IZA Institute of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17643

Degrees of Deception: How Score Manipulation Mitigates Temperature's Impact on Student Performance

Rosario Maria Ballatore
Alessandro Palma
Daniela Vuri

JANUARY 2025

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17643

Degrees of Deception: How Score Manipulation Mitigates Temperature's Impact on Student Performance

Rosario Maria Ballatore
Bank of Italy

Alessandro Palma
GSSI and Tor Vergata University

Daniela Vuri
Tor Vergata University, IZA and CESifo

JANUARY 2025

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Degrees of Deception: How Score Manipulation Mitigates Temperature's Impact on Student Performance*

Using Italian data on the universe of mandatory tests conducted in a low-stakes setting without air conditioning, we investigate the effect of temperature on student performance, with a focus on how manipulation distorts causal estimates of temperature effects on test scores. While high temperatures adversely affect students' performance, we find that score manipulation also increases with temperature within a specific range. Leveraging the random assignment of inspectors to schools as a natural experiment, we estimate the effect of temperature on test scores net of manipulation. We find that achievement declines at lower temperature thresholds when manipulation is accounted for, implying a larger number of affected students than previously estimated. Additionally, individual survey responses collected during the tests indicate that very high temperatures induce shifts in students' emotional states, affecting self-esteem and anxiety levels.

JEL Classification: J21, J24, Q54, O15

Keywords: student performance, temperature, manipulation, cognitive ability, emotional stress

Corresponding author:
Daniela Vuri
Department of Economics and Finance
University of Rome "Tor Vergata"
Via Columbia, 2
00133 Rome
Italy
E-mail: daniela.vuri@uniroma2.it

* We are grateful to Joshua Angrist, Joshua Graff Zivin, Ludovica Gazzé, Giulia Bovini, Luca Citino, Marzio Galeotti, Guido De Blasio, Marco Tonello and Francesco Vona for useful suggestions. We also thank participants to the Bank of Italy webinars of the project "The effects of climate change on the Italian economy", Fondazione ENI Enrico Mattei invited seminar, 11th IAERE Annual Conference, SKILS 2025 Conference. We are also grateful to Giuliano Marco Federico Rolle for excellent research assistance. We are solely responsible for any and all errors. The views expressed in this paper are those of the authors and do not reflect those of the Bank of Italy.

1 Introduction

A relatively recent but influential body of literature has focused on the impact of temperature fluctuations on cognitive demanding tasks, such as standardized assessments at school (Cho, 2017; Park et al., 2020; Park, 2022; Garg et al., 2020; Graff Zivin et al., 2018; Zivin et al., 2020). A common finding in these works is that higher temperatures are related to lower cognitive performance in standardized tests, especially in math-related subjects and for certain minority groups.[1]

While these studies have significantly contributed to understanding the effect of temperature on cognitive performance at school, they overlooked a critical aspect of the relationship between these two factors: the potential for score adjustment strategies to mitigate the impact of temperature on student achievement.[2] When cognitive functioning declines as temperatures rise during standardized tests, students or teachers might use compensatory practices to contrast the negative effects of high temperatures. These strategies may include activities like copying from other classmates, teachers suggesting correct answers to students or inflating scores during the test grading (Battistin, 2016; Lucifora and Tonello, 2015). Failing to consider these potential mechanisms could result in a significant underestimation of temperature's true impact on students' cognitive performance.

We fill this gap by showing that manipulation occurs within specific temperature ranges, creating a wedge between the *observed* and *true* effect of temperature on student's performance. We analyze this aspect using the universe of mandatory examination records in Italy between school years 2011-12 and 2016-17, provided by the National Institute for the Evaluation of the Education System (INVALSI), together with data on temperatures on the days of the tests at the municipality level.

The Italian national assessment setting offers a unique opportunity to study the relationship among temperatures, cognitive performance and score manipulation. First,

[1]Besides the impact on cognitive ability, temperature has proved to raise mortality rate and disease burden (Deschênes and Greenstone, 2011; Huang et al., 2012; Karlsson and Ziebarth, 2018; Banerjee and Maharaj, 2020; Lee and Li, 2021), increase the risk of mental illness and suicide rates (Obradovich et al., 2018; Mullins and White, 2019; Burke et al., 2018; Martinelli and Palma, 2024), and reduce labor supply (Deschênes, 2014; Graff Zivin and Neidell, 2014), as well as agricultural income and nutrition (Deschênes and Greenstone, 2007; Shah and Steinberg, 2017) and consumption behavior (Lee and Zheng, 2022).

[2]An important exception is Park (2022)'s study, which represents a first attempt to quantify the manipulation of test results. Exploiting the test score threshold, Park uses bunching estimates as a measure of manipulation and finds evidence of upward grade manipulation. The effects are quite pronounced and provide ample motivation for further research on this aspect. However, Park identifies manipulation activities only by teachers as a form of ex post compensation in exams conducted on hotter days. In our study we directly observe a statistical measure of manipulation, which can be performed by both students and teachers.

students or teachers cannot control the timing of the tests, making temperature shocks unrelated to schools or students' characteristics. Second, unlike most of the other countries analyzed in previous studies, air conditioning penetration is very low in Italian schools, which make our estimates unbiased from this potential confounding factor. Third, and most important for our purpose, our data provides a measure of manipulation in classes. Classes with likely manipulated scores are identified using a statistical model that detects unusually high average scores, low within-class variability, and implausible data patterns (see Section 3.2 for details).

These unique features lead to our empirical strategy. Since the evaluation tests are scheduled at the national level several months in advance, we exploit the quasi-random variation in temperature across test dates in subsequent academic years within school-by-grade cells. To net out the potential attenuating effect of any in-test compensating behaviors and estimate the true effect of temperature students' performance, we explicitly control for the degree of manipulation at the class level. However, we cannot simply compare the outcomes of classes whose test scores were manipulated with those whose test scores were left unmanipulated, as these two groups could differ in terms of observable and unobservable factors as well as cognitive performance during the examination. To address this issue, we leverage a unique feature of the examination procedure in Italy: the random assignment of external monitors to schools during exam administration. In the same spirit of Angrist et al. (2017) in their study of class size and achievement, we use this natural experiment to frame the analysis in an IV setting where score manipulation is orthogonal to student's and school's characteristics.

Providing clean estimates of the effect of temperature shocks on cognitive performance is important for several reasons. As far as students are concerned, the short-run effects on performance could indicate a decline in students' learning and skills and potentially lead to negative labor market outcomes and overall economic growth in the long-run (Deschênes, 2014; Graff Zivin et al., 2018). In addition, school assessments are often used to compare different geographic areas and formulate policies to address regional disparities. Thus, it is crucial to fully understand how environmental factors interact with educational outcomes and contribute to exacerbate regional disparity through the channel of climate inequality (Park et al., 2021b; Hallegatte and Rozenberg, 2017). Lastly, cognitive performance plays a critical role in numerous aspects of our life, such as competitive examinations, college admissions, and financial decision-making. Any evidence of reduced cognitive function at high temperatures could have substantial implications for scheduling cognitively demanding tasks optimally (Graff Zivin and Neidell, 2013).

Our OLS estimates (without accounting for manipulation) show that higher temperatures adversely affect students' math test performance, with negative effects observed already at 23°C, and a peak decline of approximately 0.05 standard deviations (s.d.) at temperatures over 31°C. No significant effects of temperature are instead observed on Italian test scores. At the same time, manipulation increases up to 30°C but then declines sharply beyond this threshold, becoming negligible afterwards. The onset of negative effects at relatively moderate temperatures (23-30°C) and a gradual increase with rising temperatures aligns with findings from previous studies that analyze different outcomes.[3] However, the decline of manipulation for temperatures above 30°C is relatively new. We provide two explanations for this highly non-linear trend. First, manipulation is an activity that requires cognitive effort and, as such, is affected by high temperatures. This implies that when it is too hot, students or teachers struggle to access their cognitive functions, which are also used for engaging in manipulation. The second mechanism relates to the quality of the manipulation, as its effectiveness may not be granted. Very high temperatures may indeed interfere with the quality of the manipulation, reducing its effectiveness and consequently the impact on the final test score.

When we account for compensatory behavior during the test, our IV estimates reveal that the impact of temperature on scores starts at a lower temperature threshold (23°C-26°C), reaching a peak reduction of approximately 0.08 s.d. between 27°C and 30°C (compared to -0.022 in the same bin of the OLS estimation), and it remains relatively stable afterwards. The size of the effects is not negligible as it corresponds to a reduction of about 8 percent in earning 7 years after high school, especially for women and men with low initial test scores (Rose, 2006). The occurrence of negative effects at warm, but not extreme, temperatures carries significant implications, as it means

[3] In particular, regarding students' cognitive abilities, Krebs (2022) and Park (2022) find a reduction in test scores when the temperature exceeds approximately 22-25°C, while Park et al. (2021b) estimate a significant reduction in PSAT scores on school days with maximum temperatures below 20°C. For other outcomes potentially connected to attention deficits and cognitive capacity, Filomena and Picchio (2024), as well as Park et al. (2021a), find that workplace accidents significantly increase with temperatures higher than 20-22°C.

that a greater number of students are impacted. In fact, the number of days with non-extreme temperatures throughout the year, as well as the geographic regions affected, are significantly larger. As before, we find no significant impact on language tests.

We also provide suggestive evidence on the role of emotional status when students are exposed to high temperatures, and find an increase of anxiety and a reduction of self-esteem while attending the test. This result is consistent with experimental studies showing that exposure to extreme heat affects neurotransmitter levels in the brain, including those responsible for regulating emotional states such as anxiety (Nakagawa and Ishiwata, 2021), and causes inflammation in the hippocampus, affecting cognitive capacity (Chauhan et al., 2021; Lee et al., 2015).

Our paper provides several contributions to the existing literature. First, we expand the significant body of work investigating the relationship between exposure to extreme heat and human capital formation (Park et al., 2021b). This investigation is crucial as global average temperatures continue to rise, with prolonged peaks of extreme temperatures occurring earlier in the season and becoming increasingly common (WMO, 2023). Moreover, despite score manipulation has being recognized as a crucial factor in determining school accountability (Mansfield and Slichter, 2021; Battistin, 2016; Angrist et al., 2017), there have been relatively few studies examining its external determinants (Persico and Venator, 2021). In this regard, we are the first to offer a large-scale study that provides a clean estimate of the temperature's effect on students' cognitive performance. We achieve this by estimating how temperature influences manipulation and how the final score is affected by temperature itself, net of manipulation. The focus on manipulation therefore expands our understanding of how environmental factors affect not only score manipulation, but also levels of accountability within the education system. Second, our research speaks directly to the literature investigating the mechanisms in human behavior under heat stress. This includes not only the physiological literature, which has shown that prolonged exposure to an excessively hot environment disrupts core cognitive abilities (Taylor et al., 2016), including memory (Gaoua et al., 2011; Lee et al., 2015) and decision-making (Froom et al., 1993; Coehoorn et al., 2020), but also the economic literature analyzing the impacts of extreme heat on human behavior such as changes in temperament and expressed sentiment (Baylis, 2020), mental health disorders (Basu et al., 2018; Martinelli and Palma, 2024), or even more severe consequences such as increased suicides and children maltreatment (Burke et al., 2018; Evans et al., 2023). Analyzing data from more than seven million students, which include information on test perceptions such as anxiety, this study examines emotional disruption as a potential mechanism influencing students' cognitive performance when faced with high temperatures. Lastly, our paper is also related to the literature on academic performance and allocations to educational resources, encompassing both overall school resources (Jackson et al., 2015; Lafortune et al., 2018; Jackson and Mackevicius, 2024) and those specifically devoted to school infrastructure (Cellini et al., 2010; Park et al., 2020). While this literature often struggles to isolate the impact of school air conditioning from other aspects of school facilities, with the notable exception of Park et al. (2020), our analysis is conducted in a setting where air conditioning is rarely used in school buildings. This enables us to evaluate the impact of temperature stress without being affected by the controlled environment created by air conditioning. Furthermore, the uniformity of school programs and calendars throughout Italy ensures that we are comparing students' performance based on nearly identical study programs and equal time spent in school, thereby keeping educational inputs constant.

The rest of the paper is organized as follows. Section 2 presents a simple conceptual framework that links cognitive performance, temperature and score manipulation to guide the empirical analysis. Section 3 describes the data and institutional context and presents key summary statistics. Section 4 shows the effect of temperature on manipulation and test score, while Section 5 presents evidence of the effect of temperature on test score net of manipulation. Section 6 tests the robustness of our results and Section 7 explores one of the potential mechanisms that could explain our findings. Section 8 concludes.

