D Selm-Orr, B Coyle, and T Short. Patient literacy and the readability of written cancer educational materials. *Oncol. Nurs. Forum*, 22(9): 1345–1351, October 1995.

Douglas W. Dockery, C. Arden Pope, Xiping Xu, John D. Spengler, James H. Ware, Martha E. Fay, Jr. Ferris, Benjamin G., and Frank E. Speizer. An association between air pollution and mortality in six u.s. cities, Dec 1993. URL http://dx.doi.org/10.1056/NEJM199312093292401.

Matthew Gentzkow, Bryan Kelly, and Matt Taddy. Text as data, Sep 2019. URL http://dx.doi.org/10.1257/jel.20181020.

John Green. The agricultural revolution: Crash course world history 1, Jan 2012. URL https://www.youtube.com/watch?v=Yocja_N5s1I&list=PL8dPuuaLjXtOJN5C_ex76lbX0uVsKDgGf&index=1.

Rema Hanna and Paulina Oliva. The effect of pollution on labor supply: Evidence from a natural experiment in mexico city, Feb 2015. URL http://dx.doi.org/10.1016/j.jpubeco.2014.10.004.

Joshua Roesslein "Harmon and Other Contributors". "tweepy", 2022. URL 'https://www.tweepy.org/'.

James J. Heckman. Sample selection bias as a specification error, Jan 1979. URL http://dx.doi.org/10.2307/1912352.

S. E. Hinton. *The outsiders*. Viking, an imprint of Penguin Random House, 1967.

C. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text, May 2014. URL http://dx.doi.org/10.1609/icwsm.v8i1.14550.

Timothy Jay Janschewitz and Kristin. "the science of swearing", Apr 2012. URL https://www.psychologicalscience.org/observer/the-science-of-swearing.

J. Peter Kincaid, Jr. Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel, 1979. URL https://stars.library.ucf.edu/istlibrary/56/.

Christian Lamprecht. Meteostat: The weather's record keeper, Jan 2023. URL https://meteostat.net/en/.

Ioannis Manisalidis, Elisavet Stavropoulou, Agathangelos Stavropoulos, and Eugenia Bezirtzoglou. Environmental and health impacts of air pollution: A review, Feb 2020. URL http://dx.doi.org/10.3389/fpubh.2020.00014.

A. A. Milne. *Winnie the Pooh*. Five Mile Press, 1926.

Tamar Mitts. Countering violent extremism and radical rhetoric, May 2021. URL http://dx.doi.org/10.1017/s0020818321000242.

Brennan Lee Mulligan. Yes or no - season 2, 2021. URL https://www.dropout.tv/game-changer/season:2/videos/yes-or-no.

III Pope, C. Arden, Majid Ezzati, and Douglas W. Dockery. Fine-particulate air pollution and life expectancy in the united states, Jan 2009. URL http://dx.doi.org/10.1056/NEJMsa0805646.

Salman Rushdie. *Midnight's children*. Vintage, 1981.

Wolfram Schlenker and W. Reed Walker. Airports, air pollution, and contemporaneous health, Oct 2015. URL http://dx.doi.org/10.1093/restud/rdv043.

Seuss. *Oh, the places you'll go!* HarperCollins Children's Books, 1990.

Mieczyslaw Szyszkowicz, Brian Rowe, and Ian Colman. Air pollution and daily emergency department visits for depression, Jan 2009. URL http://dx.doi.org/10.2478/v10001-009-0031-6.

Tariq Luqmaan Trotter. Black thought freestyles on flex — freestyle087, Dec 2017. URL https://www.youtube.com/watch?v=prmQgSpV3fA.

month=Jan United States Environmental Protection Agency, year=2023. Air data: Air quality data collected at outdoor monitors across the us. URL https://www.epa.gov/outdoor-air-quality-data.

Mohammad Javad Zare Sakhvidi, Antoine Lafontaine, Emeline Lequy, Claudine Berr, Kees de Hoogh, Danielle Vienneau, Marcel Goldberg, Marie Zins, Cédric Lemogne, and Bénédicte Jacquemin. Ambient air pollution exposure and depressive symptoms: Findings from the french constances cohort, Dec 2022. URL http://dx.doi.org/10.1016/j.envint.2022.107622.

Joshua Graff Zivin and Matthew Neidell. The impact of pollution on worker productivity, Dec 2012. URL http://dx.doi.org/10.1257/aer.102.7.3652.

Appendix A.

