Hot Temperature and High Stakes Cognitive Assessments*

R. Jisung Park
University of California, Los Angeles
jisungpark@luskin.ucla.edu

First draft: August 10, 2016;
This draft: September 4, 2018

Abstract

Despite the prevalence of high stakes cognitive assessments – and the growing likelihood of heat exposure during such assessments – the effect of temperature on high stakes cognitive performance has not yet been studied. Using student-level administrative data for the largest public school district in the United States, I provide the first estimates of the impact of hot temperature on high-stakes exam performance and subsequent educational attainment. Hot days reduce performance by up to 15% and lead to persistent impacts on high school graduation status, despite what appears to be compensatory responses by teachers.

Keywords: temperature, human capital, education
JEL Codes: I21, I26, J24, O18, Q51, Q54, Q56

*The author would like to thank Larry Katz, Andrei Shleifer, Robert Stavins, Joe Aldy, Geoffrey Heal, Raj Chetty, Claudia Goldin, Edward Glaeser, Melissa Dell, Michael Kremer, Josh Goodman, Jonah Rockoff, Jeff Miron, Max Auffhammer, Olivier Deschenes, Patrick Behrer, Sarah Reber and numerous seminar participants at Harvard, Columbia, UCLA, Duke, Maryland, UCSB, UC Berkeley, Oxford, IZA, Seoul University, the NYC Department of Health, the NBER Summer Institute and the Bill and Melinda Gates Foundation for valuable comments and feedback, as well as to the NYC Department of Education for data access, and to Nicolas Cerkez and Rodrigo Leal for excellent research assistance. All remaining errors are my own. Funding from the Harvard Environmental Economics Program, the National Science Foundation, and the Harvard Climate Change Solutions Fund are gratefully acknowledged.

1 Introduction

This study investigates the effect of hot temperature on high stakes cognitive performance. Using linked administrative data from the nation's largest public school district, I find that students who take high school exit exams on a hot day perform substantially worse than they otherwise would have, and that these transitory shocks to cognition can have lasting consequences for educational attainment. Given the economic stakes of these exams, it seems likely that these effects are not driven by reduced effort, suggesting substantial cognitive impacts associated with heat exposure. Consistent with a view that random shocks to cognition arising from adverse environmental conditions are not reflective of one's stock of human capital, I find evidence of compensatory grade manipulation by teachers ex post: a phenomenon that is far more pronounced when students experience unusually hot exam conditions.

This investigation is of interest for three reasons. First, cognitive assessments such as school examinations or job interviews are a ubiquitous fact of life in modern economies, due in large part to the increasing importance of cognitive skills in the workplace (Autor et al., 2003; Goldin and Katz, 2009; Acemoglu and Autor, 2011; Hanushek and Woessmann, 2012). Such assessments often take place in "high stakes" environments, which can involve considerable physical and temporal constraints, and under conditions that are often not at the discretion of the individuals being assessed.[1] In many countries, students who perform worse than expected on their college entrance exams must wait up to an entire year to take them again, potentially creating high opportunity costs of having to retake the exam. When the stakes are high, even small perturbations in realized cognitive performance may have lasting educational and labor market consequences for the individual – and potential allocative inefficiencies for society – making it important to understand the effects of environmental conditions on cognition (Ebenstein et al., 2016).[2] Of particular policy concern is the potential for disparities in test-taking conditions across demographic groups, due for instance to residential sorting on local environmental amenities (Tiebout, 1956; Roback, 1982), and well-documented correlation between income and appliances such as air conditioners (Gertler et al., 2016).

Second, the link between temperature and cognitive performance is of heightened policy relevance due to the global externality associated with greenhouse gas emissions. Episodes of acute heat exposure are becoming more frequent in many parts of the world, and are predicted to increase at an accelerating rate (Stocker, 2014). Importantly, much of this warming will occur in

[1] The list of high stakes standardized examinations that determine degree eligibility or impose hurdles to further schooling is long, and include the SAT, ACT, LSAT, MCAT, GRE in the United States, the GCSE in the United Kingdom, the NCEE in China,and the CSAT in South Korea, among others. Similarly, job interviews are often conducted over the course of several hours in one day, or at most several days, and often involve a high degree of coordination which make rescheduling costly.

[2] Empirical evidence of such effect persistence operating through educational institutions is rare, despite much suggestive evidence from studies of in-utero exposure (Currie and Hyson, 1999; Isen et al., 2017). The only paper that documents this type of effect is Ebenstein et al. (2016), which studies Israeli college entrance exams and finds large negative impacts of air pollution on exam performance, which result in earnings losses later in life. Possibly due to mandatory air conditioning of test centers, they find no evidence for temperature-driven effects.

places and during times of year that do not currently feature such temperature extremes, meaning that local institutions may not be efficiently adapted to new expected climate distributions – whether in the timing of examinations or policies regarding protective built infrastructure (e.g. air conditioning).

Third, in assessing the impact of temperature on labor market outcomes, it is important to account for behavioral responses: in particular, endogenous changes in effort. Due to the well-documented relationships between metabolic rate and core body temperature it is likely that the marginal disutility of effort increases with elevated ambient temperature (Hocking et al., 2001; Bouchama and Knochel, 2002; Lim et al., 2008; USDHHS, 2010). The existing literature on temperature and cognitive performance, however, has assessed this link primarily in voluntary or experimental settings where the stakes are not economically meaningful (Graff Zivin et al., 2017; Garg et al., 2017).[3] It is therefore unclear whether documented relationships between hot temperature and task performance arise from reduced effort or residual impacts on direct cognitive capacity.[4]

To explore the impact of hot temperature on high stakes cognitive performance, I combine local daily weather data with test scores of 1 million US public high school students taking synchronized high stakes exams, and link this data to administrative data on subsequent educational attainment. This is, to my knowledge, the most comprehensive dataset assembled to date aimed at assessing the effect of temperature on cognitive performance. Student fixed effects regressions identify the causal impact of heat exposure on exam performance and eventual educational attainment by exploiting a unique institutional setting that effectively shuts down the extensive margin response (i.e. absenteeism), and results in quasi-random variation in temperature for an individual student across multiple exam dates and times. Causal identification rests on a simple premise: that within-student variations in day-to-day temperature are not correlated with unobserved determinants of educational performance. I also assess potential longer-run implications by linking individual exam records to high school graduation status. Importantly, the high stakes setting allows an assessment of the residual impact of hot temperature on cognition net of compensatory responses.

The first main finding is that hot temperature reduces cognitive performance substantially,

[3] A growing literature explores defensive investments such as air conditioning as well as avoidance behavior in the form of time allocation decisions (Graff Zivin and Neidell, 2014; Barreca et al., 2016). However, effort responses on the intensive margin remain relatively understudied. Notable exceptions are Graff-Zivin and Neidell (2012) and He et al. (2017), which attempt to account for potential effort responses, but in the context of air pollution. Neither study finds strong evidence for interactions between air pollution and effort levels, possibly due to the lack of clear metabolic pathways.

[4] This distinction matters for welfare and policy. Educational and labor market institutions may implicitly operate on quasi-procedural norms of fairness, such that changes in performance due to reductions in effort are viewed differently from "unavoidable" physiological impacts. If residual cognitive effects are meaningfully large, evaluations based on standardized assessments administered across a wide spectrum of geographic or economic contexts may need to be adjusted to account for environmental conditions, and/or the built environment that mediates these conditions. Due to informational asymmetries between test takers and administrators (or coordination problems among test takers), it may be difficult without impartial evidence for the private market to provide the information necessary for such adjustments.

and that this effect is likely not due to reductions in effort alone. Each student takes a series of mandatory exams in June which are spread over the course of two weeks and feature harmonized timing and pre-determined testing sites. Because I am able to link multiple exam records for each student and school location, and to match these records to local ambient temperature on the day of each subject exam, the analyses presented here likely identify the causal impact of hot temperature on contemporaneous cognition. Hot temperature during an exam results in reduced exam performance: a decline of approximately -0.2 percentiles per °F above room temperature (70°F). This implies that taking an exam on a 90°F day reduces performance by 14 percent of a standard deviation relative to a more optimal 70°F day. For a sense of magnitude, the within-school Black-White achievement gap is approximately 25 percent of a standard deviation. At least 18% of the students in the study sample experience an exam with ambient temperatures exceeding 90°F.

The second finding is that transitory changes in test-taking conditions can lead to permanent impacts on educational attainment. Consistent with inflexible exam administration (i.e. no rescheduling) and high opportunity or stigma costs of retaking, I find that hot temperature during a test reduces a student's likelihood of graduating from high school. For the median student, taking an exam on a 90°F day leads to a 10.9% lower likelihood of passing a particular subject (e.g. Algebra), which in turn reduces the probability of graduation. A one standard deviation increase in average exam-time temperature reduces a student's likelihood of graduating on time by roughly 2.5 percentage points. This is despite the fact that students are able and often encouraged to retake failed exams during the ensuing summer and following school years.

Consistent with these persistent consequences and a pedagogical view that transitory shocks to cognition do not reflect underlying human capital, I find evidence of compensatory responses by teachers who selectively upward manipulate grades, especially for students who experienced hot exam days. Using a subject, school, and date-specific bunching estimator at pass-fail cutoffs adapted from previous work (Dee et al., 2016), and relating the extent of bunching to temperature on the day of an exam, I show that teachers manipulated grades more frequently for hot exam takes. The amount of excess bunching is beyond what would result from mechanical correlation between temperature-induced performance declines and an increase in the proportion of scores in the manipulable zone, suggesting that teachers are responding to hot exam-day temperature. While it is difficult to infer teachers' intentions, these patterns are consistent with a view that transitory shocks to cognition are not reflective of underlying human capital, and that the resulting educational and economic consequences would be inefficient and/or unfair.

The primary contribution of this paper is that it is the first to study the contemporaneous impact of hot temperature on exam performance in a setting where the stakes are economically meaningful. It builds on a large literature that examines the causal impact of hot temperature on economic outcomes such as health and labor supply, and a smaller but growing literature on temperature and cognitive outcomes.[5] The results provide much stronger evidence than existing

[5]For reviews of the economic literature on weather fluctuations on economic activity and heat exposure on labor-

studies that there are meaningful links between temperature and cognition net of potential endogenous changes in effort. It is also the first to document persistent impacts of heat exposure in school settings on longer-term educational outcomes, and the first to document ex post compensatory responses to acute environmental shocks.[6] The finding that transitory environmental conditions during high stakes exams can have persistent educational and economic consequences echoes findings from Ebenstein et al. (2016), who study air pollution in Israel. This study however is the first to link short-run heat exposure during exams to educational attainment, which has distinct implications for optimal carbon policy and education policy.

