Hot Temperature and High-Stakes Performance

R. Jisung Park

Journal of Human Resources, Volume 57, Number 2, March 2022, pp.
400-434 (Article)

Published by University of Wisconsin Press

Hot Temperature and High-Stakes Performance

R. Jisung Park

ABSTRACT

Despite the prevalence of high-stakes assessments—and the growing likelihood of heat exposure during such assessments—the effect of temperature on performance has not yet been studied in such settings. Using student-level administrative data for the largest public school district in the United States, I provide the first estimates of temperature's impact on high-stakes exam performance and subsequent educational attainment. Hot temperature reduces performance by up to 13 percent of a standard deviation and leads to persistent impacts on high school graduation status, despite compensatory responses by teachers, who selectively upward manipulate grades after hotter exams.

R. Jisung Park is an assistant professor at the University of California, Los Angeles (jisungpark@luskin.ucla.edu). *The author thanks Lawrence Katz, Andrei Shleifer, Robert Stavins, Joseph Aldy, Geoffrey Heal, Raj Chetty, Claudia Goldin, Edward Glaeser, Melissa Dell, Joshua Goodman, Jonah Rockoff, Jeff Miron, Max Auffhammer, Olivier Deschenes, Sarah Reber, Joshua Graff-Zivin, Matt Neidell, and numerous seminar participants at Harvard, Columbia, UCLA, Duke, Maryland, UC Berkeley, Oxford, IZA, Seoul University, the NYC Department of Education, the NBER Summer Institute, and the Bill and Melinda Gates Foundation for valuable comments, as well as two anonymous referees for detailed and highly constructive comments. Nicolas Cerkez, Rodrigo Leal, and Paul Stainier provided excellent research assistance. All remaining errors are the author's own. Funding support from the Harvard Environmental Economics Program, the National Science Foundation, and the Harvard Kennedy School Environment and Natural Resources Program are gratefully acknowledged. The author has no potential conflicts of interest to disclose. IRB approval was obtained from the Harvard University Institutional Review Board (IRB) (Application Number: F19008-104). This paper uses confidential data from the NYC Department of Education. The data can be obtained by filing a request directly with the NYC Department of Education (RPSGresearch@schools.nyc.gov).*

[Submitted June 2018; accepted January 2020]; doi:10.3368/jhr.57.2.0618-9535R3

JEL Classification: I21, I26, J24, O18, Q51, Q54, and Q56

ISSN 0022-166X E-ISSN 1548-8004 © 2022 by the Board of Regents of the University of Wisconsin System

Supplementary materials are freely available online at: http://uwpress.wisc.edu/journals/journals/jhr-supplementary.html

Jisung Park https://orcid.org/0000-0001-5065-5927

I. Introduction

Cognitively intensive assessments such as college entrance exams or job interviews have become a routine fixture of modern economies, due in large part to the increasing importance of cognitive skills in the workplace.[1] Such assessments often take place in high-stakes environments, where performance over a relatively short window can have lasting educational and economic consequences and where rescheduling may be costly due to coordination costs or other frictions.[2] Given potential welfare consequences, it is important to understand whether the physical conditions under which such assessments take place can affect realized performance. This is especially true if the playing field may not be level or may be changing over time.

Temperature stress may pose a particular concern for such assessments—and cognitively intensive tasks generally—given well-documented correlations between national income and historical climate (Dell, Jones, and Olken 2012), personal income and air-conditioning (Biddle 2008; Gertler et al. 2016), and income and the predicted increase in warming due to the global climate externality (Carleton et al. 2019). Within the United States, individuals face vastly different risk of heat exposure, and such risk appears to be changing differently over time. For instance, prior to 1990, Los Angeles, Houston, and Minneapolis averaged 36, 105, and 7 days per year above 90°F, respectively. While Los Angeles and Houston have each added approximately 10–15 days per year above 90°F since 1990, Minneapolis has actually cooled slightly during that period.[3]

Evidence from the lab suggests that temperature may affect realized human performance on both cognitive and physical tasks (Seppanen, Fisk, and Lei 2006). However, it is unclear to what extent these effects generalize to economically meaningful environments, in part because disutility of effort and concentration are likely increasing in temperature stress and because the stakes in existing studies are relatively low.[4] On the one hand, existing findings could be indicative of significant impacts on performance even in settings where individuals have an incentive to minimize realized impacts of temperature through behavioral responses. On the other hand, it is possible that observed effects are largely due to diminished attention or effort, given artificial environments that bear little resemblance to more welfare-relevant settings. If the latter is the case, simply

[1] For reviews of the literature on the returns to cognitive skill in the workplace, see Goldin and Katz (2009), Acemoglu and Autor (2011), and Hanushek and Woessmann (2012).

[2] The list of standardized exams that determine degree eligibility or impose hurdles to further schooling is long and includes 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. 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 or stigma costs of having to retake the exam. Similarly, job interviews or athletics trials 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 that makes rescheduling costly.

[3] Climate models predict substantially different rates of warming across regions over the coming decades, both within the United States and internationally, potentially widening existing differences in test-taking conditions (Chambwera et al. 2014).

[4] This is true even in the few studies that assess cognitive performance outside of experimental settings. For instance, Graff Zivin, Hsiang, and Neidell (2018) use data from ten-minute voluntary survey assessments conducted as part of the NLSY in the United States. Garg, Jagnani, and Taraz (2017) uses data from a similarly low-stakes survey measure in India. Work on temperature in schools, such as Durán-Narucki (2008), is often cross-sectional or case study based, making causal attribution difficult.

extrapolating existing dose–response relationships to real-world policy environments may mischaracterize welfare and policy implications.

In seminal work, Ebenstein, Lavy, and Roth (2016) study the effect of a different environmental hazard—air pollution—on student outcomes in Israel and demonstrate that the effects on exam performance and subsequent labor market outcomes can be substantial. However, because of their setting—namely, a unique institutional environment that mandated all test centers in Israel to have air-conditioning—they are unable to document effects of temperature. It seems highly unlikely that all high-stakes assessments globally take place in similarly climate-controlled environments.[5] Temperature and air quality effects may entail distinct policy responses, due to, for instance, differences in the scale of the externality that causes them (local versus global) or differences in potential adaptive responses (feasibility of rescheduling based on forecasts, availability and cost of adaptive technologies). Moreover, to the extent that cognitively demanding activities extend beyond formal assessments, understanding the influence of temperature on realized human performance may be important for understanding the potential magnitude of the climate externality.

To circumvent these difficulties, this study combines local daily weather data with test scores of one million U.S. students taking synchronized high-stakes exams over the course of several days each year. These student-level exam records are linked to data on subsequent educational attainment in the form of high school graduation and diploma status. Drawing on administrative records from the largest public school district in the United States (New York City), this represents to my knowledge the most comprehensive data set assembled to date aimed at assessing the effect of temperature on student performance.[6] Student fixed-effects regressions identify the causal impact of hotter temperature on exam performance and eventual educational attainment by exploiting quasi-random variation in temperature for an individual across multiple exams. Causal identification therefore rests on the premise that within-student variations in day-to-day temperature are not correlated with unobserved determinants of educational performance.

To fix ideas, I present a simple model that captures students' marginal disutility of effort as well as the returns to performance as a function of temperature, nesting the two potential drivers of a decline in performance. The key implication is that, as the stakes of a given assessment are raised, it becomes less likely that observed associations between performance and temperature are driven solely by reductions in effort, suggesting that hotter temperature may affect realized performance even in highly welfare-relevant contexts. The high-stakes empirical setting studied here effectively shuts down the extensive margin response (that is, absenteeism), limiting potential selection bias, and makes it far more likely than previous studies that the observed effect is not simply a function of reduced effort or concentration.

[5] Even in the United States, air-conditioning is far from complete, particularly in schools, where classroom air-conditioning penetration can be below 50 percent in some districts (Park et al. 2020). In poorer developing countries such as India, the rate of air-conditioning penetration can be below 10 percent (Davis and Gertler 2015). Numerous media reports suggest test centers can lack air-conditioning. For instance, see https://www.theatlantic.com/education/archive/2016/02/who-benefits-from-the-new-summer-sat/459972/ (accessed August 18, 2021).

[6] Unlike Ebenstein, Lavy, and Roth (2016), I document that air-conditioning is far from complete in this setting. Based on available building-level engineering audits, I estimate that fewer than 70 percent of NYC public schools had working air-conditioning during the study period (1998–2012).

The first main finding is that hot temperature reduces student performance substantially. Students take a series of mandatory exams in June, which are spread over the course of two weeks and feature harmonized timing and predetermined testing sites. Because I am able to link multiple exam records for each student and school location, and I can 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 student performance. A one standard deviation increase in exam-time temperature (+6.2°F) reduces performance by approximately 5.5 percent of a standard deviation. This implies that taking an exam when outdoor temperatures are 90°F reduces performance by approximately 13 percent of a standard deviation relative to a temperature of 75°F.

Second, I find that hot temperature during an exam can lead to persistent impacts on educational attainment. Consistent with inflexible exam administration (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 average student, taking an exam on a 90°F day results in a roughly 10 percent lower likelihood of passing a particular subject (for example, Algebra), which in turn reduces the probability of graduation. A one standard deviation increase in average exam-time temperature (+4.4°F) reduces a student's likelihood of graduating on time by approximately three percentage points, or 4.5 percent relative to a mean on-time graduation rate of 68 percent.

Consistent with these persistent consequences, I find evidence of compensatory responses by teachers who appear to selectively manipulate grades upward for students who experienced hot exams. Using a bunching estimator at pass–fail cutoffs adapted from previous work (Dee et al. 2019), I show that grade manipulation is significantly more frequent for exams taken under hot conditions. The amount of excess bunching is beyond what would result from mechanical correlation between temperature—which shifts more scores toward the manipulable zone—and the proportion of manipulable scores that are actually manipulated, suggesting that, consciously or not, teachers are responding to hot exam-day temperatures. These patterns are consistent with the high-stakes setting; they are also consistent with benevolently motivated teachers who attempt to compensate for the suboptimal adaptation investment by engaging in second-best responses.

This work is the first to examine the contemporaneous impact of hot temperature on exam performance in a setting where the stakes are economically meaningful. It builds on a growing literature that examines the causal impact of temperature on economic outcomes, such as health and labor supply (Deschênes and Greenstone 2011; Graff Zivin and Neidell 2014; Barreca et al. 2016), and a smaller literature on temperature and student outcomes (Garg, Jagnani, and Taraz 2017; Graff Zivin, Hsiang, and Neidell 2018).[7] Such findings provide more compelling evidence than previous studies that the effects of temperature on human performance are not driven by reductions in effort or concentration and that the net welfare impacts of elevated

[7] For reviews of the economic literature on weather fluctuations on economic activity and heat exposure on labor-related outcomes, see Dell, Jones, and Olken (2014) and Heal and Park (2016), respectively. In a related paper, Park et al. (2020) explore the effect of cumulative heat exposure during the school year (that is, hotter weekdays during the school year) on realized learning, as opposed to the effects of temperature on test scores.

temperature may be substantial. It is also the first to document persistent impacts of heat exposure in school settings on longer-term educational outcomes and among the first to document ex post adaptation responses to temperature shocks.[8]

Potential implications for welfare and policy are discussed in greater detail in the conclusion. In brief, the findings suggest that students taking standardized exams across varying climates may not be on an equal playing field. Given well-documented correlations between climate and income, it is possible that lower-income students may be subtly but systematically penalized on standardized assessments, such as the SAT or ACT. Such equity concerns are likely of heightened policy relevance given the additional correlation between expected extreme heat events and income or race.[9] While it may be possible for the timing of exams to be adjusted in response to climatic conditions, administering standardized exams separately may involve nontrivial administrative costs, due to, for instance, concerns over cheating. An alternative possibility may be that providing uniformly climate-controlled test centers represents a relatively low-cost means of further reducing longstanding achievement gaps across race and income.