2 Cognitive performance, temperature and score manipulation

We provide a simple conceptual framework that highlights how the effect of temperature on cognitive performance can be distorted by score manipulation.

Let us define P^O as the observed score on the day of the assessment. The easiest way to see this variable is as the number of correct answers to a standardized test, with one component linked to students' *true* cognitive performance (P), reflecting their skills and knowledge, and another component related to potential score manipulation (C), like copying from peers or receiving answers from teachers.

For the sake of simplicity we assume that the return to manipulation in terms of effect on the observed score is the same as cognitive performance.[4] This means that the correct answers obtained through manipulation contribute similarly to those provided by students themselves, implying that manipulation is always effective in enhancing the observed score. This assumption seems reasonable within our context, as an extensive literature documents that in Italy manipulation during assessment stems not only from students but also from teachers, either by suggesting correct answers to students, or inflating scores during grading, or allowing them use materials and collaborate (Bertoni et al., 2013; Angrist et al., 2017; Lucifora and Tonello, 2020). The observed score at the assessment thus takes the form:

$$P^O = f(P, C) = P + C \tag{1}$$

Let us ignore other potential factors and focus on the idea that cognitive performance changes according with the temperature during the test day $P = g(T)$, with $\frac{dg}{dT} < 0$, as emphasized by Graff Zivin et al. (2018) and Park (2022), among others. The crucial assumption in this basic framework is that manipulation depends on cognitive performance $C = h(P)$, as students or teachers may try to compensate for low performance on standardized tests, or refrain from manipulating scores when there is no need to do so.

As a simple formulation, in this model we assume that temperature influences manipulation only through its effect on cognitive performance, with no direct effects as follows:

$$P^O = g(T) + h(g(T)) \tag{2}$$

Deriving P^O with respect to T we obtain:

[4] For a generalization of the conceptual framework see Appendix A.

$$\frac{dP^O}{dT} = \frac{dg}{dT}\left(1 + \frac{dh}{dg}\right) \tag{3}$$

Equation (3) indicates that when manipulation is driven by temperature-induced variation in cognitive performance (P), the observed effect of temperature ($\frac{dP^O}{dT}$) differs from the *true* impact of temperature on cognitive performance ($\frac{dg}{dT}$), which poses challenges for the empirical identification when researchers only observe the overall assessment score (P^O), i.e. the *observed* performance. The key finding of this basic framework is that the extent of distortion depends on how students or teachers adjust the level of manipulation when cognitive performance varies ($\frac{dh}{dg}$).

As long as students or teachers use manipulation as a compensation when cognitive performance decreases, we have that $\frac{dh}{dg} \leq 0$, and the following predictions hold:

$$\frac{dP^O}{dT} = \frac{dg}{dT} \quad \text{if} \quad \frac{dh}{dg} = 0 \tag{4a}$$

$$\frac{dP^O}{dT} = 0 \quad \text{if} \quad \frac{dh}{dg} = -1 \tag{4b}$$

$$\left|\frac{dP^O}{dT}\right| < \left|\frac{dg}{dT}\right| \quad \text{if} \quad -1 < \frac{dh}{dg} < 0 \tag{4c}$$

$$\left|\frac{dP^O}{dT}\right| > \left|\frac{dg}{dT}\right| \quad \text{if} \quad \frac{dh}{dg} < -1 \tag{4d}$$

This implies that: (i) if there is no compensation when cognitive performance decreases ($\frac{dh}{dg} = 0$), the *observed* effect is exactly the effect of temperature on cognitive performance, i.e. the *true* effect; (ii) when the compensation is perfect (e.g. compensation occurs exactly for each items of the test for which the student does not know the correct answer: $\frac{dh}{dg} = -1$) we would observe no effect of temperature, even if the *true* effect on cognitive performance is different from zero; (iii) each time there is no perfect compensation (e.g. not all items of the test for which the answer is unknown are compensated: $-1 < \frac{dh}{dg} < 0$), the *observed* effect is a lower bound of the *true* effect of temperature on cognitive performance; iv) when there is overcompensation (e.g. the compensation is more than proportional to the number of items of the test for which the answer is not known: $\frac{dh}{dg} < -1$), the *observed* effect is larger than the *true* effect on cognitive performance.[5]

While this framework does not claim to explain every aspect of manipulation during assessments, it provides a basic theoretical mechanism that illustrates the interplay between cognitive performance, temperature, and score manipulation. This helps set up the empirical analysis and interpret our results. To empirically address the issue of bias, we use our institutional setting to exploit a measure of score manipulation and a natural experiment that breaks the link between manipulation and cognitive performance during assessments.

3 Institutional setting and data

3.1 The standardized test in the Italian school system

Italian schools have long used matriculation exams for tracking and placement in the transition from elementary to middle school and throughout high school, but starting from academic year 2009-10 standardized testing for evaluation purposes has become compulsory for all schools and students. The National Students' Assessment Survey (SNV) conduced by INVALSI is designed to assess students' achievement at different points of their school career and it is held on an annual basis. The assessment focuses on language and mathematics competencies of students attending grades 2$^{\text{nd}}$, 5$^{\text{th}}$, 8$^{\text{th}}$ and 10$^{\text{th}}$ by means of a standardized testing procedure.[6] Students are asked to answer a series of questions of different difficulties aimed at testing different skills: reading comprehension, grammar and lexical competences for the language test, and problem solving and logical skills for mathematics.[7] SNV tests include multiple choice questions and open-response items, for which some grading is required.

The SNV evaluations administered to students in 2$^{\text{nd}}$, 5$^{\text{th}}$ and 10$^{\text{th}}$ grades have low-stake nature. Indeed, the outcomes of these tests hold no bearing on students' future career paths, nor do they influence the allocation of school resources or the salaries of

[5] Although theoretically possible, we exclude the latter case as manipulation is a risky and costly activity.

[6] Grades 2$^{\text{nd}}$ and 5$^{\text{th}}$ correspond to ISCED level 1 (primary schools), grade 8$^{\text{th}}$ to ISCED level 2 (lower secondary), 10$^{\text{th}}$ corresponds to ISCED level 3 (upper secondary school). Starting from school year 2018-19 also grade 13$^{\text{th}}$ takes part to the national assessment.

[7] Starting from school year 2018-19 also foreign language skills are assessed.

teachers. The SNV test for 8$^{\text{th}}$ grade is instead considered as high-stake because it is part of the final examination and contributes to the final mark.[8]

In this paper, we focus on low-stake grades for two reasons. First, the tests for 2$^{\text{nd}}$, 5$^{\text{th}}$ and 10$^{\text{th}}$ grades are carried out every year in the first ten days of May and therefore students of different grades belonging to the same municipality are comparable in terms of temperature. Differently, the test for 8$^{\text{th}}$ grade is carried out between the second and the third week of June when students are already exposed to high temperatures and may exhibit very different responses to temperature than students in low-stake grades. Second, given the different nature of the test (low-stake vs high-stake), students in 8$^{\text{th}}$ grade may employ different strategies to manipulate the results compared to those in low-stake grades. This would require to run separate analyses for students in low-stake and in high-stake grades. Unfortunately, we can identify an exogenous source of variability in score manipulation only for low-stake grades and this motivates the decision to restrict our analysis to 2$^{\text{nd}}$, 5$^{\text{th}}$ and 10$^{\text{th}}$ grade students (see Section 5 for details).

Crucially for our analysis, the days of SNV assessments are the same for the whole national territory and cannot be manipulated by schools or regions. The dates are set centrally at the beginning of each school year, making it impossible to predict weather conditions on the day of the test. There is a difference between grades in the scheduling of language and math assessments: they take place within the same day for grade 10$^{\text{th}}$ and on two different days for grades 2$^{\text{nd}}$ and 5$^{\text{th}}$ (see Table 1).[9]

Although the tests take place in a controlled environment by the teachers, previous literature has shown that score manipulation is widespread also in low-stakes settings (Angrist et al., 2017; Lucifora and Tonello, 2015, 2020). Score manipulation indicates any dishonest or unfair action implemented by the students or teachers in order to obtain any profit or advantage in the evaluation of the performance. This could take place before the test (alteration of the pool of students; Figlio (2006)), during the test (students copying from one another or teachers telling the students the answers or

[8]Following an intense public debate on the opportunity to include the SNV test result in the final average grade of the middle school exam, from school year 2017-18 the SNV test for 8$^{\text{th}}$ grade was moved to April and became a prerequisite for accessing the final exam without contributing to the final grade anymore.

[9]The tests are administered following a protocol set by INVALSI, according to which proctoring is done by teachers from the same school but not from the same class and specialized in a subject different from the one being tested. In addition, teachers are expected to grade and then copy students' original responses onto machine-readable answer sheets (called *scheda risposta*) for submission to INVALSI.

lowering monitoring standards; Lazear (2006); Neal and Schanzenbach (2010); Angrist et al. (2016)), or after the test (unfair grading; Angrist et al. (2017); Jacob (2005); Dee et al. (2019); Diamond and Persson (2016); Park (2022)). However, the timing of the score manipulation is not an issue in our context, and we do not explore this aspect. Indeed, even manipulation after the examination could reflects a compensation of teachers for the temperature related deterioration of performance during the test.

In an effort to reduce score manipulation, INVALSI randomly assigns external monitors to institutions, and to specific classes within institution.[10] Monitors supervise test administration, encouraging compliance with INVALSI testing standards. They are also responsible for score sheet transcription in a sample of selected classes within the monitored schools. Regional education offices select monitors from a pool consisting of retired teachers and principals who have not worked in the past two years in the towns or at the schools they are assigned to monitor. The presence of an external inspector establishes a "non-cheating environment, where the possibility of manipulation on the part of both students and teachers, both during and after the test, is remarkably reduced" (Bertoni et al., 2013).[11]

3.2 INVALSI data

For each grade and subject, slightly less than 500,000 students take the SNV test every year. Scores indicate the percentage of correct answers. For the ease of interpretation, we standardized these by subject, year of survey, and grade to have zero mean and unit variance. Data on test scores are matched to administrative information describing institutions, schools, classes, and students. Students' data include gender, citizenship, parental employment status and educational background.

These data are collected as part of test administration and meant to be provided by school staff when scores are submitted. Additional individual-level information are collected through the Student Questionnaire, which is taken by 5th and 10th grade students after finishing the test. The Student Questionnaire contains information on

[10] In Italy an institution is the main administrative unit of the educational system. An institution is administered by a principal and it includes one or more schools.

[11] Classes in 8th grade are exempted by the assignment of an external monitor because an internal committee made by all the teachers of the class chaired by the school principal is in charge of proctoring and grading all the tests. The lack of an external monitor in 8th grade makes it impossible to causally estimate the effect of temperature on test scores net of manipulation and justifies the exclusion of 8th grade classes from the analysis.

students' perceptions while taking the test, such as anxiety, feeling of performing badly or feeling fine during the assessment, among other information. Importantly, the data also include anonymous school identifiers, which make it possible to follow schools over time. This is crucial for our empirical strategy which uses school fixed effects.

INVALSI has adopted a statistical procedure, developed by Quintano et al. (2009) to detect ex-post classes with manipulation. This variable is class and subject specific and can be interpreted as the part of the score that is achieved through manipulation. It is computed through four within-class statistics of class response behavior: (1) average class score, (2) within class variability, (3) level of heterogeneity in responses to each individual item of the questionnaire across all students in the class and (4) rate of missing data. In addition, in the data, we can also distinguish between classes for which the test is proctored and marked by an external inspector (monitored classes), and classes where the test is proctored by local school staff (not-monitored classes). In our sample, approximately 16% of institutions are assigned an external monitor. This information can be used as an exogenous source of variability for manipulation, as in Angrist et al. (2017) (see Section 5).

Although test scores data are available from school year 2009-10, we limit our analysis to the six consecutive test waves from 2011-12 to 2016-17. Two reasons lie behind this restriction. First, the manipulation variable has been computed from INVALSI only from academic year 2011-12. Second, from the academic year 2017-18 the assessment procedure is computer-based and is carried out on multiple days, making it impossible to retrieve the exact day of the test.[12] Additionally, we exclude the academic year 2014-15 for the language records only as the manipulation variable contains errors that we were not able to fix.