Specific implications for welfare and policy are discussed in greater detail in the conclusion. In brief, the findings suggest that educational and labor market institutions that do not adapt assessment timing to a shifting climate distribution (e.g. final exams in early summer months for much of the Northeast US and Europe) may experience reduced allocative efficiency due to the increased likelihood of hot temperature episodes during high stakes assessments: unless investments in built infrastructure can effectively offset these adverse effects.[7] It suggests furthermore that sorting on local amenities may give rise to inequities in standardized exam performance across race and income groups, even if air conditioning is an effective adaptation (Tiebout, 1956; Roback, 1982). Such equity concerns are of particular policy relevance given the growing prevalence of standardized exams used for student advancement or teacher evaluations, and the fact that such exams are administered centrally across wide areas, potentially leading to vastly different test-taking conditions for any given test administration.

The rest of this paper is organized as follows. Section 2 presents relevant stylized facts and a simple conceptual framework that guides the empirical analysis. Section 3 describes the data and institutional context and presents key summary statistics. Section 4 presents the main results and various sensitivity analyses. Section 5 presents results on longer-run educational attainment, and Section 6 presents evidence consistent with compensatory effort by students and teachers. Section 7 discusses implications and concludes.

2 Background and Conceptual Framework

2.1 Temperature and Human Welfare

That individuals experience direct disutility from extreme temperature is well documented in market transactions such as housing or energy demand (Auffhammer and Mansur, 2014; Albouy

related outcomes, see Dell et al. (2014) and Heal and Park (2016) respectively. For a review of the experimental literature on temperature and task productivity, see Seppanen et al. (2006).

[6]There are two studies that document longer-run consequences of heat exposure on human capital-related outcomes. Isen et al. (2017) looks at heat shocks in-utero and finds negative impacts on wages later in life, andCho (2017) explores the effect of summertime heat exposure on exam performance in November. This study complements these findings in that it is the first to assess the consequences of heat exposure in school on educational attainment, suggesting a wide range of overlapping mechanisms that may be operating in the so-called "missing middle".

[7]Conversely, climate change may reduce the likelihood of adverse impacts from cold, though existing hedonic literature suggest that the marginal WTP to avoid heat is non-linear, whereas it is linear for avoiding cold (Albouy et al., 2016), and that countries generally seem to adopt space heating more quickly than air conditioning.

et al., 2016; Sinha et al., 2017). It is also well-known that physical activity and mental exertion both raise metabolic rates, implying that marginal disutility of effort is likely rising in ambient temperature (Lim et al., 2008). Consistent with this phenomenon, time-use decisions are sensitive to temperature, with evidence from the United States suggesting that workers reduce time spent working outdoors when temperatures reach above 80°F, with imprecisely estimated impacts of cold temperature (Graff Zivin and Neidell, 2014). Importantly, estimates from the hedonic literature suggest that the revealed preference optimal temperature is between 65°F and 70°F, and that marginal WTP to avoid heat is likely non-linear, whereas marginal WTP to avoid cold is approximately linear (Albouy et al., 2016).

Existing studies of the effect of temperature on cognition fall into two categories. They consist either of observational (cross-sectional) analyses and case studies, where causal attribution is difficult (Durán-Narucki, 2008), or take place in low-stakes survey-based settings, where it is unclear whether the observed effect is due to reduction in effort or actual residual cognitive decline (Mackworth, 1946; Seppanen et al., 2006). The two studies closest in spirit to this analysis are Graff Zivin et al. (2017) and Garg et al. (2017).[8] In pioneering work, Graff Zivin et al. (2017) study voluntary cognitive assessments of children surveyed as part of the NLSY for roughly 8,000 households in the United States. While they find that hot temperature on the day of the survey reduces math (but not reading) performance, these assessments carry little if any economic weight, making reductions in effort a potential driver of the results.[9] Garg et al. (2017) study the effect of temperature on voluntary cognitive assessments of Indian students, which are similarly low stakes. In their setting it is likely that both effort reduction and poor nutrition may be contributing mechanisms, a possibility that is bolstered by their finding that the effects of heat are most pronounced during the growing season.[10] To my knowledge, no previous study examines the effect of temperature on high stakes cognitive assessments.

[8]Previous research has documented effects of temperature on other related outcomes, including on mortality, morbidity, and labor productivity (Hsiang, 2010; Deschênes and Greenstone, 2011; Cachon et al., 2012; Deryugina et al., 2016). Existing studies exploring the effect of temperature extremes on productivity in the workplace are unable to assess whether the realized impacts are driven by responses on the effort margin or reduced residual cognitive function. Moreover, most of these studies asses physical occupations (e.g. manufacturing) or low-skilled cognitive tasks (e.g. call center operation), which may or may not provide applicable insights for understanding the effect of environmental conditions on knowledge-intensive cognitive tasks.

[9]The fact that reading performance does not decline with temperature suggests that the outcome variable may be measured with considerable error. Graff Zivin et al. (2017) attempt to control for effort by studying changes in time to completion, but this approach is of arguably limited value in that it is still possible to reduce the intensity of effort without finishing earlier or later, and because these assessments are so short (approximately 10 minutes) to begin with.

[10]In addition to the potential for selective sorting based on unobservable student characteristics, survey-based analyses such as Graff Zivin et al. (2017) or (Garg et al., 2017) face an additional challenge due to the fact that hot temperature may lead to systematic under-reporting of data by administrators. For instance, a substantial proportion of NLSY surveys are missing cognitive (PIAT) assessments, or show incomplete reports, which may be due to heat-fatigued surveyors selectively skipping sections of the assessment. (See: https://www.nlsinfo.org/content/cohorts/nlsy97/topical-guide/education/piat-math-test).

2.2 Conceptual Framework

To motivate the empirical analysis, consider a simple model of effort and cognition under temperature stress. Denote the stock of human capital as h. This may represent general ability or a specific set of skills. Suppose that the application of this human capital to a particular task, whether answering questions on an exam or performing skill-intensive assignments on the job, depends on the level of effort expended e, as well as on ambient environmental conditions (specifically, ambient temperature) a, both of which are normalized to be $e \in (0, 1)$ and $a \in (0, 1)$. In the case of e, 1 denotes maximal effort; for a, 1 denotes a physically uninhabitable ambient temperature.[11]

Realized cognitive performance (or test score) can be expressed as:

$$y = y(e, a; h) \tag{1}$$

As can be seen in equation 1, effort and ambient environmental conditions jointly determine the realized level of cognitive performance for a given stock of human capital. Let us define *y(e, a)* such that $\frac{\partial y}{\partial e} > 0$, and $\lim_{e \to 1} y = h$: in other words, maximal effort is required to perform at one's peak capacity *h*.

Individuals derive utility from consuming some composite good X, and experience disutility from physiological stress P, such that $U(X, P), U_X > 0$, and $U_P < 0$. Importantly, suppose that P depends on a and e: $p(a, e) = ap$. This implies that not only is direct disutility from physical distress increasing in effort, but also that the marginal disutility from effort is increasing in environmental stress. The medical literature provides strong support for this assumption. For instance, core body temperature, which is the most commonly used metric of thermal stress, depends on the product of metabolic rate, an indicator of exertion, and ambient temperature (Hocking et al., 2001; Lim et al., 2008).

The representative individual's utility maximization problem is:

$$\max_{e} U(X, P) = U(X, P(e, a)) \tag{2}$$

subject to the constraint $X = I + w(y(e, a; h))$, where I denotes endowment income and w denotes wage income. Wages depend on realized cognitive performance, either because they are determined by the labor market's assessment of relevant human capital stock, h, which is signaled by performance on a formal assessment y, or because the individual is paid a piece-rate contract and productivity depends on cognition.

For simplicity, I abstract away from specific changes to the built environment (e.g. air conditioning) that may reduce experienced temperature, and take ambient environmental conditions during a given assessment as beyond the individual's control. This seems to correspond to most high stakes exam or job interview settings. While I present a simple static framework, the intu-

[11] There is evidence that both extreme heat and extreme cold can have adverse physiological impacts. Conceptually, a can be thought of as representing absolute deviations from thermoregulatory optimum. As discussed above, it seems likely that disutility of effort is increasing with hot temperature but not cold.

ition extends naturally to settings where incremental changes in realized cognitive performance at a particular point in time can have persistent ramifications for wages in many subsequent periods. Finally, I do not model potential time reallocation or rescheduling decisions, and focus on settings where cognitive performance is required within some externally imposed temporal constraints.[12]

Substituting $w(y(e, a; h)$ for X and differentiating with respect to e yields the familiar first order condition:

$$\frac{U_P}{U_X} = \frac{-w' \frac{\partial y}{\partial e}}{\frac{\partial P}{\partial e}} \tag{3}$$

The individual maximizes utility by exerting effort in a way that balances the tradeoff between marginal disutility of physical discomfort and marginal utility of cognitive performance, which operates through the labor market.

Equation 3 implicitly defines optimal effort e^*, as a function of environmental conditions and other parameters. This allows us to express the total derivative of cognitive performance with respect to ambient environmental conditions as:

$$\frac{dy}{da} = \frac{\partial y}{\partial e}\frac{de^*}{da} + \frac{\partial y}{\partial a} \tag{4}$$

This expression shows that the realized change in cognitive performance can be decomposed into two terms: $\frac{\partial y}{\partial e}\frac{de^*}{da}$, which is the change in performance due to changes in effort, arising from increased disutility of effort in the context of hotter temperature, and $\frac{\partial y}{\partial a}$, which describes the direct effect of elevated temperature on cognitive performance. Any empirical estimates of $\frac{dy}{da}$, even when conducted in experimental settings or utilizing exogenous variation in a, will be a combination of these two effects, making it difficult to assess whether the residual is driven by changes in effort ($\frac{\partial y}{\partial e}\frac{de^*}{da}$) or reduced cognition ($\frac{\partial y}{\partial a}$).

Totally differentiating equation 3 with respect to a, it is possible to obtain an expression for $\frac{de^*}{da}$, which can be shown to depend on the intensity of the economic stakes. That is:

1. If $w' = 0$, then $\frac{de^*}{da}$ will be strictly negative.
2. If $w' > 0$, then $\frac{de^*}{da}$ is increasing in the curvature of $w(y)$, and possibly even positive.
3. If $\frac{de}{da} > 0$, then non-positive changes in y as a function of a imply $\frac{\partial y}{\partial a} < 0$.

In other words, in low/no-stakes environments where wages do not vary with realized cognitive performance ($w' = 0$), there will likely be a mechanical relationship between realized performance and ambient environmental stress, due to the expected reduction in effort. As the stakes are raised ($w' > 0$), $\frac{de^*}{da}$ becomes less negative, and can potentially become positive, as individuals attempt to compensate for the deleterious consequences of transitory environmental conditions

[12]For a detailed treatment of avoidance behavior and defensive investments in the context of environmental stressors, see Graff Zivin and Neidell (2013), Behrer and Park (2017) and Barreca et al. (2016). For a discussion of time use reallocation in response to heat exposure in the labor market see Graff Zivin and Neidell (2014).

on current or future consumption. For instance, consider the example of a college entrance exam. If adverse test-taking conditions can nudge a student on the margin of qualifying for a high school diploma to being ineligible, and if employers use degree status as a signal of worker ability, the result may be a reduction in expected wages for many future periods. Even if the student is able to retake the exam, the time/opportunity costs of preparing for and taking the exam again, as well as potential stigma in the eyes of future employers, will weigh on the student's effort decision. Unless the individual is highly myopic, one would expect some degree of compensatory increases in effort, at least for the duration of the high stakes assessment.