The rest of this paper is organized as follows. Section II presents relevant stylized facts and a simple conceptual framework that guides the empirical analysis. Section III describes the data and institutional context and presents key summary statistics. Section IV presents the results for short-run exam performance and sensitivity analyses. Section V presents results on longer-run educational attainment, and Section VI presents evidence consistent with compensatory investments by teachers. Section VII discusses implications and concludes.

II. Background and Conceptual Framework

A. 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 et al. 2016). 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, Byrne, and Lee 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 75°F (Albouy et al. 2016).

[8] There are two studies that document longer-run consequences of heat exposure on human capital-related outcomes. Isen, Rossin-Slater, and Walker (2017) looks at heat shocks in utero and finds negative impacts on wages later in life, and Cho (2017) explores the effect of summertime heat exposure on exam performance in November. The finding that transitory environmental conditions during exams can have persistent educational and economic consequences echoes findings from Ebenstein, Lavy, and Roth (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.

[9] Lower-income individuals and racial minorities are less likely to have air-conditioning at home or at school (Biddle 2008; Park et al. 2020) and are more likely to live in areas with fewer environmental amenities, such as urban greenspace, which reduce heat island effects, for instance due to residential sorting (Tiebout 1956; Christensen and Timmins 2018).

Existing studies of the effect of temperature on cognitively demanding tasks fall broadly 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 (quasi-)experimental settings, where external validity of observed dose–response relationships is unclear (Mackworth 1946; Seppanen, Fisk, and Lei 2006).[10] For instance, Graff Zivin, Hsiang, and Neidell (2018) study in-home survey data for children in roughly 8,000 U.S. households who took part in the NLSY. While they find that hot temperature on the day of the survey reduces math (but not reading) performance, these short assessments carry little if any weight, making it difficult to know whether the effects generalize to more economically meaningful environments, such as the classroom or the workplace. Garg, Jagnani, and Taraz (2017) study the effect of temperature on similarly low-stakes cognitive assessments administered to Indian primary and secondary school students. In their setting, it is likely that both effort reduction and poor nutrition may be contributing mechanisms, a possibility that is bolstered by the finding that heat's effects are most pronounced during the growing season.[11]

Empirical evidence of persistent impacts of transitory temperature shocks on educational attainment in school settings—where policy responses may be more directly applicable—is limited, despite suggestive evidence from studies of in-utero exposure (Currie and Hyson 1999; Isen, Rossin-Slater, and Walker 2017).

B. *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 specific skills. Suppose that the application of 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 $e \in [0,1]$, as well as on ambient temperature $a \in [0,1]$. In the case of e, 1 denotes maximal effort. For a, 1 denotes a physically uninhabitable ambient temperature, and 0 denotes the ideal temperature. 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.

[10] Previous research has documented effects of temperature on other related outcomes, including on mortality, morbidity, and labor productivity (Hsiang 2010; Deschênes and Greenstone 2011; Deryugina and Hsiang 2014; Somanathan et al. 2018). 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. Moreover, most of these studies asses physical occupations (for example, manufacturing) or low-skilled cognitive tasks (for example, call center operation), which may or may not provide applicable insights for understanding the effect of environmental conditions on knowledge-intensive cognitive tasks.

[11] In addition to the potential for selective sorting based on unobservable student characteristics, survey-based analyses, such as Graff Zivin, Hsiang, and Neidell (2018) or Garg, Jagnani, and Taraz (2017), face an additional challenge due to the fact that hot temperature may lead to systematic biases in reporting. 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 (accessed August 18, 2021). While they attempt to measure effort by looking at time to survey completion, it is possible for respondents to vary intensity of exertion—which is far more difficult to measure—without varying time to survey completion.

Performance on cognitively demanding tasks can be expressed as $y(e;a,h)$, such that effort and ambient environmental conditions jointly determine the realized performance for a given stock of human capital. Let us define $y(e;a,h)$ so that $\partial y/\partial e > 0$ and:

$$(1) \quad \lim_{\substack{e\to 1\\ a\to 0}} y(e;\ a,\ h) = h$$

In other words, maximal effort and ideal thermal conditions are required to perform at one's peak capacity h.

Individuals derive utility $U(x,p)$ from consuming some composite good x, and they experience disutility from physical discomfort $p : U_x > 0$ and $U_p > 0$. Importantly, suppose that physical discomfort p is increasing in effort e and increasing in thermal stress a. The representative individual's utility maximization problem is:

$$(2) \quad \max_e \ U[x, p(e, a)] \ni x = w[y(e; a, h)]$$

where w denotes wage income. Wages depend on realized performance, either because they represent the return to human capital in the labor market (where y is used as a signal of h) or because workers are paid a piece rate depending on marginal product of labor. In the context of formal assessments, w is always increasing in y—a good test score helps labor market returns. But different assessments have different w functions depending on the stakes involved; $\mathrm{d}w/\mathrm{d}y$ may be steeper for a college entrance exam compared to a short quiz. This implies that the marginal disutility of effort is increasing in environmental stress.[12]

For simplicity, I abstract away from investments that may reduce experienced temperature (for example, air-conditioning) 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 presented as a static framework for simplicity, the intuition of the model extends naturally to settings where incremental changes in performance in one period can have persistent ramifications for wages in many subsequent periods.

Let e^* be the level of effort that maximizes U. Substituting $w[y(e;a,h)]$ for x and setting $\mathrm{d}U/\mathrm{d}e = 0$ yields the following first-order condition:

$$(3) \quad \frac{\mathrm{d}U}{\mathrm{d}e^*} = U_x \frac{\mathrm{d}w}{\mathrm{d}y}\frac{\partial y}{\partial e^*} + U_p \frac{\partial p}{\partial e^*} = 0$$

The individual chooses effort to balance the trade-off between marginal utility of realized performance, which operates through the labor market, and marginal disutility of physical discomfort—subject to the economic stakes involved ($\mathrm{d}w/\mathrm{d}y$).

Equation 3 implicitly defines optimal effort e^* as a function of environmental conditions and other parameters. We can also express the total derivative of performance with respect to ambient environmental conditions as:

$$(4) \quad \frac{\mathrm{d}y}{\mathrm{d}a} = \frac{\partial y}{\partial e^*}\frac{\mathrm{d}e^*}{\mathrm{d}a} + \frac{\partial y}{\partial a}$$

[12] 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, Byrne, and Lee 2008).

The realized change in performance is thus a combination of two terms: $\partial y/\partial a$, which describes the direct effect of elevated temperature on cognitive performance, and $\frac{\partial y}{\partial e^*}\frac{\mathrm{d}e^*}{\mathrm{d}a}$, which is the change in performance due to changes in effort. Equation 4 suggests that empirical estimates of $\mathrm{d}y/\mathrm{d}a$, even when utilizing exogenous variation in ambient temperature a, will be a combination of these two effects. Because it is often very difficult to measure $\frac{\partial y}{\partial e^*}\frac{\mathrm{d}e^*}{\mathrm{d}a}$ and $\partial y/\partial a$ separately, it is important to understand how the empirically identifiable object $\mathrm{d}y/\mathrm{d}a$ may depend on the setting in which it is estimated.

Rearranging Equation 3, we get:

$$(5) \qquad \frac{\partial y}{\partial e^*} = \frac{-U_p \dfrac{\partial p}{\partial e^*}}{U_x \dfrac{\mathrm{d}w}{\mathrm{d}y}}$$

which is positive, since all terms on the right-hand side are positive except U_p. Substituting Equation 5 into Equation 4, we can see that:

$$(6) \qquad \frac{\mathrm{d}y}{\mathrm{d}a} = \frac{-U_p \dfrac{\partial p}{\partial e^*}}{U_x \dfrac{\mathrm{d}w}{\mathrm{d}y}} \frac{\mathrm{d}e^*}{\mathrm{d}a} + \frac{\partial y}{\partial a}$$

Equation 6 suggests that the empirically observed reduced form variation in test performance, $\mathrm{d}y/\mathrm{d}a$, will depend on the marginal returns to effort, which are a function of the economic stakes. All else equal, $\mathrm{d}y/\mathrm{d}a$ will be weakly higher (less negative) in high-stakes settings. Formally we can show that $\mathrm{d}\frac{\mathrm{d}y}{\mathrm{d}a}\Big/\mathrm{d}\frac{\mathrm{d}w}{\mathrm{d}y} > 0$ as long as (i) $\mathrm{d}e^*/\mathrm{d}a < 0$ or (ii) $\mathrm{d}e^*/\mathrm{d}a < 0$ and $\mathrm{d}e^*/\mathrm{d}a$ is increasing elastically in $\mathrm{d}w/\mathrm{d}y$.[13] As the stakes of any given assessment are raised, the effect of ambient temperature on realized performance will likely be less negative, as long as the stakes are high enough to override the direct disutility cost of extra effort.

C. Implications for Empirical Analyses

One implication of the model is that, from the perspective of welfare and policy analysis, it is important to estimate the responsiveness to temperature in settings where individuals face meaningful economic incentives. Uncovering $\mathrm{d}y/\mathrm{d}a$ in a setting where the

[13] The first condition refers to situations where the net effort response to hotter temperature is negative and suggests that raising the stakes will lead to less effort reduction. The second condition alternatively refers to situations where the stakes are already high enough that effort responses to increased temperature are positive and states that $\mathrm{d}\frac{\mathrm{d}y}{\mathrm{d}a}\Big/\mathrm{d}\frac{\mathrm{d}w}{\mathrm{d}y} > 0$ as long as the net effect of higher stakes on effort is not offset by myopic sensitivity to increasing disutility of effort under hotter temperatures. Both conditions seem plausible in most settings. Consider, for instance, a college entrance exam. If adverse test-taking conditions can nudge a student on the margin of qualifying to a top-tier university to a second-tier group, and if employers use university rankings 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 that the individual's tradeoff between future consumption and current disutility at the margin would not be declining in the importance of the exam, at least for the duration of the assessment.

stakes are similar to or higher than the median school or workplace setting would likely provide a conservative estimate of the average net-of-effort-reallocation effect and provide the policymaker with more confidence that the underlying impacts are welfare relevant. Moreover, if an empirical analysis of high-stakes settings finds $dy/da < 0$, this might 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.[14] Of course, there may be important institutional differences in the relative costs of and constraints to defensive investments, which would ideally be taken into account, but for which there is as yet limited research.