3.3 Weather data

We use information on geographic location of schools to match our data with climate conditions on the days of the assessment at the municipality level using information taken from Agri-4-Cast dataset published by the Joint Research Centre of the European Commission. This data contains observations from weather stations interpolated

[12]Grades 2$^{\text{nd}}$ and 5$^{\text{th}}$ make an exception and continue with the paper-based examination. Although we could in principle extend the analysis to the most recent academic years for these grades only, we choose to include grade 10$^{\text{th}}$ and limit our analysis to time span 2011-12 to 2016-17 to consider a homogeneous set of low-stakes tests.

on a 25×25 km grid on minimum, maximum and average temperatures (in Celsius degrees), as well as on total precipitation (mm) and wind speed (m/s). Appendix Figure B1 displays the grid of meteorological data overlaid on the boundaries of the Italian municipalities. In our analysis, we focus on the maximum daily temperature rather than its average as the tests take place in a time slot in which external temperatures are close to its maximum (around noon). We also collect data on relative humidity from ERA-5, available on a regular grid of 0.1×0.1 degrees (about 11×11 km).

Temperature recorded at the municipality level reflects ambient outdoor conditions, which may differ from the temperature experienced by students during exams in the classroom. This discrepancy might introduce significant measurement error in temperature assessment. Moreover, the presence of air conditioning in classroom might strongly exacerbate this issue by potentially mitigating the impact of temperature on test scores. Although the first issue is not easily resolved as data on classroom temperature readings do not exist, if we assume a classical measurement error scenario, this would tend to attenuate the coefficient estimates, pulling them towards zero.[13] The second issue is not relevant in our context since official data collected by the Ministry of Education and Research (MIUR) indicate that less than 2% of school buildings are equipped with air conditioning during the academic year 2020-2021, and arguably, this figure is even lower in previous years.[14] Although we cannot measure temperature directly in schools during the test, this setup allows us to estimate the clean effect from the presence of devices that artificially alter indoor temperatures during the months in which the tests are administered.[15]

Figure 1 shows the maximum temperature in the Italian municipalities on the days of the assessment in the relevant years of our analysis. The picture highlights a marked geographical heterogeneity. In addition to the typically warmer areas concentrated in the Southern regions and the islands, we observe temperatures exceeding 30°C in

[13]Park (2022) shares a similar problem and has performed two spatial and temporal imputation procedures to reduce measurement errors, showing that the direction and overall magnitude of the results are not sensitive to either of these corrections. Differently from Park (2022), we do not encounter a temporal issue where some tests are administered in the morning while others in the afternoon, as the SNV tests are always conducted in the morning. Unfortunately we cannot perform the spatial correction because we do not know the exact location of the school within the municipality as in Park (2022) and in our study all the schools belonging to the same municipality are assigned the same value of the weather variables.

[14]See `https://dati.istruzione.it/opendata/opendata/catalogo/elements1/leaf/?area=Edilizia%20Scolastica&datasetId=DS0176EDITIPORISCSTA2021`

[15]The tests are carried out in the months of May, when winter heating is turned off in the vast majority of municipalities. However, the presence of heating is not a problem in our setting because it only mitigates the effect of very low temperatures.

much of the Central-Northern area known as the Po Valley, where peaks can even reach values above 36°C. These temperatures appear significantly anomalous compared to the typically moderate ones observed in the middle of the spring season, particularly in the northern areas of the country. Figure 2 provides a more accurate representation of the large test-to-test variation in temperature across municipalities in our sample. The absolute variation in the maximum temperature ranges from approximately -30 to 30 degrees in both math and language test samples.

3.4 Summary statistics

The final working datasets consist of about 8 million exam records for math test, and of about 7,6 million exam records for language test. Both datasets include approximately 24,000 schools and 6,700 municipalities (on a total of approximately 7,900 municipalities). Table 2 presents summary statistics for the key outcome variables of the analysis. Although our statistical analyses use standardized scores, the score means reported in Table 2 give the class average percent correct. Scores are lower in math than in language. The table also shows averages for an indicator of score manipulation. Similarly to test scores, manipulation rate is higher in math. Regarding the weather variables, in math tests, the average maximum temperature is about 22°C, with peaks reaching 35.2°C in some locations, displaying very similar values for language tests. These figures indicate a significant variation in temperature on the assessment dates for both math and language tests. Other weather variables also appear very similar between the two subjects, with minor differences attributable to the assessment procedure in primary school being conducted on two separate days.

In both math and language tests, controls for students' characteristics point to an almost perfect gender balance. Approximately 10% of students are foreign, 1.5% are early enrolled, nearly 7% are retained, and the average class size is approximately 19 students.

4 Effect of temperature on manipulation and test score

Figure 3 provides a graphical representation of the relationship between performance and temperature, as well as between manipulation and temperature, motivating the analysis that follows. In math test (Panels a and b) we clearly observe that tests taken on warmer days are associated with noticeably lower scores and a higher manipulation, while in language test this relationship is much less pronounced (Panels c and d).

To identify the true causal impact of temperature on test scores and manipulation, we exploit the quasi-random variation in temperature across test dates within school-by-grade cells. While we can exclude the possibility of students selecting themselves into different temperature treatments, as the days of the tests are scheduled months in advance and temperature is exogenous to student behavior, time-varying unobservables might still be correlated with weather variables. For instance, if the math test is scheduled always later in the morning after the language test for 10$^{\text{th}}$ grade, or the language test is earlier in the week while the math test is toward the end of the week for 2$^{\text{nd}}$ and 5$^{\text{th}}$ grades (e.g., Thursday as opposed to Monday), there might be a mechanical correlation between temperature and test scores or between temperature and manipulation, unrelated to the actual causal effect of temperature on student cognition. To account for this and other confounding factors we include multiple fixed effects in our baseline specification as follows:

$$y_{icgsht}^{f} = \alpha_0 + \sum_{k=1}^{8} \alpha_1^k T_{ht}^k + \alpha_2 W_{ht} + \alpha_3 Z_{icgsht} + \tau_t + \sigma_{gs} + \theta_w + \pi_{rt} + \varepsilon_{icgsht} \qquad (5)$$

where, y denotes either the test score or manipulation in subject $f \in$ {language,math} of student i attending class c in grade g in school s in municipality h in the school year t. T_{ht}^k are a series of indicators for whether the maximum outdoor temperature in the municipality h at time t falls into temperature bin k from 1 to 8 aimed to capture the non-linearity of heat exposure. We deploy eight bins, i.e. lower than 7°C, higher than 31°C, and six 4°C-wide bins in between and we assume the bin 19-22°as the reference category.[16] We also control for weather conditions at the municipal level (W_{ht}) such as

[16] As seen in other studies, the optimal range for obtaining better cognitive performance is around 22°C (Cedeño Laurent

rain, wind and relative humidity (reported in ten bins), for a vector of individual variables (Z_{icgsht}) that includes dummies for sex, immigrant status, anticipated enrollment, repeating student, and class size.[17]

Our specification also includes school year (τ_t), grade-by-school (σ_{gs}) as well as day-of-the-week (θ_w) fixed effects. Controlling for annual fixed-effects help mitigate spurious correlations between secular performance improvements and the increased probability of hotter days attributed to climate change. Day-of-the-week fixed effects account for systematic differences across days of the week, and grade-by-school fixed effects allow for exploiting the variation of interest, that is test-to-test changes in temperatures within schools. In addition, since the school system is managed to a small extent at the regional level, we control for a region specific non-linear time trend (π_{rt}), to capture time-varying factors common at the region level that may be correlated with temperature and may influence performance at the same time, such as specific school calendars. Standard errors are clustered at the municipality level to solve three potential issues: arbitrary spatial correlation across municipalities, autocorrelation in test scores over time and assignment of the same temperature to several children. Since the days of the tests are assigned several months in advance, the temperature fluctuations can be considered as good as random. It is therefore reasonable to assume that this variation is orthogonal to the determinants of cognitive test scores. Therefore, conditioning on the set of fixed-effects listed above the key parameters α_1^k identify the causal effects of interest.

Table 3 shows the results based on our baseline specification for the two outcomes of interest in maths and language. The results on test score, reported respectively in columns 1 and 3, indicate that very high temperatures lead to a statistically significant decrease in performance. In particular, if we consider changes from comfort temperatures of 19-22°C to extreme temperatures of > 31°C observed in May, the child's math score decreases by 0.047 of a standard deviation, while there is no significant effect on language score. These effects are in line with those estimated by Graff Zivin et al. (2018), Krebs (2024), Park et al. (2020), and Park (2022).[18] In addition, the much

et al., 2018; Hancock and Vasmatzidis, 2003). At this temperature, the ability to carry out tasks is slightly better than in situations with a greater intensity of heat. Therefore, to obtain better performance at school or at work it is useful to maintain a room temperature around 20°C.

[17]The early enrollment in primary school is allowed for children who turn six years old by April 30th of the relevant school year. We cannot use variables such as parental education and occupation as controls since they are not present for all grades and school years.

[18]Considering the two studies most similar to ours, Graff Zivin et al. (2018) estimate a 0.12 s.d. reduction in math test score as temperatures increases from 20-22°C to 30-32°C, while Park (2022)'s estimated impacts range from -0.085

less extensive and significant effects in the language test appear consistent with previous scientific evidence explaining how heat stress impacts differentiated areas of the brain, particularly the prefrontal cortex, the main seat of logical-mathematical reasoning (Hocking et al., 2001; Graff Zivin et al., 2018).[19] Figure 4 plots the corresponding estimates of column 1 for math in Panel A and of column 3 for language in Panel B. It clearly shows that the decline in performance is flat and not significantly different from zero for language, while for math temperatures exert a negative and significant effect on score, which is much more pronounced at very high temperatures.

However, this relationship may be affected by the attenuating effect of score manipulation. Indeed, previous work has documented grade manipulation by teachers (Angrist et al., 2017; Diamond and Persson, 2016; Dee et al., 2019; Park, 2022) and students (McCabe, 2005; Lucifora and Tonello, 2015; Carrell et al., 2008) as compensatory behavior. For example, Dee et al. (2019) explicitly suggests that grade manipulation in NYC public schools was primarily driven by teachers who wanted to prevent students from long-term negative consequences of having experienced a bad-day test. In our case, a bad-day test could definitely be a hot day test which could lead teachers and students to engage in compensatory behavior, altering test results.

To empirically assess the link between manipulation and temperature, we estimate equation 1 using manipulation as an outcome. Point estimates are reported in columns 2 and 4 of Table 3 and in Figure 5. They show that the extent of manipulation is related to temperature in the day of the test but the impact is only relevant for math, while there is no obvious association between temperature exposure and manipulation for language. In math (Panel A of Figure 5), at low temperatures (below 22°C), we notice a flat and non-significant trend. As the temperature rises, we observe an increase in manipulation, reaching a positive peak of 0.012 p.p. at around 30°C. The implied magnitude is nontrivial since it represents an increase of approximately 25% w.r.t the sample mean. Above 31°, the coefficient loses significance and becomes virtually zero. As previously hypothesized, a plausible explanation for this behavior is that manipulation, like the concentration required to fairly tackle the test, is an activity that demands cognitive effort, and is thus influenced by temperature.

to 0.12 z-scores for temperatures higher than 90°F (approximately 32.2°C).

[19] As in Graff Zivin et al. (2018), it is unlikely that this difference is explained by increased fatigue because language and math tests are taken in different days, at least for grades 2^{nd} and 5^{th}.

We can interpret these results in light of the simple conceptual framework outlined in sections 2 and section 8 (Appendix A). When temperature rises cognitive performance starts to deteriorate and the need for compensation increases ($\frac{dh}{dg} \leq 0$). At high but not extreme temperature range manipulation is still effective, increasing the final score and creating a wedge between the *observed* and the *true* effect of temperature on performance. As temperature further increases reaching extreme values, compensation itself could become either difficult to implement ($\frac{dh}{dg} = 0$, e.g. students or teachers do not try to compensate) or noneffective ($\frac{df}{dh} = 0$, e.g. they try to compensate without improving the final score).

5 The effect of temperature on test score net of manipulation

In Section 4, we proved that temperature has an effect on both test scores and manipulation. This means that to estimate the *true* effect of temperature on student's performance, it is necessary to take into account the variation in scores due to manipulation. To the best of our knowledge, this represents an empirical challenge that has not been fully addressed by previous studies, likely due to a lack of data enabling direct measurement of manipulation and the need for causal settings capable of breaking the endogenous link between manipulation, performance, and temperature.