2.3 Implications for Empirical Analyses

One implication of the model is that it is difficult to draw conclusions regarding the link between temperature and cognition without measuring $\frac{dy}{da}$ in settings where the stakes are economically meaningful. If, however, an empirical analysis of high stakes settings finds $\frac{dy}{da} < 0$, this could imply physical limits to the capacity of students or workers in compensating for exogenous (particularly unexpected) deterioration in environmental conditions, even at levels that are not life-threatening.[13] The model also suggests that welfare implications depend on how policymakers or labor market institutions evaluate the different components of equation 4. If, in practice, assessments of human capital – often proxied using performance on high stakes assessments – implicitly account for effort, then information on residual cognitive impacts may be important for the design and implementation of efficient and equitable policy.

In summary, the primary insights of the model are that (i) hot ambient temperature can drive a wedge between realized cognitive performance (e.g. exam score) and underlying human capital, (ii) effort levels will, in addition to being endogenous to ambient temperature, also depend on the economic stakes involved, and that (iii) empirical analyses that uncover residual impacts of hot temperature on exam performance in high stakes environments will likely imply direct (residual) impacts on cognition, given the high likelihood of compensatory increases in effort (as opposed to reductions in effort).

3 Institutional Setting, Data, and Summary Statistics

3.1 New York City High Schools: High Stakes Exams

The New York City public school system (NYCPS) is the largest in the United States, with over 1 million students as of 2012. Each June, these students take a series of high-stakes exams called "Regents exams", which are standardized subject assessments administered by the New York State Education Department (NYSED).

[13] Most of the literature on physical limits to heat exposure in the workplace has focused on very extreme temperatures: for instance, wet-bulb globe temperatures (WBGT) of 32°C or above (Kjellstrom and Crowe, 2011). One implication of this paper's findings is that elevated temperature has an effect on cognition even at levels well below such life-threatening extremes.

Regents exams carry important consequences. Students are required to meet minimal proficiency status – usually a scale score of 65 out of 100 – in five "core" subject areas to graduate from high school.[14] Many local universities including City University of New York (CUNY) use strict Regents score cutoffs in the admissions process as well: for instance, requiring that students score above 75 on English and Math simply to apply. These exams are therefore pivotal for the median student in determining high school diploma eligibility and college admissions.

The average 4-year graduation rate, at 68%, is comparable to other large urban public school districts, and suggest that standardized high school exit exams are a binding constraint for a large number of students. System-wide averages mask considerable discrepancies in achievement across neighborhoods. Schools in predominantly Black or Hispanic sub-districts have four-year graduation rates as low as 35% per year (Figure 3a and 3b).

The vast majority of students take their Regents exams during a pre-specified two-week window in mid-to-late June each year. The dates, times, and locations for each of these Regents exams are fixed over a year in advance by the state education authority (NYSED), and synchronized across schools in the NYC public school system to prevent cheating. Each exam is approximately 3 hours long, with morning and afternoon sessions each day, and are taken at the student's home school, unless they required special accommodations which were not available at their home school. Students who fail their exams (or are deemed unready by their teachers to progress to the next grade) are required to attend summer school, which occurs in July and August. Figures 1a and 1b provide a sample exam schedule and cover sheet.

3.2 Student Data

I obtain individual exam-level information from the New York City Department of Education (NYC DOE). This includes records for the universe of NYC public high school students who took one or more Regents exams over the period 1999 to 2011. Information on exam dates comes from archived Regents exam schedules, which provide date and time information for each subject by year and month of administration. Graduation status by student is available in a separate file, which can be linked to exam records using unique 10-digit student identifiers. These records include cohort and school information, as well as graduation status up to 6 years post-matriculation. A detailed description of the matching procedures and subsequent sample restrictions are provided in the online appendix.

All exams are written by the same state-administered entity and scored on a 0-100 scale, with scaling determined by subject-specific rubrics provided by the NYSED in advance of the exams each year. All scores are therefore comparable across schools and students within years, and the scaling designed in such a way that is not intended to generate a curve based on realized scores. I use standardized performance at the subject level as the primary measure of exam performance

[14]The core subject areas are English, Mathematics, Science, U.S. History and Government, and Global History and Geography. The passing threshold is the same across all core subjects. Students with disabilities take separate RCT exams, and are evaluated on more lenient criteria.

in this study, though the results are robust to using scale scores. While centrally administered, exams were locally graded by committees of teachers in the students' home schools, usually on the evening of the associated subject exam.

3.3 Weather Data

Weather data comes from the National Oceanic and Atmospheric Administration's Daily Global Historical Climatology Network, which provides daily min, max, and mean temperatures, precipitation and dew point information from a national network of several thousand weather stations over the period 1950-2014. I take daily minimum and maximum temperature as well as daily average precipitation and dewpoint readings from the 5 official weather stations in the NYC area that provide daily data for the entirety of the sample period (1998-2011). I match schools to the nearest weather station, one for each of the five boroughs: The Bronx, Brooklyn, Manhatten, Queens, Staten Island.[15] Given existing evidence on the impact on air quality on student performance, I include controls for pm2.5 and ozone, taken from EPA monitoring data from Manhattan.[16]

3.4 Summary Statistics

The final working dataset consists of 4,509,102 exam records for 999,582 students. It includes data from 91 different exam sessions pertaining to the core Regents subjects over the 13 year period spanning the 1998-1999 to 2010-2011 school years.

Table 4 presents summary statistics for the key outcome variables that form the basis of this analysis. The student body is 40% Latino, 31% African American, 14% Asian and 13% White, and approximately 78% of students qualify for federally subsidized school lunch. Students take on average 7 June Regents exams over the course of their high school careers, and are observed in the Regents data set for roughly 2 years, though some under-achieving students are observed for more than 4 years, as they continue to retake exams upon failing.

Fewer than 0.2% of students are marked as having been absent on the day of the exam, corroborating the high-stakes, compulsory nature of these exams. The median student scores just around the passing cutoff, with a score of 66 (sd = 17.9), though there is considerable heterogeneity by neighborhood as well as demographic group.

[15] To account for spatial heterogeneity in outdoor temperature due to urban heat island effects, I assign spatial correction factors generated by satellite reanalysis data. I impute test-time temperature – for instance, average outdoor temperature between 9:15am to 12:15pm for morning exams – by fitting a fourth-order polynomial in hourly temperature. Further details regarding these corrections are presented in the online appendix. The primary results reported below are not sensitive to either of these corrections. The corrections reduce standard errors but leave implied point estimates relatively unchanged.

[16] Air pollution in NYC during this period is relatively low, compared, for instance, to the levels found to affect Israeli student performance (Ebenstein et al., 2016). The maximum recorded value of pm2.5 in my data is 38 micrograms per cubic meter, compared to readings that regularly went above 120 micrograms per cubic meter in Ebenstein et al. (2016). The air quality controls used here are nevertheless crude, especially for localized pollutants such as ozone. Given the focus of the study, the relatively low levels of particulate matter during the sample period, and the high correlation between ozone and summertime temperature, I run analyses with and without controls for air quality but do not attempt to separately identify or interpret causal effects of fine particulates or ozone.

Figure 2 illustrates the source of identifying variation for short-run temperature impacts, with temperatures weighted by exam observation and school location. Outdoor temperature during exams range from a low of 60°F to a high of 98°F. Day-to-day variation within the June exam period can be considerable, as suggested by Figure 2b, which shows the variation in outdoor temperature by school and exam take across two consecutive test dates within the sample period.

4 Effect of Temperature on High Stakes Exam Performance

Figure 5a presents a visual depiction of performance and temperature that motivates the analysis that follows. It shows a binned scatterplot of standardized exam score by percentile of observed exam-day temperature, plotting residual variation after controlling for school fixed effects and average differences across subjects and years. Exams taken on hot days clearly exhibit lower scores.

To further isolate the causal impact of short-run temperature fluctuations on student performance, I exploit quasi-random variation in day-to-day temperature across days within student-month-year cells, focusing on the main testing period in June. While it is unlikely that temperature is endogenous to student behavior, nor is it likely for students to select into different temperature treatments given the rigidity of exam schedules, time-varying unobservables may still be correlated with weather realizations. For instance, if certain subjects tend to be scheduled more often in the afternoon when students are relatively fatigued (as in Sievertsen et al. (2016)) or toward the end of the exam period (Thursday as opposed to Monday), we may expect mechanical correlation between temperature and test scores that is unrelated to the causal effect of temperature on student cognition. This motivates a baseline specification that includes year, time of day, and day of week fixed effects:

$$Y_{ijsty} = \gamma_{iy} + \eta_s + \beta_1 T_{jsty} + X_{jsty}\beta_2 + \beta_3 Time_{sty} + DOW_{sty}\beta_4 + \epsilon_{ijsty} \quad (5)$$

Here, Y_{ijsty} denotes standardized exam performance for student i taking an exam in subject s in school j on date t in year y. The terms γ_{iy} and η_s denote student-by-year and subject fixed effects respectively. T_{jsty} is the outdoor temperature in the borough of school j during the exam (subject s on date t, year y). X_{jsty} is a school- and date-specific vector of weather and air quality controls, which include precipitation, dewpoint, and ozone. $Time_{sty}$ represents a dummy for time of day (morning versus afternoon, $Time$=1 denotes an afternoon exam), and DOW_{sty} represents a vector of fixed effects for each day of the week in which exams were taken.

Student-by-year fixed effects ensure that I am comparing the performance of the same student across different exam sittings within the same testing window, some of which may be taken on hot days, others not, leveraging the fact that the average student takes 7 June Regents exams over the course of their high school career (between 3 and 4 per year). Subject fixed effects control for persistent differences in average difficulty across subjects. Year fixed effects control for pos-

sible spurious correlation between secular performance improvements and likelihood of hotter exam days due, for instance, to climate change. To the extent that temperature variation within student-month-year cells are uncorrelated with unobserved factors influencing test performance, one would expect the coefficient β_1 to provide an unbiased estimate of the causal impact of temperature on exam performance, subject to attenuation bias due to measurement error in weather variables as well as downward bias from positive grade manipulation.

Table 5b presents the results from running variations of equation (4) for the subset of students who take at least 2 exams. As suggested by the first row of columns (1)-(4), exam-time heat stress exerts a significant causal impact on student performance. The estimates are robust to allowing for arbitrary autocorrelation of error terms within boroughs and test dates, which is the level of exogenous temperature shock recorded in the data.