III. Institutional Setting, Data, and Summary Statistics

A. New York City High Schools: High-Stakes Exams

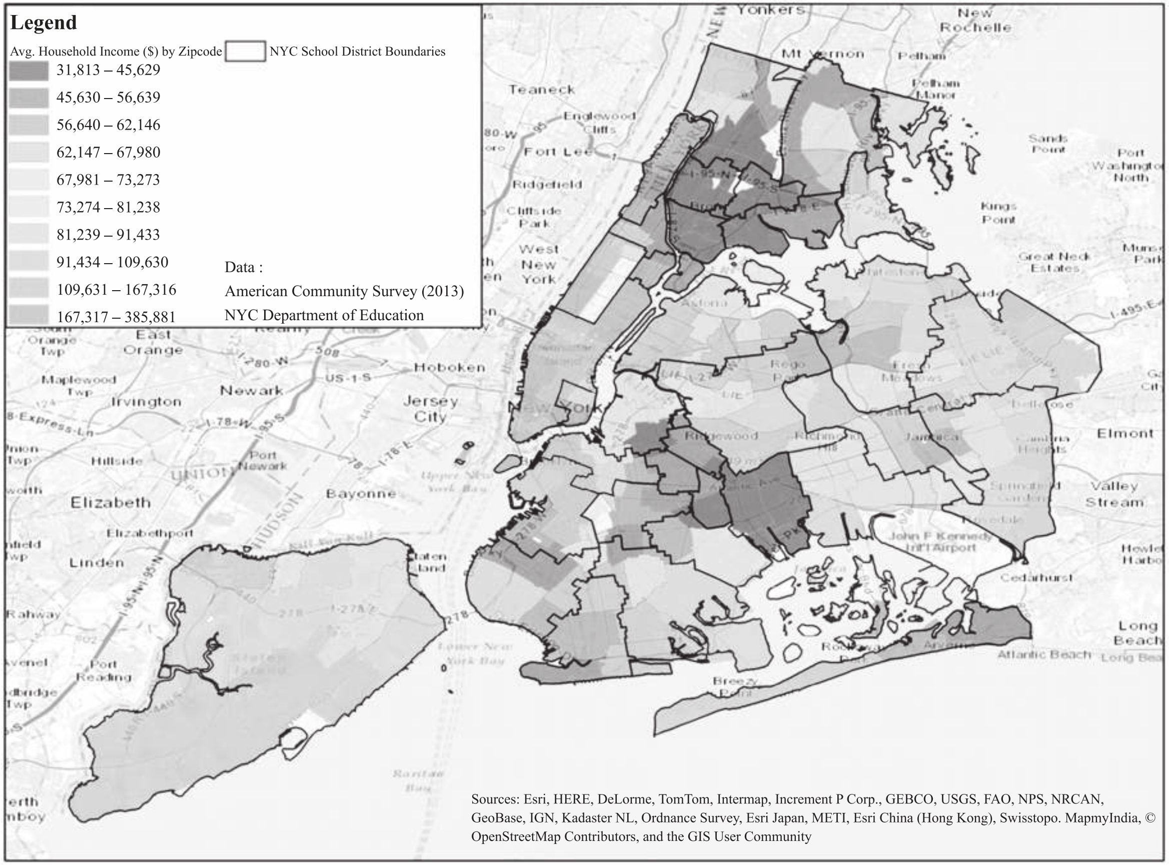
The New York City public school system is the largest in the United States, with more than one million students as of 2012. Each June, these students take a series of high-stakes exams called "Regents Exams," which are standardized subject assessments administered by the New York State Education Department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

Regents Exams can 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. The core subject areas are English, Mathematics, Science, U.S. History and Government, and Global History and Geography. In addition, local universities including City University of New York (CUNY), use strict Regents score cutoffs in the admissions process—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 four-year graduation rate, at 68 percent, is comparable to other large urban public school districts and suggests 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 subdistricts have four-year graduation rates as low as 35 percent per year (Figure 1).

The vast majority of students take their Regents Exams during a prespecified two-week window at the end of June. The dates, times, and locations for each of these exams are fixed more than a year in advance by the state education authority (NYSED) and synchronized across schools in the New York City (NYC) public school system to prevent cheating. Each assessment is approximately three hours long, and exams are administered either in the morning at 9:15 a.m. or in the afternoon beginning at 1:15 p.m. All exams are taken at the student's home school unless they require special accommodations. Students who fail their exams are required to attend summer school, which occurs in July and August. Figure 2 provides a sample exam schedule and cover sheet.

[14] 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 (89.6°F) 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.

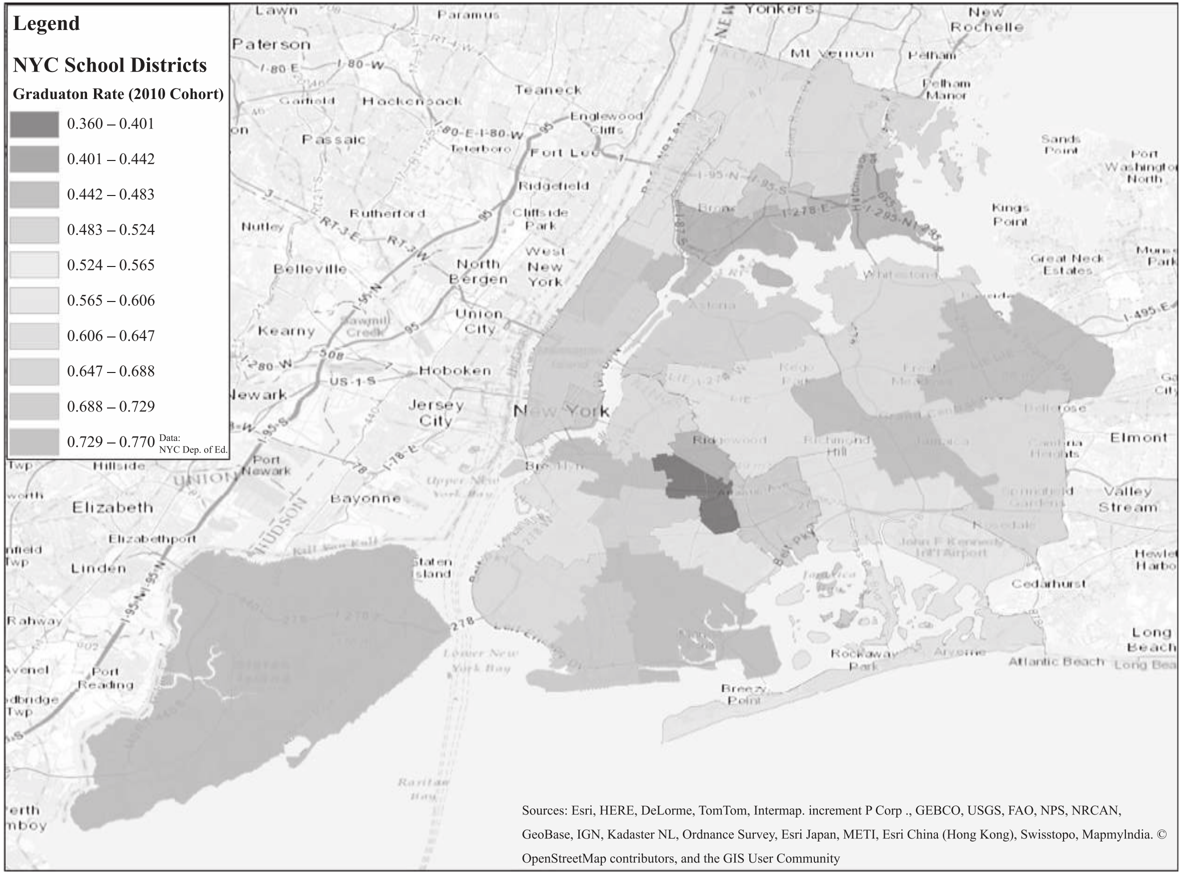


Sources: Esri, HERE, DeLorme, TomTom, Intermap, Increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), Swisstopo. MapmyIndia, © OpenStreetMap Contributors, and the GIS User Community

Figure 1

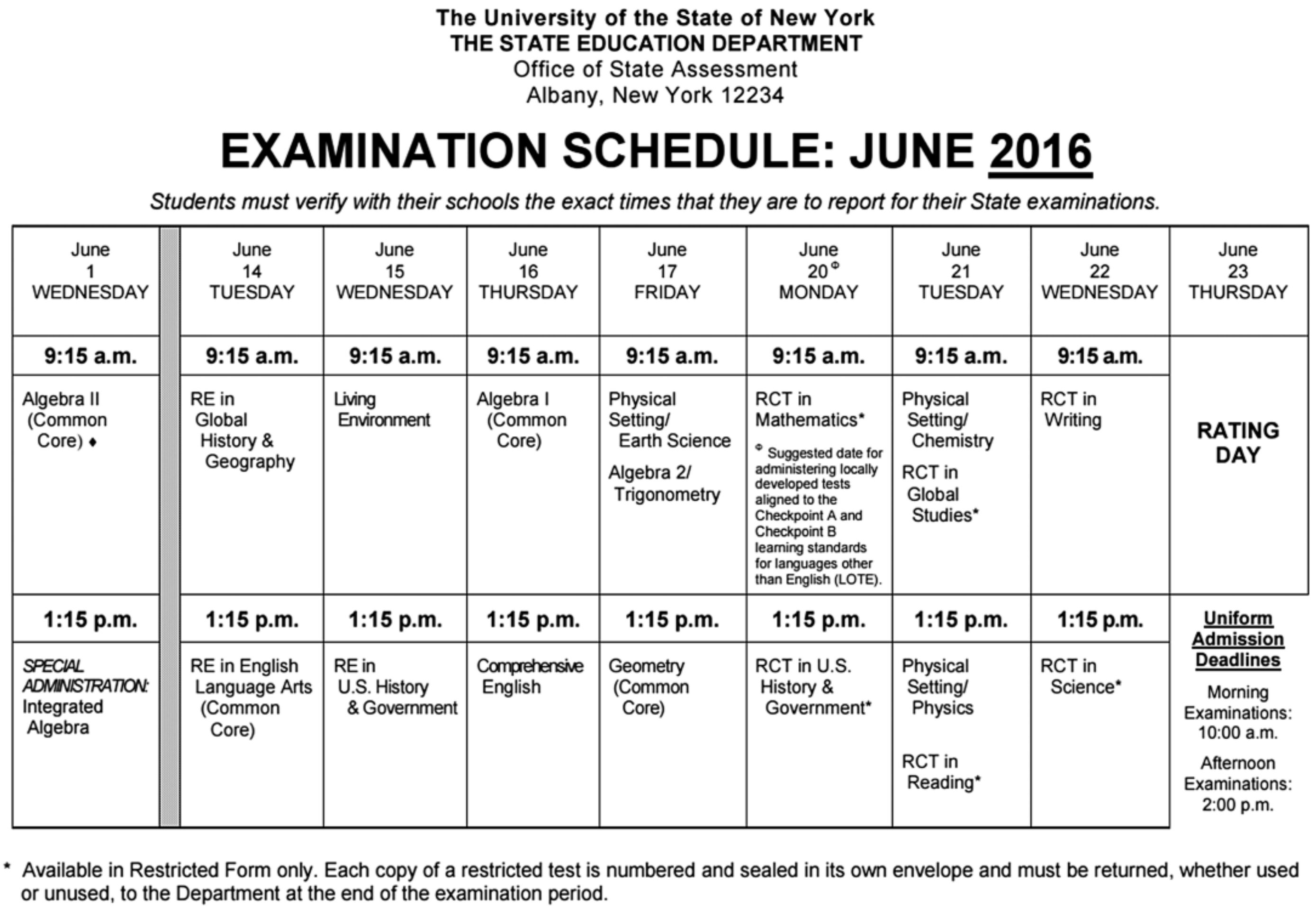
Average Household Income and High School Graduation Rates

Notes: The top panel presents average household income in 2010 by zip code, with New York City Public School subdistricts superimposed. The bottom panel presents average four-year high school graduation rates of students by subdistrict within the New York City Public Schools system.



