Journal of Environmental Economics and Management 104 (2020) 102365

Contents lists available at ScienceDirect

Journal of Environmental Economics and Management

journal homepage: www.elsevier.com/locate/jeem

Temperature and high-stakes cognitive performance: Evidence from the national college entrance examination in China[☆]

Joshua Graff Zivin [a], Yingquan Song [b], Qu Tang [c], Peng Zhang [d,*]

[a] School of Global Policy and Strategy and Department of Economics, University of California, San Diego, and NBER, USA
[b] The China Institute for Educational Finance Research, Peking University, China
[c] Institute for Economic and Social Research, Jinan University, China
[d] School of Management and Economics, The Chinese University of Hong Kong, Shenzhen, Shenzhen Finance Institute, Shenzhen Research Institute of Big Data, China

ARTICLE INFO

Article history:
Received 28 March 2019
Received in revised form 17 July 2020
Accepted 29 July 2020
Available online 9 August 2020

JEL classification:
Q54
I23
I24

Keywords:
Cognitive performance
Temperature
Climate change
Standardized test

ABSTRACT

We provide the first nation-wide estimates of the effects of temperature on high-stakes cognitive performance in a developing country using data from the National College Entrance Examination (NCEE) in China. The NCEE is one of the most important institutions in China and affects millions of families. We find that a one-standard-deviation increase in temperature during the exam period within counties (2 °C/3.6 °F) decreases the total test score by 0.68%, or 5.83% of a standard deviation, with effects concentrated on the highest performing students. This suggests that temperature plays an important role in high-stakes cognitive performance and has potentially far-reaching impacts for the careers and lifetime earnings of students.

1. Introduction

The planet is expected to warm considerably over the coming century as a result of climate change, driving up average temperatures and shifting the climate toward greater and more frequent temperature extremes. The threat of global warming has spawned a sizable corpus of economic research that explores the impacts of temperature on a wide range of outcomes.[1] One area that has been comparatively underexplored, but which touches many aspects of our everyday lives, is the impacts on cognitive performance.

☆ All authors contributed equally and are ordered alphabetically.
* Corresponding author.
E-mail addresses: jgraffzivin@ucsd.edu (J. Graff Zivin), songyingquan@pku.edu.cn (Y. Song), qutang@jnu.edu.cn (Q. Tang), jumpersdu@gmail.com (P. Zhang).

[1] For example, see Mendelsohn et al. (1994); Schlenker et al. (2006); Deschênes and Greenstone, 2007; Burke et al. (2009); Hsiang and Solomon, 2010; Deschênes and Greenstone, 2011; Dell et al. (2012); Graff Zivin and Neidell, 2014; Ranson (2014); Barreca et al. (2016).

https://doi.org/10.1016/j.jeem.2020.102365
0095-0696/

In principal, environmental conditions could impact cognitive performance vis-à-vis a number of inter-related channels. The brain's chemistry, electrical properties and function are all temperature sensitive (Bowler and Tirri, 1974; Schiff and Somjen, 1985; Deboer, 1998; Yablonskiy et al., 2000; Hocking et al., 2001). Moreover, exposure to heat has been shown to diminish attention, memory, information retention and processing, and the performance of psycho-perceptual tasks (Hyde et al., 1997; Hocking et al., 2001; Vasmatzidis et al., 2002). The impacts of thermal stress on working memory performance are especially relevant as cognitively challenging tasks rely heavily on the working memory for multi-step processing.

In this paper, we provide the first nation-wide estimates of the impacts of temperature on high-stakes cognitive performance in a developing country using data from the National College Entrance Examination (NCEE), or *gaokao*, in China. The NCEE offers a useful means of examining the effect of heat on cognitive performance for several reasons. It is one of the most important institutional features of admissions to post-secondary education in China and affects the lives of millions of families (Bai et al., 2014; Chen and Onur Kesten, 2017; Jia and Li, 2017; Cai et al., 2019). Each year, around 9 million students take the exam to compete for admission to around 2300 colleges and universities. Unlike other countries which rely upon standardized tests along with other factors such as high-school GPA, extracurricular activities, and recommendation letters to determine college admissions, the NCEE is almost the sole determinant for college admission in China,[2] making it an extremely high-stakes exam. This is especially true for those aiming for the top-tier universities, as graduates can expect, on average, to earn 40% more per month than their counterparts from lesser universities (Jia and Li, 2017). The competition is fierce. Though the overall admission rate of test takers to college or university has been around 75% in recent years, the admission rate for the roughly 100 first-tier universities in China is only 12% (China Education Online, 2016).

Several other features of the NCEE make it particularly well-suited for measuring the causal effects of temperature on cognitive performance. First, the date of the NCEE is fixed, on June 7th and 8th, making self-selection on test dates impossible. Second, because the NCEE is held only once a year, the cost of retaking the exam is quite high, essentially requiring students to repeat an additional year of high school. Third, during our sample period of 2005–2011, students were required to take the exam in the same county as their household registration (*hukou*). Therefore, self-selection on exam locations is heavily regulated. Finally, air conditioning is not available at testing facilities,[3] thereby eliminating a potentially endogenous adaptation strategy, and providing a better simulacrum of the conditions under which cognitive tasks are performed throughout the developing world where air conditioning penetration is quite low.

To examine the impact of temperature on the NCEE performance, we obtained a unique dataset that covers the universe of students from 2227 counties who were admitted into college between 2005 and 2011 across China, yielding more than 14 million observations. The dataset reports the exam scores (ranging from 0 to 750) and exam counties for each student. We then match this dataset with daily weather data on temperature, precipitation, relative humidity, wind speed, sunshine duration, pressure, and visibility from 752 weather stations spread across the entire country.

We find both economically and statistically significant negative effects of temperature on test scores. In particular, a one-standard-deviation increase in temperature during the exam period within counties (2 °C/3.6 °F) decreases total test scores by 0.68%, or approximately 6 percent of one standard deviation in test performance. The effects are roughly linear in the temperature range found in China during early June – mean temperature during the exam period is 23.21 °C (73.78 °F). Given the significant negative effect of temperature on exam scores, we then turn our attention to the effects of temperature on whether a student's score is above the cutoff for the first-tier universities.[4] Since we do not have data on college admission, we proxy top-tier university admissions based on obtaining a score higher than the cutoff. We find that a one-standard-deviation increase in temperature within counties decreases the probability of getting into first-tier universities by 1.2 percent. Together, these results indicate that temperature plays an important role in high-stakes cognitive performance and has potentially far-reaching impacts for the careers and lifetime earnings of students.

In the Chinese context, hotter regions may be unfairly penalized by the current system, and climate change is expected to exacerbate these inequalities. We believe that one policy response to remedy this injustice is to install and use air conditioning in the exam rooms to help protect against the harmful effects of heat and level the playing field across regions which vary considerably in their average summertime temperatures.

This paper builds upon a small and recent body of economics literature that examines the impacts of temperature on cognitive performance in a developed country context in which air conditioning is far more prevalent and the ramifications from underperforming on a test are significantly less consequential. Graff Zivin et al. (2018) find modest and statistically significant impacts of warmer temperatures on low-stakes test performance administered in U.S. homes as part of the National Longitudinal Survey of Youth. Park (forthcoming) exploits data from a higher-stakes environment – New York City high

[2] Less than 0.1% of students can gain admission to college without taking the NCEE (Bai et al., 2014). They usually take the exams administered by the university itself, or they are waived from having to take the NCEE because of special talent, such as the winners of National High-School Olympic Competitions.

[3] In regions where air conditioning is available, its use is prohibited during the test period to ensure fair competition with regions in which AC is not available (Sina, 2007, 2014).

[4] In China, only students whose scores are above a pre-specified cutoff are eligible to apply for first-tier universities. Approximately 75% of students are admitted into first-tier universities if their scores are above the cutoff (Jia and Li, 2017).

school exit exams — and finds similar results.[5] Our results from China, which imply that a 1 °C increase in temperature during the exam period decreases total test scores by 2.91% of a standard deviation, is approximately twice as large as the impacts estimated by these studies based in the U.S. Our study also complements recent work by Garg et al. (2018), which finds that increases in annual temperature exposure in India can impair test performance largely through impacts on agricultural yields and nutrition.

Thus, the key contribution of our work is the focus on high-stakes testing (arguably the highest in the world) on a national sample in a developing country at a temporal scale that allows us to disentangle direct cognitive impairments from other potential channels. Our findings also have important implications for the study of standardized test performance more generally. While these tests are often viewed as the gold standard for assessing the academic competence of students (Koretz and Deibert, 1995; Robelen, 2002; US Legal, 2014), recent studies have shown that the time the test is given as well as local air pollution can impact performance (Sievertsen et al., 2016; Ebenstein et al., 2016). Temperature appears to be another important factor to add to this list.

2. Empirical background

The NCEE is a prerequisite for entrance into almost all higher education institutions at the undergraduate level in China. It is held annually, and is generally taken by students in their last year of high school. The NCEE has undergone continuous reform since 1978. It was once uniformly designed by the Ministry of Education (MOE) such that all the students across the country took exactly the same examination. In the early 2000s, the MOE launched the "unified examination, provincial proposition" reform (Zhu and Lou, 2011). Provinces and municipalities were allowed to customize their own exams independently, while the MOE continued to provide a national exam that could be used by provinces not employing independent exams. In 2011, 16 out of 31 provinces created customized exams while the others adopted national exam versions.

The most common examination format across provinces during our study period (from 2005 to 2011) is the two-day exam, which takes place annually on June 7th and 8th, and is scored on a 0–750 scale based on the "3 + X" subjects system.[6] In the "3 + X" subjects system, "3" refers to the three compulsory subjects: Chinese, Mathematics, and a foreign language usually English (each accounting for 150/750 of the total score) and "X" refers to the combination of science subjects (biology, chemistry, and physics) for students on the science track, or the combination of art subjects (geography, history, and political science) for students on the art track (accounting for 300/750 of the total score).[7]

The NCEE is an extremely high-stakes exam. It is almost the sole determinant for higher education admission in China. Every year, around 9 million students in China take the exam to compete for admission to approximately 2300 colleges and universities. These colleges are divided into two hierarchical categories: regular colleges and universities that are degree-granting and academically oriented; and advanced vocational colleges that certify students based on the attainment of practical and occupational skills. Though the overall admission rate of exam takers to both forms of higher education ranges from 57% to 72% during our study period, the admission rate for the former category is only around 30%.[8] The regular colleges and universities can further be classified into three tiers according to the recruitment process. Tier 1 universities recruit before Tier 2 and Tier 3 universities within each province and require a much higher cut-off score for admission, according to provincial education authorities.