Since temperature affects both manipulation and test score, a naïve regression of test score on temperature controlling for manipulation would be biased as manipulation would enter the model as a "bad control". To properly address this issue, we employ a natural experiment provided by the the random assignment of external monitors sent to schools to supervise test administration. Our strategy is similar to the one employed by Angrist et al. (2017) who use external monitor as an instrument for manipulation when studying class size effects on learning. As discussed in Section 2, this strategy allows us to estimate the *true* effect of temperature on student's performance.[20] Our 2SLS model with multiple fixed effects is similar to the one used in Section 4:

[20] If the presence of the monitor in the classroom completely eliminates any manipulation, we could estimate the true impact of temperature on student performance by focusing exclusively on monitored classes. Unfortunately, in our case, the presence of the monitor diminishes manipulation but does not entirely eliminate it. This circumstance justifies the use of an instrumental variable approach.

$$m^f_{icgsht} = \lambda_0 + \sum_{k=1}^{8} \lambda_1^k T^k_{ht} + \lambda_2 Monitor_{cgst} + \lambda_3 W_{ht} + \lambda_4 Z_{icgsht} + \tau_t + \sigma_{gs} + \theta_w + \pi_{rt} + \epsilon_{icgsht} \tag{6}$$

$$y^f_{icgsht} = \beta_0 + \sum_{k=1}^{8} \beta_1^k T^k_{ht} + \beta_2 \widehat{m^f}_{icgsht} + \beta_3 W_{ht} + \beta_4 Z_{icgsht} + \tau_t + \sigma_{gs} + \theta_w + \pi_{rt} + \xi_{icgsht} \tag{7}$$

where m is a variable ranging from 0 to 1 and denoting the amount of manipulation computed at the class level, and $Monitor$ is a dummy variable indicating classes at institutions with randomly assigned monitors.

Table 4 presents the effects of monitoring on manipulation on math (column 1) and language (column 2) tests. This first-stage effect is large in magnitude and strongly significant: class monitoring reduces the fraction of manipulation by 0.041 percentage points in math test and 0.036 percentage points in language test. The F-statistic is 1002.87 for math test and 1053.40 for language test, well above the threshold adopted in the most recent research on valid IV inference (Lee et al., 2022), indicating that our instrument is valid and our first-stage estimates do not suffer from weak identification issues.

Figure 6 provides a graphical representation of the effect of temperature on test score net of manipulation (in red), while including also the coefficients reported in Table 5 (in green). In math test, we observe a virtually zero and non-significant effect on test score at comfort and lower temperatures, while test score starts dropping substantially when the maximum temperature becomes warmer (23-26°C) up to -0.76 s.d. for temperature between 27°C and 30°C. We then observe a slightly less pronounced, yet still negative effect at higher temperatures (>31°C), where the effect on test scores is about 0.055 s.d. We also observe a negative impact of temperatures above 27°C on test score for language, even though our point estimates are almost never significant, except between 27°C and 30°C where the effect is only weakly significant.

To better evaluate the effect of temperature net of manipulation in math test scores, we compare the results in Figure 6 with the effect on test scores without controlling for manipulation, as shown in Figure 4, and the effect of temperature on manipulation, displayed in Figure 5. When the temperature rises (≥23°C), the student's *true* performance (in red) declines more rapidly than the student's observed performance (in green), until it reaches a negative peak at 27-30 °C. This pattern mirrors, but in the opposite direction, that of manipulation displayed in Figure 5, where manipulation increases with temperature up to a peak and then decreases as the temperature rises above 31°C. This is consistent with score manipulation acting as a compensation for temperatures between 23-30 °C. Within this interval, the observed temperature effect on performance is lower than its true effect (e.g. $\left|\frac{dP^O}{dT}\right| < \left|\frac{dg}{dT}\right|$ in terms of our simple framework in Section 2). However, as a cognitively demanding task, manipulation collapses at extreme temperatures, leading the observed and true effects on cognitive performance to become aligned again ($\frac{dP^O}{dT} = \frac{dg}{dT}$), where the red and the green lines overlap.

6 Robustness checks

Avoidance behavior – A debated issue when estimating the effect of temperature using test-to-test variation among different academic years relates to the possibility that students or schools learn from past tests' exposure to warm temperatures and engage in potential compensatory behaviors in subsequent assessments. This is what the literature refers to as *avoidance behavior*. In our setting, it could be that students put more effort into studying for the test when they assume the day of assessment is going to be hot. Similarly, teachers could act to compensate for the disruption of performance when they know, from their past experience, that extremely high temperatures affect students' performance. In our identification strategy we already control for in-test compensating behaviors like score manipulation but it is possible that such actions also take place between tests. If the time span between one test and another is large (e.g. a year), it is possible that teachers or students have time to adopt strategies to better deal with the test even in stressful situations (e.g. teaching to the test or other similar strategies). To address this concern, we exploit variation between subjects, as the assessment of language and math tests in 2$^{\text{nd}}$ and 5$^{\text{th}}$ grades takes place on two distinct but very close days. We run a regression where we control for student-by-grade-by-school year fixed effects as in Park et al. (2020), leveraging exogenous variation in temperatures observed in two close days between subjects to identify the effect of interest. As the time span between the two tests is very short (approximately two days), it is very unlikely that avoidance behaviors take place, since students or teachers have little time to put an avoidance strategy in place. Figure 7 displays the non-linear estimates using this identification strategy for grades 2^{nd} and 5^{th} (we report full estimates in the Appendix Table B1).

The results follow the same pattern shown in Figure 6, with a significant and permanent drop in performance when the temperature exceeds 27°C and no effect for lower temperatures. Although estimates are smaller in magnitude compared to the ones presented in Figure 6 for math tests (-0.055 vis-à-vis -0.042 at $T \geq 31$°C), this is not surprising since in this model we also include language tests, for which the effects are much smaller and almost never significant at conventional levels. Overall, we take this as suggestive evidence that avoidance behavior does not represent an issue in our framework.

Falsification test – Since test dates are set several months in advance at the national level without any possibility for endogenous scheduling, temperatures the day of the assessment can be considered as good as random. To further highlight this point, in this section we perform a falsification test for which we expect to find non-significant results. We reshuffle temperatures on different days within the same municipality, school year and grade and report mean coefficients and standard errors of estimates based on 50 iterations. Table B2 shows very small and non-significant coefficients for both math and language tests. Overall, this evidence provides further validation for the causal interpretation of our results.

7 A potential mechanism: emotional disruption

We still know very little about the mechanisms driving the effects of temperature on cognitive outcomes such as student performance. Experimental evidence utilizing mice as exposed subjects has recently provided some useful clues. One potential mechanism is that brief exposure to extreme heat may impact neurotransmitter levels in the brain, including those responsible for regulating emotional states such as anxiety (Nakagawa and Ishiwata, 2021). Additionally, heat stress can have detrimental effects on cognitive functions, such as memory, caused by inflammation in the hippocampus (Chauhan et al., 2021; Lee et al., 2015). This research suggests that our understanding of these effects can be viewed through the lens of both physiological and emotional responses. In this respect, our study is the first that connects the results of these experimental findings obtained on laboratory ceilings to human behavior.

We do this by exploring the emotional perception data contained in the individual Student Questionnaire administered to 5^{th} and 10^{th} grade students after completing both tests.[21] This outcome data allows to observe the student's status perception after completing the test, and precisely: i) being worried before the tests; ii) feeling anxiety during the tests; iii) feeling confident during the tests; iv) feeling the tests are not going well. Questions iii) and iv) mirror each other and can be considered as a double check on the accuracy of the students' answers. These are categorical variables taking four values ranging from "strongly agree" to "strongly disagree" to the questions mentioned above. We transform these variables into dummy indicators, e.g. the variable "anxiety" is equal to one if the student answers "strongly agree" or "agree" to the question "feeling anxiety during the test". Table 6 displays summary statistics for these emotional variables.[22] We use equation 1 as a linear probability model to estimate the effect of temperature on the emotional outcomes controlling for the same set of variables W_{ht} and Z_{icgsht} plus school, year, weekday and region-by-year fixed effects.

To capture any age-related differences, this analysis is conducted separately for 5^{th} grade and 10^{th} grade students, both for mathematics and language.[23] As in previous sections, the results are presented in graphical format to better appreciate the nonlinear effect, while the complete estimation tables are provided in the Appendix B (Table B3 and Table B4). For the math test, Figure 8 and Figure 9 show, respectively for grade 5^{th} and 10^{th}, estimates for each of the four emotional outcome variables. Results show a deterioration in the student's emotional state as temperature increases above 19-22°C only for grade 5^{th}, while the patter is rather flat for grade 10^{th} and not significantly different from zero for most of the coefficients, with few exceptions like feeling bad above 31°C. For example, moving from comfortable temperatures (19-22°C) to temperatures

[21] 2^{nd} grade students are not interviewed because they are considered too young.

[22] For grade 10^th we do not observe these variables for school years 2015-16 and 2016-17.

[23] In principle, emotional variables are not subject-specific. However, for 5^{th} grade the test for math and language the tests are run in two different days with observed temperature. For this reason we explore the correlation between temperature and emotional variables separately for the two subjects. In 10^{th} grade, although the two tests are run in the same day, the two samples diverge because we exclude a.a. 2014-2015 for language, as mentioned in section 3

above 31°C for grade 5th, we notice a marked increase in the predicted probability of being worried before the test by 1.5 percentage points (p.p.) (Panel a), experiencing anxiety by 0.7 p.p. (Panel b) as well as feeling that the test is going badly by 2.2 p.p. (Panel d), and a simultaneous decrease in the probability of feeling confident by about 3.5 p.p. (Panel c). Similarly, we observe mostly the same pattern of effects in the case of the language test, shown in Figure 10 and Figure 11, respectively for grade 5th and 10th.

From these results, we can draw two main conclusions. The first is that we find evidence of a worsening of the sensations perceived by students when the external temperature becomes high, above 27°C. Such students' emotional distress is consistently signaled by a deterioration in all our indicators of emotional sensation. This evidence aligns with recent experimental findings that demonstrate how exposure to heat alters important neurotransmitter hormones such as noradrenaline, dopamine, and serotonin, which regulate our physiological functions and influence cognition and emotional states (Nakagawa and Ishiwata, 2021; Suri et al., 2015; Nakagawa et al., 2020). Therefore, this represents a plausible channel to explain the decline in performance during the test for students exposed to extreme heat. Secondly, the mechanism of emotional distress is much more pronounced in younger students. This could be due to both their greater physical vulnerability and a less developed capacity to adapt to severe environmental conditions.

8 Conclusions

In this paper, we explore the effect of temperature on student performance using Italian administrative data from mandatory language and mathematics assessment tests taken by students in low stake grades from school years 2011-12 to 2016-17 matched with meteorological data. We find that increases in temperature lead to statistically significant decreases in cognitive performance in math (but not in language) beyond 27-30 °C. Additionally, we find that temperature influences score manipulation. Controlling for this aspect when estimating the effect of temperature on school performance, we find significant negative effects that are larger and emerge at lower interval ranges. Therefore, failing to account for the role of manipulation could result in inaccurate estimates of the effect of temperature on cognitive performance in national assessments. The occurrence of negative effects at high, but not extreme, temperatures carries significant implications, as it means that a larger number of students are impacted. In fact, the number of days with non-extreme temperatures throughout the year, as well as the geographic regions affected, are considerably greater.

The (net of manipulation) causal link between heat exposure and cognitive performance holds significant policy relevance given the alarming trend of global warming and the widespread lack of access to air conditioning for much of the world's population (WMO, 2023; Allen et al., 2018). Consequently, our findings have significant and direct policy implications. First, our findings could help policy makers design effective strategies to circumvent the negative effects of extreme heat to make school assessments more even, mitigating the impacts of external factors that differently affect individuals who live in different places or who take the tests in the most at risk periods. For instance, many countries, including Italy, exhibit significant regional variations in temperature, with southern areas experiencing notably higher temperatures compared to the rest of the country. This climatic disparity suggests that students residing in hotter regions may face disadvantages relative to their peers in cooler areas, raising important concerns regarding equitable peers' comparison when looking at school national assessment. In this regard, our analysis stimulates the debate about the quality standard of school facilities considering that school buildings in many advanced economies are seldom equipped with air conditioning.

Second, apart from school context, cognitive performance plays a critical role in various aspects of our life. Common examples are competitive examinations (e.g. public competition), college admissions or any financial decision-making. Our evidence of reduced cognitive functioning at high temperatures show that there is room for the optimally scheduling of cognitively demanding tasks.

References

Allen, M. R., Babiker, M., Chen, Y., de Coninck, H., Connors, S., van Diemen, R., Dube, O. P., Ebi, K. L., Engelbrecht, F., Ferrat, M., et al. (2018). Summary for policymakers. In *Global Warming of 1.5: An IPCC Special Report on the impacts of global warming of 1.5\C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty.* IPCC.