Sources: Esri, HERE, DeLorme, TomTom, Intermap. increment P Corp ., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), Swisstopo, MapmyIndia. © OpenStreetMap contributors, and the GIS User Community

Figure 1 (*continued*)



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

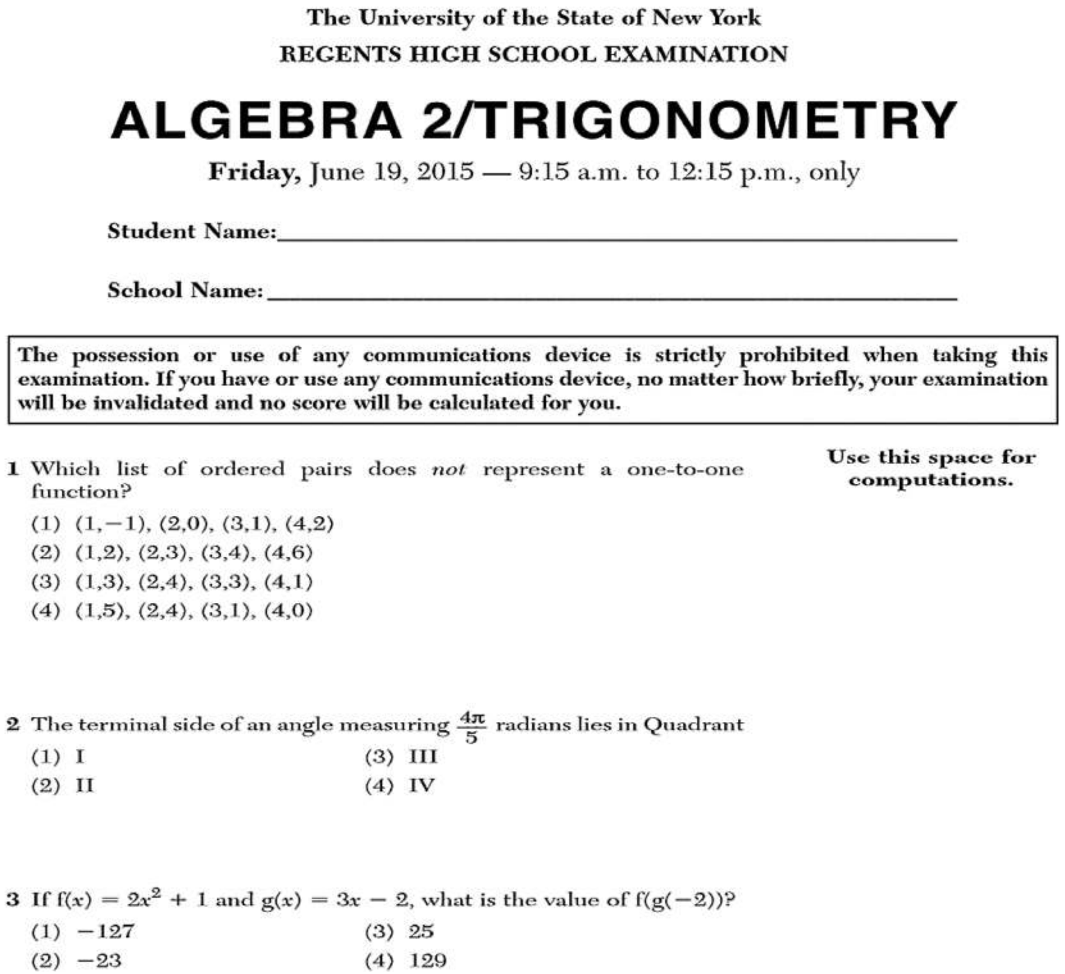
EXAMINATION SCHEDULE: JUNE 2016

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 ᶲ 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.**	**RATING DAY**
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* ᶲ 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	
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.

Figure 2
Sample Regents Schedule and Cover Page



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

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

(1) −127 (3) 25
(2) −23 (4) 129

Figure 2 (*continued*)

B. Student Data

I obtain individual exam-level information from administrative data provided by the New York City Department of Education (NYC DOE). These include records for the universe of NYC public high school students eligible to take Regents Exams over the period 1999–2011. Information on exam dates comes from a web scrape of archived 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 ten-digit student identifiers. These records include cohort and school information, as well as graduation and diploma status.

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 is designed in such a way that is not intended to generate a curve based on realized scores. I use standardized performance across all June Regents Exams as the primary measure of exam performance in this study, though the results are robust to using scale scores, or scores standardized by subject and year. While centrally administered, exams were locally graded by committees of teachers from the students' home schools, usually on the evening of the associated subject exam.

C. Weather Data

Weather data come from the National Oceanic and Atmospheric Administration's Daily Global Historical Climatology Network, which provides daily temperature, precipitation, and dew point information from a national network of weather stations over the period 1950–2014. I take daily minimum and maximum temperature as well as daily average precipitation and dew point readings from the five official weather stations in the NYC area that provide daily data for the entirety of the sample period (1998–2011).

While the true explanatory variable of interest is the classroom temperature experienced by students during an exam, what is recorded in the station data are ambient outdoor temperatures at weather stations that can in some cases be a few miles away. Given potential variation in urban microclimates (Rosenzweig, Solecki, and Slosberg 2006), there may be substantial measurement error in temperature. If classical, such measurement error would attenuate the coefficient estimates toward zero.

In an attempt to reduce measurement error, I perform two spatial and temporal imputation procedures. First, I impute test-time temperature—for instance, average outdoor temperature between 9:15 a.m. and 12:15 p.m. for morning exams—by fitting a fourth-order polynomial in hourly temperature, using diurnal temperature gradients implied by nighttime minimum and daytime maximum temperatures on consecutive days. This allows exams taken in the afternoon to receive a different temperature treatment from those taken in the morning.

Second, I match schools to the nearest weather station (one for each of the five boroughs) and use satellite reanalysis data to assign spatial correction factors by school. The latter procedure allows a school in the heart of Hell's Kitchen, which likely experiences additional urban heat island effects due to the density of structures and paved surfaces, to receive a different temperature treatment from schools in the same borough (Manhattan) that border large bodies of water, such as the Hudson river, or green-space, such as Central

Table 1
Summary Statistics

	Score	Pass	Proficiency	Previous *z*-Score
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)

Notes: Table 1 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 3–8.

Park. The direction and overall magnitude of the results reported below are not sensitive to either of these corrections, though the corrections appear to reduce standard errors slightly. Given existing evidence on the impact on air quality on student performance, I also include controls for $PM_{2.5}$ and ozone, taken from EPA monitoring data from Manhattan.

D. Summary Statistics

The final working data set 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 1 presents summary statistics for the key outcome variables that form the basis of this analysis. The student body is 40 percent Latino, 31 percent African American, 14 percent Asian, and 13 percent white, with approximately 78 percent of students qualifying for federally subsidized school lunch. On average, students take seven June Regents Exams over the course of their high school careers and are observed in the Regents data set for roughly two years, though some underachieving students are observed for more than four years, as they continue to retake exams upon failing.

Fewer than 0.2 percent 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

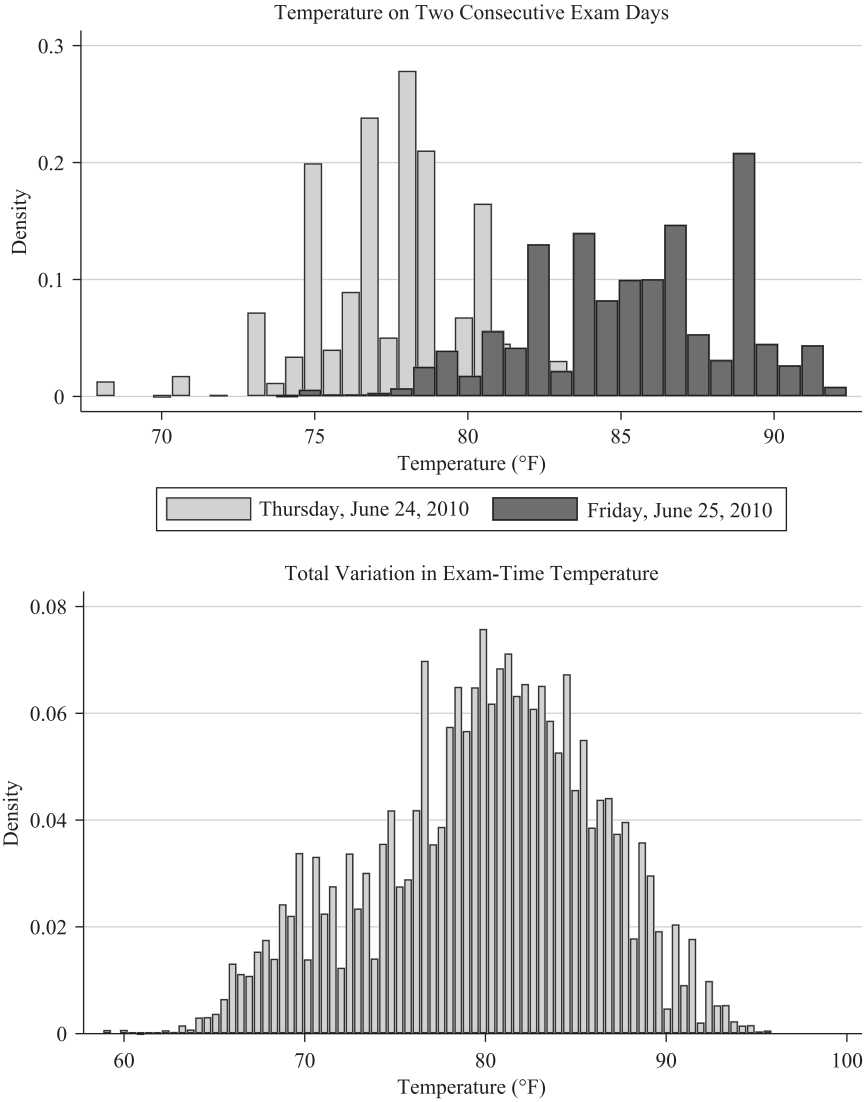


Figure 3

Temperature Variation

Notes: This figure illustrates the source of identifying variation in exam-time temperature. The first panel presents realized temperatures for two consecutive days within an exam period—Thursday, June 24, 2010, and Friday, June 25, 2010—inclusive of spatial and temporal temperature corrections, weighted by the number of observations. The second panel presents the distribution of all exam-time temperatures within the study sample (1998–2011), weighted by the number of observations.

student scores just around the passing cutoff, with a score of 66 (SD = 17.9), though there is considerable heterogeneity by neighborhood and demographic group.

Figure 3 illustrates the source of identifying variation for short-run temperature impacts, with temperatures weighted by exam observation and school location. Outdoor temperature during exams ranges 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 the top panel of Figure 3, which shows the variation in outdoor temperature by school and exam-take across two consecutive test dates within the sample period.