The cut-off score for each tier is the minimum qualifying score for students to apply to universities for that tier and varies annually across provinces and subject tracks. It is determined by the Provincial Admission Offices based on each year's admission quota and the distribution of student scores within the province (Chen and Onur Kesten, 2017). As such, it is important to keep in mind that any allocative inefficiencies that arise from exposure to extreme temperatures will result from variation in temperatures across counties within a given province. Since western provinces are larger and exhibit greater variability in weather than their eastern counterparts, these within province inefficiencies can also translate to inequality across regions.

Throughout our research period, Tier 1 universities include all elite universities as well as other important institutions affiliated with the Ministry of Education and other national ministries. Admission rates for Tier 1 universities in recent years has hovered around 12% and was even lower in earlier years (China Education Online, 2016). The higher the NCEE score the greater the chance that a student can attend an elite university, which is highly correlated with future life opportunities and earning potential (Jia and Li, 2017).

[5] While the relationship between long-run temperature exposure and changes in test scores have also been examined, these changes in test scores (controlling for weather during the test) reflect the impacts of weather on learning, as opposed to performance. Graff Zivin et al. (2018) find no such effects on learning, while recent work by Goodman et al. (forthcoming) find evidence of very small effects that are completely offset by access to air conditioning.

[6] Six provinces take a three-day exam on Jun 7th, 8th and 9th, including Shanghai, Jiangsu and Guangdong from 2005 to 2011, Hainan and Shandong from 2007 to 2011, and Zhejiang from 2009 to 2011. The exam on the 9th in most provinces only occurred in the morning and accounted less than 10% of total scores. Four provinces use a scale rather than 0-750 marks, including Hainan 0–900 from 2005 to 2011, Guangdong 0–900 from 2005 to 2006, Jiangsu 0–440 in 2008, and Shanghai 0–630 from 2005 to 2011. We normalize the scale to 750 marks for these four provinces.

[7] There are also a small number of specialized tracks, which include sports, art (music, painting, dancing), military, and pedagogical. These constitute less than 10% of all track specializations in our data. All students choose their track of study prior to the start of their second year of high school.

[8] See annual statistical data from the MOE: http://old.moe.gov.cn//publicfiles/business/htmlfiles/moe/moe_1651/index.html.

Table 1
Summary statistics.

Variable	Mean	SD	Min	Max
Panel A: Score (0–750)				
Full sample	518.96	60.40	60.00	750.00
Art track	512.66	57.24	74.00	749.17
Science track	521.20	62.56	60.00	750.00
Panel B: Proportion of students above cutoff for the first-tier universities				
Full sample	0.29	0.45	0.00	1.00
Art track	0.20	0.40	0.00	1.00
Science track	0.32	0.47	0.00	1.00
Panel C: Weather				
Temperature (°C)	23.21	3.29	2.55	31.96
DD $\geq$ 14 (degree days)	9.23	3.24	0.00	18.00
DD < 14 (degree days)	0.01	0.21	0.00	11.45
Precipitation (cm)	0.54	1.01	0.00	15.42
Relative humidity (%)	69.20	15.34	13.56	99.74
Wind speed (m/s)	2.30	0.86	0.26	16.22
Sunshine duration (hour)	5.77	3.73	0.00	14.17
Pressure (hpa)	965.33	53.67	581.45	1014.39
Visibility (km)	13.32	5.99	0.27	29.76

Notes: The NCEE data covers all students enrolled in college during 2005–2011. The observations for the full sample: 14,042,417. The observations for the art track: 3,699,915. The observations for the science track: 8,972,856. There are 2227 counties in total. The sum of observations between the art track and the science track is not equal to the observations for all tracks due to the existence of a small number of specialized tracks. The score scale is 0–750 for most provinces. We normalize the score scale to 750 for provinces that are not using the same scale. All students need to take three compulsory subjects: Chinese, mathematics, and a foreign language (typically English). The students in the art track need to take one combined subject comprising politics, history, and geography, and the students in the science track need to take one combined subject comprising physics, chemistry, and biology. The data on the cutoff of the first-tier universities are only available for the art and science tracks. The Tier 1 cut-off score is the minimum qualifying score for students to apply to Tier 1 universities. It is determined by the Provincial Admission Offices based on each year's admission quota and the distribution of student scores within the province and track. It varies annually across provinces and subject tracks. The weather variables are averaged using daily values on June 7th and 8th, when the NCEE is held.

3. Data

We obtain the NCEE data from the China Institute for Educational Finance Research at Peking University, which reports the total score and ID for the universe of students enrolled into college from 2227 counties during 2005–2011. This dataset includes observations for roughly 2 million students each year. The student ID contains a six-digit code for county of residence, which we use to match with weather data. The ID also reports the specific track, allowing us to explore heterogeneity across the science and art tracks. Unfortunately, we do not have data on scores by specific subject. If a student retakes the exam, he/she will receive a new exam ID. Thus, we cannot distinguish between new and old exam takers. We also do not have data on students' demographics and high school information. Data on the cut-off scores that determine eligibility to apply to first-tier universities for each province-year-track are obtained from a website specialized for the exam: gaokao.com.

The weather data are obtained from the China Meteorological Data Service Center, which is an affiliate of the National Meteorological Information Center of China. The data report daily maximum, minimum and average temperatures, precipitation, relative humidity, wind speed, sunshine duration, and atmospheric pressure for 752 weather stations in China. Thus, each county has 0.34 weather stations on average. Data on visibility are obtained from the National Oceanic and Atmospheric Administration of the U.S. We extract weather data during the exam time and then convert from station measures to county measures using the inverse-distance weighting (IDW) method (Deschênes and Greenstone, 2007, 2011). The basic algorithm calculates weather for a given county based on weighted averages of all weather station observations within a 200 km radius of the county centroid, where the weights are the inverse distance between the weather station and the centroid.

Panel A of Table 1 presents the summary statistics of the exam score, which ranges from 0 to 750. There are more than 14 million observations in total, with a mean of 518.96, and a standard deviation of roughly 60 marks. Approximately 26% students were in the art track, and 64% students were in the science track, with the remaining 10% corresponding to students in specialized tracks.

In panel B, we define a dummy variable which is equal to one if a student's score is above or equal to the cutoff for the first-tier universities, as a proxy for admission into a first-tier university. Approximately 75% of students with a score above the cutoff are admitted into first-tier universities (Jia and Li, 2017). That corresponds to approximately 30% of the students in our sample.[9] Admission rates for the science track are higher than the art track using this proxy, a results that is consistent with actual admission patterns at top-tier universities.

[9] This rate is higher than the 10% admission rate for the entire population of high school graduates since our sample only includes students who enrolled into an institution of higher learning.

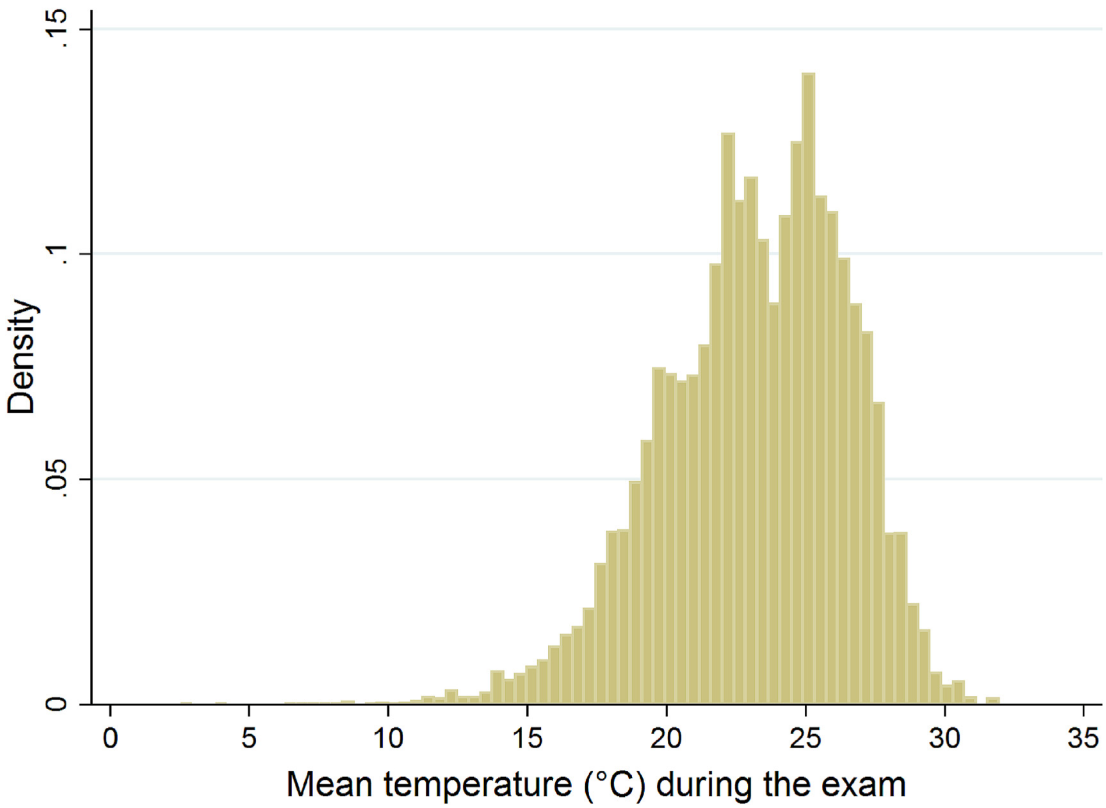

Fig. 1. Histogram of mean temperature (°C) during the exam. *Notes:* Mean temperature over this two-day period is defined as the average of the daily average temperature on June 7th and 8th over 2005–2011. As is standard practice, the daily average temperature is the average of the daily maximum and minimum temperatures.