Taking an exam on a hot day reduces performance by approximately -0.0075 standard deviations (se=0.002) per °F. This amounts to -5.2 percent of a standard deviation in performance per standard deviation increase in temperature, or -15 percent of a standard deviation if a student takes an exam on a 90°F day as opposed to a more optimal 70°F day.[17] The effect of a 90°F day is thus comparable in magnitude to roughly 1/4 of the Black-White score gap, or 3/4 of the within-school Black-White score gap.

This effect is slightly larger than the impacts on mathematical reasoning found by Graff Zivin et al. (2017), who find a 90°F day to reduces NLSY math scores by approximately -0.12 standard deviations, and substantially smaller than the test-day impacts of -0.30 standard deviations documented by Garg et al. (2017) in India. They are similar in magnitude with effects from laboratory experiments (Seppanen et al., 2006), which generally find effects on the order of 1% decline per °F increase in temperature above the optimum of approximately 70°F. This is consistent with substantial residual cognitive impacts, as well as with measurement error in Graff Zivin et al. (2017) and nutrition and health being significant confounders in Garg et al. (2017). These results provide strong evidence that temperature in the learning environment plays an important role in determining student outcomes, and that whatever compensatory effort is exerted by students due to the high stakes nature of some exams may not be enough to offset the physiological impacts of temperature on cognitive performance.

A series of robustness checks, including models that replace student-by-year fixed effects with student- or school-by-year fixed effects, are presented in columns (3) and (4) of table 5b, as well as in the online appendix. The point estimates using the school-by-year fixed effects specification are slightly larger (more negative) on average, and remain statistically significant. Also presented in the appendix are heterogeneity analyses by gender and ethnicity. I find relatively little evidence of heterogeneity by demographic groups, though it is possible that adaptive responses by teachers

[17] Precipitation has a slightly positive effect, and ozone has a negative but insignificant effect, with a 1 standard deviation increase in ozone corresponding to a point estimate roughly 1/5th the size of a 1 standard deviation temperature effect. Despite previous literature documenting adverse impacts of pm2.5 in Israel (Ebenstein et al., 2016), I find little evidence for that here, perhaps because average concentrations of pm2.5 are much lower in NYC than in Israel, as well as the fact that the performance impacts documented by Ebenstein et al. (2016) are highly non-linear, driven mostly by heavily polluted days with pm2.5 above 100 micrograms per cubic meter.

are offsetting impacts disproportionately for certain subgroups.

Running versions of equation 5 that replace standardized exam scores with a dummy variable for whether or not students scored a passing grade, I find that hot temperature substantially reduces the likelihood of passing. A one standard deviation increase in temperature results in a 2.1% lower probability of passing a given subject (a 0.31 (se=0.12) percentage point decline per °F, relative to a mean likelihood of 0.57). Interestingly, the effect is more negative for a dummy for "proficiency status", at -3.5% per standard deviation increase in temperature. In other words, a 90° day results in a 9.7% lower chance of passing a given exam, and a 17.4% lower probability of achieving proficiency status for the average student. The latter is a merit that is useful for some but not all college-bound students, while the former is *a regula* related to graduation eligibility for all students. This is consistent with higher reduction in effort for students who are at a higher segment of the ability distribution, for whom these exams carry less economic weight. It is possible however that this discrepancy arises from differences in the extent of teacher grade manipulation, as discussed below.

5 Persistent Impacts on Educational Attainment

If such changes in cognitive performance are simply transitory shocks due to changing environmental conditions, and do not reflect underlying stock of human capital, one might expect few if any long-term consequences, especially if retaking is possible. However, if the opportunity costs or stigma associated with retaking are high, one might observe persistent, long-run impacts on educational attainment. Similarly, if there are dynamic complementarities in the education production function, whereby students, parents, and/or teachers use test scores as signals of ability or potential (marginal returns to effort), then even transitory shocks might have persistent consequences.

Figure 6a plots variation in 4-year graduation status against average exam-time temperature, and provides suggestive evidence of such persistent impacts. To account for the possibility that the temperature experienced by a student during exams may be mechanically correlated with the number of exams taken (due to mean-reversion in daily temperatures), I compare the difference in graduation likelihood between students who, conditional on the number of draws from the climate distribution, experience different amounts of heat stress. Specifically, I collapse the data to the student level and estimate variations of the following model:

$$g_{ijcn} = \alpha_0 + \alpha_1 \overline{T_{ij}} + X_{ij}\alpha_2 + \chi_j + \theta_c + Z_i\alpha_3 + exams_n\alpha_4 + \epsilon_{ijc} \tag{6}$$

Here, g_{ijcn} is a dummy denoting whether student i in school j and entering cohort c who takes n June Regents exams over the course of her high school career has graduated 4, 5 or 6 years after matriculation. $\overline{T_{ij}}$ denotes the average temperature experienced by student i while taking June Regents exams in school j, up through her senior year. X_{ij} is a vector of weather controls averaged at the student-by-school level. χ_j denotes school fixed effects. θ_c denotes cohort fixed

effects. Z_i is a vector of student-level controls including race, gender, federally subsidized school lunch eligibility, and where applicable scores from previous standardized exams. $exams_n$ denotes a vector of fixed effects for the number of June exam takes.

The parameter of interest is α_1, which captures the impact of an additional degree of heat exposure during exams on the likelihood of graduating on time. School fixed effects account for potential omitted variable bias due to unobserved determinants of graduation rates being correlated with average temperature in the cross-section (e.g. if urban heat island effects are stronger in poorer neighborhoods). Cohort fixed effects in graduation rates allow for the possibility that heat exposure and graduation rates are correlated due to secular trends in both variables – though warming trends and average improvements in NYC schools would suggest this effect to lead to downward rather than upward bias in the estimate of α_1.

Table 6a presents the results from running variations of equation 6 with and without school and cohort fixed effects, as well as flexible controls for the number of exams. Standard errors are clustered at the borough by date and time level, based on the intuition that this conservatively approximates the level of quasi-random temperature variation, though the results are once again robust to alternative levels of clustering.

Columns (1)-(3) suggest that a 1 degree F increase in average exam-time temperatures is associated with a 0.71 (se=0.17) to 0.76 (se=0.22) percentage point decline in the likelihood of graduating on time. A a one standard deviation in average exam-time temperature (+4.4°F) leads to a 3.12 to 3.34 percentage point decline in the likelihood of on-time graduation, or 4.59% to 4.91% decline relative to a mean on-time graduation rate of 68 percent.

These effects are large. Even without correcting for adaptive grading by teachers, I estimate that, over the period 1998 to 2011, upwards of 510,000 exams that otherwise would have passed received failing grades, affecting the on-time graduation prospects of at least 90,000 students. These estimates are described in greater detail in the appendix. This is consistent with the high stakes nature of these exams, suggesting non-trivial economic and psychic costs of hot temperature during inflexibly administered high stakes exams.

6 Compensatory Responses

Given the high stakes, we would expect compensatory effort (or at least less effort reduction) by students. We might also expect responses by other principals or agents who have a stake in the students' welfare: for instance, by teachers or parents. This would be especially true if either group has a view that idiosyncratic shocks to cognitive performance due to environmental conditions are somehow inefficient or unfair. Such compensatory investments are potentially important determinants of overall welfare impacts (Deschenes et al., 2017), but have received relatively little study, in part due to data constraints. The unique institutional features of NYC public schools during the study period allow an indirect assessment of teacher compensation, and provide the first available evidence of ex post compensation in response to hot temperature events.

6.1 Teacher Responses

Up until 2013, Regents exams in NYC were administered centrally by state authorities but graded locally by home school teachers. Exploiting this feature and a similar data set for slightly different years, Dee et al. (2016) find evidence consistent with grade manipulation by teachers of scores that were just below pass/fail cutoffs, and that these are not due to teacher incentives. It seems possible that part of this might be in response to adverse test-taking conditions, especially given the temporal constraints and the substantial economic stakes for the median student.[18]

A hot test day may be viewed as a bad test day, particularly if air conditioning is inadequately provided. In that case, it seems possible for discretionary grade manipulation to have been a response to perceived performance impacts of heat stress. Teachers may be able to observe or at least intuit the disruptive impacts of elevated classroom temperatures on test day, especially since exams are taken in students' home schools and graded by a committee of teachers from that school. If benevolently motivated, they may be inclined to engage in more grade manipulation precisely for those exams that took place under unusually hot conditions. Even if teachers do not actively intend to offset heat-related performance impacts, it is possible that such manipulation may in effect blunt some of the idiosyncratic effects of transitory cognitive shocks on longer-run educational outcomes.

Consistent with this hypothesis, I find that teachers selectively upward manipulated grades around pass/fail cutoffs, and that the extent of manipulation is highly correlated with temperature on the day of the exam. Importantly, I find that this correlation is not due to a mechanical shift in the grade distribution, but is consistent with a higher rate of bunching in response to hotter temperature.

6.2 Estimating Teacher Grade Manipulation

Figure 7 provides a histogram of Regents scale scores in all core subjects prior to 2011. As is clearly visible in the graph, there is substantial bunching at the passing kinks, especially at scores of 65 and 55, suggesting upward grade manipulation. We would expect any form of grade manipulation for students who initially score just below the passing cutoff, even indiscriminate grade manipulation uncorrelated with exam-time temperature, to downward attenuate the estimates of heat-related performance impacts uncovered above.

To assess the presence and magnitude of "compensatory grading", I estimate a bunching esti-

[18] Dee et al. (2016) find that most of the manipulating behavior occurred at or around the passing margin of 65 and that, while varied in magnitude across schools and student types, such manipulation was a near-universal phenomenon within the NYCPS system. Upon careful analysis of competing explanations, the authors suggest the most likely explanation to be the goodwill of teachers who seek to offset the impact of "a bad test day". Importantly, such manipulation was likely not a byproduct of teacher incentive pay programs, since teacher pay was not tied to these performance metrics during the study period. To quote the authors: *In sum, these estimates suggest that manipulation was unrelated to the incentives created by school accountability systems, formal teacher incentive pay programs, or concerns about high school graduation. Instead, it seems that the manipulation of test scores may have simply been a widespread "cultural norm" among New York high schools, in which students were often spared any sanctions involved with failing exams, including retaking the test or being ineligible for a more advanced high school diploma (pg 27).*

mator by school, subject, month, and year: in effect, the level of exam-time temperature variation. Starting with the student-exam level data, I calculate the fraction of observations in each 1 point score bin from 0 to 100 by core Regents subject. I then fit a polynomial to these fractions by subject, excluding data near the proficiency cutoffs with a set of indicator variables, using the following regression:

$$F_{ks} = \Sigma_{i=0}^{q} \psi_{ismyj} \dot{(Score)}^i + \Sigma_{i \in -M_{cs}, +M_{cs}} \lambda_{ismyj} \dot{1}[Score = i] + \epsilon_{ksmyj} \tag{7}$$

where F_{ks} denotes the fraction of observations with score k for subject s (e.g. ELA), q is the order of the polynomial, and $-M_{cs}, +M_{cs}$ represent manipulable ranges below and above the passing thresholds. The subscripts m, y and j denote, month, year, and school respectively.