IV. Effect of Temperature on High-Stakes Exam Performance

A. Empirical Specification

Figure 4 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 and year fixed effects. Each dot represents approximately 40,000 observations. 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 exams within student–month–year cells. While it is unlikely that temperature is endogenous to student behavior, nor is it likely that students 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, Gino, and Piovesan 2016) or toward the end of the exam period (Thursday as opposed to Monday), one might 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:

$$(7)\quad Y_{ijsty} = \gamma_{iy} + \eta_s + \beta_1 T_{jsty} + X_{jsty}\beta_2 + \beta_3 Time_{sty} + DOW_{sty}\beta_4 + \epsilon_{ijsty}$$

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 vicinity 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, dew point, and ozone. $Time_{sty}$ represents a dummy for time of day (morning versus afternoon, where $Time = 1$ 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 exams within the same testing window, where some exams may be taken on hot days and others not, leveraging the fact that the average student takes seven June Regents Exams over the course of their high school career (roughly three to four per year). Subject fixed effects control for persistent differences in average difficulty across subjects and the possibility that some subjects tend to be scheduled during earlier

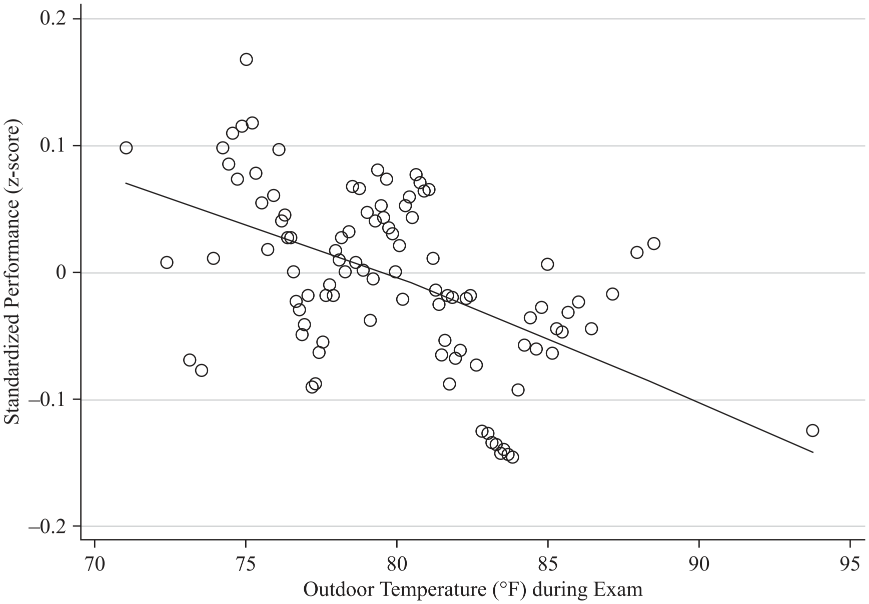


Figure 4

Temperature and High-Stakes Exam Performance

Notes: This figure presents a binned scatterplot of exam performance by quantile of the exam-time temperature distribution, controlling for school and year fixed effects in addition to controls for student demographic characteristics (gender, ethnicity, and subsidized school lunch status). Each dot represents approximately 40,000 exam observations.

or later dates during the two-week window. Year fixed effects control for possible spurious correlation between secular performance improvements and likelihood of hotter exam days due to climate change.

To the extent that temperature variation within student–month–year cells is uncorrelated with unobserved factors influencing test performance, the coefficient β_1 represents the causal impact of temperature on exam performance, subject to potential downward attenuation bias due to measurement error in weather variables.

B. Primary Results

Table 2 presents the results from running variations of Equation 4 for the subset of students who take at least two exams. As suggested by the first column, exam-time heat stress exerts a significant causal impact on student performance. The estimates are robust to allowing for arbitrary autocorrelation of error terms at the level of school and date, though as described below, this result is robust to alternative clustering, including at the level of weather station (borough) and year (Online Appendix Table A.1).

Taking an exam under hotter conditions reduces performance by –0.009 standard deviations (SE = 0.003) per degree Fahrenheit. This amounts to –5.5 percent of a

Table 2
Temperature and High-Stakes Exam Performance

	(1)	(2)	(3)	(4)
Temperature (°F)	–0.009*** (0.003)	–0.007** (0.003)	–0.010*** (0.004)	–0.011*** (0.004)
Afternoon	–0.030* (0.018)	–0.033* (0.017)	–0.018 (0.022)	–0.016 (0.020)
N	3,581,934	3,581,934	3,581,934	3,581,934
Fixed effects				
Student × year	X			
Subject	X	X	X	X
Day of week	X	X	X	X
Student		X		
Year		X	X	
School			X	
School × year				X

Notes: Robust standard errors clustered by school and date in parentheses ($*p<0.10$, $**p<0.05$, $***p<0.01$). Coefficients in each column and panel come from a regression of Regents *z*-scores on the variables shown. The sample comprises all students in the New York City public high school system who took Regents Exams during the years 1998–2011. All regressions include controls for daily dew point, precipitation, ozone, and $PM_{2.5}$.

standard deviation in performance per standard deviation increase in exam-time temperature (+6.2°F), or −13 percent of a standard deviation if a student takes an exam when it is 90°F outside as opposed to a more optimal 75°F.[15]

This effect is roughly equivalent in magnitude to the impacts on mathematical reasoning found by Graff Zivin, Hsiang, and Neidell (2018), who find a day with temperature between 86°F and 89.6°F reduces NLSY math scores by approximately 11 percent of a standard deviation compared to a day with temperatures in the 70–78°F range, and smaller than the test-day impacts of 30 percent of a standard deviation documented by Garg, Jagnani, and Taraz (2017) in India for a 90°F day. They are similar in magnitude to effects from laboratory experiments (Seppanen, Fisk, and Lei 2006), which generally find effects on the order of 1 to 2 percent decline per degree Fahrenheit increase in temperature above the optimum of 70–74°F. These results provide strong evidence that temperature can affect cognition even in economically meaningful settings where effort reduction is unlikely to be the only mechanism.

[15] Precipitation has a slightly positive effect, and ozone has a negative but insignificant effect, with a one standard deviation increase in ozone corresponding to a point estimate roughly one-fifth the size of a one standard deviation temperature effect. Despite previous literature documenting adverse impacts of $PM_{2.5}$ in Israel (Ebenstein, Lavy, and Roth 2016), I find little evidence for that here, perhaps because average concentrations of $PM_{2.5}$ are much lower in NYC than in Israel, as well as the fact that the performance impacts documented by Ebenstein, Lavy, and Roth (2016) are highly nonlinear, driven mostly by heavily polluted days with $PM_{2.5}$ above 100 micrograms per cubic meter.

C. Robustness Checks

A series of robustness checks are presented in Columns 2–4 of Table 2 and in Online Appendix Tables A.1 and A.2. For instance, running models that replace student-by-year fixed effects with student and year or school-by-year fixed effects are presented in Columns 2 and 3 of Table 2 and suggest significantly negative point estimates across specifications. As shown in Table A.1, the main effect does not appear to be sensitive to alternative clustering of standard errors, including by school, or by weather sensor and year.

Building on previous work that finds evidence for differences in temperature effects across math and reading subjects (Graff Zivin, Hsiang, and Neidell 2018), I assess the effect of temperature on quantitative (for example, algebra, physics) and verbal (for example, English language and arts, world history) subjects separately. The results are presented in Online Appendix Table A.2. While the point estimate on quantitative subjects is more negative, and the coefficient on verbal is statistically indistinguishable from zero, the effects on quantitative and verbal subjects are not significantly different from each other. The standard errors are larger in both cases due to the fact that most of the identifying variation is coming from day-to-day variation across exams within a more limited number of subjects. Unlike Graff Zivin, Hsiang, and Neidell (2018), I do not find evidence that temperature has significantly different impacts on mathematical versus verbal reasoning. While it is possible that temperature affects various parts of the brain differently, given previous work that finds simple verbal assessments to be noisier measures of student achievement (Kraft, Blazar, and Hogan 2018; Kraft 2019), an alternative explanation may be that the lack of impact on reading performance in Graff Zivin, Hsiang, and Neidell (2018) is driven in part by measurement error and the type of verbal assessment used in the NLSY.

D. Linear versus Nonlinear Impacts

To assess the potential for nonlinear effects, Online Appendix Table A.3 presents results that specify temperature as a series of indicator variables corresponding to several 5 or 10°F bins, where coefficients can be interpreted as impacts of exams with temperatures in a given bin relative to an optimal omitted bin (70–80°F). Online Appendix Table A.4 presents quadratic specifications in exam-time outdoor temperature. In both cases, the coefficients provide little evidence for nonlinear impacts in extreme heat. For instance, focusing on Column 2 of Table A.3, the coefficients for temperature in the 80–90°F range appear to be significantly different from those in the 70–80°F range ($F = 7.57$, $p = 0.007$), but I cannot reject the null that days with temperature above 90°F are significantly different from days in the 70s or 80s ($p = 0.47$ and $p = 0.36$, respectively).

Several explanations seem plausible. First, this could be due to the fact that, within the study sample, extremely hot temperatures are relatively rare, and thus effects at the higher temperature range are downward attenuated due to measurement error given the fixed effects (for instance, as in Ashenfelter and Krueger 1994). If there are relatively few days in the 90°F and hotter bins, even small amounts of measurement error could downward-bias the coefficient. In the data, exams with temperatures above 90°F are 1/15th as likely as exams with temperatures between 80 and 90°F. Second, it is possible that, due to compensatory responses by teachers, the most extreme performance impacts are being

Table 3
Temperature and Likelihood of Passing Exam

	(1)	(2)	(3)	(4)
Temperature (°F)	–0.004** (0.001)	–0.003** (0.001)	–0.005****** (0.002)	–0.005*** (0.002)
Afternoon	–0.013* (0.008)	–0.014* (0.008)	–0.007 (0.009)	–0.006 (0.009)
N	3,581,934	3,581,934	3,581,934	3,581,934
Fixed effects				
Student × year	X			
Subject	X	X	X	X
Day of week	X	X	X	X
Student		X		
Year		X	X	
School			X	
School × year				X

Notes: Robust standard errors clustered by school and date in parentheses (*$p < 0.10$, **$p < 0.05$, ***$p < 0.01$). Coefficients in each column and panel come from a regression of a dummy variable for passing an exam on the variables shown. The sample comprises all students in the New York City public high school system who took Regents Exams during the years 1998–2011. All regressions include controls for daily dew point, precipitation, ozone, and $PM_{2.5}$.

partially muted in this setting, particularly if the disruptive influences of heat are more salient on extremely hot days. Finally, it is possible that the effects of temperature on performance are in fact roughly linear in this range of outdoor temperatures (70–95°F).

E. Temperature and Pass/Proficiency Status

Running versions of Equation 7 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 any given subject exam. A one standard deviation (6.2°F) increase in temperature results in a 2.4 percent lower probability of passing (SE = 0.13, Column 1 of Table 3). This amounts to a 0.7 percentage point decline per degree Fahrenheit, relative to a mean likelihood of 57 percent—in other words, taking an exam when it is 90°F outside results in a 10 percent lower chance of passing a given exam relative to a 75°F day. These results are presented in Table 3. Columns 2–4 probe the robustness of this finding to alternative specifications and show similar point estimates across models that replace student-by-year fixed effects with school and year or school-by-year fixed effects.

Online Appendix Table A.5 provides a similar analysis for "mastery" or "proficiency" status, which occurs at a threshold score of 75 or 85 depending on the subject and year. This merit is useful for some college-bound students, since it is often used in college admissions decisions. Similarly to pass rates, I find that elevated temperature reduces the likelihood of achieving mastery status (Columns 1–4 of Table A.5). Both results suggest that short-run environmental conditions may affect longer-run outcomes, including educational attainment.

V. Persistent Impacts on Educational Attainment

While the evidence presented above suggests that hotter temperature can reduce realized performance, these short-run shocks presumably do not reduce the stock of human capital. If the stakes are high enough, however, one might expect even transient temperature shocks to affect longer-term educational attainment. Specifically, if the opportunity or stigma costs associated with retaking exams are high, or 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, then even transitory shocks might generate persistent consequences in future periods. In this section, I use linked administrative data on student-level graduation status to assess whether short-run temperature shocks during high-stakes cognitive assessments may result in persistent consequences.

A. Empirical Specification

Figure 5 plots variation in four-year graduation status against average exam-time temperature, and provides suggestive evidence of such persistent impacts. Students who experienced hotter temperatures during exams on average tend to be less likely to graduate high school.