We report the summary statistics of weather variables in panel C. The average mean temperature during the exam period is 23.21 °C. The histogram of average temperature during the 2-day exam period is plotted in Fig. 1. This figure reveals a great heterogeneity, with temperatures ranging from 10 °C to 30 °C, and a peak around 25 °C. To measure the non-linear effects of temperature, we construct two measures. The first is degree days (DD), which is a piece-wise linear function that measures the number of degrees above and below a threshold. We deploy a threshold of 14 °C, as it is within the temperature range that is associated with the highest test scores when we use the binned approach. As can be seen in panel C of Table 1, the average degree days above or equal 14 °C ($DD \geq 14$) is 9.23, and the average below 14 °C ($DD < 14$) is 0.01, consistent with the skewed distribution of temperature seen in Fig. 1. The second measure we deploy to capture non-linear effects is a series of indicators of 2 °C bins (Deschênes and Greenstone, 2011; Barreca et al., 2016; Graff Zivin and Neidell, 2014; Graff Zivin and Neidell, 2018), with the lowest bin including all temperatures below 12 °C and the highest bin including all temperatures above 28 °C due to data sparseness at the extremities of the distribution. Figure A1 in the online appendix plots the percentage of days that fall into each bin, while Figure A2 illustrates the large cross-sectional variation in temperature across counties during our study period.[10]

It is important to note that the exams are graded one to two weeks after the exams are completed by professionals (trained teachers) in hotels (typically with air conditioning) in each of the respective provincial capitals (often quite far from the counties where students took their exam).[11] In addition, each grader only grades one question, and each question is graded by two graders and then cross validated. Thus, the effect we estimate on NCEE scores should be further minimized by any potential impacts of temperature on grader behavior.[12]

[10] For completeness, the appendix also includes maps of the distribution of NCEE scores, the ratio of students whose scores are above the cutoff for the first-tier universities, and prefecture-level average GDP per capita during our sample period. See Figures A3, A4, and A5 respectively.

[11] Recent figures from China Family Panel Studies suggest that AC penetration in provincial capitals is 62.55 percent as compared to 22.08 percent in non-capital cities.

[12] Note that two recent papers have found evidence of grade manipulation on high-stakes tests in Sweden and the U.S., respectively (Diamond and Persson, 2016; Dee et al., 2019). If graders are not fully insulated from environmental extremes, temperatures could influence grader behavior by altering their demand for manipulation.

We are also not terribly concerned about student selection. While more savvy students may wish to defer their exams in anticipation of punishing test-taking conditions, the costs of this deferral are very high. Students would need to wait an entire year to retake the NCEE (Muthanna and Sang, 2015).

4. Empirical strategy

In order to assess the effect of temperature on students' performance, we estimate the following equation:

$$Y_{ict} = \alpha_0 + \beta_1 T_{ct} + \beta_2 W_{ct} + \gamma_c + \eta_t + \in_{ict},$$

where i denotes an individual student, c denotes the county in which the exam was taken, and t denotes the year the exam was taken. We have two measures for Y_{ict}. The first is the logarithm of the exam score. The logarithm specification was chosen to facilitate interpretation, since point estimates correspond to the semi-elasticity of exam scores with respect to temperature. As we will show later, our results are also robust to specifying exam scores in levels. The second is a dummy variable which is equal to one if a student's score is equal to or higher than the cutoff for first-tier universities and zero otherwise. Both specifications are estimated using OLS, although our results for admission to elite universities remain unchanged when we use a logit specification. We use T_{ct} to denote the average of daily mean temperature (the average between daily maximum and minimum temperatures) on June 7th and 8th. We do not include temperature in each day separately because of the strong serial correlation in temperature across days. To explore the non-linearity of temperature, we use degree day measures and a series of 2 °C bins as described earlier.

The variable W_{ct} denotes a vector of weather variables, including precipitation, relative humidity, wind speed, sunshine duration, atmospheric pressure, and visibility. As with our temperature variable, all of these are averaged across the two-day exam period. We use γ_c to denote county fixed effects, which controls for any county-specific time-invariant characteristics, such as geography or cultural and demographic features that are stable over our study period. We use η_t to represent year fixed effects, to control for any nation-wide policy or economic shocks that could differ by year but affect test takers equally across all counties. Since the weather variables are grouped at the county level, the standard errors may be biased downward (Moulton, 1986), so we cluster the error terms $\in_{ict}$ by county.

In the end, our identifying variation is based on county deviations from the mean after we adjust for common shocks for the whole country in a given year. We refrain from using province-year fixed effects since this absorbs most of the variation in our data (see Table A1 in the Online Appendix) and because the random nature of short-term temperature fluctuations within a given county (Deschênes and Greenstone, 2007) makes its correlation with exam difficulty unlikely. Nevertheless, we conduct a robustness check using province-specific time trends and find similar results.

The coefficient of interest is β_i. Under our linear measure of temperature, this coefficient measures the percentage change in total score (or the probability change of admission to first-tier universities) when temperature during the exam increases by 1 °C. When we use degree days, the coefficient of DD $\geq$ 14 (DD < 14) measures the percentage change in total score (or the probability change of admission to first-tier universities) if temperature increases (decreases) by 1 °C conditional on temperature being above (below) 14 °C. The non-linear binned approach has a slightly different interpretation. Here the coefficient of each bin measures the percentage change in total score (or the probability change of admission to first-tier universities) when temperature falls into that bin relative to the reference bin of 12–14 °C, which was chosen as it is associated with the highest exam scores.

5. Results

5.1. Main results

Table 2 presents the main regression results, where outcomes are defined as the logarithm of the total test score. The total test score is the summation of scores from three compulsory subjects, including Chinese, mathematics, and foreign language (typically English) with 150 marks each plus scores from one combined subject with 300 marks comprising politics, history, and geography for the art track and physics, chemistry, and biology for the science track. Unfortunately, the data does not report the score for each specific subject. We report results for all students in columns (1) and (2), only students in the art track in columns (3) and (4) and only those in the science track in columns (5) and (6).

In columns (1), (3), and (5), temperature is measured using the average of daily mean temperature during June 7th and 8th. All the estimates are negative and statistically significant at the 1% significance level. The coefficient of temperature in column (1) suggests that a 1 °C increase in temperature decreases the total test score by 0.34%, or 1.76 marks evaluated at the mean level (mean = 518.96). To better place these figures in context, it is helpful to situate them relative to the weather variability in our dataset. A one-standard-deviation increase in temperature (3.29 °C) decreases total test scores by 1.12%, or 9.59% of a standard deviation (standard deviation = 60.40). Since our model includes county fixed effects, we can also calibrate the magnitudes using within-county standard deviations in temperature, which is 2 °C. It suggests that a one-standard-deviation increase in temperature within counties (2 °C) decreases total test scores by 0.68%, or 5.83% of a standard deviation.

In columns (2), (4), and (6), we relax the assumption of linearity by specifying temperature in terms of degree days as described above. As can be seen in column (2), the effect of DD $\geq$ 14 is significantly negative, and the magnitude is almost

Table 2
Effect of temperature on log of exam score.

	Dependent variable: Log of exam scores					
	All track		Art track		Science track	
	(1)	(2)	(3)	(4)	(5)	(6)
Temperature	−0.0034***	–	−0.0036***	–	−0.0018***	–
	(0.0004)	–	(0.0004)	–	(0.0004)	–
DD ≥ 14	–	−0.0034***	–	−0.0036***	–	−0.0018***
	–	(0.0004)	–	(0.0004)	–	(0.0004)
DD < 14	–	0.0014	–	0.0018	–	−0.0023
	–	(0.0018)	–	(0.0014)	–	(0.0025)
Precipitation	−0.0008	−0.0009*	0.0002	0.0002	−0.0004	−0.0004
	(0.0005)	(0.0005)	(0.0006)	(0.0006)	(0.0006)	(0.0006)
Humidity	0.0000	0.0000	−0.0002**	−0.0002**	0.0003***	0.0003***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Wind	0.0039***	0.0039***	0.0008	0.0008	0.0024***	0.0024***
	(0.0007)	(0.0007)	(0.0008)	(0.0008)	(0.0007)	(0.0007)
Sunshine	0.0025***	0.0025***	0.0018***	0.0018***	0.0023***	0.0023***
	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)
Pressure	−0.0000	−0.0001	0.0004***	0.0003***	−0.0006***	−0.0006***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Visibility	0.0001	0.0001	0.0004*	0.0004*	−0.0002	−0.0002
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
Observations	14,042,417	14,042,417	3,699,915	3,699,915	8,972,856	8,972,856
R-squared	0.2697	0.2697	0.4035	0.4035	0.2738	0.2738

Notes: The dependent variable is the log of the exam score. All students need to take three compulsory subjects: Chinese, mathematics, and foreign language (typically English). Students in the art track need to take one combined subject comprising politics, history, and geography, and students in the science track need to take one combined subject comprising physics, chemistry, and biology. The observations for all tracks does not equal the sum of observations from the art and science tracks, due to the existence of a small number of specialized tracks. Regression models also include county fixed effects and year fixed effects. Degree days (DD) ≥14 (<14) is the number of degrees above (below) 14 °C. Standard errors are clustered at the county level and reported in parentheses. ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.

identical to the linear effect in column (1). This is largely an artefact of exam timing. June temperatures in China tend to be quite high, with a mean temperature of 22.78 °C, which lies above the degree-day threshold. In contrast, the effect of DD < 14 is statistically insignificant.

When we run subsample analyses for each track separately (see columns (3)–(6)), we find that the negative effect of temperature is much larger for students in the art track than those in the science track. For example, a 1 °C increase in temperature decreases the score for the art track by 0.36%, but only by 0.18% for the science track. Evaluated at the mean level, this is equivalent to 1.85 marks for the art track (mean = 512.66) and 0.94 marks for the science track (mean = 521.20). One possible explanation for this difference is sample composition. The art track is disproportionately female relative to the science track and recent research suggests that female test performance in China may be more stress-dependent (Cai et al., 2019).

In addition to temperature, we also include precipitation, relative humidity, wind speed, sunshine duration, pressure, and visibility in the regression model. We find a significantly positive effect of wind speed, consistent with the notion that higher wind speeds reduce perceived temperature – the effect of so-called wind chill.[13] The effect of sunshine duration is also significantly positive, as many studies find that sunshine induces good mood and happiness (Schwarz and Clore, 1983; Guven, 2012) and further increases labor productivity (Oswald et al., 2015). The effect of precipitation, humidity, pressure, and visibility are either weakly significant or statistically insignificant.

Fig. 2 plots the coefficients (in blue) as well as 95% confidence intervals (in grey) under our non-parametric binned approach when the dependent variable is the log of exam score. As noted earlier, the 12–14 °C bin is omitted as the reference group, so all other estimates are relative to it. We find that the coefficient decreases monotonically for all bins hotter than 12–14 °C. The magnitude here is also comparable to column (1) in panel A of Table 2. For example, the estimated coefficient for the above 28 °C bin is −0.0553. Since the difference between bins above 28 °C and 12–14 °C is approximately 15 °C, each 1 °C increase in temperature decreases a score by 0.0037 (0.0553/15) log points (as compared to −0.0034 log points in column (1) of Table 2).

5.2. Learning vs. cognitive performance

In the previous section, we found a significantly negative effect of temperature on students' cognitive performance. In this section, we add temperature for the whole year prior to the exam in the regression model. This serves two purposes: 1) to

[13] http://www.nws.noaa.gov/om/cold/wind_chill.shtml.