Angrist, J. D., Battistin, E., and Vuri, D. (2017). In a small moment: Class size and moral hazard in the italian mezzogiorno. *American Economic Journal: Applied Economics*, 9(4):216–249.

Angrist, J. D., Cohodes, S. R., Dynarski, S. M., Pathak, P. A., and Walters, C. R. (2016). Stand and deliver: Effects of bostonâs charter high schools on college preparation, entry, and choice. *Journal of Labor Economics*, 34(2):275–318.

Banerjee, R. and Maharaj, R. (2020). Heat, infant mortality, and adaptation: Evidence from india. *Journal of Development Economics*, 143:102378.

Basu, R., Gavin, L., Pearson, D., Ebisu, K., and Malig, B. (2018). Examining the association between apparent temperature and mental health-related emergency room visits in california. *American journal of epidemiology*, 187(4):726–735.

Battistin, E. (2016). How manipulating test scores affects school accountability and student achievement. *IZA World of Labor*.

Baylis, P. (2020). Temperature and temperament: Evidence from twitter. *Journal of Public Economics*, 184:104161.

Bertoni, M., Brunello, G., and Rocco, L. (2013). When the cat is near, the mice won't play: The effect of external examiners in italian schools. *Journal of Public Economics*, 104:65–77.

Burke, M., González, F., Baylis, P., Heft-Neal, S., Baysan, C., Basu, S., and Hsiang, S. (2018). Higher temperatures increase suicide rates in the united states and mexico. *Nature climate change*, 8(8):723–729.

Carrell, S. E., Malmstrom, F. V., and West, J. E. (2008). Peer effects in academic cheating. *Journal of human resources*, 43(1):173–207.

Cedeño Laurent, J. G., Williams, A., Oulhote, Y., Zanobetti, A., Allen, J. G., and Spengler, J. D. (2018). Reduced cognitive function during a heat wave among residents of non-air-conditioned buildings: An observational study of young adults in the summer of 2016. *PLoS medicine*, 15(7):e1002605.

Cellini, S. R., Ferreira, F., and Rothstein, J. (2010). The value of school facility investments: Evidence from a dynamic regression discontinuity design. *The Quarterly Journal of Economics*, 125(1):215–261.

Chauhan, N. R., Kumar, R., Gupta, A., Meena, R. C., Nanda, S., Mishra, K. P., and Singh, S. B. (2021). Heat stress induced oxidative damage and perturbation in bdnf/erk1/2/creb axis in hippocampus impairs spatial memory. *Behavioural Brain Research*, 396:112895.

Cho, H. (2017). The effects of summer heat on academic achievement: a cohort analysis. *Journal of Environmental Economics and Management*, 83:185–196.

Coehoorn, C. J., Stuart-Hill, L. A., Abimbola, W., Neary, J. P., and Krigolson, O. E. (2020). Firefighter neural function and decision-making following rapid heat stress. *Fire safety journal*, 118:103240.

Dee, T. S., Dobbie, W., Jacob, B. A., and Rockoff, J. (2019). The causes and consequences of test score manipulation: Evidence from the new york regents examinations. Technical Report 3.

Deschênes, O. (2014). Temperature, human health, and adaptation: A review of the empirical literature. *Energy Economics*, 46:606–619.

Deschênes, O. and Greenstone, M. (2007). The economic impacts of climate change: evidence from agricultural output and random fluctuations in weather. *American economic review*, 97(1):354–385.

Deschênes, O. and Greenstone, M. (2011). Climate change, mortality, and adaptation: Evidence from annual fluctuations in weather in the us. *American Economic Journal: Applied Economics*, 3(4):152–85.

Diamond, R. and Persson, P. (2016). The long-term consequences of teacher discretion in grading of high-stakes tests. Technical report, National Bureau of Economic Research.

Evans, M. F., Gazze, L., and Schaller, J. (2023). Temperature and maltreatment of young children. National Bureau of Economic Research WP no.31522.

Figlio, D. N. (2006). Testing, crime and punishment. *Journal of Public Economics*, 90(4-5):837–851.

Filomena, M. and Picchio, M. (2024). Unsafe temperatures, unsafe jobs: The impact of weather conditions on work-related injuries. *Journal of Economic Behavior & Organization*, 224:851–875.

Froom, P., Caine, Y., Shochat, I., and Ribak, J. (1993). Heat stress and helicopter pilot errors. *Journal of Occupational Medicine*, pages 720–724.

Gaoua, N., Racinais, S., Grantham, J., and El Massioui, F. (2011). Alterations in cognitive performance during passive hyperthermia are task dependent. *International Journal of Hyperthermia*, 27(1):1–9.

Garg, T., Jagnani, M., and Taraz, V. (2020). Temperature and human capital in india. *Journal of the Association of Environmental and Resource Economists*, 7(6):1113–1150.

Graff Zivin, J., Hsiang, S. M., and Neidell, M. (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, J. and Neidell, M. (2013). Environment, health, and human capital. *Journal of Economic Literature*, 51(3):689–730.

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

Hallegatte, S. and Rozenberg, J. (2017). Climate change through a poverty lens. *Nature Climate Change*, 7(4):250–256.

Hancock, P. A. and Vasmatzidis, I. (2003). Effects of heat stress on cognitive performance: the current state of knowledge. *International Journal of Hyperthermia*, 19(3):355–372.

Hocking, C., Silberstein, R. B., Lau, W. M., Stough, C., and Roberts, W. (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):719–734.

Huang, C., Barnett, A. G., Wang, X., and Tong, S. (2012). The impact of temperature on years of life lost in brisbane, australia. *Nature Climate Change*, 2(4):265–270.

Jackson, C. K., Johnson, R. C., and Persico, C. (2015). The Effects of School Spending on Educational and Economic Outcomes: Evidence from School Finance Reforms *. *The Quarterly Journal of Economics*, 131(1):157–218.

Jackson, C. K. and Mackevicius, C. L. (2024). What impacts can we expect from school spending policy? evidence from evaluations in the united states. *American Economic Journal: Applied Economics*, 16(1):412–46.

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the chicago public schools. *Journal of public Economics*, 89(5-6):761–796.

Karlsson, M. and Ziebarth, N. R. (2018). Population health effects and health-related costs of extreme temperatures: Comprehensive evidence from germany. *Journal of Environmental Economics and Management*, 91:93–117.

Krebs, B. (2022). *Temperature and cognitive performance*. JSTOR.

Krebs, B. (2024). Temperature and cognitive performance: Evidence from mental arithmetic training. *Environmental and Resource Economics*, pages 1–31.

Lafortune, J., Rothstein, J., and Schanzenbach, D. W. (2018). School finance reform and the distribution of student achievement. *American Economic Journal: Applied Economics*, 10(2):1–26.

Lazear, E. P. (2006). Speeding, terrorism, and teaching to the test. *The Quarterly Journal of Economics*, 121(3):1029–1061.

Lee, D. S., McCrary, J., Moreira, M. J., and Porter, J. (2022). Valid t-ratio inference for IV. *American Economic Review*, 112(10):3260–90.

Lee, S. and Zheng, S. (2022). Extreme temperatures, adaptation capacity, and household retail consumption.

Lee, W., Moon, M., Kim, H. G., Lee, T. H., and Oh, M. S. (2015). Heat stress-induced memory impairment is associated with neuroinflammation in mice. *Journal of neuroinflammation*, 12:1–13.

Lee, W.-S. and Li, B. G. (2021). Extreme weather and mortality: Evidence from two millennia of chinese elites. *Journal of Health Economics*, 76:102401.

Lucifora, C. and Tonello, M. (2015). Cheating and social interactions. evidence from a randomized experiment in a national evaluation program. *Journal of Economic Behavior & Organization*, 115:45–66.

Lucifora, C. and Tonello, M. (2020). Monitoring and sanctioning cheating at school: What works? evidence from a national evaluation program. *Journal of Human Capital*, 14(4):584–616.

Mansfield, J. and Slichter, D. (2021). The long-run effects of consequential school accountability.

Martinelli, G. and Palma, A. (2024). Some (don't) Like it Hot. Persistent High Temperatures Increase Depression and Anxiety. GSSI Regional Science & Economic Geography Discussion Papers Series No.2024-02.

McCabe, D. L. (2005). Cheating among college and university students: A north american perspective. *International Journal for Educational Integrity*, 1(1).

Mullins, J. T. and White, C. (2019). Temperature and mental health: Evidence from the spectrum of mental health outcomes. *Journal of health economics*, 68:102240.

Nakagawa, H. and Ishiwata, T. (2021). Effect of short-and long-term heat exposure on brain monoamines and emotional behavior in mice and rats. *Journal of thermal biology*, 99:102923.

Nakagawa, H., Matsunaga, D., and Ishiwata, T. (2020). Effect of heat acclimation on anxiety-like behavior of rats in an open field. *Journal of Thermal Biology*, 87:102458.

Neal, D. and Schanzenbach, D. W. (2010). Left behind by design: Proficiency counts and test-based accountability. *The Review of Economics and Statistics*, 92(2):263–283.

Obradovich, N., Migliorini, R., Paulus, M. P., and Rahwan, I. (2018). Empirical evidence of mental health risks posed by climate change. *Proceedings of the National Academy of Sciences*, 115(43):10953–10958.

Park, J., Pankratz, N., and Behrer, A. (2021a). Temperature, workplace safety, and labor market inequality.

Park, R. J. (2022). Hot temperature and high-stakes performance. *Journal of Human Resources*, 57(2):400–434.

Park, R. J., Behrer, A. P., and Goodman, J. (2021b). Learning is inhibited by heat exposure, both internationally and within the united states. *Nature human behaviour*, 5(1):19–27.

Park, R. J., Goodman, J., Hurwitz, M., and Smith, J. (2020). Heat and learning. *American Economic Journal: Economic Policy*, 12(2):306–39.

Persico, C. L. and Venator, J. (2021). The effects of local industrial pollution on students and schools. *Journal of Human Resources*, 56(2):406–445.

Quintano, C., Castellano, R., and Longobardi, S. (2009). A fuzzy clustering approach to improve the accuracy of italian student data: An experimental procedure to correct the impact of outliers on assessment test scores. *Statistica and Applicazioni*, 2(4):149–71.

Rose, H. (2006). Do gains in test scores explain labor market outcomes? *Economics of Education Review*, 25(4):430–446. School-to-Work and Educational Reform Symposium.

Shah, M. and Steinberg, B. M. (2017). Drought of opportunities: Contemporaneous and long-term impacts of rainfall shocks on human capital. *Journal of Political Economy*, 125(2):527–561.

Suri, D., Teixeira, C. M., Cagliostro, M. K. C., Mahadevia, D., and Ansorge, M. S. (2015). Monoamine-sensitive developmental periods impacting adult emotional and cognitive behaviors. *Neuropsychopharmacology*, 40(1):88–112.