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 experience different amounts of heat during exams, conditioning on the number of draws from the climate distribution. Specifically, I collapse the data at the student level and estimate variations of the following model:

$$(8) \quad g_{ijcn} = \alpha_0 + \alpha_1 \overline{T_{ij}} + X_{ij}\alpha_2 + \chi_j + \theta_c + Z_i\alpha_3 + exams_n\alpha_4 + \epsilon_{ijc}$$

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 their high school career has graduated four years after matriculation. $\overline{T_{ij}}$ denotes the average temperature experienced by student i while taking June Regents Exams in school j, up through their senior year. X_{ij} is a vector of weather controls averaged at the student-by-school level. χ_j denotes school fixed effects, and θ_c denotes cohort fixed effects. Z_i is a vector of student-level controls including race, gender, and federally subsidized school lunch eligibility. The vector $exams_n$ includes fixed effects for the total number of June exams taken.

The parameter of interest is α_1, which captures the impact of an additional degree of heat exposure during all June exams on the likelihood of graduating on time. School fixed effects account for potential omitted variable bias due to correlation between unobserved determinants of graduation rates and higher average temperature in the cross-section (for example, if urban heat island effects are stronger in poorer neighborhoods). Cohort fixed

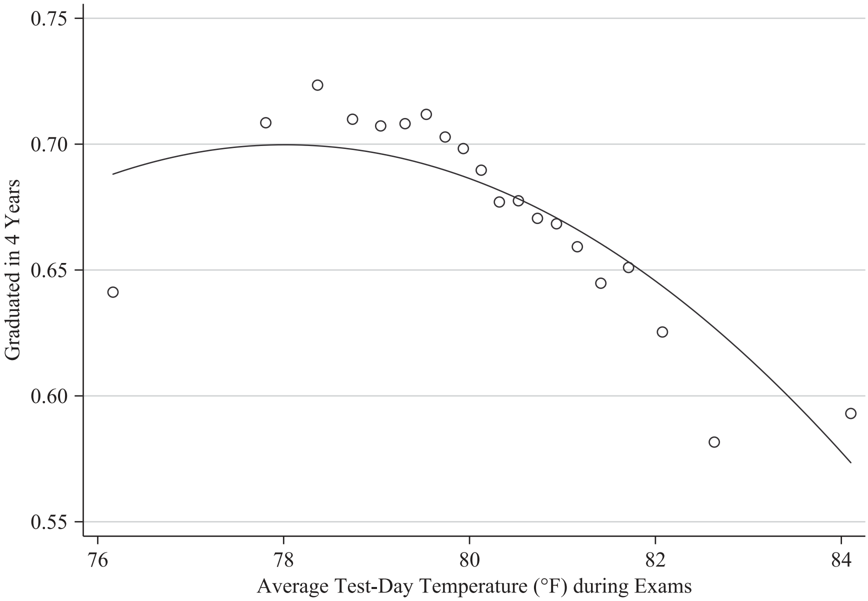


Figure 5

Persistent Impacts on Educational Attainment (High School Graduation)

Notes: This figure presents a binned scatterplot of four-year graduation status by ventile of the exam-time temperature distribution. Temperatures are averaged by student for June exam sessions up through their senior year. Residual variation after controlling for school and number of exam fixed effects, student-level observable characteristics, and weather and air quality controls. Included in the analysis are all June Regents Exams in core subjects between 1998 and 2011. Each dot represents approximately 30,000 students.

effects 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.

B. Primary Results

Table 4 presents the results from running variations of Equation 8 with and without school and cohort fixed effects, as well as flexible controls for the number of exams. Standard errors are clustered at the school level to allow for arbitrary correlation of error terms within a given school, though the results appear to be robust to alternative levels of clustering (Online Appendix Table A.6).

Columns 1–3 suggest that a 1°F increase in average exam-time temperatures is associated with a 0.7 (SE = 0.1) to 0.8 (SE = 0.1) percentage point decline in the likelihood of graduating on time. A one standard deviation in average exam-time temperature (+4.4°F) leads to approximately three percentage points lower likelihood of on-time graduation, or a 4.5 percent decline relative to a mean on-time graduation rate of

Table 4
Persistent Impacts of Temperature on Educational Attainment (High School Graduation)

	(1)	(2)	(3)
Mean temperature (°F)	–0.007*** (0.001)	–0.008*** (0.001)	–0.007*** (0.001)
Number of takes			0.193*** (0.007)
Number of takes^2			–0.015*** (0.001)
Number of takes^3			0.000*** (0.000)
N	515,199	515,199	515,199
Fixed effects			
School	X	X	X
Number of takes	X	X	
Cohort		X	X

Notes: Robust standard errors clustered by school in parentheses (*$p<0.10$, **$p<0.05$, ***$p<0.01$). Coefficients in each column come from a regression of a dummy for graduation status on the variables shown. Temperature is measured with average exam-time temperature experienced by a student (at the school level) during exams preceding senior year of high school. All regressions include controls for observable demographic characteristics including ethnicity, subsidized school lunch status, and English language learner status. The sample comprises all students in the New York City public high school system who took Regents Exams during the years 1998–2011. All regressions include controls for daily dew point, precipitation, ozone, and $PM_{2.5}$.

68 percent. These effects do not appear to be sensitive to how one controls for the possible mechanical correlation between average exam-time temperature and the number of exams taken.

These effects are economically significant. Over the period 1998–2011, upwards of 510,000 exams that otherwise would have passed likely received failing grades due to hot exam conditions, affecting the on-time graduation prospects of at least 90,000 students. This is consistent with the high-stakes nature of these exams, suggesting nontrivial economic and psychic costs of hot temperature during inflexibly administered high-stakes exams. The number of students affected would likely have been larger in the absence of compensatory teacher responses, as described in greater detail below.

To the extent that temperature during exams affects terminal degree status or subsequent human capital investment decisions, the associated lifetime earnings impacts may be substantial. Evidence suggests that the sheepskin effects of having a high school diploma alone may confer earnings advantages of up to 18 percent (Jaeger and Page 1996). While data limitations inhibit an assessment of later-life impacts on earnings or other labor market outcomes, quasi-experimental analyses such as Ebenstein, Lavy, and Roth (2016) find positive returns to high school exit exam performance. These results should be interpreted in light of such findings.

VI. Adaptive Responses

Given the importance of these exams, we might expect responses by those who have a stake in the outcomes.[16] For instance, if students are aware of these effects and allowed to engage in avoidance behaviors, they might choose to reschedule their exams for a milder day or choose testing centers that are well cooled. The institutional rigidity of this setting means that such margins of adaptation are closed off to students. Teachers and administrators, however, may in some instances have additional discretion. If teachers—as agents in the principal–agent relationship between students and educators—hold a view that idiosyncratic shocks to cognitive performance outside students' control are somehow inefficient or unfair, they might exercise this discretion in ways that act as a form of buffer. While it appears that teachers could not adjust the timing or location of these particular (highly standardized and coordinated) state exams, they do seem to have had some discretion in grading. The unique institutional features of NYC public schools during the study period allow an indirect assessment of teacher responses and provide some of the first available evidence of ex post compensation in response to environmental shocks.

A. Teacher Responses

Previous work has documented grade manipulation by teachers (Diamond and Persson 2016; Angrist, Battistin, and Vuri 2017; Dee et al. 2019), including, in the case of Dee et al. (2019), teachers in NYC public schools. In the human capital literature, such manipulation has been used to document persistent human capital consequences of exam performance. For NYC, Dee et al. (2019) suggest that grade manipulation was motivated primarily by teachers who wanted to prevent students from suffering long-term adverse consequences of having experienced, as the authors put it, "a bad test day." They assess the potential impacts of teacher incentive programs and rule out grade manipulation as a means of cheating test-based NCLB accountability standards or teacher incentives, though they do not assess the possible contribution of environmental factors.[17]

Here, I explore the possibility that compensatory grade manipulation functioned as a partial buffer between adverse test-taking conditions and persistent educational consequences. A hot test day may be viewed as a bad test day, particularly if air-conditioning is inadequately provided. While classroom-level air-conditioning data are not publicly available, an analysis of archived building condition assessment reports for 644 middle and high schools in the study sample suggests that, of these schools, only 62 percent were reported as having air-conditioning as of 2012, and among the schools with air-conditioning, nearly 40 percent had some form of defective components, consistent with highly incomplete air-conditioning penetration. Given such constraints, it seems possible for discretionary grade manipulation to have been motivated in part by exam-time temperature conditions.

Teachers may be able to observe the disruptive impacts of elevated temperatures on test day, especially since exams are taken in students' home schools and graded by a

[16] Such compensatory investments have been shown to be important determinants of overall welfare impacts, in particular in the context of the health impacts of air pollution (Deschenes, Greenstone, and Shapiro 2017).

[17] See Dee et al. (2019, p. 25–27) for details.

committee of teachers from that school. If benevolently motivated, they may engage in more grade manipulation precisely for those exams that took place under unusually hot conditions. Numerous media reports suggest that teachers are often aware of the effect of environmental conditions on student behavior and performance and that various institutional factors constrain their set of possible responses.[18] Even if teachers do not consciously identify temperature as a determinant of student performance, their attempts at "correcting" for deviations between what they perceive to be a given student's true ability and realized exam score may have the realized effect of blunting some of the longer-term consequences.

B. Estimating Teacher Grade Manipulation

Figure 6 provides a histogram of Regents Exam results in all core subjects prior to 2011. As is clearly visible in the graph, there is substantial bunching at the passing thresholds, especially at scores of 65 and 55, suggesting upward grade manipulation. The primary method by which teachers manipulated grades appears to have been through selective and coordinated leniency in the amount of partial credit granted to certain free-response questions (Dee et al. 2019).

To assess the presence and magnitude of "compensatory grading," I estimate a bunching estimator by school, subject, month, and year—in effect, the level of exam-time temperature variation. Calculating the fraction of observations in each one-point score bin from 0 to 100 by core Regents subject, I fit a polynomial to these fractions by subject, excluding data near the proficiency cutoffs with a set of indicator variables, using the following regression:

$$(9) \quad F_{ks} = \Sigma_{i=0}^{q} \psi_{ismyj} \dot{(}Score)^{i} + \Sigma_{i \in -M_{cs},\, +M_{cs}} \lambda_{ismyj} \dot{1}[Score = i] + \epsilon_{ksmyj}$$

Here F_{ks} denotes the fraction of observations with score k for subject s (for example, 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. (2019), 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. 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]

[18] For instance, "[There are] several months where the heat just saps the energy from the kids and even the teacher. If it's really hot certainly [student] engagement goes down." (Barnum 2017). In addition, teacher's unions have frequently petitioned for classroom air-conditioning. The *New York Times* quotes the president of the United Federation of Teachers as follows: "It's inhumane to subject kids and adults to schooling in this kind of heat.… If this doesn't convince people that we need to air-condition schools, then I don't know what will."

[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.

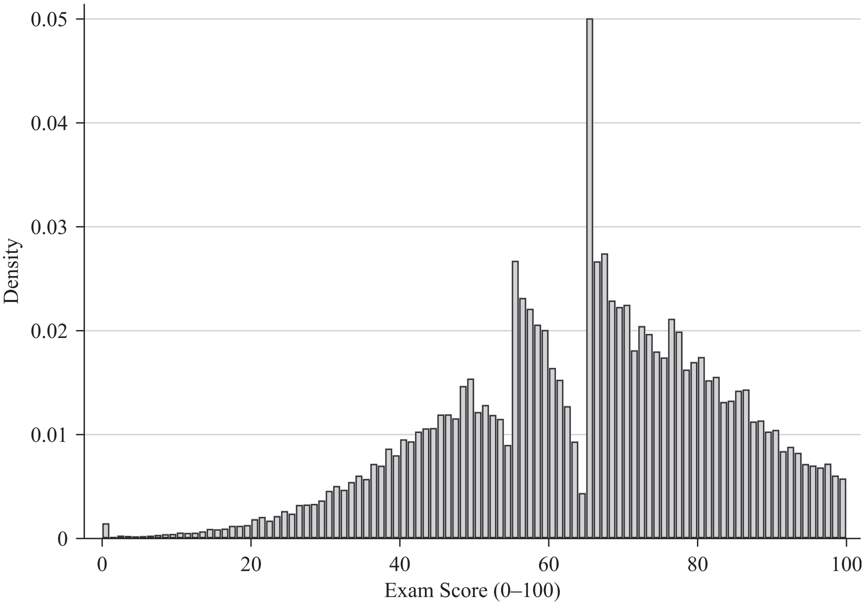


Figure 6

Evidence for Grade Manipulation by Teachers

Notes: This figure presents a histogram of exam scores for all students in the study sample (June 1998–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. See Section III for a discussion of the different cutoffs for subgroups of students and exams.