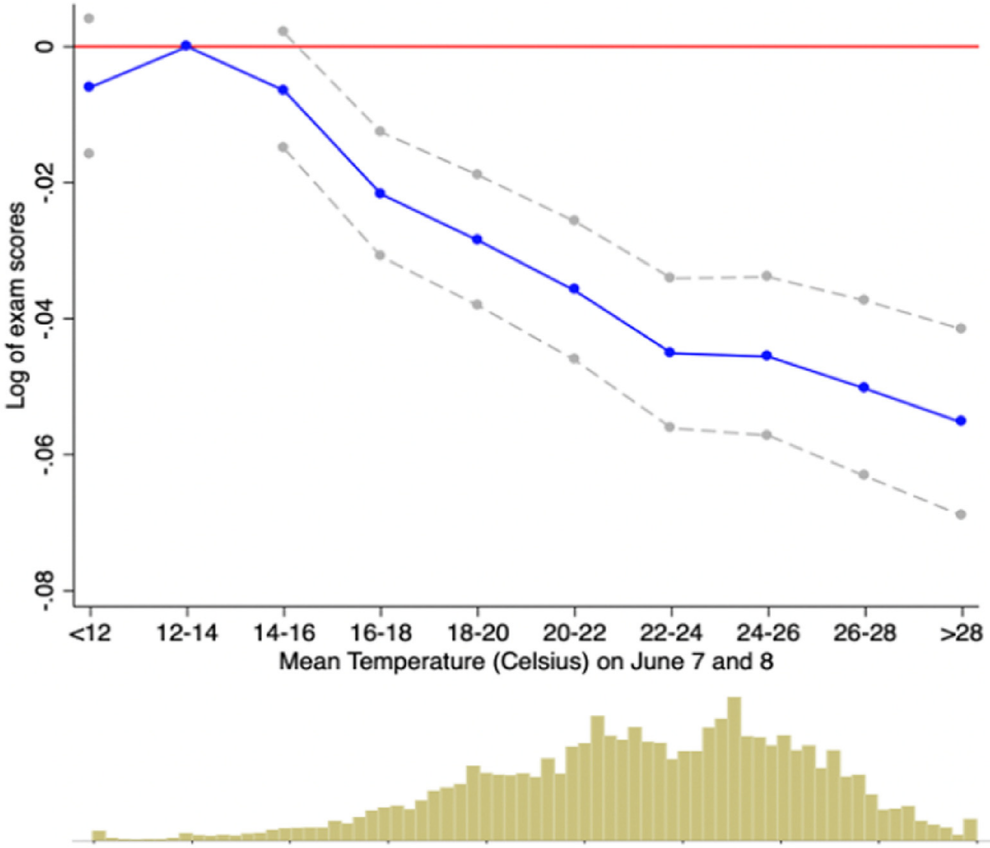

Fig. 2. Relationship between temperature and log of exam scores. *Notes:* The upper panel of the figure shows the non-linear effects of temperature on exam scores, and the lower panel shows the temperature distribution. Each point estimate represents the effect of replacing a day with temperature in the 12–14 °C interval (reference group) with a day with temperature in the corresponding interval. Control variables include: precipitation, relative humidity, wind speed, sunshine duration, pressure, visibility, county fixed effects, and year fixed effects. Whiskers denote the 95% confidence interval, after adjusting for spatial and serial correlation within each county.

make sure our short-run effects on cognitive performance are not driven by long-run effects on students' learning; 2) to detect if there are any long-run learning effects.

The results are reported in Table 3. Columns (1) and (2) repeat our baseline model: column (1) uses average temperature and column (2) uses degree days. In column (3), we add the average temperature during the whole year (i.e. 365 days) prior to the exam.[14] The effect of contemporaneous temperature remains almost the same, but the effect of prior-year temperature is small and statistically insignificant. In column (4), we explore non-linear effects using degree days as our measure of temperature. The effect of contemporaneous temperature changes little, but we now find a significantly negative effect of degree days above 14 °C.[15] These findings confirm that our short-run performance effects are not driven by long-run learning effects and also provide evidence that temperature extremes can impair learning, a result consistent with the findings by Goodman et al. (forthcoming) in the U.S.[16]

5.3. Mechanism test

Our baseline model uses daily mean temperature, i.e., the average between daily maximum and minimum temperatures. As such, our estimates could reflect two potential channels through which temperature affects exam scores: student performance may be directly impaired by heat during the exam period or performance may be indirectly impaired due to the

[14] Because of the extremely intensive competition for the NCEE, many students stay at school during the summer break between the second and third year of high school.

[15] We find a similar effect using temperature bins.

[16] These findings are similar if use a longer exposure window, such as two and three years.

Table 3
Learning vs. cognitive performance.

	Dependent variable: Log of exam scores			
	Baseline		Prior Year	
	(1)	(2)	(3)	(4)
Temperature	−0.0034***	–	−0.0035***	–
	(0.0004)	–	(0.0004)	–
DD ≥ 14	–	−0.0034***	–	−0.0030***
	–	(0.0004)	–	(0.0004)
DD < 14	–	0.0014	–	0.0006
	–	(0.0018)	–	(0.0017)
Prior Year Temperature	–	–	−0.0013	–
	–	–	(0.0013)	–
Prior Year DD ≥ 14	–	–	–	−0.0210***
	–	–	–	(0.0029)
Prior Year DD < 14	–	–	–	−0.0078***
	–	–	–	(0.0016)
Observations	14,042,417	14,042,417	14,042,417	14,042,417
R-squared	0.2697	0.2697	0.2709	0.2716

Notes: The dependent variable is the log of the exam score. Baseline regression uses the average daily mean temperature on the 2 Em day. Prior Year regression uses the average daily mean temperature one year prior to the exam. Degree days (DD) ≥14 (<14) is the number of degrees above (below) 14 °C. All regression models include county fixed effects and year fixed effects. Standard errors are clustered at the county level and reported in parentheses. ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.

Table 4
Mechanism test.

	Dependent variable: Log of exam scores				
	(1)	(2)	(3)	(4)	(5)
Mean	−0.0034***	–	–	–	−0.0034***
	(0.0004)	–	–	–	(0.0004)
Min	–	−0.0028***	–	−0.0010*	–
	–	(0.0004)	–	(0.0005)	–
Max	–	–	−0.0023***	−0.0023***	–
	–	–	(0.0004)	(0.0004)	–
Diff	–	–	–	–	−0.0015***
	–	–	–	–	(0.0005)
Observations	14,042,417	14,042,417	14,042,417	14,042,417	14,042,417
R-squared	0.2697	0.2699	0.4035	0.4037	0.2740

negative impacts that extreme heat may have exerted on students' sleep. Because daily maximum and minimum temperatures measure heat during daytime and nighttime respectively, we utilize each of these measures separately to shed light on these potential channels.[17]

Column (1) of Table 4 repeats the baseline model. In columns (2) and (3), we use daily minimum and maximum temperatures separately. We find a significantly negative effect in both models. We then include both measures together in column (4). The effect of daily maximum temperature changes little, but the effect of daily minimum temperature is much smaller and only weakly significant. This suggests that daily maximum temperature plays a more prominent role in this relationship and underscores the importance of the direct channel described above.

Our baseline model uses the average temperature across 2 Em day. This ignores the deviations between these two days. For example, two days with temperature 20 °C and 30 °C may have a very different impact than two days with temperature 25 °C due to nonlinearities in the dose-response function or difficulties adjusting to abrupt changes in weather. In column (5), we add the difference of temperatures between the 2 Em day in addition to the average. We find a significantly negative effect of temperature deviations.

5.4. Dynamics

In this section, we explore the dynamics of temperature, i.e., how lagged and lead temperatures affect exam scores. Because daily temperatures are highly serially correlated, we use averages across different exposure windows. Fig. 3 plots the

[17] An alternative approach would exploit hourly temperature to focus on the hours of the exam and perhaps those immediately preceding it. Unfortunately, we are unable to obtain this data for our study period at the spatial resolution required for our analyses.

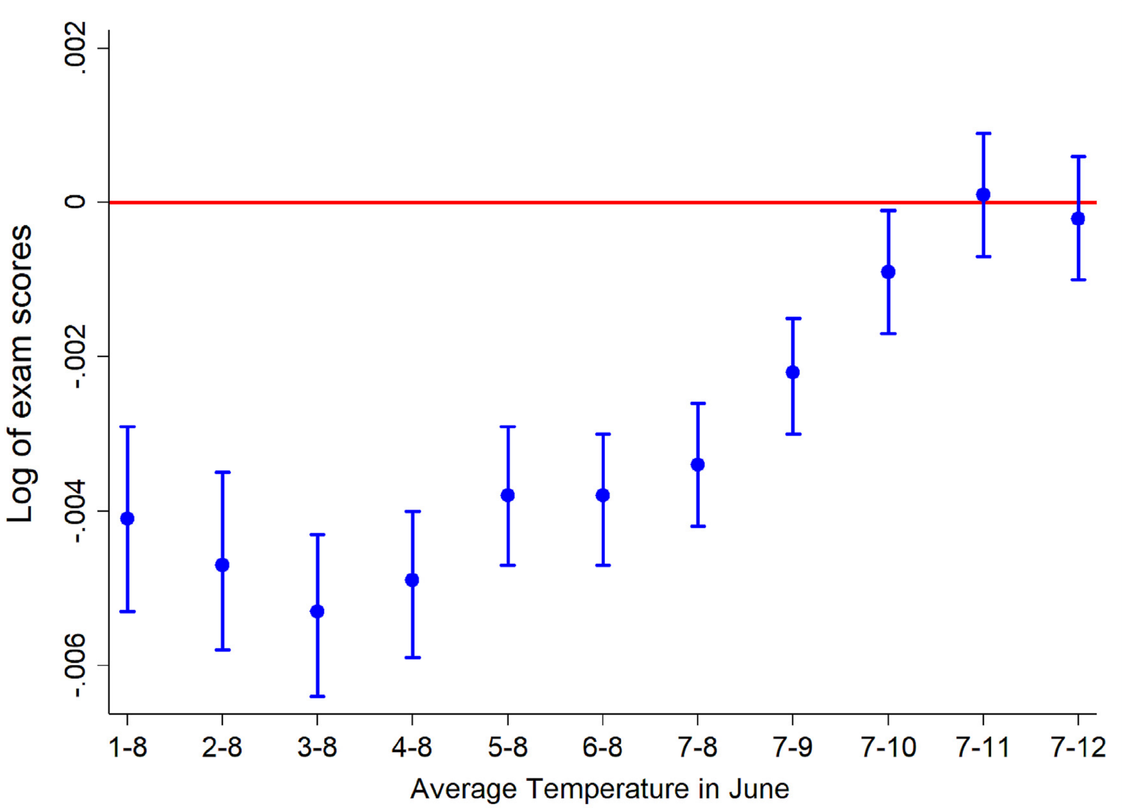

Fig. 3. Dynamic effect. *Notes:* Each point estimate represents the effect of average temperature in the corresponding exposure windows on exam scores. For example, 1–8 corresponds to the average temperature during June 1st to June 8th. Control variables include: precipitation, relative humidity, wind speed, sunshine duration, pressure, visibility, county fixed effects, and year fixed effects. Whiskers denote the 95% confidence interval, after adjusting for spatial and serial correlation within each county.

estimated coefficients as well as the 95% confidence intervals for various exposure windows. In general, we find significantly negative effect of temperatures before the exam. This is intuitive because lagged temperatures could either affect students' preparation for the exam as we saw in our analysis of long-run temperature exposure earlier. In contrast, when we include temperatures after the exam, the effect becomes smaller and eventually insignificant.