7.1 Construction of Emotional and Cognitive Outcome Variables

7.1.1 Twitter Streaming API Client

In this section, I discuss my process for obtaining the Twitter data used in this paper. The simplest way to collect Twitter data would be to purchase data from Twitter directly. However, this is unfeasible for my case, as Twitter data is sold as a per-Tweet expense, so collecting a complete sample of Tweets over a significant time period would be overly costly. As a result, I utilize the Tweet streaming functionalities from the Twitter API v1.1 through the Tweepy python package ("Harmon and Contributors", 2022). Since I utilize the most up-to-date Tweepy library (and the Twitter v1.1 API), my process differs from Baylis (2020), despite ending with the same set of data.

As of March 2023, Twitter is in the process of migrating to a new academic research API client and policies, so specific details of Twitter's Streaming Client may change. In order to get access to streaming Tweets, I first registered for API credentials through the Twitter Developers Portal.[7]. I did not use the academic research-specific license[8], so once I created a new project in accordance with Twitter licensing and with Twitter approval, I was granted a set of keys to be used to access the API in my streaming code. I used these keys to write a Python script I wrote in order to build my dataset. The script can be found on my GitHub repository for the paper here, which will work out-of-the-box after you put in your own keys. My script throws out extraneous data to save storage costs, and it saves key information like the id of the Tweet (for later access), the text itself, the user ID, the location, and the timestamp. I was able to collect Tweets without ever reaching Twitter's monthly cap on accessing Tweets, which is a function of geolocated Tweets being a small subsection of all Tweets.

Since the only way to create a dataset of Tweets without paying Twitter[9] for data is by streaming Tweets, it is necessary that the script is running during the duration of sample collection. As a result, I set my script to run on an Amazon Web Services EC2 instance– essentially a computer that can run code nonstop and store the results. At the end of my sample period, I transferred the results back to my machine for data analysis. There were

[7]Found here: https://developer.twitter.com/en/portal/dashboard

[8]I was actually rejected from it! Twitter's academic research agreements are only available to full-time researchers, master's, and Ph.D. students. Luckily, the general account had loose enough restrictions that I was able to do the work for this paper without the need for academic access.

[9]Twitter data is generally expensive and is meant for enterprise use around specific events. The example use-case on their website was a company gauging feedback by downloading all Tweets about their Superbowl ad in a 24h period.

some periods of downtime related to my EC2 instance temporarily running out of storage [10], but these periods are independent of Twitter and the date/time, so they are plausibly random.

7.1.2 VADER Sentiment Measurements

After creating a dataset of every Tweet during a sample period, I aggregate my Tweets into CBSA-days in order to compress the data without removing any information. On each CBSA-day, I calculate a variety of statistics using a script that counts the number of Tweets alongside both the mean outcome measures and the distribution of outcome measures. I've posted the script on my GitHub here, which contains a function that determines all the statistics at a CBSA-day level such that it is fully parallelizable. [11]

Sentiment analysis with VADER is considered a standard and consistently cited method for sentiment analysis, pioneered by the work of Hutto and Gilbert (2014). The benefit of VADER is that it is both accurate and relatively efficient as a result of its simple architecture. While the most accurate sentiment analysis tools involve large language models[12] or transformer models that rely on artificial intelligence networks, these more precise tools sacrifice efficiency in exchange for accuracy. Instead, VADER is strictly a rules-based process, where words are scored individually and key grammatical structures are noticed in order to determine if the meaning of those words are emphasized or de-emphasized. As a result, it is one of the best scoring methods on sample datasets for its speed. I use the Python Natural Language ToolKit (NLTK) implementation, which is the most well-used implementation of VADER in Python. It includes a compound score, ranging between [-1,1] with -1 being the most negative, and 1 being the most positive (and 0 being neutral). It also includes scores for negativity and positivity from [0,1], of which I use the negativity score where a score of 1 is the most negative. The following table presents some sample passages and their scores for the VADER compound score and VADER negative score using the same NLTK implementation.

[10] EC2 instances are much cheaper than purchasing data from Twitter, but they are not costless, and I am an undergraduate student with a $0 research budget. As a result, I purchased less storage than what ultimately was needed, creating cutoffs for hours/days a few times during the data collection process before I was able to upgrade my instance.

[11] Without parallel computing, a single-core processor working through 10 Tweets a second would still mean 35 days to process my dataset of 30M Tweets! On my machine, I found that 8 cores in parallel would work through the dataset in 2-3 days uninterrupted.

[12] These have been in the news as of late, and it's very possible to see a machine like GPT-4 outperform VADER in sentiment analysis.