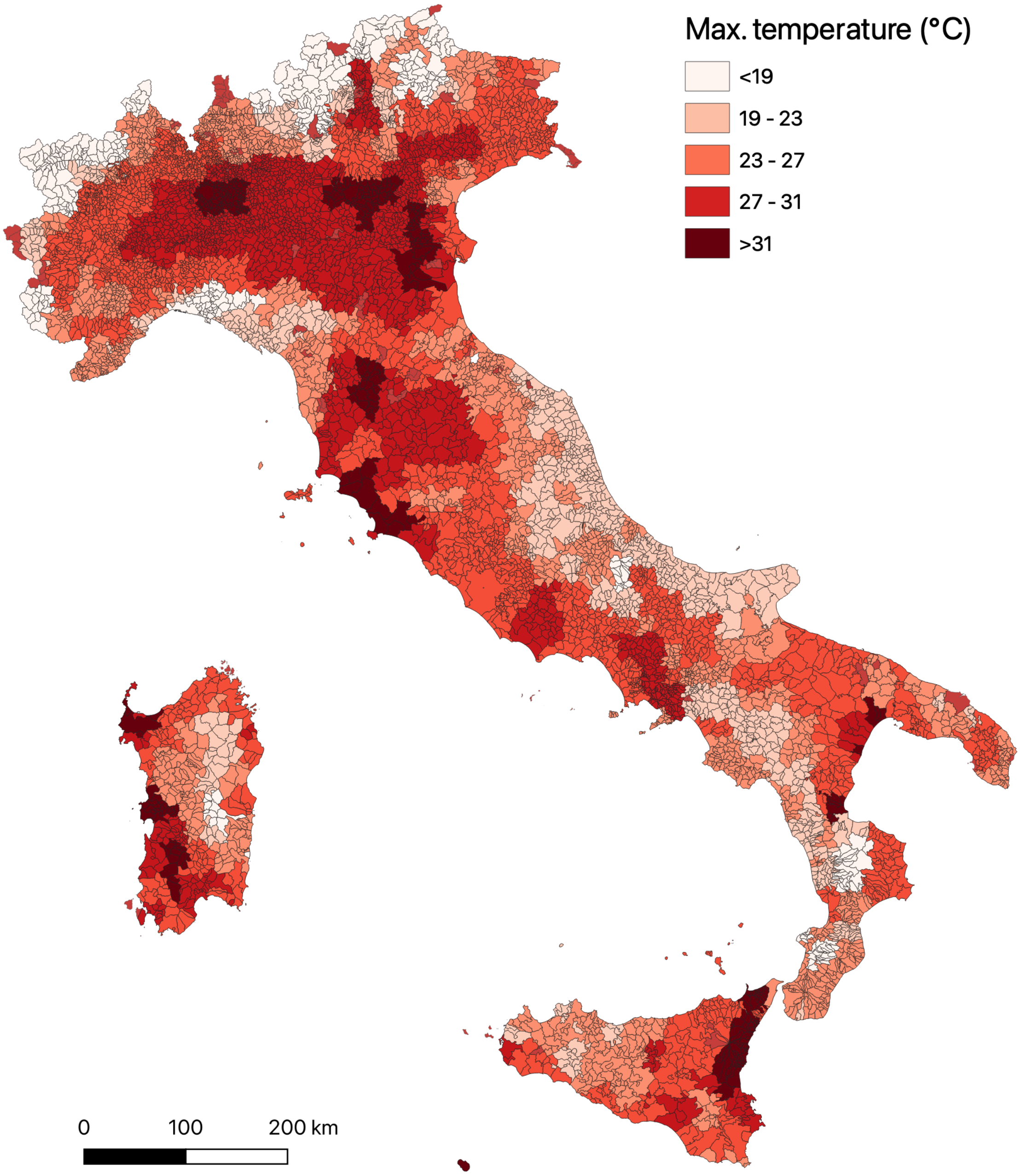
Taylor, L., Watkins, S. L., Marshall, H., Dascombe, B. J., and Foster, J. (2016). The impact of different environmental conditions on cognitive function: a focused review. *Frontiers in physiology*, 6:140292.

WMO (2023). *State of the global climate 2022*. World Meterological Organization, UN.

Zivin, J. G., Song, Y., Tang, Q., and Zhang, P. (2020). Temperature and high-stakes cognitive performance: Evidence from the national college entrance examination in china. *Journal of Environmental Economics and Management*, 104:102365.

Figures

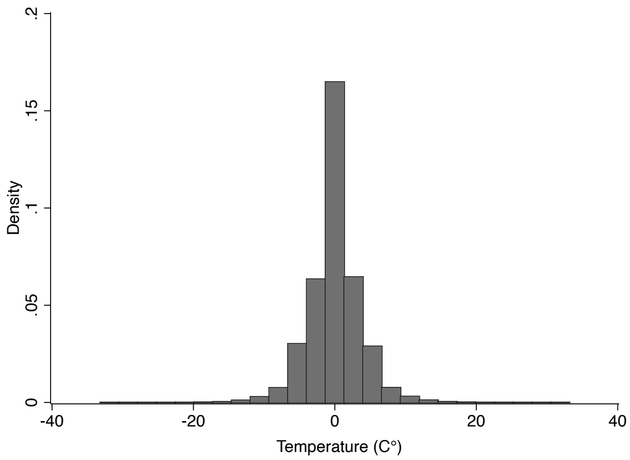
Figure 1: Maximum Temperature During the Tests



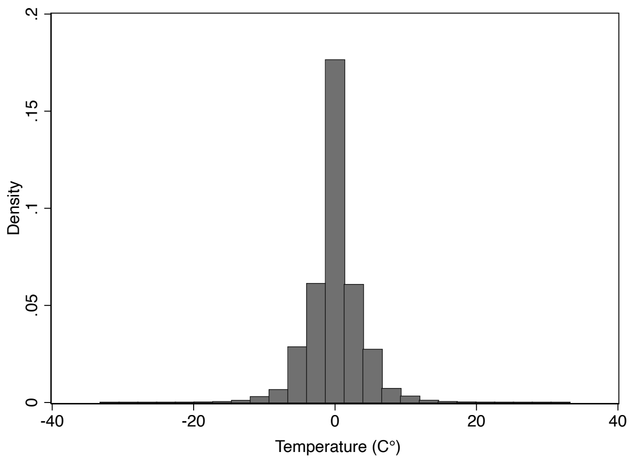
Notes: The figure displays the maximum temperature (measured in°C) in each municipality averaged over the days of the test. Pooled sample of grades 2nd, 5th and 10th in school years from 2011-12 to 2016-17.

Figure 2: Test-To-Test Temperature Variation

A. Math

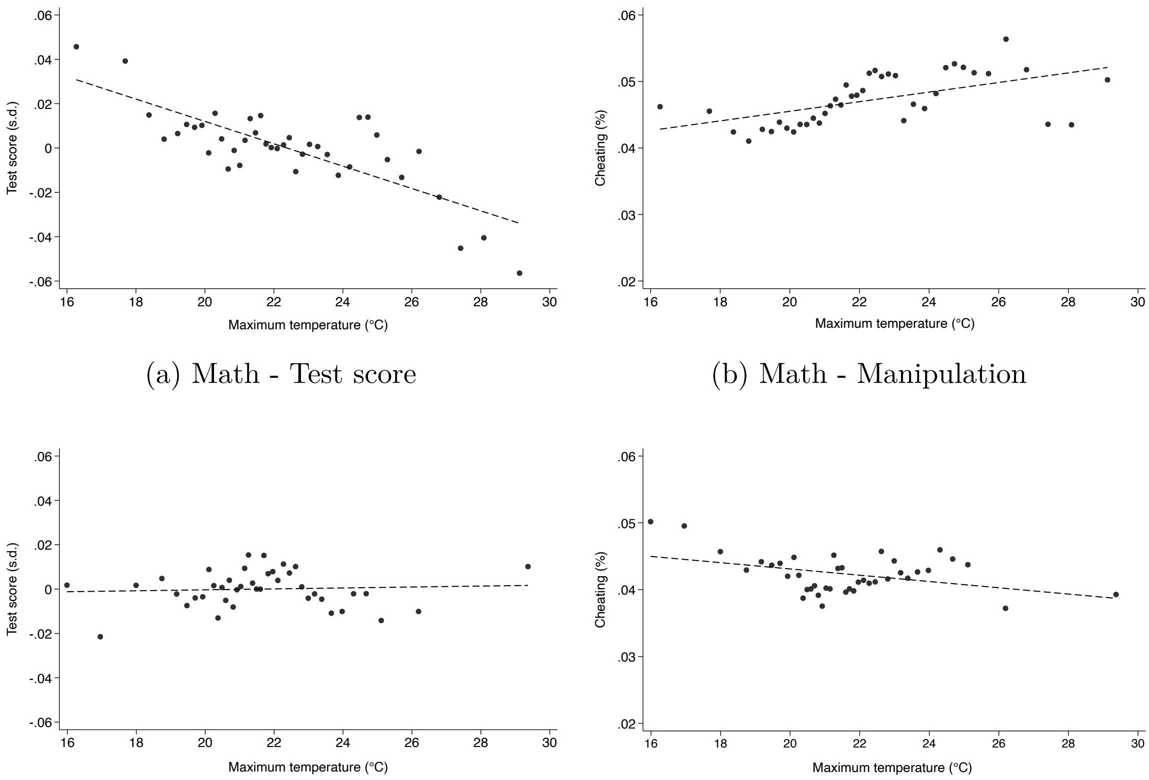


B. Language



Notes: Pooled sample of grades 2nd, 5th and 10th in school years from 2011-12 to 2016-17. Figure display the test-to-test variation between consecutive school years in the maximum temperature. Temperature is measured in Celsius degrees (°C).

Figure 3: Relationship Between Test Score, Manipulation and Temperature



(a) Math - Test score

(b) Math - Manipulation

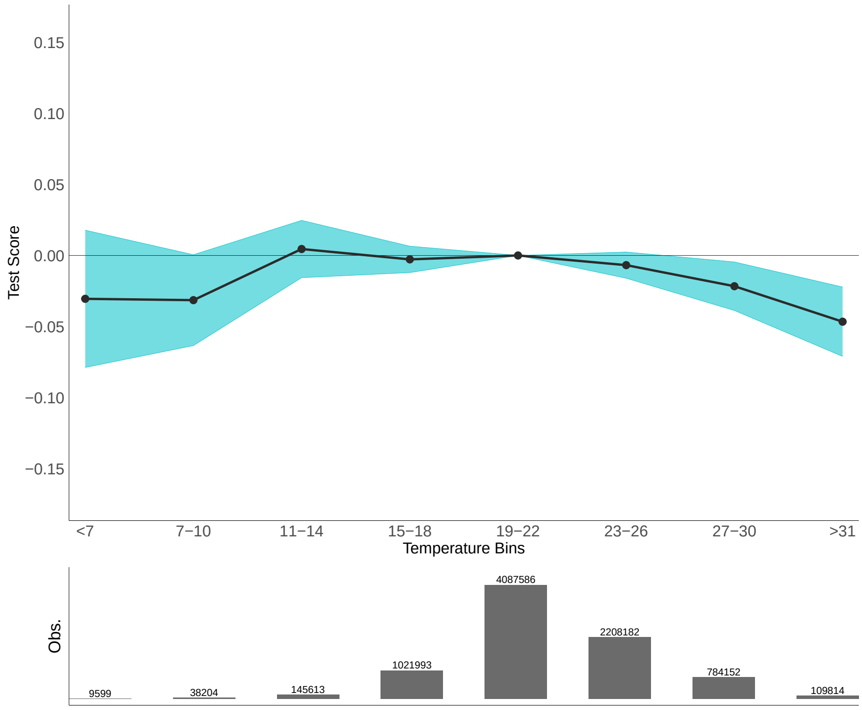
(c) Language - Test score

(d) Language - Manipulation

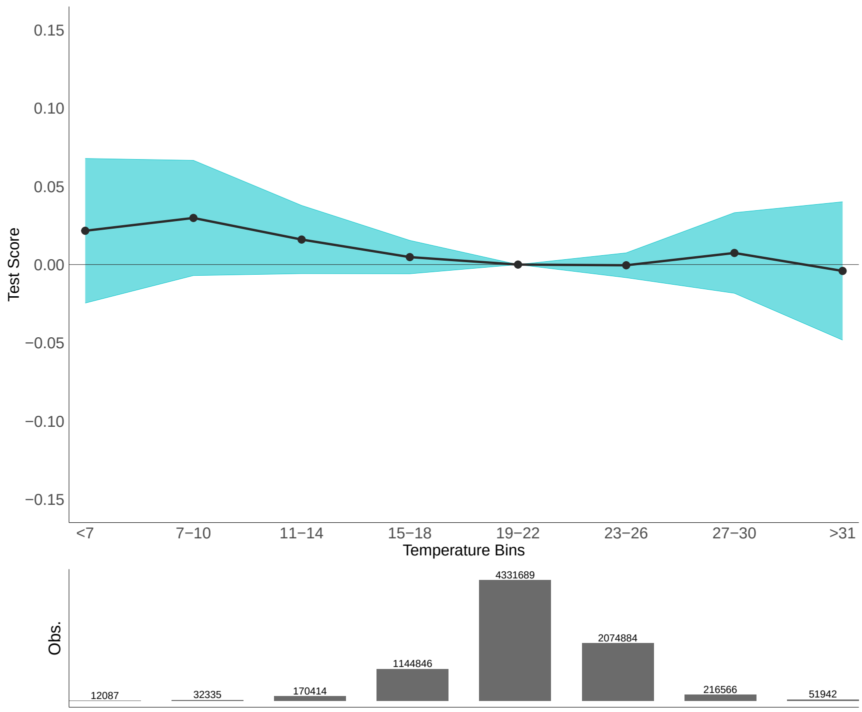
Notes: Figures display the relationship between test score and temperature during the test (panels a and c), and between manipulation and temperature (panels b and d), after controlling for school-grade and school year fixed effects.