5.5. *The role of air conditioning*

Studies show that air conditioning (AC) can protect the human body from harms due to excess heat (Barreca et al., 2016) and it seems plausible that these protective effects might also extend to cognitive performance. Unfortunately, we do not have data on the availability of AC at test facilities. Moreover, AC use is supposed to be prohibited during the NCEE to ensure fairness across regions, some of which clearly do not have AC. Nonetheless, we explore the potential role of AC indirectly, by splitting our sample into urban districts and rural counties,[18] under the assumption that cities are more likely to have AC. Table A2 in the online appendix reports these results. The effects of temperature appear larger in urban districts than rural ones, although these differences are not significant at conventional levels. Whether the lack of difference suggests a limited protective role for air conditioning, urban heat island effects, the effectiveness of the policy ban on usage, or the noisiness of our AC measure remains an open question.

5.6. *Air pollution as a possible confounder*

Since others have found that exposure to fine particulate matter less than 2.5 μm in diameter ($PM_{2.5}$) can also impair test performance (Ebenstein et al., 2016), one concern with our study is that our results may be confounded by air pollution levels in ways that are not fully captured by our controls for visibility. To examine this issue directly, we use data on the air pollution

[18] In China, districts (*qu*) and counties (*xian*) are in the same administrative level, but districts are typically located in urban cities.

Table 5
Effect of weather and air pollution on log of exam scores.

	Dependent variable: Log of exam scores		
	(1)	(2)	(3)
Temperature	−0.0034***	−0.0031**	−0.0032**
	(0.0004)	(0.0015)	(0.0016)
Precipitation	−0.0008	−0.0016	−0.0015
	(0.0005)	(0.0018)	(0.0018)
Humidity	0.0000	0.0000	0.0000
	(0.0001)	(0.0004)	(0.0004)
Wind	0.0039***	0.0034**	0.0035*
	(0.0007)	(0.0018)	(0.0018)
Sunshine	0.0025***	0.0024***	0.0025***
	(0.0003)	(0.0008)	(0.0009)
Pressure	−0.0000	0.0003***	0.0003***
	(0.0001)	(0.0001)	(0.0001)
Visibility	0.0001	−0.0003	−0.0002
	(0.0002)	(0.0005)	(0.0005)
API	–	–	0.0000
	–	–	(0.0001)
Observations	14,042,417	6,321,398	6,321,398
R-squared	0.2697	0.2252	0.2252

Notes: The dependent variable is the log of the exam score. All weather and air pollution variables are calculated using the average between June 7th and 8th. Column (1) reports the baseline estimates from Table 1, column (1). Column (2) reports results from the same specification but only for the sample of cities covered by the air pollution index (API). In column (3) we add controls for pollution as measured by the API. The regression models also include county fixed effects and year fixed effects. Standard errors are clustered at the county level and reported in parentheses. ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.

index (API) – a composite measure of pollution that ranks air quality based on its associated health risks (Ministry of Environmental Protection of China, 2006) – to examine the relationship between air quality and test performance.[19] The API is only available in major cities and thus our sample size for this analysis is greatly reduced. The estimates are reported in Table 5. Column (1) reports the baseline estimates from Table 2, column (1). Column (2) reports results from the same specification but only for the sample of cities covered by the API. In column (3) we add controls for pollution as measured by the API. Though the sample size in columns (2) and (3) is less than half of column (1), the effect of temperature remains unchanged, which suggests that air pollution is not driving our temperature results.

While the results on API are reassuring, it remains possible that $PM_{2.5}$ could be confounding our results. To further probe this possibility, we utilize data from a more recent period when those data are available to examine the correlation between $PM_{2.5}$ and temperature. These results are reported in Tables A3–A5 in the online appendix for the period 2013–2016. Regardless of functional form, the correlation coefficients are small, providing additional evidence that $PM_{2.5}$ is unlikely to explain the relationship between temperature and test performance in our setting.

5.7. Heterogeneity analysis

We conduct several heterogeneity analyses in this section. First, we split the sample into counties that are above or below the median average temperature in order to explore whether people in warmer regions are better able to cope with hotter temperatures (Heutel et al., 2017; Taraz, 2018). The results are presented in Table 6. Columns (1) and (2) display our core results for reference, while columns (3) and (4) present results only for hot counties and columns (5) and (6) only for cold counties. Although effects in hot counties are smaller than effects in cold counties, the standard errors indicate that they are statistically indistinguishable from one another. In Table A6, we further explore heterogeneity by interacting temperature with an indicator for hot counties using the full sample. The interactions are statistically insignificant, again suggesting that there is no statistical difference in temperature effects between hot and cold locations.

Next, we explore the heterogeneity by GDP per capita to assess whether wealthier regions are able to invest in infrastructure that better insulates them from the harms of extreme heat. We divide the sample by rich and poor counties based on the median of prefecture-level GDP per capita and estimate temperature effects separately for rich and poor counties.[20] As can be seen in Table 7, the estimates are slightly larger in rich counties relative to poor counties, but again, they are not statistically different from one another. The interaction between temperature and a dummy for rich counties is also insignificant (see Table A6 in the online appendix).

[19] The data on $PM_{2.5}$ are only available since 2013.

[20] We classify rich and poor counties using prefecture-level GDP per capita, because county-level data is not available. Prefecture is the administrative level between the provincial and the county level. On average, each prefecture contains 9 counties.

Table 6
Heterogeneity analysis by average temperature.

VARIABLES	Dependent variable: Log of exam scores					
	Full sample		Hot counties		Cold counties	
	(1)	(2)	(3)	(4)	(5)	(6)
Temperature	−0.0034***	–	−0.0015**	–	−0.0023***	–
	(0.0004)	–	(0.0007)	–	(0.0004)	–
DD ≥ 14	–	−0.0034***	–	−0.0015*	–	−0.0023***
	–	(0.0004)	–	(0.0007)	–	(0.0004)
DD < 14	–	0.0014	–	–	–	0.0007
	–	(0.0018)	–	–	–	(0.0022)
Observations	14,042,417	14,042,417	7,167,181	7,167,181	6,875,236	6,875,236
R-squared	0.2697	0.2697	0.2053	0.2053	0.3105	0.3105

Notes: The dependent variable is the log of the exam score. Columns (1) and (2) use the full sample. Columns (3) and (4) focus on hot counties, and columns (5) and (6) focus on cold counties. We define a county as hot or cold based on the median temperature for all counties in our sample period. Degree days (DD) ≥14 (<14) is the number of degrees above (below) 14 °C. There is no DD < 14 for hot counties. Standard errors are clustered at the county level and reported in parentheses. ***p < 0.01, **p < 0.05, *p < 0.1.

Table 7
Heterogeneity analysis by GDP per capita.

VARIABLES	Dependent variable: Log of exam scores					
	Full sample		Rich counties		Poor counties	
	(1)	(2)	(3)	(4)	(5)	(6)
Temperature	−0.0034***	–	−0.0046***	–	−0.0028***	–
	(0.0004)	–	(0.0007)	–	(0.0005)	–
DD ≥ 14	–	−0.0034***	–	−0.0046***	–	−0.0028***
	–	(0.0004)	–	(0.0007)	–	(0.0005)
DD < 14	–	0.0014	–	0.0027	–	0.0023
	–	(0.0018)	–	(0.0040)	–	(0.0027)
Observations	14,042,417	14,042,417	5,548,809	5,548,809	7,875,324	7,875,324
R-squared	0.2697	0.2697	0.2680	0.2680	0.2273	0.2273

Notes: The dependent variable is the log of the exam score. Columns (1) and (2) use the full sample. Columns (3) and (4) focus on rich counties, and columns (5) and (6) focus on poor counties. We define a county as rich or poor based on the median GDP per capita for all counties in our sample period. Degree days (DD) ≥14 (<14) is the number of degrees above (below) 14 °C. Standard errors are clustered at the county level and reported in parentheses. ***p < 0.01, **p < 0.05, *p < 0.1.

In Fig. 4, we explore heterogeneity by students' scores. In particular, we estimate the impacts of temperature at a variety of points in the performance distribution. The estimated coefficients and the 95% confidence intervals for different percentiles are reported in Fig. 4. For example, "10%" indicates students in the lowest 10% percentiles, "30%" indicates students in the lowest 30% percentiles, etc., and the "100%" is the full sample.[21] Overall, the estimated coefficients are decreasing from the lowest to the highest percentiles, suggesting that the test performance of high-ability students are more sensitive to temperature, with the lowest ability students unaffected by environmental conditions. Fortunately, this pattern of results diffuses concerns we might have had about sample selection in our study. While we only have data for students admitted to an institution of higher learning, the impacts appear to be concentrated on those most likely to fall into this category.

5.8. *Temperature effects on admission to elite institutions*

In Table 8, we turn our attention to the effects of temperature on admissions to elite universities. To be clear, our estimates are based on reaching the eligibility threshold to apply to Tier 1 universities and does not include actual admissions data. Our estimate in column (1) suggests that a 1 °C increase in temperature decreases the probability of being admitted to first-tier universities by 0.60%, or 1.33% of a standard deviation (standard deviation = 0.45). When we measure temperature using degree days in column (2), we find that a 1 °C increase in temperature above 14 °C decreases the admission probability by 0.60%. As with the linear results, we also find that the effect is larger for those students in the art track. Interestingly, the impacts of other weather variables differ under this specification, with the coefficients on precipitation, humidity, and pressure all statistically significant and small (relative to their mean values). The coefficient on visibility, a proxy measure for

[21] We do not use the separate bins, such as 10–30 percentiles, because the observations falling into each separate bins are small.