Sample Passage	VADER Compound Score	VADER Negative Score
"Don't go from written bars filled with rage to primetime television and your gilded cage. Then forget there's people in the world still enslaved. I barb-wired my wrist and let it fill the page"	-0.802	0.215
"Well I'm NOT HAVING IT. I solved your labyrinth, puzzle master! The minotaur's escaped and you're gonna get the horns, buddy! I CAN NOT WIN"	-0.601	0.222
"It includes a compound score, ranging between [-1,1] with -1 being the most negative, and 1 being the most positive (and 0 being neutral). It also includes scores for negativity and positivity from [0,1]"	-0.0258	0.176
"What day is it?', asked Winnie the Pooh. 'It's today,' squeaked Piglet. 'My favorite day,' said Pooh."	0.4588	0.0
"I mean the way your bright eyes go wild when you smile and how your laughter's so melodic it's a song. I mean the way your creativity's a compass that leads you to what you love."	0.8625	0.0

Table 4: VADER Sentiment Analysis Sample Scores

Note: Entries for this table taken in order from Trotter (2017), Mulligan (2021), this paper, Milne (1926), and Brown (2015). I enjoyed picking them out immensely.

7.1.3 Flesch-Kincaid Grade Level

To save computational time, I process the Flesch-Kincaid Grade Level of a Tweet at the same time and use the same script as when I process the sentiment of a Tweet. Since calculating the grade level using Flesch-Kincaid relies upon counting syllables in words, it is difficult to implement manually, so I use the TextStat Python implementation (tex, 2022). An added benefit of the TextStat implementation is functionality with multiple languages, including English, Spanish, German, Italian, and Russian, so the vast majority of Tweets are able to be processed correctly in their respective languages. If a language is not defined

within TextStat, I remove the observation from my dataset. The following are some example passages that are of similar length to Tweets and their corresponding Flesch-Kincaid grade levels according to TextStat.

Sample Passage	Flesch-Kincaid Grade Level (via TextStat)
"Memory's truth, because memory has its own special kind. It selects, eliminates, alters, exaggerates, minimizes, glorifies, and vilifies also; but in the end it creates its own reality, its heterogeneous but usually coherent version of events; and no sane human being ever trusts someone else's version more than his own."	14.2
"I study whether heightened air pollution leads to cognitive and emotional responses at the infra-marginal level."	11.9
"The test will last your entire life, and it will be comprised of the millions of decisions that, when taken together, will make your life yours."	9.9
"It seemed funny to me that the sunset she saw from her patio and the one I saw from the back steps was the same one. Maybe the two different worlds we lived in weren't so different. We saw the same sunset."	4.0
"You have brains in your head. You have feet in your shoes. You can steer yourself any direction you choose."	1.2

Table 5: Flesch-Kincaid Implementation Sample Scores

Note: Entries for this table taken in order from Rushdie (1981), this paper's abstract, Green (2012), Hinton (1967), and Seuss (1990). I enjoyed picking them out immensely.

7.1.4 Measuring Profanity Use

My last outcome that I measure in this paper revolves around profanity use, which I take as a measure of heightened emotion levels or adult content. To extract this measure from Tweets, I use the same methods built into my Python script for measuring sentiment and grade level. This script leverages the Better Profanity package, which is useful because it is able to extract whether a text contains profanity while accounting for common typos or alternative spellings (pro, 2021). A complete wordlist of all words counted can be found on

the better profanity GitHub here.

7.2 Sensitivity Checks for Main Results

7.2.1 Alternate Time Controls

One concern is that, in order to increase statistical power, I stray from a strict two-way-fixed-effects model into one that has looser time fixed-effects. I do this in the paper by removing day-of-sample fixed-effects and replacing them with day-of-week (and holiday) effects. This removes the assumption that all days that are the same day-of-week within a CBSA-day are similar. Here are the results which mirror Figure 5 using the stricter implementation that includes day-of-sample effects. It is clear that results are directionally similar, but lack the added power of loosening my assumption.