Figure 4: Non-Linear Effect of Temperatures on Test Score

A. Math



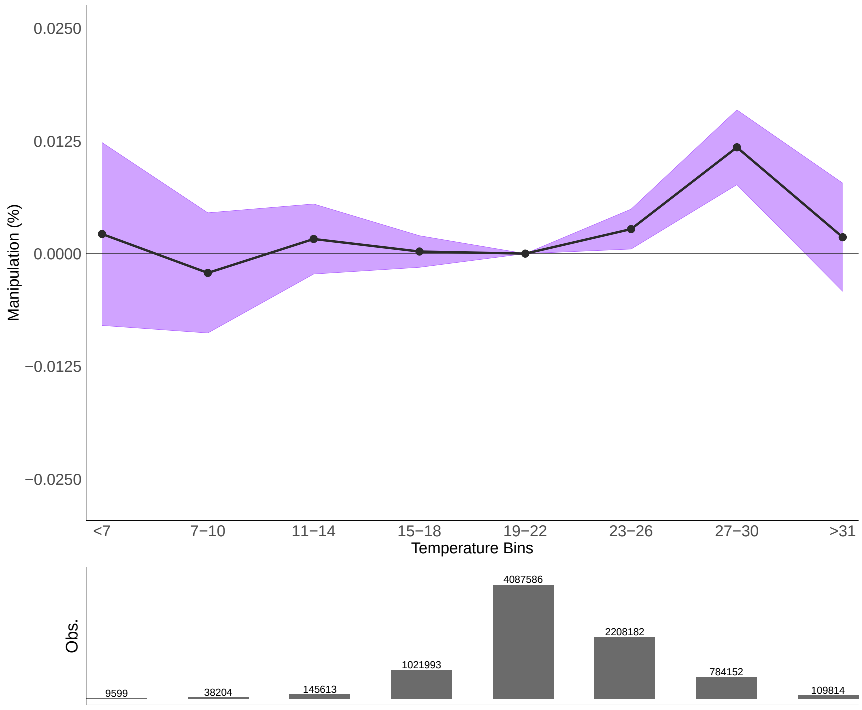
B. Language



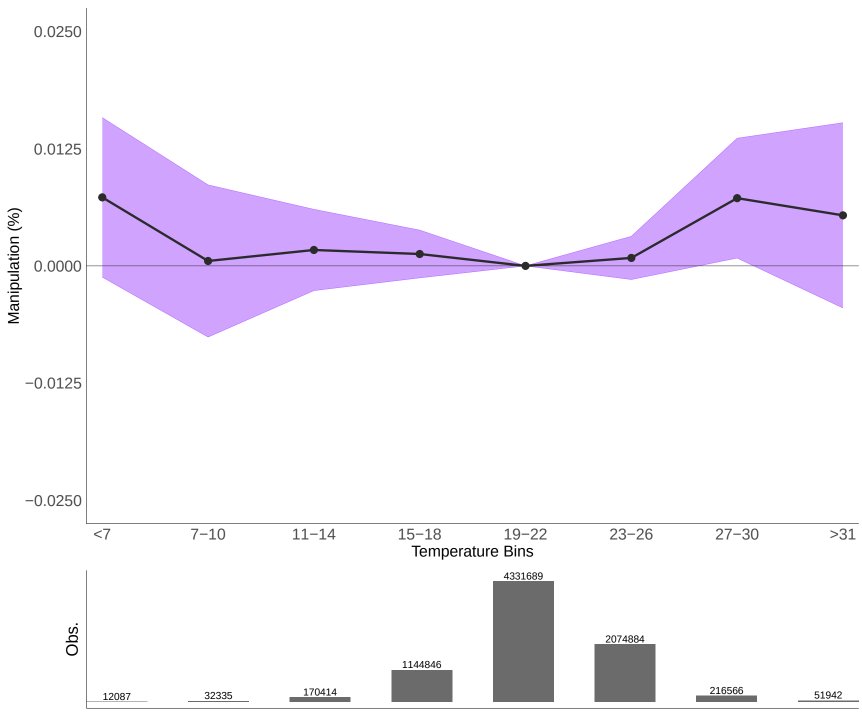
Notes: OLS estimates of standardized test score on bins of maximum temperatures observed the day of the test at the municipal level. Pooled sample of grades 2, 5 and 10 in school years from 2011-12 to 2016-17. Estimates include controls for weather (10 bins of rainfall, windspeed and humidity) at the municipal level, the day of the test. Student-level controls include female, foreign, early enrolled, retained. We also control for class size. Fixed effects include: school year, school-by-grade, day-of-week and region-by-school year. Max. temperature is in Celsius degree (°C), with reference category of 19-22°C. Confidence intervals are at 95%. The number of observations for each temperature bin (number of students) is reported in the bottom panel.

Figure 5: Non-Linear Effect of Temperatures on Manipulation

A. Math

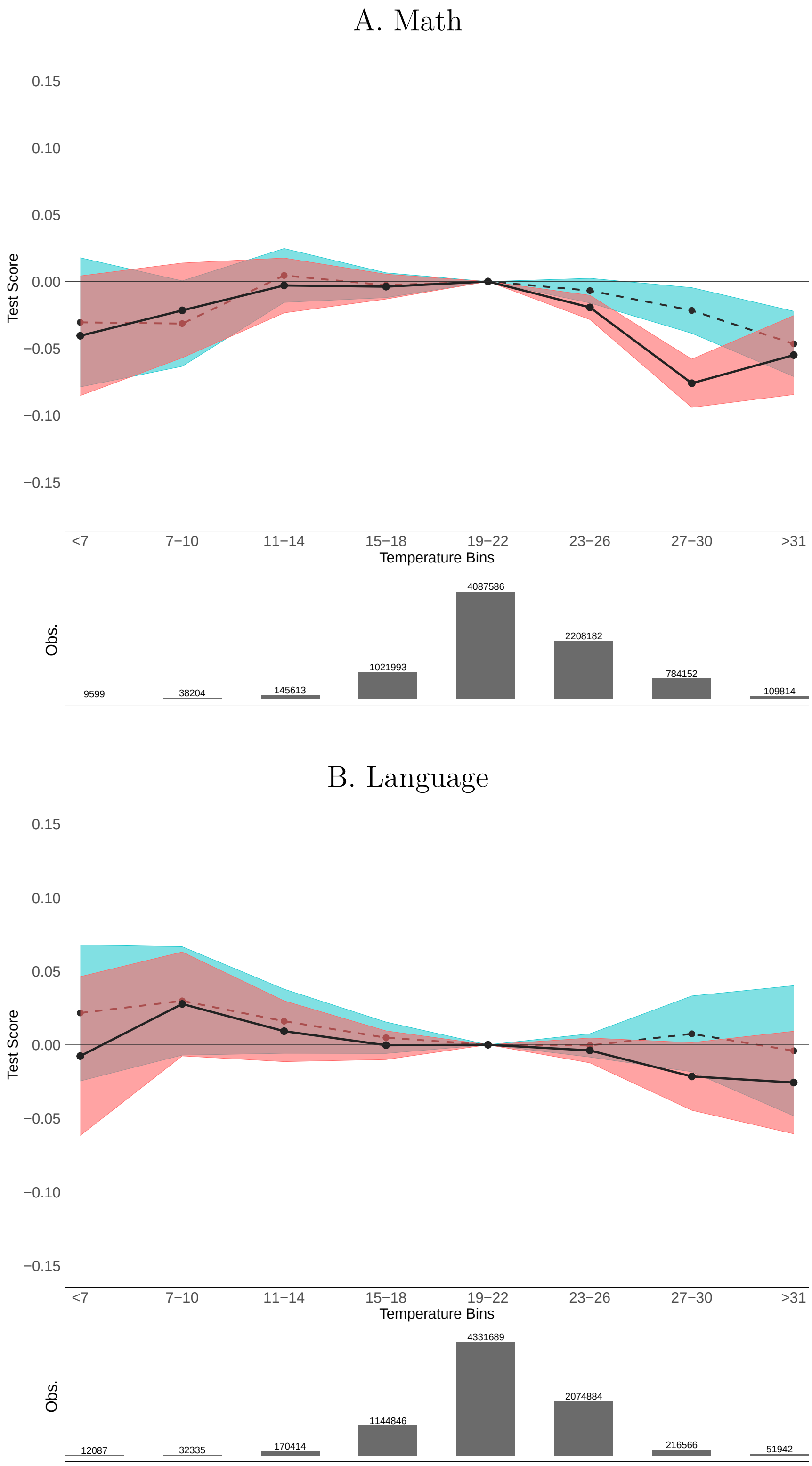


B. Language



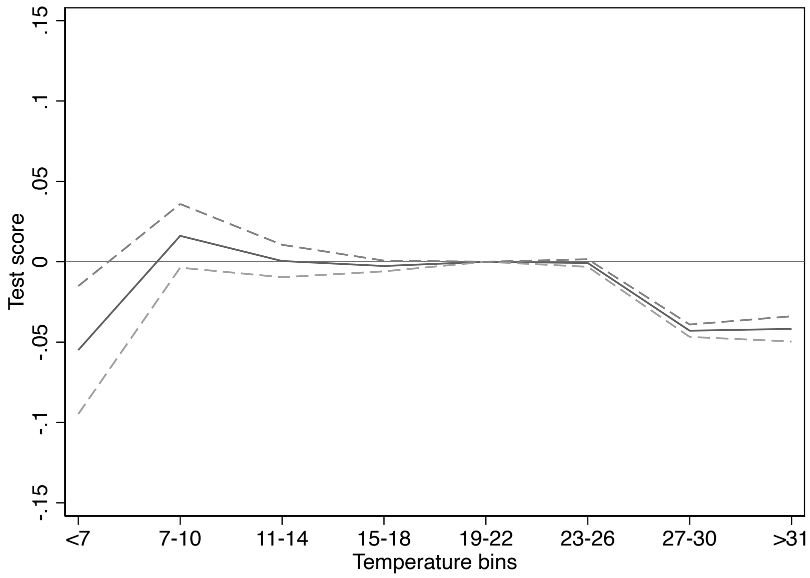
Notes: OLS estimates of score manipulation fraction on bins of maximum temperatures observed the day of the test at the municipal level. Pooled sample of grades 2, 5 and 10 in school years from 2011-12 to 2016-17. Estimates include controls for weather (10 bins of rainfall, windspeed and humidity) at the municipal level, the day of the test. Student-level controls include female, foreign, early enrolled, retained. We also control for class size. Fixed effects include: school year, school-by-grade, day-of-week and region-by-school year. Max. temperature is in Celsius degree (°C), with reference category of 19-22°C. Confidence intervals are at 95%. The number of observations for each temperature bin (number of students) is reported in the bottom panel.

Figure 6: Effect of Temperature on Test Score With and Without Manipulation



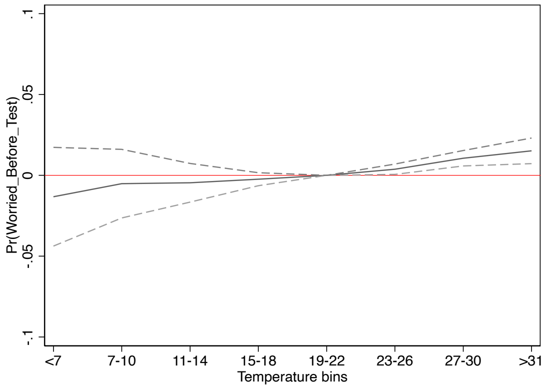
Notes: Comparison of 2SLS estimates (solid black line and c.i. in red) net of manipulation, and OLS estimates (dashed line and c.i. in green, see also Figure 4). Test scores are standardized. In the 2SLS regression, manipulation is instrumented for random class monitoring. Sanderson-Windmeijer F-statistics of excluded instrument: 1002.87, $p = 0.000$ for math test; 1053.40, $p = 0.000$ for language test). Pooled sample of grades 2, 5 and 10 in school years from 2011-12 to 2016-17. Estimates include controls for weather (10 bins of rainfall, windspeed and humidity) at the municipal level, the day of the test. Student-level controls include female, foreign, early enrolled, retained. We also control for class size. Fixed effects include: school year, school-by-grade, day-of-week and region-by-school year. Max. temperature is in Celsius degree (°C), with reference category of 19-22°C. Confidence intervals are at 95%. The number of observations for each temperature bin (number of students) is reported in the bottom panel.