Following Dee et al, (2016), I define a score as manipulable to the left of each cutoff if it is between 50 - 54 and 60 - 64, and manipulable to the right if it is between 55 - 57 and 65 - 67 as a conservative approximation of their subject-and-year-specific scale score-based rubric. In practice, I use a fourth-order polynomial (q=4) interacted with exam subject s, but constant across years for the same exam subject. Realized bunching estimates are not sensitive to changes in the polynomial order or whether one allows the polynomial to vary by year or subject.[19]

This generates a set of predicted fractions by score and subject. The average amount of bunching observed in my data is similar to that documented by Dee et al. (2016), who find that approximately 6% of Regents exams between 2003 and 2011 exhibited upward grade manipulation. I then calculate observed fractions for each score from 0 to 100 by school, month, year, and subject, and generate a measure of bunching that integrates the differences between observed and predicted fractions: that is, summing the excess mass of test results that are located to the right of the cutoff (above the predicted curve) and the gaps between predicted and observed fractions of test results to the left of the cutoff (below the predicted curve). The bunching estimator can be written as:

$$\zeta_{smyj} = \frac{1}{2} \Sigma_{i \in +M_{ck}} (F_{ks} - \hat{F}_{ksmyj}) + \frac{1}{2} |\Sigma_{i \in -M_{ck}} (F_{ks} - \hat{F}_{ksmyj})| \tag{8}$$

where ζ_{smyj} denotes the degree of bunching at the passing cutoff for subject s, month m, year y, and school j.

To assess the magnitude of this relationship controlling for school-, subject-, and/or year-level differences in the degree of manipulation that are unrelated to temperature, I run a series of regressions with ζ_{smyj} as the dependent variable:

$$\zeta_{smyj} = \delta_0 + \delta_1 T_{smyj} + \delta_2 X_{smyj} + \chi_j + \eta_s + f(Year_y) + \epsilon_{smyj} \tag{9}$$

[19] I also estimate a linear approximation of the above estimator by generating predicted fractions using a linear spline between boundary points along the distribution that are known to be outside the manipulable range by subject. I then generate an estimate of the extent of bunching by school-subject-month-year cell, taking the absolute value of the distance between observed and predicted fractions by Regents scale score. The results are similar using this simplified measure of bunching.

where T_{smyj} denotes temperature, X_{smyj} denotes precipitation and humidity, χ_j, η_s, and θ_y denote school, subject, and year fixed effects respectively, and $f(Year_y)$ denotes a cubic time trend in scores. The parameter of interest is δ_1, which represents the increase in grade manipulation due to exam-time temperature.

The amount of bunching increases by approximately 0.10 to 0.16 percentage points per degree F, or 1.7% to 2.8% per degree F hotter exam-time temperature. Coefficients are positive and significant in specifications with and without school and year fixed effects. This relationship is depicted graphically in figure 8a. The short-run performance impacts documented above will, in this sense, be net of compensatory grading.[20] The implied magnitudes are non-trivial. The difference in overall share of exams manipulated between a 90°F day and a 70°F day can be as much as 50%, suggesting temperature fluctuations may represent a large component of the variation in extent of grade manipulation throughout the period.

While it is possible that, due to the distributional properties of most Regents exams, heat-related performance impacts may lead to a mechanical increase in the number of grades that fall within the manipulable zone (which could in principle lead to a correlation between bunching behavior and exam-time temperature) running the above analysis replacing the dependent variable with the fraction of manipulable scores actually manipulated suggests that this is not driving the relationship between temperature and manipulation. Figure 8b presents a binned scatterplot of the bunching estimator and exam-time temperature by subject-month-year-school cell, adding school fixed effects to allow for arbitrary differences in the average amount of grade manipulation across schools, and expressing the bunching estimate as a proportion of scores within the manipulable zone (50-54, 60-64). It suggests a clear positive relationship between the degree of grade manipulation and the ambient temperature during the exam being graded.

These results are consistent with positive grade manipulation as a compensatory response to hot temperature in the test-taking environment. Teachers may have a longitudinal sense of whether a particular student scored below his or her "true ability", and respond by upward manipulating grades more intensively when they perceive the test-taking conditions to have been especially adverse. Irrespective of whether teachers' explicit intentions are to compensate for heat-related impacts, however, the realized effect has been for this behavior to mitigate the adverse welfare impacts associated with exam-time heat exposure.

[20] This likely results in a smaller point estimate than otherwise would have been the case. The only case in which the bias may be upward is if teachers grade differentially and punitively according to the temperatures *they* experience while grading, and temperature during the exam is correlated with temperature during grading. If hot temperatures make teachers less productive and causes more errors, this will simply add noise to the score variable. If hot temperature makes teachers irritable and more punitive in grading, then we might expect the beta coefficient to be picking up some of the correlation between test day temp and grading punitive-ness, although the most striking feature of the score distribution as described below is that the majority of grade manipulation seems to be positive in direction, making this unlikely in practice.

6.3 School Air Conditioning

Data limitations prohibit a detailed assessment of the efficacy of air conditioning in mitigating the observed effects of temperature on test performance. However, I am able to scrape a series of building condition assessment reports from online archives for 644 middle and high schools in the study sample for the year 2012. Of these 644 schools, 62% were reported as having air conditioning as of 2012, and among these, nearly 40% had some form of defective components, consistent with incomplete AC penetration. Because the data does not provide information regarding AC utilization during the dates and years of interest, they are at best only crude proxies for the true variables of interest. In the Appendix, I present the results of running the analyses above for schools with and without AC as of 2012. The results are consistent with AC having some protective effect, but even the schools that have AC by the end of the sample experience negative (albeit smaller) impacts over the period 1998-2012. Importantly, the data confirms the premise that, on hot days, indoor temperatures were likely elevated in many of the schools during much of the study period.

7 Discussion and Conclusion

This paper explores the impact of temperature on high stakes exams, and provides evidence of persistent impacts on subsequent educational attainment as well as compensatory responses to such transitory shocks. Using administrative data from the largest public school district in the United States, I find that hotter temperatures exert a causal and economically meaningful impact on student cognition which do not appear to be driven by reduced effort alone. The research design exploits quasi-random, within-student temperature variation to identify the impact of hot days on performance. These short-run impacts – which presumably do not reduce the stock of human capital – nevertheless result in persistent impacts on educational attainment as measured by high school graduation status. This is despite a pattern of what appears to be ex post compensatory behavior by teachers who upward manipulate borderline scores for exams taken under hot conditions.

Taking an exam on a 90°F day results in a 15% of a standard deviation reduction in exam performance relative to a more optimal 70°F day, controlling for student ability. A 90°F day results in a 10.9% lower probability of passing a subject, and, for the average New York City student, a 2.5% lower likelihood of graduating on time, despite the ability of students to retake failed exams. I estimate that, over the period 1998 to 2011, upwards of 510,000 exams that otherwise would have passed received failing grades due to hot temperature, affecting at least 90,000 students, possibly many more.

Teachers seem to have responded to hot exam sittings by selectively boosted grades of students just below pass/fail thresholds. Using a variant of bunching estimators developed in previous work (Dee et al., 2016), I find a pattern of grade manipulation that intensifies as exam day temperatures increase, even when controlling for potential mechanical correlation between temperature

and the fraction of manipulable scores and the possibility that teacher cognition may itself be endogenous to test-day temperature. One interpretation is that teachers view transitory shocks to cognition as not being reflective of underlying human capital, and may have thus used their discretion to offset a portion of the long-term impacts of such shocks. A possible unintended consequence of eliminating teacher discretion in New York City public schools in 2011 may have been to expose more low-performing students to climate-related human capital impacts, eliminating a protection that applied predominantly to low-achieving Black and Hispanic students.

The findings presented in this paper have several policy implications. First, they suggest that ambient temperature may be an important variable to consider when designing education policy: for instance, in determining how much weight to put on high stakes exams, either for student advancement, teacher promotion, or school funding decisions (Chetty et al., 2014; Jacob and Rothstein, 2016). Similarly, in prioritizing various policy options aimed at reducing achievement gaps, environmental conditions such as climate – or school infrastructure such as air conditioning or HVAC systems – could play a larger role than previously suggested (Cellini et al., 2010; Jackson et al., 2015; Lafortune et al., 2016).

From a distributional equity standpoint, these results suggest that students taking high stakes standardized exams across varying climates and built environment may not be on a level environmental playing field. Such fairness concerns may be especially important for nationally and sometimes internationally harmonized examinations such as the ACT, SAT, and LSAT in the United States, as well as analogous exams in other countries. Unless schools and homes in hotter regions have access to perfectly offsetting amounts of ex ante defensive investments or adaptation capital, individuals taking nationally or internationally standardized exams in a hotter region may be placed at a disadvantage relative to their peers in cooler regions.

Moreover, as the span of geographies covered by a standardized exam widens, the potential for differences in test-day temperature increases. The LSAT, for instance, is taken more or less simultaneously not only across the fifty United States but also across countries as diverse as Brazil, China, India, New Zealand, and Ukraine. Air conditioning penetration is relatively low in many developing economies, and there is a well documented relationship between income and air conditioning ownership at the household level (Davis and Gertler, 2015), as well as growing evidence of liquidity constraints in the context of energy-intensive appliance demand (Gertler et al., 2016). Such factors may also be relevant in thinking about the persistence of racial achievement gaps in the United States, given correlations between race, income, and average climate across neighborhoods (Roback, 1982; Albouy et al., 2016).

Finally, from the perspective of environmental policy, this study suggests that current social cost of carbon estimates may omit important elements of the climate damage function: especially those mechanisms, including human capital accumulation, that may affect the rate of growth as opposed to the level of economic activity (Pindyck, 2013; Heal and Park, 2016), More careful research on the impact of cumulative heat exposure on the pace of learning seems warranted, especially given the high levels of temperature exposure faced by students in tropical developing

economies.

References

Acemoglu, Daron and David Autor (2011), "Skills, tasks and technologies: Implications for employment and earnings." In *Handbook of labor economics*, volume 4, 1043–1171, Elsevier.

Albouy, David, Walter Graf, Ryan Kellogg, and Hendrik Wolff (2016), "Climate amenities, climate change, and american quality of life." *Journal of the Association of Environmental and Resource Economists*, 3, 205–246.

Annan, Francis and Wolfram Schlenker (2015), "Federal crop insurance and the disincentive to adapt to extreme heat." *The American Economic Review*, 105, 262–266.

Auffhammer, Maximilian and Erin T Mansur (2014), "Measuring climatic impacts on energy consumption: A review of the empirical literature." *Energy Economics*, 46, 522–530.

Autor, David H, Frank Levy, and Richard J Murnane (2003), "The skill content of recent technological change: An empirical exploration." *The Quarterly journal of economics*, 118, 1279–1333.

Barreca, Alan, Karen Clay, Olivier Deschenes, Michael Greenstone, and Joseph S Shapiro (2016), "Adapting to climate change: The remarkable decline in the us temperature-mortality relationship over the twentieth century." *Journal of Political Economy*, 124, 105–159.