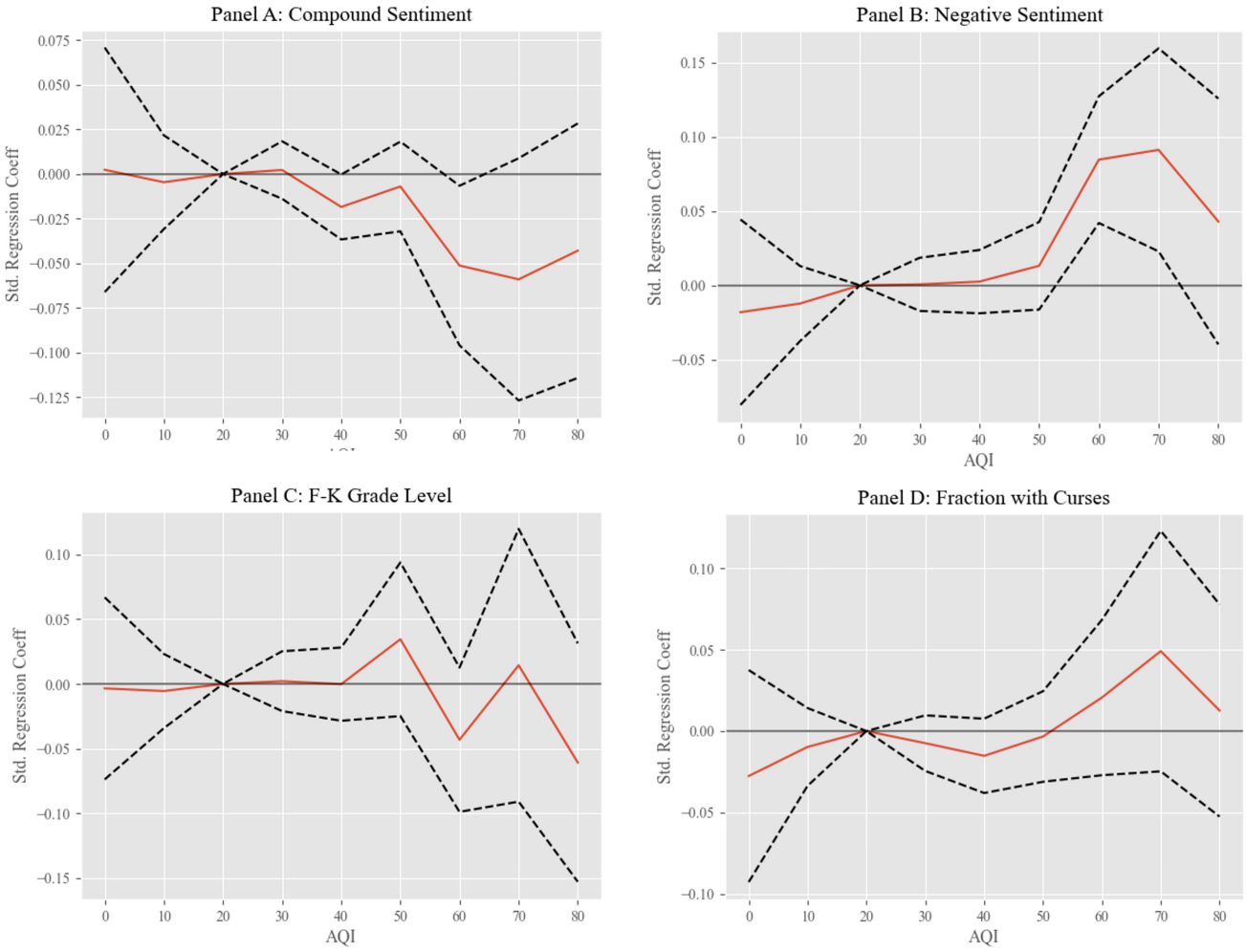

Figure 13: Nonlinear Responses to AQI

Note: Panel A, B show standardized TWFE coefficients for AQI on VADER Compound Sentiment, Negative Sentiment, respectively. Panel C shows standardized TWFE coefficients for AQI on F-K Grade Level. Panel D shows standardized TWFE coefficients for AQI on the fraction of Tweets that contain curse words. Solid lines represent the regression coefficients on AQI and represent the difference (measured in standard deviations) in CBSA-date sentiment for the AQI bin $A \in [a, b)$ relative to $[20, 30]$, controlling for maximum temperature and inches of precipitation and with fixed effects for CBSA and day of sample. Dotted lines are 95% confidence intervals. Standard errors are unclustered.

7.2.2 Alternative Temperature Controls

It is important that my results are robust to weather, which is one of the key potential confounding variables that can affect both pollution and sentiment. I state in Section 3 that my results are robust to various measurements of temperature. Here, I present results with

both minimum and average temperatures, whereas my main results were presented with maximum temperatures as controls.