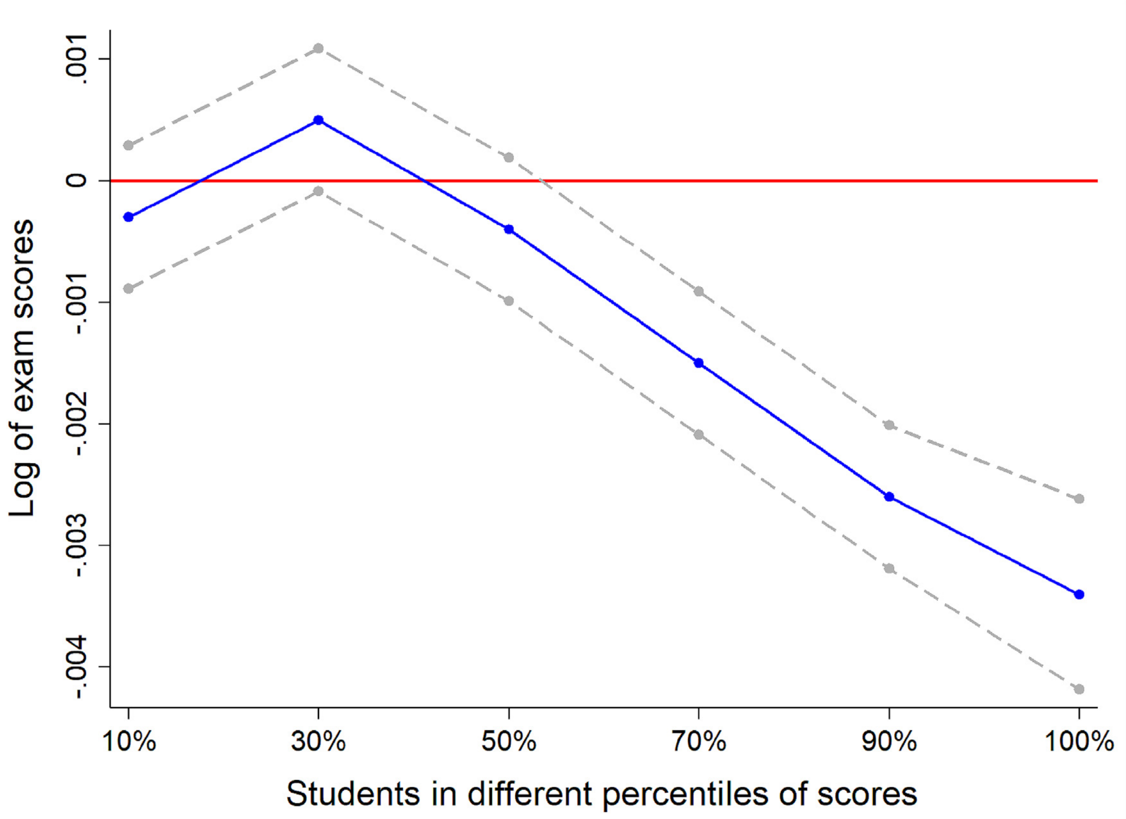

Fig. 4. Subsample analysis for students in different percentiles of scores. *Notes:* This figure presents the estimates for students in different percentiles of the performance distribution. For example, "10%" indicates students in the lowest 10% percentiles, "30%" indicates students in the lowest 30% percentiles, etc., and the "100%" is the full sample. The effect for each subgroup is estimated separately.

pollution, is also negative, statistically significant, and reasonably large. We conduct a similar non-linear exercise in Fig. 5, where the dependent variable is the dummy variable for admission to first-tier universities, and find similar results.

To address concerns about the endogeneity of province-level cutoffs to temperature, we also estimate the impact of temperature on reaching the national average cutoff during our research period (columns (1) to (3) in Appendix Table A7), or province-track specific cutoff score from 2004 (columns (4) to (6) in Appendix Table A7). Since these measures either average over space or do not coincide with our study period (our data begins in 2005), endogeneity concerns should be minimized. This approach yields slightly smaller, but qualitatively similar results.

For all of these results, it is important to recognize that exceeding the threshold to apply for an elite institution does not guarantee entry. According to Jia and Li (2017), only three-quarters of students with scores that exceed the cutoff are admitted to elite universities. Since our heterogeneity analysis suggests that the impacts of temperature are strongest at the top of the performance distribution, and the process that governs selection amongst those that are eligible for admission is unknown, it is difficult to assess the bias of our results in terms of the true effects on admissions. If the selection process is based on factors other than the scores amongst the elite performers, such as political connectedness, then our findings likely understate the true effects. Regardless, our estimates on elite university admissions should be viewed as net of this selection process.

5.9. Robustness checks

Table 9 presents robustness checks for our main results. Column (1) is the baseline model. In column (2) we cluster the standard errors by prefecture (an administrative unit between province and county), to control for spatial and serial correlation within each prefecture. The effect of temperature remains statistically significant at the 1% significance level.

In column (3), instead of using individual-level score data in the baseline model, we average scores to county-year and then estimate the regression model to reflect the fact that the weather data are only at the county-year level. Again, our

Table 8
Effect of temperature on the probability of above cutoff for first-tier universities.

	Dependent variable: Above cutoff for first-tier universities					
	All track		Art track		Science track	
	(1)	(2)	(3)	(4)	(5)	(6)
Temperature	−0.0060***	–	−0.0083***	–	−0.0052***	–
	(0.0012)	–	(0.0015)	–	(0.0011)	–
DD ≥ 14	–	−0.0060***	–	−0.0084***	–	−0.0052***
	–	(0.0012)	–	(0.0015)	–	(0.0011)
DD < 14	–	−0.0074	–	−0.0157***	–	−0.0030
	–	(0.0055)	–	(0.0039)	–	(0.0067)
Precipitation	−0.0185***	−0.0187***	−0.0220***	−0.0224***	−0.0159***	−0.0160***
	(0.0018)	(0.0018)	(0.0020)	(0.0021)	(0.0017)	(0.0018)
Humidity	−0.0010***	−0.0010***	−0.0017***	−0.0016***	−0.0007***	−0.0007***
	(0.0003)	(0.0003)	(0.0004)	(0.0004)	(0.0002)	(0.0002)
Wind	−0.0010	−0.0010	0.0023	0.0022	−0.0023	−0.0023
	(0.0016)	(0.0016)	(0.0019)	(0.0018)	(0.0016)	(0.0016)
Sunshine	0.0002	0.0003	0.0025***	0.0026***	−0.0005	−0.0005
	(0.0006)	(0.0006)	(0.0007)	(0.0007)	(0.0006)	(0.0006)
Pressure	−0.0017***	−0.0018***	−0.0022***	−0.0025***	−0.0013***	−0.0014***
	(0.0004)	(0.0004)	(0.0005)	(0.0004)	(0.0005)	(0.0005)
Visibility	−0.0033***	−0.0033***	−0.0056***	−0.0056***	−0.0021***	−0.0021***
	(0.0007)	(0.0007)	(0.0008)	(0.0008)	(0.0007)	(0.0007)
Observations	12,672,771	12,672,771	3,699,915	3,699,915	8,972,856	8,972,856
R-squared	0.0550	0.0550	0.0666	0.0667	0.0568	0.0568

Notes: The dependent variable is a dummy variable, which equals one if a student's score is above or equal to the cutoff of the first-tier universities and zero otherwise. All students need to take three compulsory subjects: Chinese, mathematics, and foreign language (typically English). Students in the art track need to take one combined subject comprising politics, history, and geography, and students in the science track need to take one combined subject comprising physics, chemistry, and biology. The data on the cutoff of the first-tier universities are only available for the art and science tracks. Regression models also include county fixed effects and year fixed effects. Degree days (DD) ≥14 (<14) is the number of degrees above (below) 14 °C. Standard errors are clustered at the county level and reported in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

results remain robust. In column (4), we use the Conley standard error (Conley, 1999) to adjust for spatial correlation based on a radius of 200 km. Since it is infeasible to run the regression using the Conley standard error at the individual level, we run the model at the county-year level. Our results are robust.[22]

In column (5) of panel A, we use the level of score, instead of the log of score in the baseline model, as the dependent variable. The point estimate is very close to the estimate when we use the log of score and evaluated at the mean level. In column (5) of panel B, we use the logit model and report the marginal effect evaluated at the mean level. The estimate is similar to the linear model. While the NCEE is held in most provinces on June 7th and 8th only, some provinces also have exams on June 9th. Note that the exam on the 9th only occurs in the morning and accounts for less than 10% of total scores in most provinces. Nevertheless, in column (6), we calculate the average of temperature on June 7th to 9th for provinces with a three-day exam. The results are robust.

Currently, the main specification is the average daily temperature in June 7th and 8th. To mitigate the attenuation bias from averaging the two days, we use the maximum of daily mean temperature on the 2 Em day in column (7) and find robust results. Lastly, in column (8), we use the wet-bulb temperature, which is a combination between temperature and humidity. Our results are again robust.

In Table A8 in the Online Appendix, we check the robustness of our findings using different fixed effects specifications. Our baseline model in column (1) includes county fixed effects and year fixed effects. We then add year-by-track fixed effects in column (2) to control for year-specific difficulty between the art and science track. Our results change little. In column (3), we add year-by-province fixed effects. As expected, our results become statistically insignificant because these fixed effects absorb a significant amount of temperature variation (Table A1 in the Online Appendix). We further add year-by-track-by-province fixed effects in column (4). Again, the results are statistically insignificant.[23] In column (5), we return to county fixed effects and year fixed effects and add province-specific linear time trends to capture any province-year level variation in exam difficulty that might be driving our empirical findings. In the last column, we replace linear time trends with quadratic time trends to capture potential non-linearities. In both cases, our results are robust.[24]

[22] These findings are also robust to alternate choices of radii, ranging from 100 to 300 km.

[23] Our results are also insignificant if we standardize the test scores by year-province-track. This is similar to the approach that uses year-province-track fixed effects, as they absorb a significant amount of variations.

[24] It is important to note that these linear time trends may reflect two distinct phenomena in our data. On the one hand, they may capture the province-year level variation in exam difficulty that is correlated with temperature as desired. Unfortunately, they will also capture the impacts of temperature on learning, at least insofar as weather and test scores are trending in a given province. This latter effect is one of interest, rather than a threat to identification, but both explanations are indistinguishable from one another using this approach.