Behrer, A Patrick and Jisung Park (2017), "Will we adapt? temperature, labor and adaptation to climate change." *Harvard Kennedy School Working Paper*.

Bouchama, Abderrezak and James P Knochel (2002), "Heat stroke." *New England Journal of Medicine*, 346, 1978–1988.

Cachon, Gerard P, Santiago Gallino, and Marcelo Olivares (2012), "Severe weather and automobile assembly productivity."

Cellini, Stephanie Riegg, Fernando Ferreira, and Jesse Rothstein (2010), "The value of school facility investments: Evidence from a dynamic regression discontinuity design." *The Quarterly Journal of Economics*, 125, 215–261.

Chetty, Raj, John N Friedman, and Jonah E Rockoff (2014), "Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood." *The American Economic Review*, 104, 2633–2679.

Cho, H (2017), "Effect of summer heat on test scores: A cohort analysis." *Journal of Environmental Economics and Management*, 83, 185–196.

Currie, Janet and Rosemary Hyson (1999), "Is the impact of health shocks cushioned by socioeconomic status? the case of low birthweight." *American Economic Review*, 89, 245–250.

Davis, Lucas W and Paul J Gertler (2015), "Contribution of air conditioning adoption to future energy use under global warming." *Proceedings of the National Academy of Sciences*, 112, 5962–5967.

Dee, Thomas S, Will Dobbie, Brian A Jacob, and Jonah Rockoff (2016), "The causes and consequences of test score manipulation: Evidence from the new york regents examinations." *NBER Working Paper*.

Dell, Melissa, Benjamin F Jones, and Benjamin A Olken (2014), "What do we learn from the weather? the new climate–economy literature." *Journal of Economic Literature*, 52, 740–798.

Deryugina, Tatyana, Garth Heutel, Nolan H Miller, David Molitor, and Julian Reif (2016), "The mortality and medical costs of air pollution: Evidence from changes in wind direction." Technical report, National Bureau of Economic Research.

Deschênes, Olivier and Michael Greenstone (2011), "Climate change, mortality, and adaptation: Evidence from annual fluctuations in weather in the us." *American Economic Journal: Applied Economics*, 3, 152–185.

Deschenes, Olivier, Michael Greenstone, and Joseph S Shapiro (2017), "Defensive investments and the demand for air quality: Evidence from the nox budget program." *American Economic Review*, 107, 2958–2989.

Durán-Narucki, Valkiria (2008), "School building condition, school attendance, and academic achievement in new york city public schools: A mediation model." *Journal of environmental psychology*, 28, 278–286.

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

Garg, Teevrat, Maulik Jagnani, and Vis Taraz (2017), "Human capital costs of climate change: Evidence from test scores in india."

Gertler, Paul J, Orie Shelef, Catherine D Wolfram, Alan Fuchs, et al. (2016), "The demand for energy-using assets among the world's rising middle classes." *American Economic Review*, 106, 1366–1401.

Goldin, Claudia Dale and Lawrence F Katz (2009), *The race between education and technology*. Harvard University Press.

Graff Zivin, Joshua, Solomon M Hsiang, and Matthew Neidell (2017), "Temperature and human capital in the short-and long-run." *Journal of the Association of Environmental and Resource Economists*.

Graff-Zivin, Joshua and Matthew Neidell (2012), "The impact of pollution on worker productivity." *The American Economic Review*, 102, 3652–3673.

Graff Zivin, Joshua and Matthew Neidell (2013), "Environment, health, and human capital." *Journal of Economic Literature*, 51, 689–730.

Graff Zivin, Joshua and Matthew Neidell (2014), "Temperature and the allocation of time: Implications for climate change." *Journal of Labor Economics*, 32, 1–26.

Hanushek, Eric A and Ludger Woessmann (2012), "Do better schools lead to more growth? cognitive skills, economic outcomes, and causation." *Journal of economic growth*, 17, 267–321.

He, Jiaxiu, Haoming Liu, and Alberto Salvo (2017), "Severe air pollution and labor productivity: Evidence from industrial towns in china."

Heal, Geoffrey and Jisung Park (2016), "Temperature stress and the direct impact of climate change: A review of an emerging literature." *Review of Environmental Economics and Policy*, 10, 347–362.

Hocking, Chris, Richard B Silberstein, Wai Man Lau, Con Stough, and Warren Roberts (2001), "Evaluation of cognitive performance in the heat by functional brain imaging and psychometric testing." *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology*, 128, 719–734.

Hsiang, Solomon M (2010), "Temperatures and cyclones strongly associated with economic production in the caribbean and central america." *Proceedings of the National Academy of sciences*, 107, 15367–15372.

Isen, Adam, Maya Rossin-Slater, and Reed Walker (2017), "Relationship between season of birth, temperature exposure, and later life wellbeing." *Proceedings of the National Academy of Sciences*, 201702436.

Jackson, C Kirabo, Rucker C Johnson, and Claudia Persico (2015), "The effects of school spending on educational and economic outcomes: Evidence from school finance reforms." *The Quarterly Journal of Economics*, 131, 157–218.

Jacob, Brian and Jesse Rothstein (2016), "The measurement of student ability in modern assessment systems." *Journal of Economic Perspectives*, 30, 85–108.

Kahn, Matthew E (2005), "The death toll from natural disasters: the role of income, geography, and institutions." *The Review of Economics and Statistics*, 87, 271–284.

Kjellstrom, Tord and Jennifer Crowe (2011), "Climate change, workplace heat exposure, and occupational health and productivity in central america." *International Journal of Occupational and Environmental Health*, 17, 270–281.

Lafortune, Julien, Jesse Rothstein, and Diane Whitmore Schanzenbach (2016), "School finance reform and the distribution of student achievement." Technical report, National Bureau of Economic Research.

Lim, Chin Leong, Chris Byrne, and Jason KW Lee (2008), "Human thermoregulation and measurement of body temperature in exercise and clinical settings." *Annals Academy of Medicine Singapore*, 37, 347.

Mackworth, Norman H (1946), "Effects of heat on wireless operators." *British journal of industrial medicine*, 3, 143.

Pindyck, Robert S (2013), "Climate change policy: What do the models tell us?" *Journal of Economic Literature*, 51, 860–872.

Roback, Jennifer (1982), "Wages, rents, and the quality of life." *Journal of political Economy*, 90, 1257–1278.

Seppanen, Olli, William J Fisk, and QH Lei (2006), "Effect of temperature on task performance in office environment." *Lawrence Berkeley National Laboratory*.

Sievertsen, Hans Henrik, Francesca Gino, and Marco Piovesan (2016), "Cognitive fatigue influences students' performance on standardized tests." *Proceedings of the National Academy of Sciences*, 113, 2621–2624.

Sinha, Paramita, Martha L Caulkins, and Maureen L Cropper (2017), "Household location decisions and the value of climate amenities." *Journal of Environmental Economics and Management*.

Stocker, Thomas (2014), *Climate change 2013: the physical science basis: Working Group I contribution to the Fifth assessment report of the Intergovernmental Panel on Climate Change*. Cambridge University Press.

Tiebout, Charles M (1956), "A pure theory of local expenditures." *Journal of political economy*, 64, 416–424.

USDHHS (2010), "State indicator report on physical activity." *Washington, DC: US Department of Health and Human Services*.

Tables and Figures

The University of the State of New York
THE STATE EDUCATION DEPARTMENT
Office of State Assessment
Albany, New York 12234

EXAMINATION SCHEDULE: JUNE 2016

Students must verify with their schools the exact times that they are to report for their State examinations.

June 1 WEDNESDAY	June 14 TUESDAY	June 15 WEDNESDAY	June 16 THURSDAY	June 17 FRIDAY	June 20 ° MONDAY	June 21 TUESDAY	June 22 WEDNESDAY	June 23 THURSDAY
9:15 a.m.	**9:15 a.m.**	**9:15 a.m.**	**9:15 a.m.**	**9:15 a.m.**	**9:15 a.m.**	**9:15 a.m.**	**9:15 a.m.**	
Algebra II (Common Core) ♦	RE in Global History & Geography	Living Environment	Algebra I (Common Core)	Physical Setting/ Earth Science Algebra 2/ Trigonometry	RCT in Mathematics* ° Suggested date for administering locally developed tests aligned to the Checkpoint A and Checkpoint B learning standards for languages other than English (LOTE).	Physical Setting/ Chemistry RCT in Global Studies*	RCT in Writing	**RATING DAY**
1:15 p.m.	**1:15 p.m.**	**1:15 p.m.**	**1:15 p.m.**	**1:15 p.m.**	**1:15 p.m.**	**1:15 p.m.**	**1:15 p.m.**	**Uniform Admission Deadlines** Morning Examinations: 10:00 a.m. Afternoon Examinations: 2:00 p.m.
SPECIAL ADMINISTRATION: Integrated Algebra	RE in English Language Arts (Common Core)	RE in U.S. History & Government	Comprehensive English	Geometry (Common Core)	RCT in U.S. History & Government*	Physical Setting/ Physics RCT in Reading*	RCT in Science*	

* Available in Restricted Form only. Each copy of a restricted test is numbered and sealed in its own envelope and must be returned, whether used or unused, to the Department at the end of the examination period.

(a) Sample Regents Exam Schedule

The University of the State of New York
REGENTS HIGH SCHOOL EXAMINATION

ALGEBRA 2/TRIGONOMETRY

Friday, June 19, 2015 — 9:15 a.m. to 12:15 p.m., only

Student Name:_____

School Name:_____

The possession or use of any communications device is strictly prohibited when taking this examination. If you have or use any communications device, no matter how briefly, your examination will be invalidated and no score will be calculated for you.

Use this space for computations.