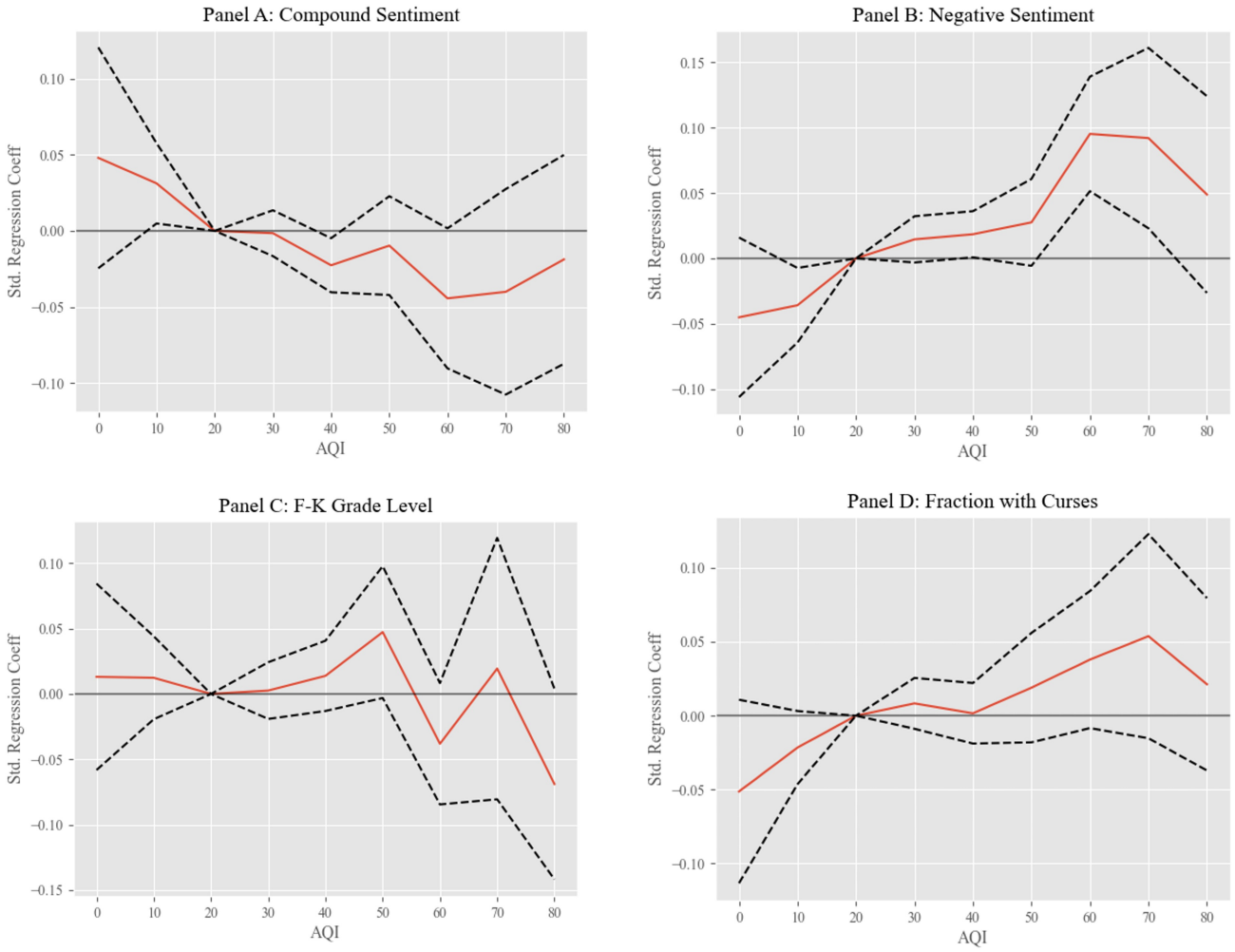

Figure 14: Nonlinear Responses to AQI

Note: Panel A, B show standardized TWFE coefficients for AQI on VADER Compound Sentiment, Negative Sentiment, respectively. Panel C shows standardized TWFE coefficients for AQI on F-K Grade Level. Panel D shows standardized TWFE coefficients for AQI on the fraction of Tweets that contain curse words. Solid lines represent the regression coefficients on AQI and represent the difference (measured in standard deviations) in CBSA-date sentiment for the AQI bin $A \in [a, b)$ relative to $[20, 30]$, controlling for minimum temperature and precipitation and with fixed effects for CBSA, day-of-week, and holiday. Dotted lines are 95% confidence intervals. Standard errors are unclustered.