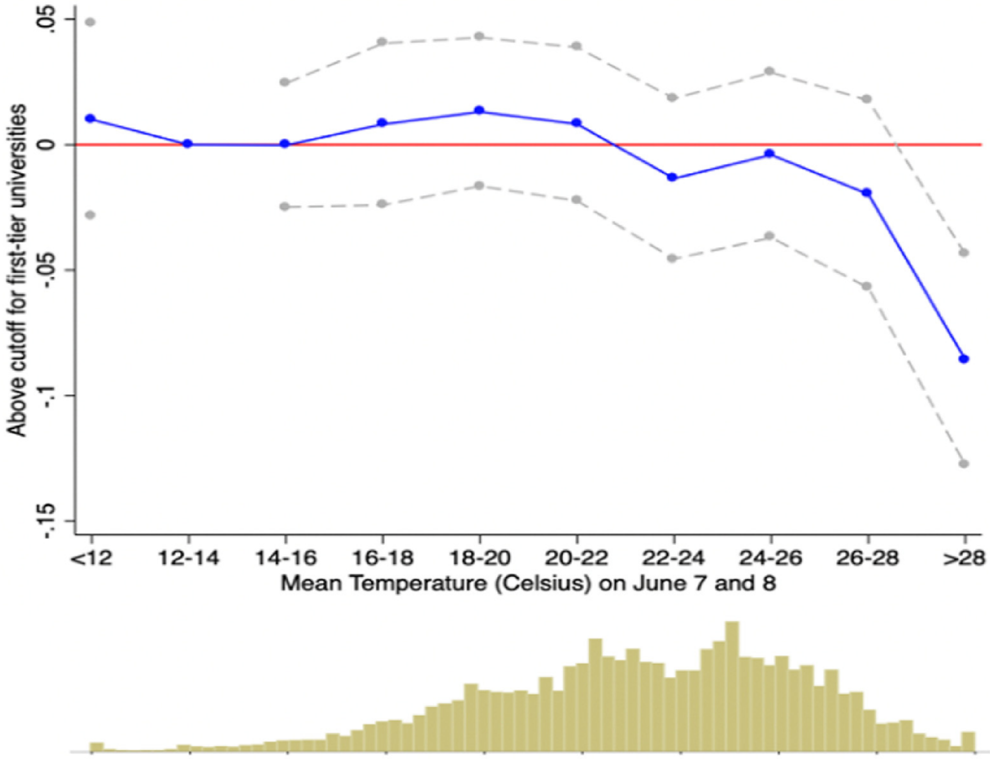

Fig. 5. Relationship between temperature and probability of getting into first-tier universities. *Notes:* The upper panel of the figure shows the non-linear effects of temperature on the probability of admission in first-tier universities, and the lower panel shows the temperature distribution. Each point estimate represents the effect of replacing a day with temperature in the 12–14 °C interval (reference group) with a day with temperature in the corresponding interval. Control variables include: precipitation, relative humidity, wind speed, sunshine duration, pressure, visibility, county fixed effects, and year fixed effects. Whiskers denote the 95% confidence interval, after adjusting for spatial and serial correlation within each county.

5.10. Climate prediction

In this section, we present a stylized forecast of the implied impact on test scores under climate change. This is largely an illustrative exercise, with the important caveat that the slow evolution of climate change will afford many opportunities for adaptation that could alter the relationship estimated in this paper (see Dell et al. (2014) for a more detailed discussion). To begin, we download the average projection from 39 downscaled climate models from the Coupled Model Intercomparison Project 5 (CMIP5) (Taylor et al. 2012). We focus on multiple climate models to account for uncertainties within each climate model (Burke et al., 2015). For simplicity, we center our analyses on four Representative Concentration Paths (RCPs): 2.6, 4.5, 6.0, and 8.5, in which RCP 2.6 is the slowest warming scenario and RCP 8.5 is the fastest. The CMIP5 reports global monthly temperature predictions at a spatial resolution of 2.5° (longitude) *2.5° (latitude) for each year until 2099.

Armed with these downscaled estimates, we calculate the temperature difference in June between two periods: 2005–2011 and 2070–2099 for each grid point within China. We then assign temperature changes to each county by inverse distance weighting all grid points within a 200 km radius. Lastly, we multiply the county-level temperature changes by our estimated coefficient of −0.0034 to estimate the change in scores for each county under climate change. Fig. 6 illustrates the changes in each county under RCP 8.5. Figures A6–8 in the online appendix illustrate the changes under other RCPs. As can be seen from the figures, counties in the west are expected to experience larger and more variable drops in scores, potentially exacerbating inequality *within* provinces and misallocative efficiencies *across* provinces under climate change.

6. Discussion and conclusion

In this paper, we show that temperature plays an important role in high-stakes cognitive performance using data from the NCEE, the most important academic examination in China. In particular, a 1 °C increase in temperature during the exam period decreases total test scores by 2.91% of a standard deviation. We compare our estimates to two papers that have studied temperature effects on exam scores in the United States. Park (forthcoming) finds that a 1 °C increase in temperature during the exam period decreases total test scores by 1.60% of a standard deviation for similarly aged students in New York City. Graff Zivin et al. (2018) find that a 1 °C increase in temperature during the exam period decreases total test scores by 1.2% of a

Table 9
Robustness checks.

Panel A	Dependent variable: Log of exam scores							
	Baseline	Clustering prefecture	County-year	Conley SE	Level of score	Three days	Max temp	Wet-bulb temp
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Temperature	−0.0034***	−0.0034***	−0.0026***	−0.0025***	−1.7453***	−0.0022***	−0.0034***	−0.0024**
	(0.0004)	(0.0009)	(0.0003)	(0.0011)	(0.2062)	(0.0004)	(0.0004)	(0.0002)
Observations	14,042,417	14,042,417	14,177	14,177	14,042,417	14,042,417	14,042,417	14,042,417
Panel B	Dependent variable: Above cutoff for first-tier universities							
	Baseline	Clustering prefecture	County-year	Conley SE	Logit	Three days	Max temp	Wet-bulb temp
Temperature	−0.0060***	−0.0060**	−0.0078***	−0.0078***	−0.0289***	−0.0029***	−0.0033***	−0.0030**
	(0.0012)	(0.0026)	(0.0009)	(0.0032)	(0.0058)	(0.0010)	(0.0010)	(0.0006)
Observations	12,672,771	12,672,771	14,177	14,177	12,672,771	12,672,771	12,672,771	12,672,771

Notes: In panel A, the dependent variable is the log of exam score except for column (3), where the dependent variable is the level of score. In panel B, the dependent variable is a dummy variable which equals to one if the student's score is above or equal to the eligibility cutoff for the first-tier universities and zero otherwise. Column (1) is the baseline model. In column (2), we cluster standard errors by prefecture, to control for serial and spatial correlation within prefecture. Note that prefecture is an administrative unit between province and county. In column (3), we collapse observations by county-year, and estimate the model using count-year observations. In column (4), we use the Conley standard error (Conley, 1999) to adjust for spatial correlation based on a radius of 200 km. Since it is infeasible to use the Conley standard error at the individual level, we run the model at the county-year level. In column (5) of panel A, we use the level of the score as the dependent variable. In column (5) of panel B, we use the logit model and report the marginal effects evaluated at the mean level. In column (6), we include temperature on June 9th for provinces with exams held on June 7th–9th. In column (7), temperature is measured using the maximum of daily mean temperature on the 2 Em day, June 7th and June 8th. In column (8), temperature is measured as wet-bulb temperature. Standard errors are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

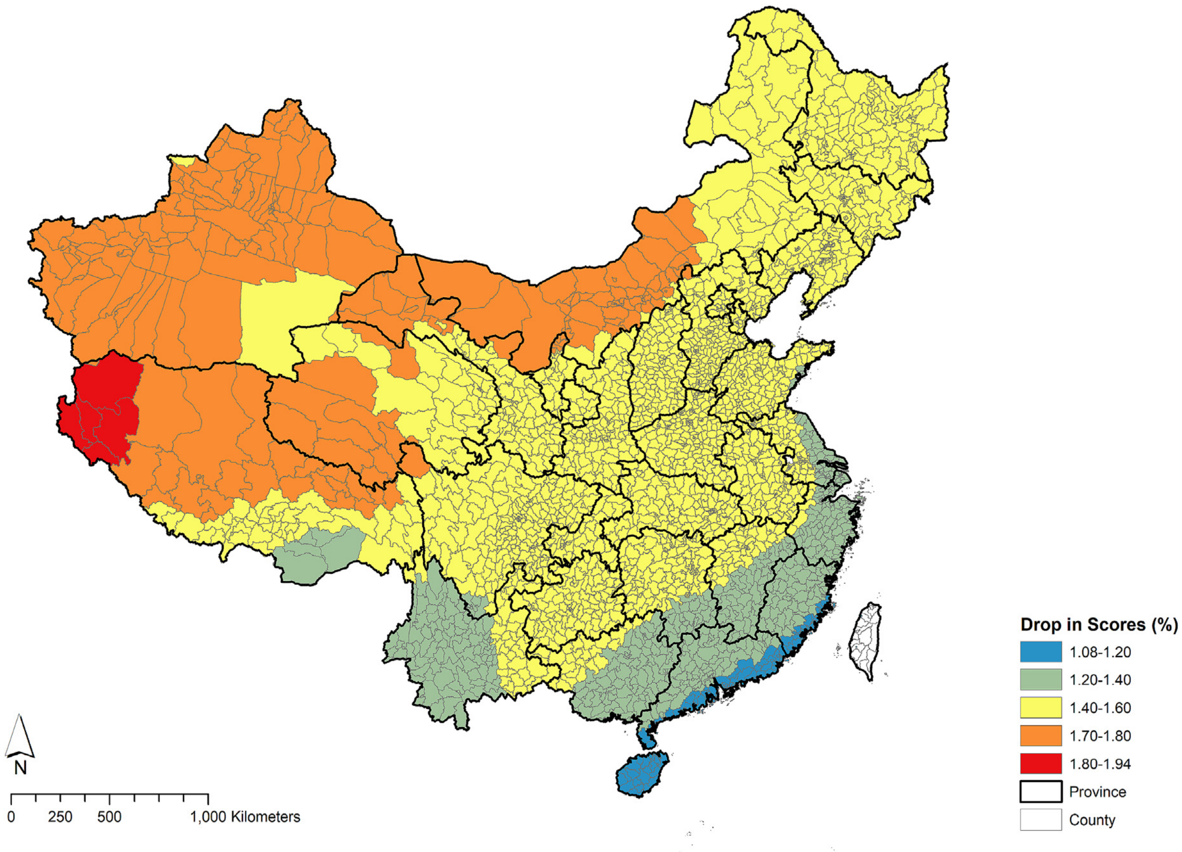

Fig. 6. The impacts of climate change (2070–2099) on NCEE scores under RCP 8.5. *Notes:* This map presents the impacts of climate change on NCEE scores in each county under RCP 8.5. Impacts are measured in percentage points.

standard deviation for young children across the United States. The larger magnitude in our setting may be a reflection of the higher-stakes environment, the limited access to air conditioning, or fundamental differences in our study populations.