This generates a set of predicted fractions by score and subject. The average amount of bunching observed in my data (5.8 percent) is similar to that documented by Dee et al. (2019), who find that approximately 6 percent of Regents Exams between 2003 and 2011 exhibited upward grade manipulation. I 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, 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:

$$(10) \quad \zeta_{smyj} = \tfrac{1}{2}\Sigma_{i\in+M_{ck}}\left(F_{ks}-\hat{F}_{ksmyj}\right) + \tfrac{1}{2}\left|\Sigma_{i\in-M_{ck}}\left(F_{ks}-\hat{F}_{ksmyj}\right)\right|$$

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

This bunching estimate is likely measured with error. To account for the possibility that this may bias the implied precision of subsequent analyses that use ζ_{smyj} and functions of ζ_{smyj} as the dependent variable, I replicate the above procedure 100 times using bootstrap resampling. The original data are cluster-bootstrap resampled at the

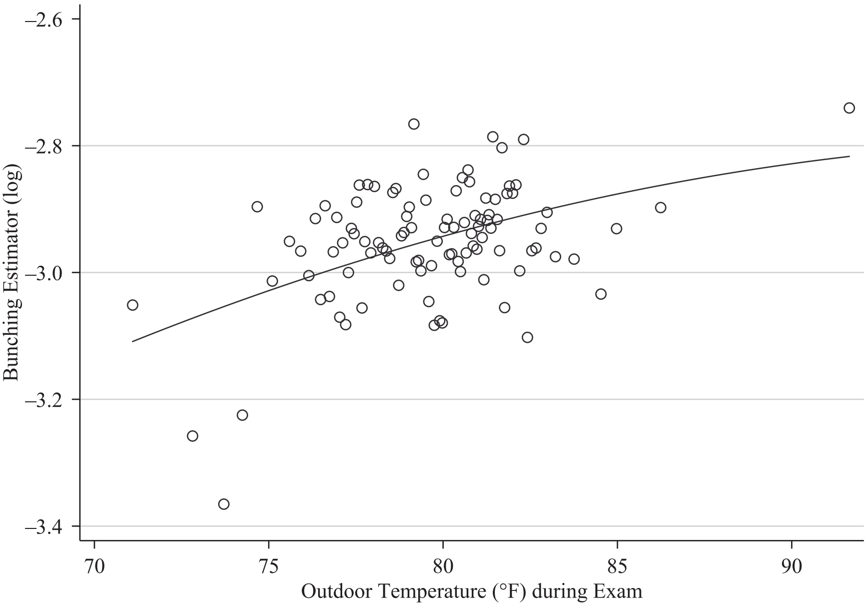


Figure 7

Exam-Time Temperature and Frequency of Grade Manipulation

Notes: This figure presents a binned scatterplot of the natural log of the fraction of all exams that exhibit upward grade manipulation by quantile of the exam-time temperature distribution, net of subject, year, and school fixed effects, as well as daily weather and air quality controls. Manipulation is estimated within school–subject–date cells using a cutoff rule described in Section VI. Included in the analysis are all June Regents Exams in core subjects between 1998 and 2011.

level of school and date, and all of the ensuing analyses are replicated for each resample. In these cases, the standard deviation of the resulting coefficient estimate are reported instead of estimated standard errors.

Using this measure, I find that the extent of manipulation varies significantly across schools, years, and subjects. For some schools in some years, particularly those schools where most students are near the pass–fail margin, up to 20 percent of all exams in a particular subject may have exhibited some form of upward grade manipulation. Figure 7 suggests that the magnitude of such manipulation may vary systematically with exam-time temperature.

C. Exam-Time Temperature and Grade Manipulation

We are interested in how the extent of manipulation may or may not have been related to temperature during the test. 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 ln(ζ_{smyj}) as the dependent variable:

$$\text{(11)} \quad \ln(\zeta_{smyj}) = \delta_0 + \delta_1 T_{smyj} + X_{smyj}\delta_2 + \chi_j + \eta_s + DOW_{sty}\delta_3 + \epsilon_{smyj}$$

Here, T_{smyj} denotes temperature, and X_{smyj} denotes a vector of other environmental factors (precipitation, ozone, $PM_{2.5}$). χ_j, η_s, and θ_y denote school, subject, and year fixed effects, respectively, and DOW_{sty} is a vector of fixed effects for each day of the week. The parameter of interest is δ_1, which, given the log transformation represents the approximate percentage change in grade manipulation due to exam-time temperature. Similarly to the analysis in Section IV, the effect of temperature on grade manipulation δ_1 can be interpreted as causal as long as variation in weather is uncorrelated with unobserved determinants of grade manipulation. While the identifying assumption for causal impacts is similar, the assumptions required for interpreting this coefficient as compensating behavior is slightly more restrictive, requiring temperature to affect bunching only through student performance rather than directly through teacher cognition.[20]

As shown in Table 5, the amount of bunching increases by approximately 0.015 log points (Column 1, SD = 0.0025) per degree Fahrenheit hotter exam-time temperature, again reporting standard deviations of the bootstrapped resampling distribution. This suggests that teachers manipulated grades significantly more frequently for exams that were taken under hot conditions. Columns 2–4 assess the robustness of this finding to alternative specifications, including regressions that replace school and year fixed effects with school-by-year fixed effects. Column 5 assesses the sensitivity to potential outliers in the form of a handful of particularly hot exam days. The point estimates appear to be remarkably stable and suggest a significant positive relationship between temperature during exams and the extent of upward grade manipulation.[21]

The implied magnitudes are nontrivial. The difference in overall share of exams manipulated between a 90°F and a 75°F exam session averages approximately 22 percent and can be as much as 40 percent, suggesting temperature fluctuations represent a large component of the variation in extent of grade manipulation throughout the period.

D. Robustness to Mechanical Increase in Manipulable Scores

One important factor to consider in this analysis is the possibility for mechanical correlation between temperature and the number of manipulable exams (scores in the 50–54 and 60–64 range), due to the properties of the score distribution. Because the modal score often lies above passing thresholds, hotter temperature may mechanically increase the number of exams in the manipulable zones. While, from a realized welfare standpoint, it may matter little whether teachers were consciously responding to hotter temperatures or not, since the achieved effect on hot days may be similar regardless of motive, we may want to go a step further to assess whether teachers are more likely to exercise their discretion when temperatures were unusually hot during exams.

[20] It seems unlikely though not impossible that teachers would engage in more grade manipulation (which, in this context, requires substantial coordination across multiple graders and thus more effort) when they themselves are heat stressed, especially given that most grading occurred either in the evening following an exam or at a later date.

[21] Online Appendix Figure A.5 provides the full sampling distributions of the estimates of δ_1 across the various specifications of Equation 11 (corresponding to Columns 1–4 of Table 5) over the 100 cluster bootstraps.

Table 5
Temperature and Teacher Grade Manipulation: Bunching Estimate

	(1)	(2)	(3)	(4)	(5)
Temperature (°F)	0.015	0.015	0.016	0.015	0.015
Bootstrapped SDs	0.00235	0.00256	0.00259	0.00256	0.00261
Observations	3,676,927	3,676,927	3,676,719	3,676,927	3,517,453
Fixed effects					
Subject	X	X	X	X	X
School	X	X		X	X
Year	X	X		X	X
Day of week		X	X	X	X
School × year			X		
Demographic controls				X	X
Dropping outliers					X

Notes: Standard deviations of the associated sampling distributions of beta are based on replications of the regressions across 100 cluster bootstrap resamples, clustered by school and date. Coefficients in each column come from a regression of the bunching estimator—which is the natural log of the excess mass above the passing threshold by school and exam-take—on the variables shown. Temperature is measured at the school level by exam date and time. All regressions include controls for daily precipitation, dew point, ozone, and $PM_{2.5}$. The sample comprises all students in the New York City public high school system who took Regents Exams during the years 1998–2011. Column 5 reports results after dropping observations with temperature readings above 90°F.

To explore this possibility, for each exam-take and school I compute the fraction of manipulable exams manipulated, as opposed to the raw amount of excess mass as above. By this measure, it appears that approximately 38 percent of manipulable exams were upward manipulated in this way, with a median of 27 percent and standard deviation of 34 percent. That is, if 100 exams had a score between 50–54 or 60–64, approximately 38 of these were upward manipulated on average.

Online Appendix Figure A.3 presents a binned scatterplot of this alternate bunching estimator and exam-time temperature by school–subject–year cell. It suggests a clear positive relationship between the degree of grade manipulation and the ambient temperature during the exam being graded. Online Appendix Figure A.4 presents the same figure, dropping observations with temperatures above 90°F to probe robustness to outliers.

Online Appendix Table A.7 presents results from running Equation 11 with the adjusted measure of grade manipulation as the dependent variable. The results suggest that the frequency of teacher grade manipulation increased by approximately 0.017 log points (SD = 0.003) per degree Fahrenheit increase in exam-time temperature. Column 5 provides results omitting potential outlier observations with temperatures above 90°F. Online Appendix Figure A.6 provides the full sampling distributions of the estimates of δ_1 across the various specifications of Equation 11 (corresponding to Columns 1–4 of Online Appendix Table A.7) over the 100 cluster bootstraps. The estimates suggest that hotter temperature during exams appears to increase the amount of upward grade manipulation above and beyond what would be expected due to a mechanical increase in the number of scores that could potentially be manipulated. While it is impossible to infer intentions on the basis of this analysis, it appears that the realized effect of teacher discretion in this setting was to blunt some of the adverse consequences of "bad test days," particularly those brought about by hotter temperatures.

In summary, I find evidence of ex post compensatory investment by teachers, which acts as a form of buffer between idiosyncratic exam conditions and educational attainment. These results are consistent with anecdotal evidence and media reports that suggest that teachers are often aware of the effect of environmental conditions on student performance, as well as the institutional barriers to optimal temperature control. Irrespective of whether teachers are cognizant of the effect of temperature on exam performance, their grading behavior had the realized effect of mitigating the adverse consequences of hot temperature.

VII. Discussion and Conclusion

This work explores the impact of temperature on high-stakes performance. 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 achievement. The research design exploits quasi-random, within-student variation in test-taking conditions to identify the impact of hot temperature on realized 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. Consistent with these persistent consequences, I also document what appears to be ex post compensatory behavior by teachers, who upward manipulate borderline scores for exams taken under hot conditions.