1 Which list of ordered pairs does *not* represent a one-to-one function?

(1) (1,−1), (2,0), (3,1), (4,2)
(2) (1,2), (2,3), (3,4), (4,6)
(3) (1,3), (2,4), (3,3), (4,1)
(4) (1,5), (2,4), (3,1), (4,0)

2 The terminal side of an angle measuring $\frac{4\pi}{5}$ radians lies in Quadrant

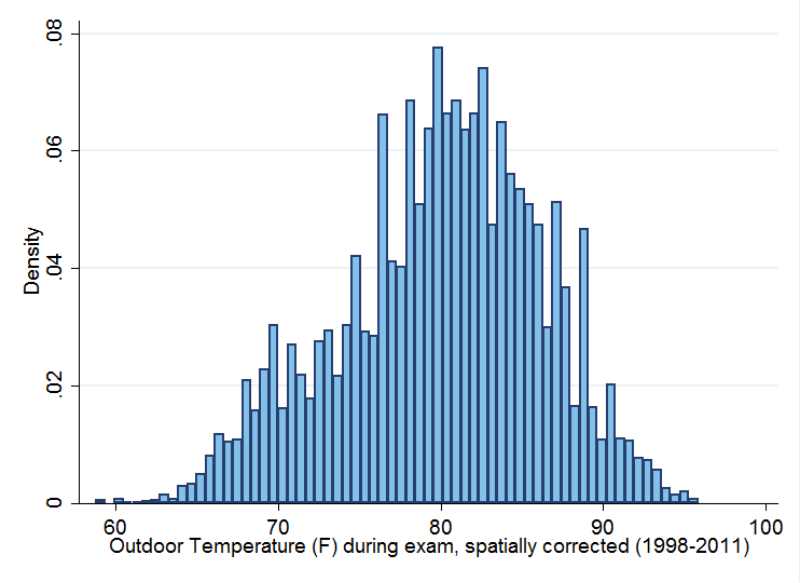
(1) I
(2) II
(3) III
(4) IV

3 If $f(x) = 2x^2 + 1$ and $g(x) = 3x - 2$, what is the value of $f(g(-2))$?

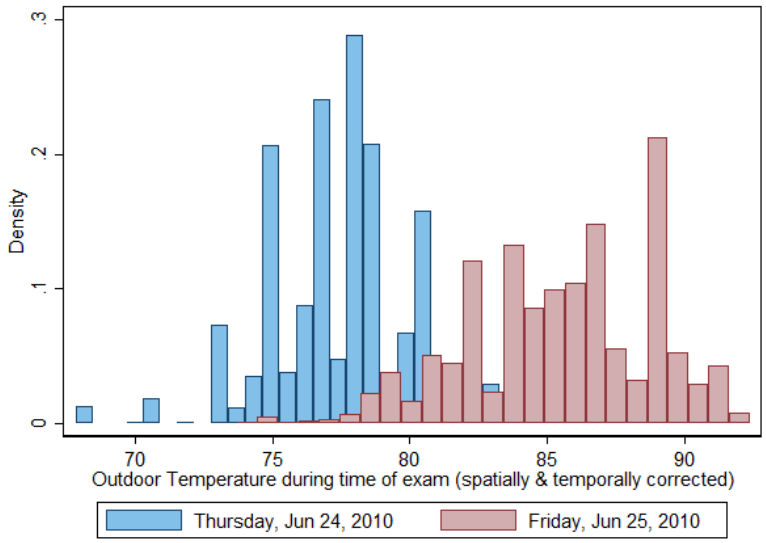
(1) −127
(2) −23
(3) 25
(4) 129

(b) Sample Regents subject exam cover sheet and questions

Figure 1: Sample Exam Schedule and Cover Page



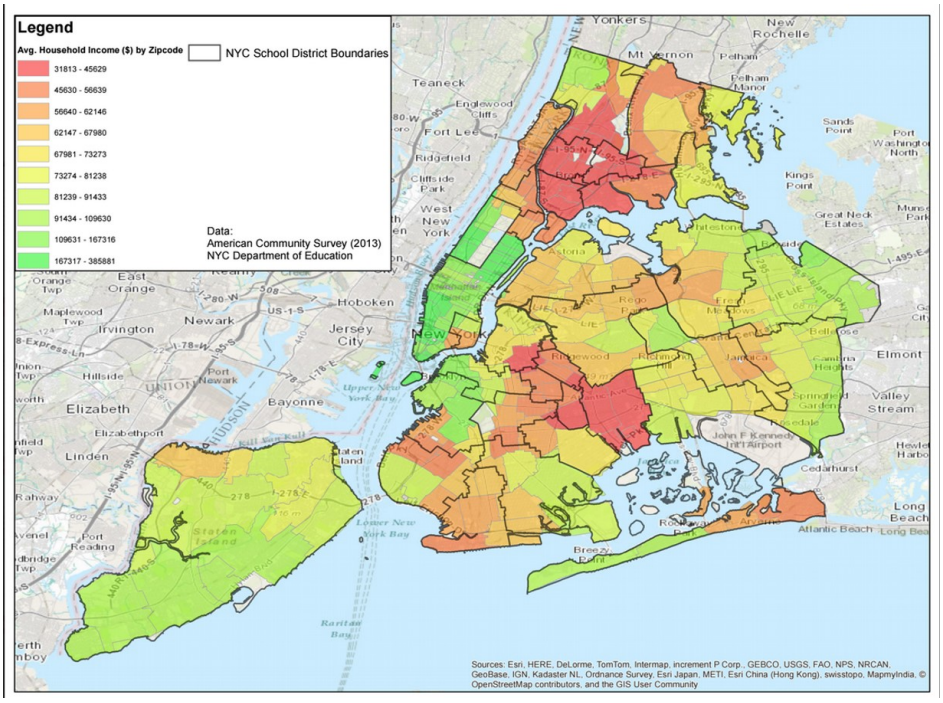
(a) Total variation in exam-time temperature



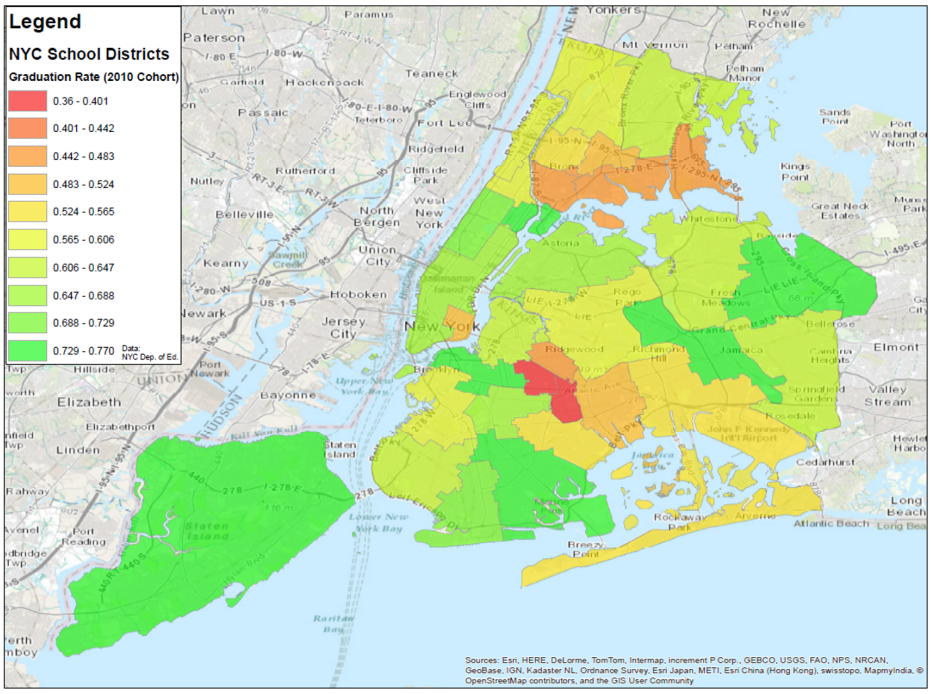
(b) Temperature on two consecutive test days

Figure 2: Short-Run Identifying Variation in Temperature

Notes: This figure illustrates the source of identifying variation for short-run performance impacts of heat exposure. It presents realized exam-time temperatures for (a) all June Regents exams (1999-2011) and (b) for two subsequent days within a Regents exam period – Thursday, June 24th, 2010, and Friday, June 25th, 2010 – inclusive of spatial and temporal temperature corrections. Temperatures are measured at the school level, weighted by number of exam observations by date and time.



(a) Average household income by zip code



(b) Graduation rates by sub-district

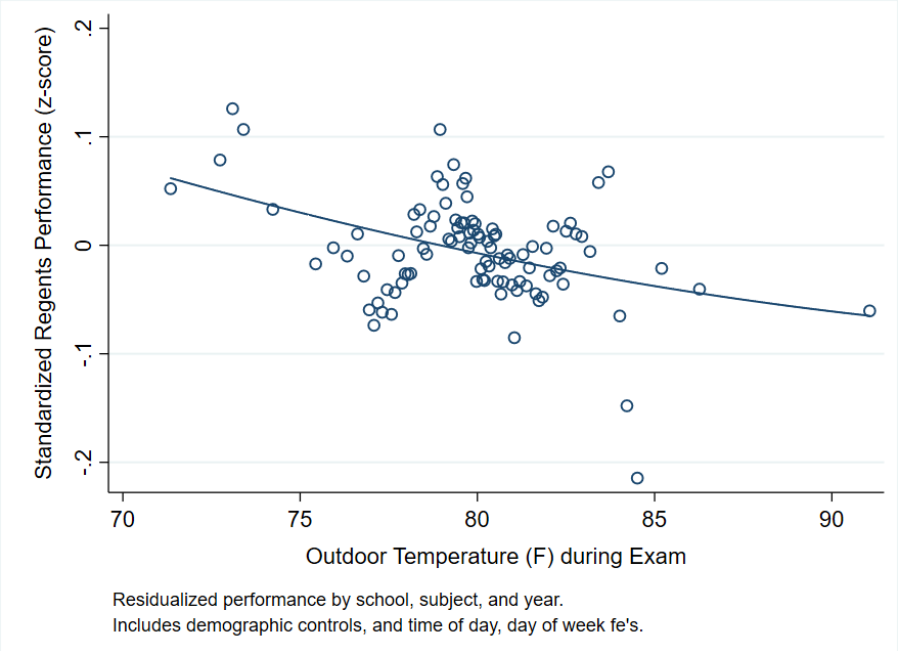
Figure 3: Average Household Income (top) and High School Graduation Rates (bottom)

Notes: Panel (a) presents average household income in 2010 by zip code, with New York City Public School sub-districts super-imposed. Panel (b) presents the average 4-year high school graduation rate of students by sub-district within the New York City Public Schools system.

Ethnicity	Score	Pass	Proficiency	Previous Ability
Asian	74.73	0.78	0.57	0.98
	(16.80)	(0.41)	(0.49)	(1.54)
Black	61.21	0.50	0.23	-0.18
	(17.05)	(0.50)	(0.42)	(1.34)
Hispanic	61.49	0.51	0.24	-0.16
	(17.23)	(0.50)	(0.42)	(1.32)
Multiracial	69.65	0.69	0.44	0.34
	(17.44)	(0.46)	(0.50)	(1.26)
Native American	61.96	0.51	0.26	-0.22
	(18.08)	(0.50)	(0.44)	(1.45)
White	72.92	0.75	0.52	1.02
	(16.78)	(0.43)	(0.50)	(1.56)
Total	64.86	0.57	0.32	0.16
	(17.92)	(0.49)	(0.47)	(1.42)

Figure 4: Summary Statistics by Ethnicity

Notes: Table (4) presents summary statistics for student performance variables. Standard deviations are in parentheses. 'Pass" and "Proficiency" denote the fraction of scores above passing and college proficiency thresholds. Previous ability is measured as average z-scores from standardized math and verbal assessments in grades 3 through 8.