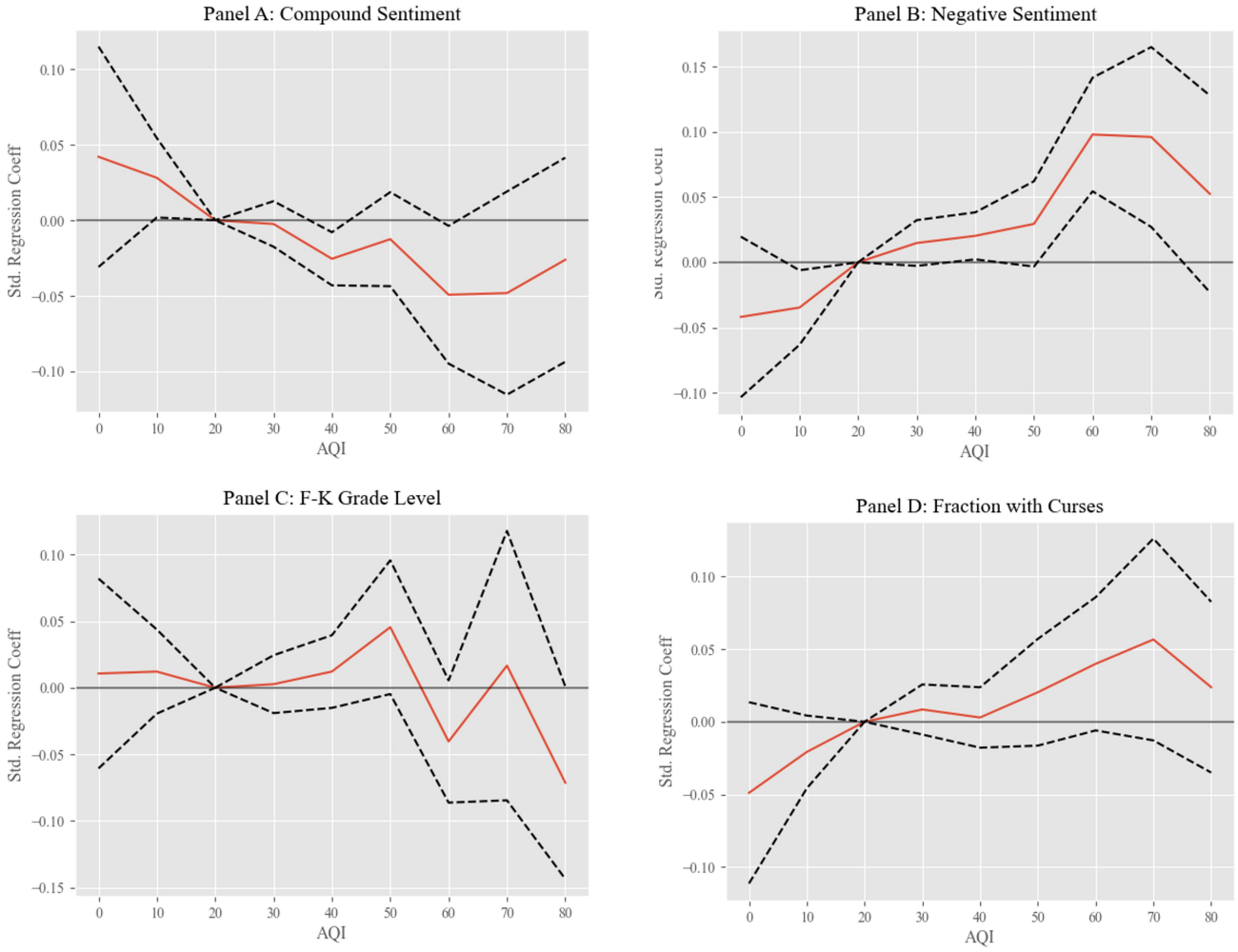

Figure 15: Nonlinear Responses to AQI

Note: Panel A, B show standardized TWFE coefficients for AQI on VADER Compound Sentiment, Negative Sentiment, respectively. Panel C shows standardized TWFE coefficients for AQI on F-K Grade Level. Panel D shows standardized TWFE coefficients for AQI on the fraction of Tweets that contain curse words. Solid lines represent the regression coefficients on AQI and represent the difference (measured in standard deviations) in CBSA-date sentiment for the AQI bin $A \in [a, b)$ relative to $[20, 30]$, controlling for average temperature and precipitation and with fixed effects for CBSA, day-of-week, and holiday. Dotted lines are 95% confidence intervals. Standard errors are unclustered.

7.2.3 Bin Width

One concern is that my bin width choices to estimate equation (2) are cherry-picked to make results appear that are not robust across various choices of bin widths. I present results with smaller bins of 5, and bigger bins of 20. My results shift in power, but are still significant

and similar in magnitude, alleviating this concern.