Our results also imply that students in hotter regions may have disadvantages compared with their peers in cooler regions within each province, highlighting potentially important concerns about equitable access to higher education within China under the NCEE system. We believe one effective policy is to install and use air conditioning in the exam rooms. Ironically, some regions prohibit the use of air conditioning to enhance the fairness to regions where air conditioning is not available, which misses the important point that some regions are always hotter than others and that the use of air conditioning may help level the playing field across regions.

It is worth emphasizing that our focus is short run cognitive performance. If our estimates have implications for performance outside of the standardized testing environment, then repeated exposure to heat may well retard performance in the classroom and thus the accumulation of human capital in the long run. Even absence such long-run impacts on human capital attainment, our results have profound long-run distributional implications. The NCEE is an extremely high-stakes exam that governs access to institutions of higher learning and ultimately professional success. Its sensitivity to random temperature shocks, generates an inefficient allocation of students to universities and ultimately to the workplace (Ebenstein et al., 2016). Under climate change, those distributional impacts will exacerbate existing east-west inequalities absent significant policy interventions.

Though our empirical setting is China, our results have important implications for other developing countries that utilize standardized testing to govern access to institutions of higher learning or access to particular professions. Whether these results generalize to a developed country setting, where air conditioning is more prevalent, remains an open question. Nonetheless, the significant effect of temperature on cognitive performance suggests another potential channel through which future climate change may affect economic well-being.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jeem.2020.102365.

Notes: The dependent variable is the log of the exam score. Column (1) reports the baseline estimates, which uses the average of daily mean temperature during the exam period. Columns (2) and (3) uses the average of daily minimum and maximum temperatures respectively. Column (4) includes both. Column (5) returns to daily mean temperature and further adds the difference of temperatures between 2 Em day. All regression models include other weather controls, county fixed effects and year fixed effects. Standard errors are clustered at the county level and reported in parentheses. ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.

References

Bai, Chong-en, Chi, Wei, Qian, Xiaoye, 2014. Do college entrance examination scores predict undergraduate GPAs? A tale of two universities. China Econ. Rev. 30, 632–647.

Barreca, Alan, Clay, Karen, Olivier, Deschênes, Greenstone, Michael, Shapiro, Joseph S., 2016. Adapting to climate change: the remarkable decline in the US temperature-mortality relationship over the twentieth century. J. Polit. Econ. 124 (1), 105–159.

Bowler, K., Tirri, R., 1974. The temperature characteristics of synaptic membrane ATPases from immature and adult rat brain. J. Neurochem. 23 (3), 611–613.

Burke, Marshall B., Miguel, Edward, Satyanath, Shanker, Dykema, John A., Lobell, David B., 2009. Warming increases the risk of civil war in Africa. Proc. Natl. Acad. Sci. Unit. States Am. 106 (49), 20670–20674.

Burke, Marshall, Dykema, John, Lobell, David B., Miguel, Edward, Satyanath, Shanker, 2015. Incorporating climate uncertainty into estimates of climate change impacts. Rev. Econ. Stat. 97 (2), 461–471.

Cai, Xiqian, Lu, Yi, Pan, Jessica, Zhong, Songfa, 2019. Gender gap under pressure: evidence from China's national college entrance examination. Rev. Econ. Stat. 101 (2), 249–263, 2019.

Chen, Yan, Onur Kesten, 2017. Chinese college admissions and school choice reforms: a theoretical analysis. J. Polit. Econ. 125 (1), 99–139.

China Education Online (2016). Retrieved from http://www.eol.cn/html/g/report/2016/report1total.shtml on March 30, 2020.

Conley, Timothy G., 1999. GMM estimation with cross sectional dependence. J. Econom. 92 (1), 1–45.

Deboer, Tom, 1998. Brain temperature dependent changes in the electroencephalogram power spectrum of humans and animals. J. Sleep Res. 7 (4), 254–262.

Dee, Thomas S., Dobbie, Will, Jacob, Brian A., Rockoff, Jonah, 2019. The causes and consequences of test score manipulation: evidence from the New York regents examinations. Am. Econ. J. Appl. Econ. 11, 382–423.

Dell, Melissa, Jones, Benjamin F., Olken, Benjamin A., 2012. Temperature shocks and economic growth: evidence from the last half century. Am. Econ. J. Macroecon. 4 (3), 66–95.

Dell, Melissa, Jones, Benjamin F., Olken, Benjamin A., 2014. What do we learn from the weather? The new climate-economy literature. J. Econ. Lit. 52 (3), 740–798.

Deschênes, Olivier, Greenstone, Michael, 2007. The economic impacts of climate change: evidence from agricultural output and random fluctuations in weather. Am. Econ. Rev. 97 (1), 354–385.

Deschênes, Olivier, Greenstone, Michael, 2011. Climate change, mortality, and adaptation: evidence from annual fluctuations in weather in the US. Am. Econ. J. Appl. Econ. 3 (4), 152–185.

Diamond, Rebecca, Persson, Petra, 2016. "The Long-Term Consequences of Teacher Discretion in Grading of High-Stakes Tests." No. W22207. National Bureau of Economic Research.

Ebenstein, Avraham, Lavy, Victor, Roth, Sefi, 2016. The long-run economic consequences of high-stakes examinations: evidence from transitory variation in pollution. Am. Econ. J. Appl. Econ. 8 (4), 36–65.

Garg, Teevrat, Jagnani, Maulik, Taraz, Vis, 2018. Temperature and Human Capital in India.

Goodman, Joshua, Michael Hurwitz, Jisung Park, and Jonathan Smith (forthcoming). "Heat and Learning." *American Economic Journal-Economic Policy*.

Graff Zivin, Joshua, Neidell, Matthew, 2014. Temperature and the allocation of time: implications for climate change. J. Labor Econ. 32 (1), 1–26.

Graff Zivin, Joshua, Solomon M. Hsiang, Neidell, Matthew, 2018. Temperature and human capital in the short and long run. Journal of the Association of Environmental and Resource Economists 5 (1), 77–105.

Guven, Cahit, 2012. Reversing the question: does happiness affect consumption and savings behavior? J. Econ. Psychol. 33 (4), 701–717.
Heutel, Garth, Miller, Nolan H., Molitor, David, 2017. Adaptation and the Mortality Effects of Temperature across US Climate Regions. National Bureau of Economic Research. No. w23271.
Hocking, Chris, Silberstein, Richard B., Lau, Wai Man, Stough, Con, Roberts, Warren, 2001. Evaluation of cognitive performance in the heat by functional brain imaging and psychometric testing. Comp. Biochem. Physiol. Mol. Integr. Physiol. 128 (4), 719–734.
Hsiang, Solomon, M., 2010. Temperatures and cyclones strongly associated with economic production in the Caribbean and Central America. Proc. Natl. Acad. Sci. Unit. States Am. 107 (35), 15367–15372.
Hyde, Dale, Thomas, John R., Schrot, John, Taylor, W.F., 1997. *Quantification Of Special Operations Mission-Related Performance: Operational Evaluation of Physical Measures*. No. NMRI-97-01. Naval Medical Research Institute, Bethesda MD.
Jia, Ruixue, Li, Hongbin, 2017. "Access to Elite Education, Wage Premium, and Social Mobility: the Truth and Illusion of China's College Entrance Exam.
Koretz, Daniel, Deibert, Edward, 1995. Setting standards and interpreting achievement: a cautionary tale from the National Assessment of Educational Progress. Educ. Assess. 3 (1), 53–81.
Mendelsohn, Robert, Nordhaus, William D., Shaw, Daigee, 1994. The impact of global warming on agriculture: a Ricardian analysis. Am. Econ. Rev. 753–771.
Ministry of Environmental Protection of China (2006). Retrieved from http://jcs.mep.gov.cn/hjzl/200604/t20060428_76218.shtml on March 30, 2020.
Moulton, Brent R., 1986. Random group effects and the precision of regression estimates. J. Econom. 32 (3), 385–397.
Muthanna, Abdulghani, Sang, Guoyuan, 2015. "Undergraduate Chinese students' perspectives on Gaokao examination: strengths, weaknesses, and implications. International Journal of Research Studies in Education 5 (2), 3–12.
Oswald, Andrew J., Proto, Eugenio, Sgroi, Daniel, 2015. Happiness and productivity. J. Labor Econ. 33 (4), 789–822.
Park, Jisung (forthcoming). "Hot temperature and high stakes cognitive performance", J. Hum. Resour..
Ranson, M., 2014. Crime, weather, and climate change. J. Environ. Econ. Manag. 67 (3), 274–302.
Robelen, E.W., 2002. An ESEA primer, 21. Educ Week Feburary.
Schiff, Steven J., Somjen, George G., 1985. The effects of temperature on synaptic transmission in hippocampal tissue slices. Brain Res. 345 (2), 279–284.
Schlenker, Wolfram, Michael Hanemann, W., Fisher, Anthony C., 2006. The impact of global warming on US agriculture: an econometric analysis of optimal growing conditions. Rev. Econ. Stat. 88 (1), 113–125.
Schwarz, Norbert, Clore, Gerald, 1983. Mood, misattribution, and judgments of well-being: informative and directive functions of affective states. J. Pers. Soc. Psychol. 45, 513–523.
Sievertsen, Hans Henrik, Gino, Francesca, Piovesan, Marco, 2016. "Cognitive fatigue influences students' performance on standardized tests. Proc. Natl. Acad. Sci. Unit. States Am. 113 (10), 2621–2624.
Sina (2007). Retrieved from http://news.sina.com.cn/c/2007-06-07/152711978182s.shtml on March 30, 2020.
Sina (2014). Retrieved from http://news.sina.com.cn/c/2014-06-05/070830296473.shtml on March 30, 2020.
Taraz, V., 2018. Can farmers adapt to higher temperatures? Evidence from India. World Dev. 112, 205–219.
Us Legal (2014). "Standardized test [education] law and legal definition", Retrieved from https://definitions.uslegal.com/s/standardized-test-education on March 30, 2020.
Vasmatzidis, Ioannis, Schlegel, Robert E., Hancock, Peter A., 2002. An investigation of heat stress effects on time-sharing performance. Ergonomics 45 (3), 218–239.
Yablonskiy, Dmitriy A., Ackerman, Joseph JH., Raichle, Marcus E., 2000. Coupling between changes in human brain temperature and oxidative metabolism during prolonged visual stimulation. Proc. Natl. Acad. Sci. Unit. States Am. 97 (13), 7603–7608.
Zhu, Hong Zhen, Lou, Shiyan, 2011. Development and Reform of Higher Education in China. Elsevier, 2011.