(a) Residualized variation in test performance

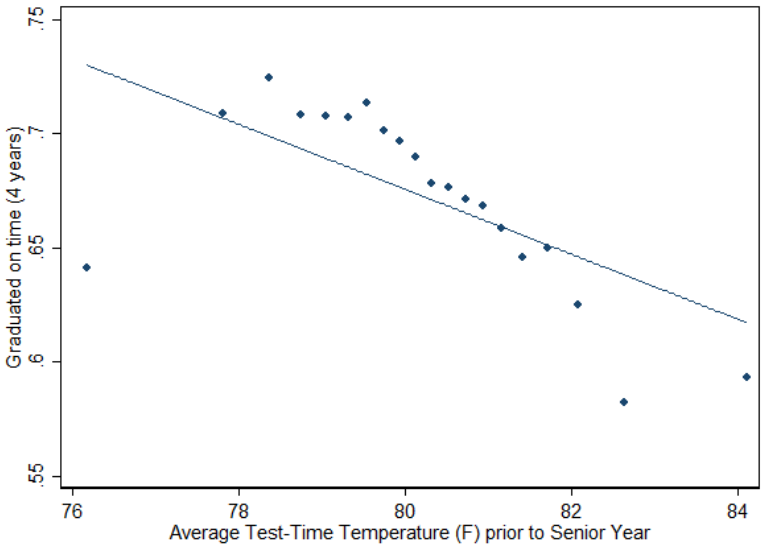
	(1) Z-score	(2) Z-score	(3) Z-score	(4) Z-score
Temperature (F)	-0.00850***	-0.00736***	-0.0102***	-0.0108***
	(0.00231)	(0.00207)	(0.00233)	(0.00226)
Afternoon	-0.0297*	-0.0334**	-0.0180	-0.0156
	(0.0130)	(0.0119)	(0.0142)	(0.0127)
Fixed Effects				
Student by Year	X			
Subject	X	X	X	X
Time of Day, Day of week	X	X	X	X
Student		X		
Year		X	X	
School			X	
School by Year				X
N	3581933	3581933	3581933	3581933
r2	0.774	0.717	0.252	0.271

Robust standard errors in parentheses, clustered at the borough (station) by date-time level.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

(b) Dependent variable is standardized performance by subject

Figure 5: Effects of Heat Exposure on High Stakes Exam Performance

Notes: Panel (a) presents a binned scatterplot of residualized exam performance by percentile of the temperature distribution controlling for school, subject, and year fixed effects. Each dot represents approximately 220,000 exam observations. Panel (b) presents the main regression results. Fixed effects are suppressed in output, and 919,067 singleton observations are dropped. All regressions include controls for daily dewpoint, precip, ozone, and pm2.5.



(a) 4-year graduation status and exam-time temperature

	(1) Graduated	(2) Graduated	(3) Graduated
Avg Exam-Time Temp (°F)	-0.00712***	-0.00758***	-0.00733**
	(0.00173)	(0.00223)	(0.00231)
Number of June exams			0.193***
			(0.00688)
Number of June exams^2			-0.0151***
			(0.000809)
Number of June exams^3			0.000312***
			(0.0000225)
Fixed Effects			
School	X	X	X
Number of June exams	X	X	
Cohort		X	X
N	515192	515192	515192
r2	0.232	0.238	0.236

Robust standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

(b) Impacts on graduation status by regression specification

Figure 6: Persistent Impacts of Short-Run Heat Exposure: Graduation Status

Notes: Panel (a) presents a binned scatterplot of 4-year graduation status by quantile of exam-time temperature distribution. Temperatures are averaged by student for June exam sessions up through senior year. Residual variation after controlling for school and number of exam fixed effects, student-level observables, and weather/air quality controls. In panel (b), the dependent variable is a dummy for whether or not student graduated in four years. All regressions include controls for daily precipitation, ozone, and dewpoint. Fixed effects are suppressed in output.

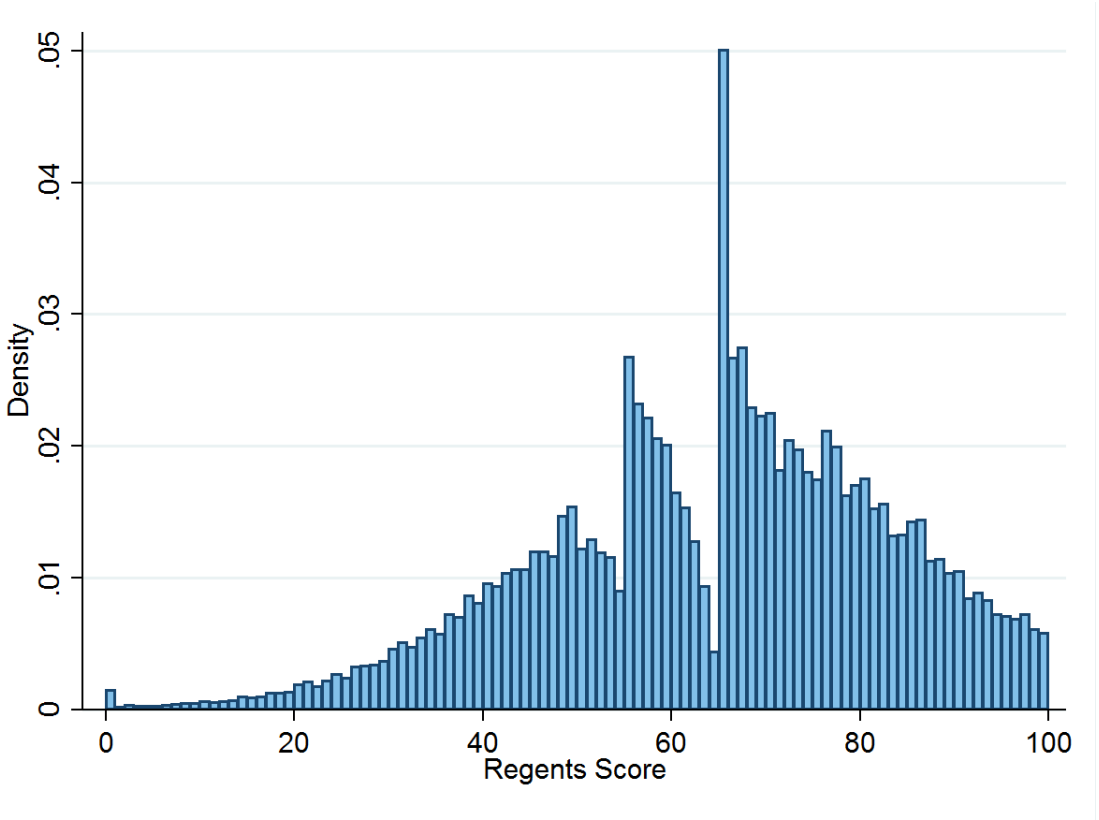
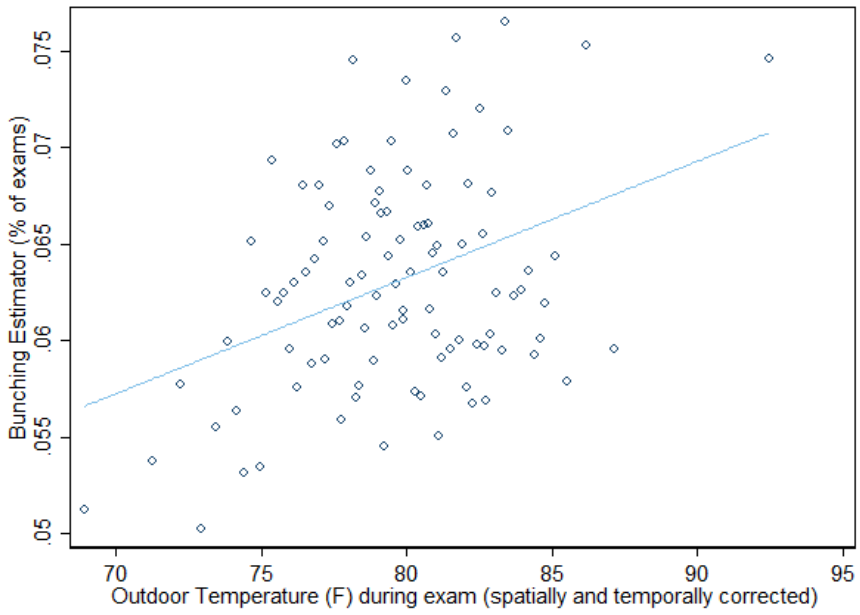
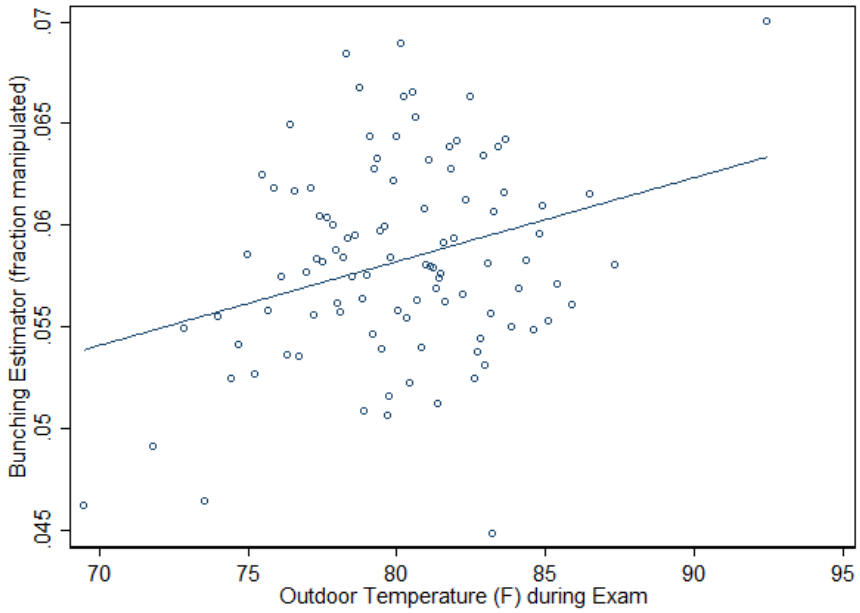


Figure 7: Exam scores exhibit bunching at pass/fail cutoffs, suggesting upward grade manipulation

Notes: This figure presents a histogram of Regents exam scores from June 1999 to June 2011. A large number of observations bunch at the pass/fail cutoffs, scores of 55 and 65 for local and Regents diploma requirements respectively.



(a) Grade Manipulation varies with exam-time temperature by subject, school, and take.



(b) Grade Manipulation expressed as a fraction of scores in manipulable range.

Figure 8: Evidence of Compensatory Responses: Grade Manipulation by Teachers

Notes: Panel (a) presents a binned scatterplot of bunching at the school-subject-date level by quantile of the exam-time temperature distribution, controlling for subject and year fixed effects and daily weather and air quality controls. Panel (b) expresses bunching as a fraction of manipulable scores, to account for potential mechanical correlation between temperature and the number of scores falling in the manipulable zone. Included in the analysis are all June Regents exams in core subjects between 1999 and 2011.