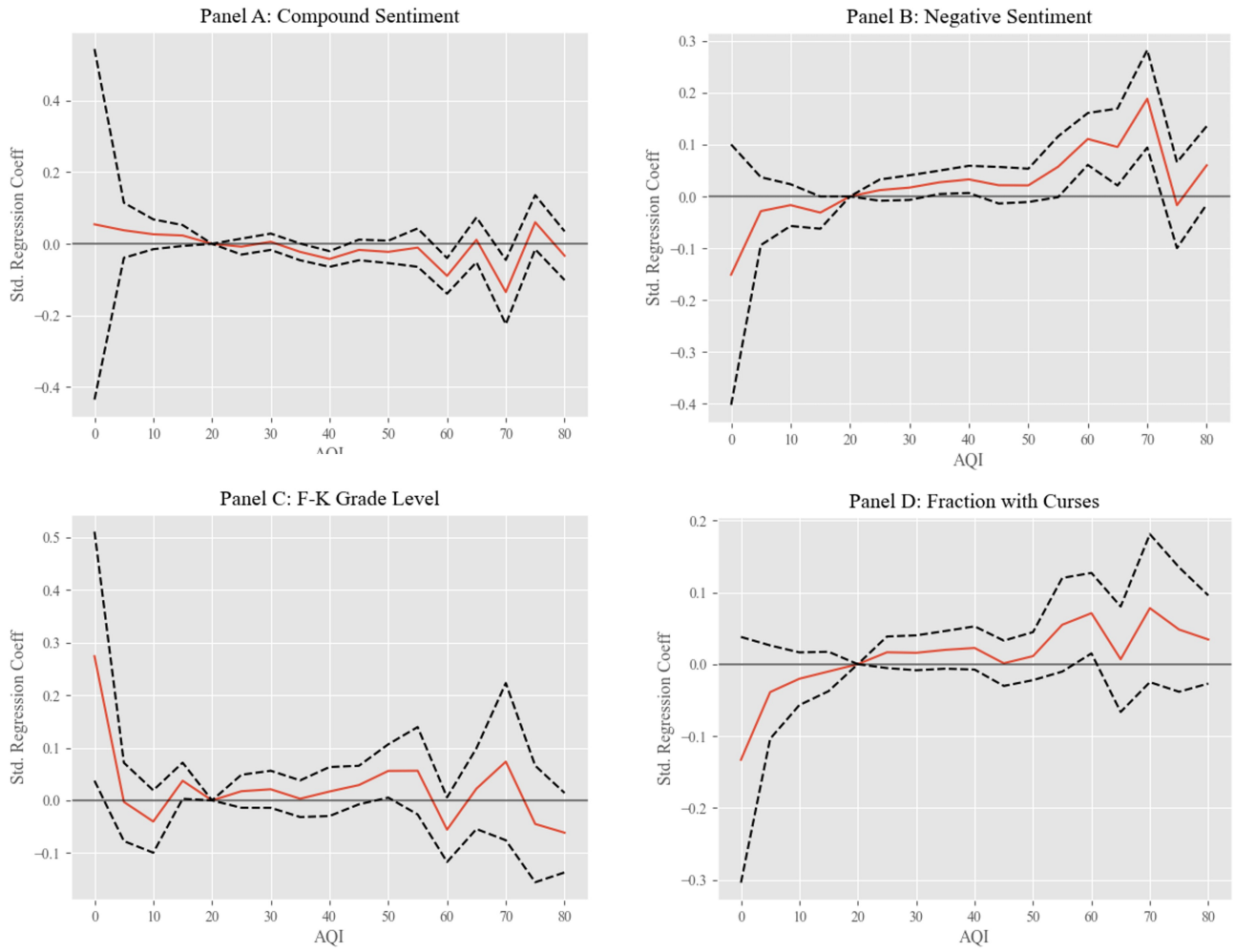

Figure 16: Nonlinear Responses to AQI

Note: Panel A, B show standardized TWFE coefficients for AQI on VADER Compound Sentiment, Negative Sentiment, respectively. Panel C shows standardized TWFE coefficients for AQI on F-K Grade Level. Panel D shows standardized TWFE coefficients for AQI on the fraction of Tweets that contain curse words. Solid lines represent the regression coefficients on AQI and represent the difference (measured in standard deviations) in CBSA-date sentiment for the AQI bin $A \in [a, b)$ relative to $[20, 25)$, controlling for average temperature and precipitation and with fixed effects for CBSA, day-of-week, and holiday. Dotted lines are 95% confidence intervals. Standard errors are unclustered.