A key advantage of the study setting is that the analysis takes place in a testing environment where the outcome is likely to be economically meaningful, making the resulting estimates more relevant for policy. Taking an exam when the temperature is 90°F results in 13 percent of a standard deviation lower exam performance relative to a more optimal 75°F, controlling for student ability. For the median New York City high school student, this results in a 10 percent lower probability of passing a subject and a significantly lower likelihood of graduating on time. Roughly 18 percent of the students in the study sample experience at least one exam with ambient temperatures exceeding 90°F, and far more experience damaging temperatures in the upper 80s. I estimate that, for the period 1998–2011, 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 suboptimal test-taking conditions by selectively boosting grades of students just below pass–fail thresholds. I find a pattern of upward grade manipulation that intensifies as exam day temperatures increase, and this pattern persists even when controlling for potential mechanical correlation between temperature and the fraction of manipulable scores. One plausible interpretation is that teachers are aware of adverse test-taking conditions and view transitory shocks to cognition as not reflecting underlying human capital. It appears they may have used their limited 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 of this study have several implications. First, they suggest that ambient environmental conditions including temperature may be important variables to consider when designing education or human resource policies. For instance, in determining how to administer high-stakes exams or interviews, administrators could account for the potential impact of environmental conditions such as temperature ahead of time. This might mean rescheduling to cooler months when possible or making sure that such assessments take place in air-conditioned buildings if scheduled during summer months. Similarly, in assessing various policy options aimed at reducing achievement gaps, improving school facilities might offer greater improvements than previously suggested (Coleman 1968; Hanushek 2006), consistent with emerging quasi-experimental findings (Jackson, Johnson, and Persico 2015; Lafortune, Rothstein, and Schanzenbach 2018). The magnitude of these effects imply that standardized exams provide noisier indicators of underlying human capital particularly for lower-income and underserved students, which has implications for efficiency in labor market sorting, as in Ebenstein, Lavy, and Roth (2016).

These results also suggest that students taking standardized exams across varying climates may not be on a level environmental playing field. Such concerns may be especially important for nationally and sometimes internationally harmonized examinations, such as the International Baccalaureate (IB), ACT, SAT, or LSAT. Given the present geographic distribution of students by race, the average Black or Hispanic student faces a 28 percent chance that their June SAT exam is taken on a day where outdoor temperatures are above 90°F. The corresponding likelihood for white and Asian students is approximately 18 percent.[22] This suggests that a small but nontrivial component—using the point estimates from this study, approximately 3–4 percent—of average racial achievement gaps in performance on that exam could be attributable to differences in the likelihood of adverse temperature conditions.[23]

As the span of geographies covered by a standardized exams widens, the potential for differences in test-day temperature increases. The SAT, for instance, is taken more or less simultaneously across the 50 United States and across countries as diverse as Brazil, India, New Zealand, South Korea, and the Ukraine. Adjusting the scaling of scores based on geography could help, but may not be sufficient if test-taking conditions vary within a given region. Recent nationwide school air-conditioning estimates from the United States suggest that racial minorities and lower-income students are substantially less likely to have classroom air-conditioning even within a given local climate (Park et al. 2020), consistent with evidence of highly localized disparities in funding for school facilities and maintenance (Filardo 2016).[24] The consequences for educational

[22] Alternatively, the difference in expected average June temperature between these two groups is 3°F (mean monthly temperature of 82.4 and 79.4, respectively).

[23] Black–white and Hispanic–white achievement gaps in the SAT were approximately 0.8 to 1 standard deviations, respectively, in 2015.

[24] Air-conditioning penetration is relatively low in many developing economies. Available evidence suggests a strong relationship between income and air-conditioning ownership at the household level (Biddle 2008; Davis and Gertler 2015) and binding liquidity constraints in the context of energy-intensive appliance demand in developing countries (Gertler et al. 2016)

attainment may be magnified if lower-income and racial minority students face higher barriers—financial or otherwise—to retaking high-stakes exams. For instance, Goodman, Gurantz, and Smith (2018) find that underrepresented groups are 9 percent less likely to retake the SAT than white students, despite clear benefits of retaking in terms of performance and eventual college quality. Such factors may be relevant for researchers interested in exploring the persistence of achievement gaps within and across countries or the distributional implications of climate change.

References

Acemoglu, Daron, and David Autor. 2011. "Skills, Tasks and Technologies: Implications for Employment and Earnings." In *Handbook of Labor Economics*, Volume 4, Part B, ed. David Card and Orley Ashenfelter, 1043–171. New York: 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(1):205–46.

Angrist, Joshua D., Erich Battistin, and Daniela Vuri. 2017. "In a Small Moment: Class Size and Moral Hazard in the Italian Mezzogiorno." *American Economic Journal: Applied Economics* 9(4):216–49.

Ashenfelter, Orley, and Alan Krueger. 1994. "Estimates of the Economic Return to Schooling from a New Sample of Twins." *American Economic Review* 84(5):1157–73.

Auffhammer, Maximilian, and Erin T. Mansur. 2014. "Measuring Climatic Impacts on Energy Consumption: A Review of the Empirical Literature." *Energy Economics* 46:522–30.

Barnum, Matthew. 2017. "Exclusive: Too Hot to Learn: Records Show Nearly a Dozen of the Biggest School Districts Lack Air Conditioning." https://www.the74million.org/article/exclusive-too-hot-to-learn-records-show-nearly-a-dozen-of-the-biggest-school-districts-lack-air-conditioning/ (accessed August 18, 2021).

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(1):105–59.

Biddle, Jeff. 2008. "Explaining the Spread of Residential Air Conditioning 1955–1980." *Explorations in Economic History* 45(4):402–23.

Carleton, Tamma, Michael Delgado, Michael Greenstone, Trevor Houser, Solomon Hsiang, Andrew Hultgren, Amir Jina, Robert E. Kopp, Kelly McCusker, Ishan Nath, James Rising, Ashwin Rode, Hee Kwon Seo, Justin Simcock, Arvid Viaene, Jiacan Yuan, and Alice Tianbo Zhang. 2019. "Valuing the Global Mortality Consequences of Climate Change Accounting for Adaptation Costs and Benefits." Working Paper 2018-51. Chicago, IL: Becker Friedman Institute for Economics, University of Chicago.

Chambwera, M., G. Heal, C. Dubeux, S. Hallegatte, L. Leclerc, A. Markandya, B.A. McCarl, R. Mechler, and J.E. Neumann. 2014. "Economics of Adaptation." In *Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, ed. C.B. Field et al., 945–77. Cambridge, UK: Cambridge University Press.

Cho, Hyunkuk. 2017. "The Effects of Summer Heat on Academic Achievement: A Cohort Analysis." *Journal of Environmental Economics and Management* 83:185–96.

Christensen, Peter, and Christopher Timmins. 2018. "Sorting or Steering: Experimental Evidence on the Economic Effects of Housing Discrimination." NBER Working Paper 24826. Cambridge, MA: NBER.

Coleman, James S. 1968. "Equality of Educational Opportunity." *Integrated Education* 6(5):9–28.

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(2):245–50.

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* 112:5962–67.

Dee, Thomas S., Will Dobbie, Brian A. Jacob, and Jonah Rockoff. 2019. "The Causes and Consequences of Test Score Manipulation: Evidence from the New York Regents Examinations." *American Economic Journal: Applied Economics* 11(3):382–423.

Dell, Melissa, Benjamin F. Jones, and Benjamin A. Olken. 2012. "Temperature Shocks and Economic Growth: Evidence from the Last Half Century." *American Economic Journal: Macroeconomics* 4(3):66–95.

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(3): 740–98.

Deryugina, Tatyana, and Solomon M. Hsiang. 2014. "Does the Environment Still Matter? Daily Temperature and Income in the United States." NBER Working Paper 20750. Cambridge, MA: NBER.

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(4):152–85.

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(10):2958–89.

Diamond, Rebecca, and Petra Persson. 2016. "The Long-Term Consequences of Teacher Discretion in Grading of High-Stakes Tests." NBER Working Paper. Cambridge, MA: NBER.

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* 28:278–86.

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(4):36–65.

Filardo, Mary. 2016. "State of Our Schools: America's K–12 Facilities 2016." Washington, DC: 21st Century School Fund.

Garg, Teevrat, Maulik Jagnani, and Vis Taraz. 2017. "Human Capital Costs of Climate Change: Evidence from Test Scores in India." Unpublished.

Gertler, Paul J., Orie Shelef, Catherine D. Wolfram, and Alan Fuchs. 2016. "The Demand for Energy-Using Assets among the World's Rising Middle Classes." *American Economic Review* 106(6):1366–401.

Goldin, Claudia Dale, and Lawrence F. Katz. 2009. *The Race between Education and Technology.* Cambridge, MA: Harvard University Press.

Goodman, Joshua, Oded Gurantz, and Jonathan Smith. 2018. "Take Two! SAT Retaking and College Enrollment Gaps." NBER Working Paper 24945. Cambridge, MA: NBER.

Graff Zivin, Joshua, Solomon M. Hsiang, and Matthew Neidell. 2018. "Temperature and Human Capital in the Short and Long Run." *Journal of the Association of Environmental and Resource Economists* 5(1):77–105.

Graff Zivin, Joshua, and Matthew Neidell. 2014. "Temperature and the Allocation of Time: Implications for Climate Change." *Journal of Labor Economics* 32(1):1–26.

Hanushek, Eric A. 2006. "School Resources." In *Handbook of the Economics of Education*, Volume 2, ed. E. Hanushek and F. Welch, 865–908. New York: Elsevier.

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.

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* 10:347–62.

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(4):719–34.

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(35):15367–72.

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* 114(51):13447–52.

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.” *Quarterly Journal of Economics* 131(1):157–218.

Jaeger, David A., and Marianne E. Page. 1996. “Degrees Matter: New Evidence on Sheepskin Effects in the Returns to Education.” *Review of Economics and Statistics* 78(4):733–40.

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(3):270–81.

Kraft, Matthew A. 2019. “Teacher Effects on Complex Cognitive Skills and Social-Emotional Competencies.” *Journal of Human Resources* 54(1):1–36.

Kraft, Matthew A., David Blazar, and Dylan Hogan. 2018. “The Effect of Teacher Coaching on Instruction and Achievement: A Meta-Analysis of the Causal Evidence.” *Review of Educational Research* 88(4):547–88.

Lafortune, Julien, Jesse Rothstein, and Diane Whitmore Schanzenbach. 2018. “School Finance Reform and the Distribution of Student Achievement.” *American Economic Journal: Applied Economics* 10(2):1–26.

Lim, Chin Leong, Chris Byrne, and Jason K.W. Lee. 2008. “Human Thermoregulation and Measurement of Body Temperature in Exercise and Clinical Settings.” *Annals Academy of Medicine Singapore* 37(4):347–53.

Mackworth, Norman H. 1946. “Effects of Heat on Wireless Operators.” *British Journal of Industrial Medicine* 3(3):143–58.

Park, R. Jisung, Joshua Goodman, Michael Hurwitz, and Jonathan Smith. 2020. “Heat and Learning.” *American Economic Journal: Economic Policy.* 12(2):306–39.

Rosenzweig, Cynthia, William Solecki, and Ronald Slosberg. 2006. “Mitigating New York City’s Heat Island with Urban Forestry, Living Roofs, and Light Surfaces.” Report to the New York State Energy Research and Development Authority.

Seppanen, Olli, William J. Fisk, and Q.H. Lei. 2006. “Effect of Temperature on Task Performance in Office Environment.” Berkeley, CA: 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(10):2621–24.

Somanathan, E., Rohini Somanathan, Anant Sudarshan, and Meenu Tewari. 2018. “The Impact of Temperature on Productivity and Labor Supply: Evidence from Indian Manufacturing.” Working Paper 2018-69. Chicago, IL: Becker Friedman Institute, University of Chicago.

Tiebout, Charles M. 1956. “A Pure Theory of Local Expenditures.” *Journal of Political Economy* 64(5):416–24.