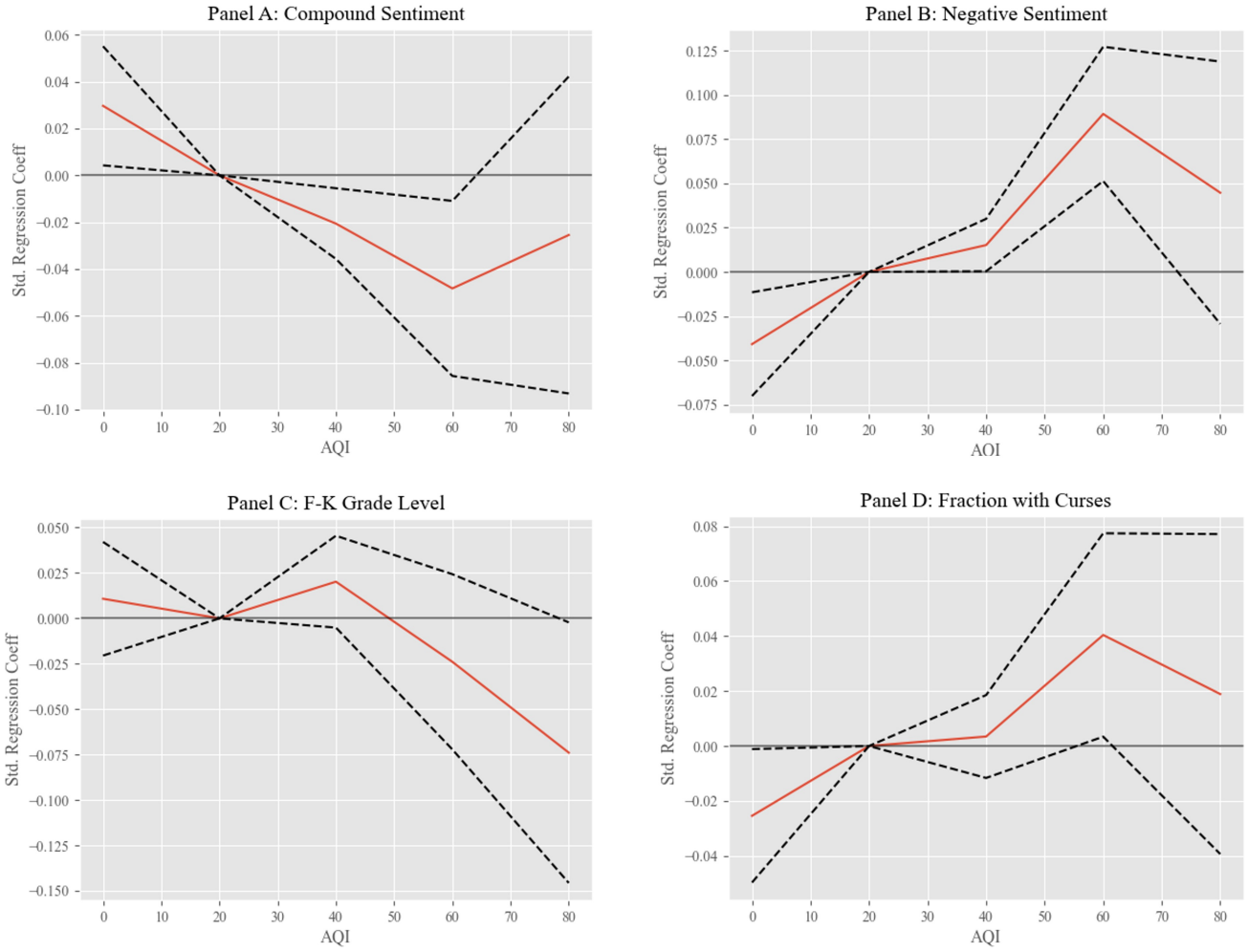

Figure 17: Nonlinear Responses to AQI

Note: Panel A, B show standardized TWFE coefficients for AQI on VADER Compound Sentiment, Negative Sentiment, respectively. Panel C shows standardized TWFE coefficients for AQI on F-K Grade Level. Panel D shows standardized TWFE coefficients for AQI on the fraction of Tweets that contain curse words. Solid lines represent the regression coefficients on AQI and represent the difference (measured in standard deviations) in CBSA-date sentiment for the AQI bin $A \in [a, b)$ relative to $[20, 40)$, controlling for average temperature and precipitation and with fixed effects for CBSA, day-of-week, and holiday. Dotted lines are 95% confidence intervals. Standard errors are unclustered.