Figure 7: Effect of Temperature on Test Score Using Between Subjects Specification

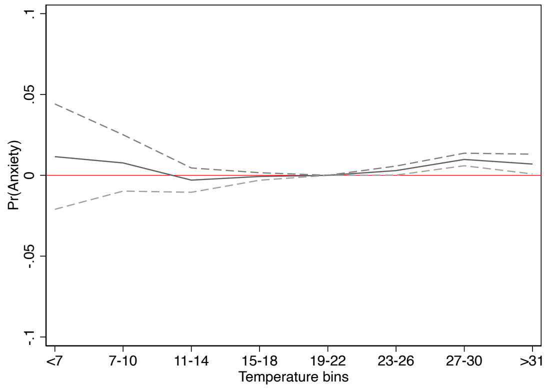


Notes: 2SLS estimates of test score on bins of maximum temperatures observed the day of the test at the municipal level. Pooled sample of subjects math and language, and grades 2, 5 in school years from 2011-12 to 2016-17. Estimates include controls for weather (10 bins of rainfall, windspeed and humidity) at the municipal level, the day of the test. Student-level controls include female, foreign, early enrolled, retained. We also control for class size. Fixed effects include: student-by-grade-year, subject, day-of-week and region-by-school year. Max. temperature is in Celsius degree (°C), with reference category of 19-22°C. Confidence intervals are at 95%.

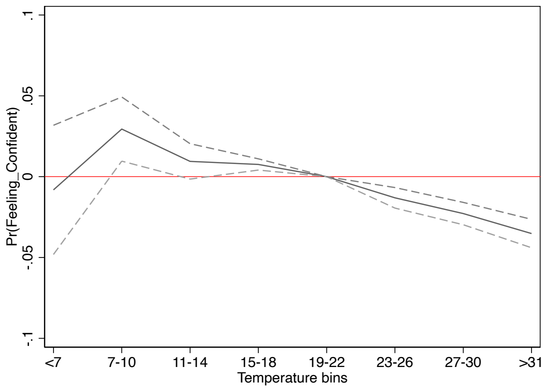
Figure 8: Effect of Temperatures During Math Test on Emotional Outcomes - Grade 5^{th}



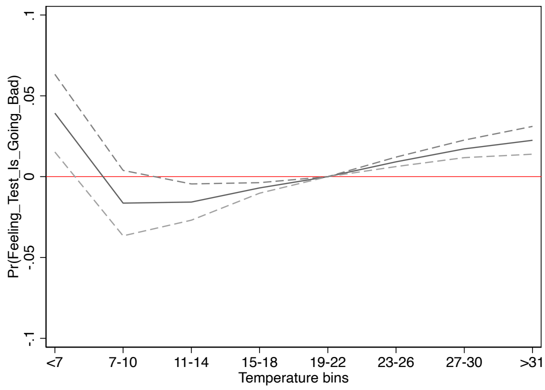
(a) Worried before test



(b) Anxiety



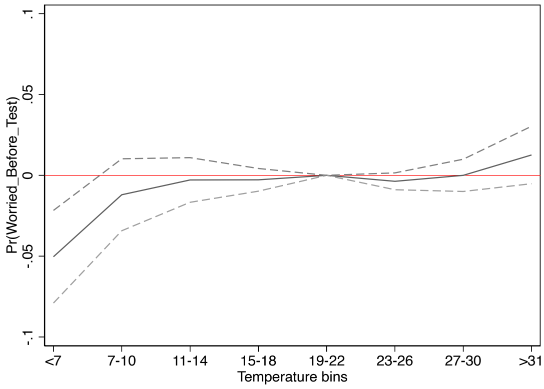
(c) Feeling confident



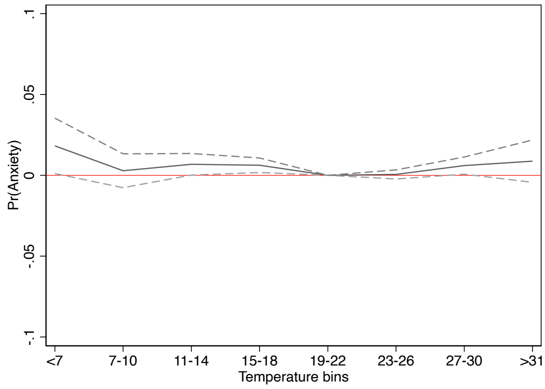
(d) Feeling test is going bad

Notes: Pooled sample of grade 5^{th} in school years 2011-12 to 2016-17. Figures display non linear estimates of the effect of temperatures in the day of math test on students emotional outcomes. Dependent variables are dummies for each emotional perceptions retrieved from students' questionnaire. Estimates include controls for weather (10 bins of rainfall, windspeed and humidity) at the municipal level, the day of the test. Student-level controls include female, foreign, early enrolled, retained. We also control for class size. Fixed effects include: school year, school, day-of-week and region-by-school year. Max. temperature is in Celsius degree (°C), with reference category of 19-22°C. Confidence intervals are at 95%.

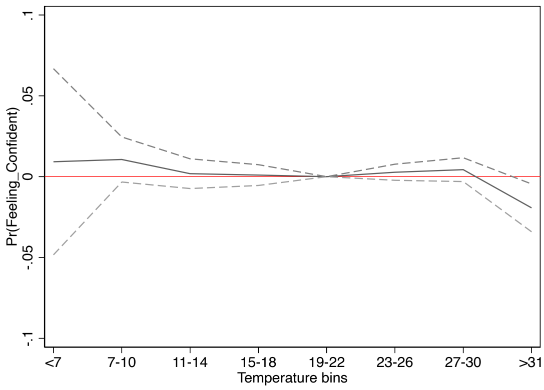
Figure 9: Effect of Temperatures During Math Test on Emotional Outcomes - Grade 10th



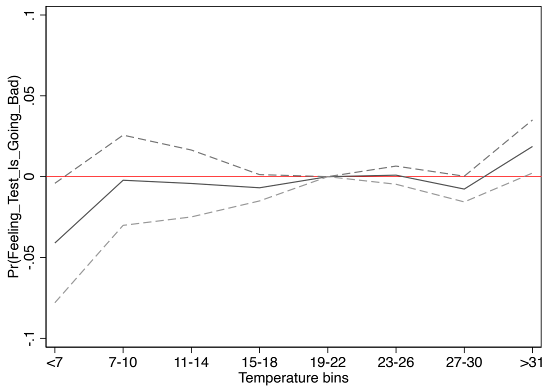
(a) Worried before test



(b) Anxiety



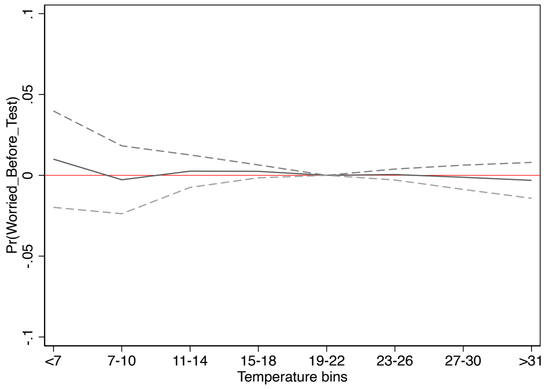
(c) Feeling confident



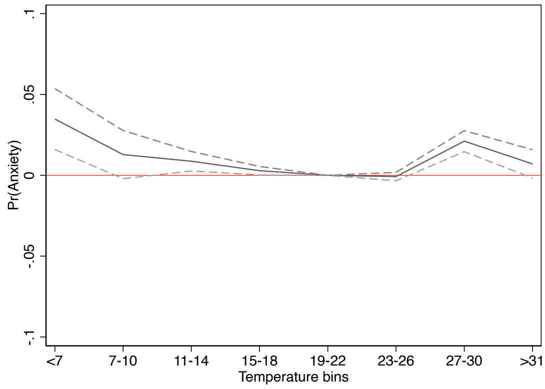
(d) Feeling test is going bad

Notes: Pooled sample of grade 10th in school years 2011-12 to 2014-15. Figures display non linear estimates of the effect of temperatures in the day of math test on students emotional outcomes. Dependent variables are dummies for each emotional perceptions retrieved from students' questionnaire. Estimates include controls for weather (10 bins of rainfall, windspeed and humidity) at the municipal level, the day of the test. Student-level controls include female, foreign, early enrolled, retained. We also control for class size. Fixed effects include: school years, school, day-of-week and region-by-school year. Max. temperature is in Celsius degree (°C), with reference category of 19-22°C. Confidence intervals are at 95%.

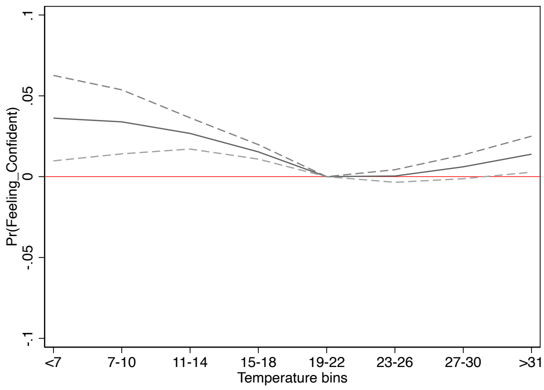
Figure 10: Effect of Temperatures During Language Test on Emotional Outcomes - Grade 5^{th}



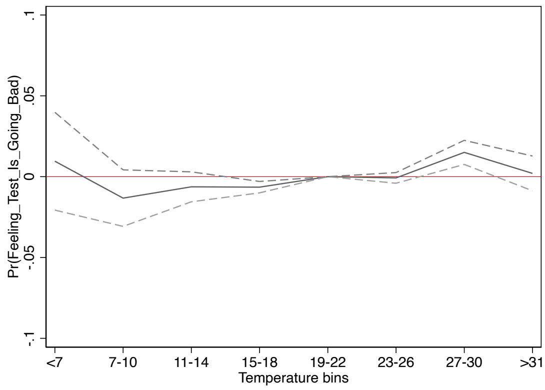
(a) Worried before test



(b) Anxiety



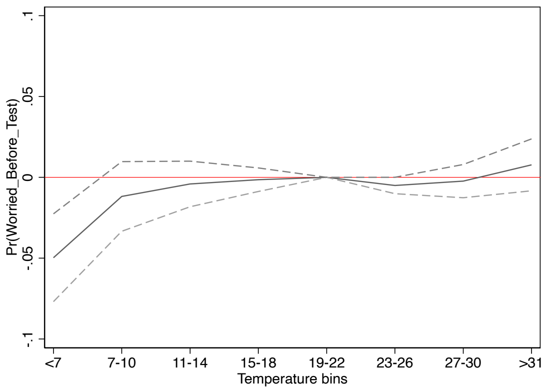
(c) Feeling confident



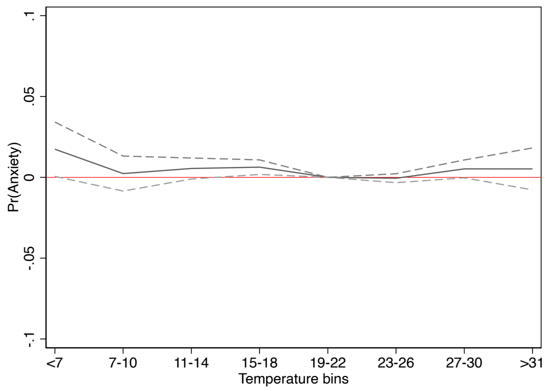
(d) Feeling test is going bad

Notes: Pooled sample of grade 5^{th} in school years 2011-12 to 2016-17. Figures display non linear estimates of the effect of temperatures in the day of language test on students emotional outcomes. Dependent variables are dummies for each emotional perceptions retrieved from students' questionnaire. Estimates include controls for weather (10 bins of rainfall, windspeed and humidity) at the municipal level, the day of the test. Student-level controls include female, foreign, early enrolled, retained. We also control for class size. Fixed effects include: school year, school, day-of-week and region-by-school year. Max. temperature is in Celsius degree (°C), with reference category of 19-22°C. Confidence intervals are at 95%.

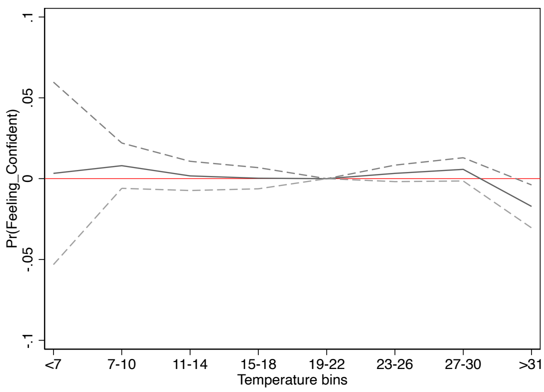
Figure 11: Effect of Temperatures During Language Test on Emotional Outcomes - Grade 10^{th}



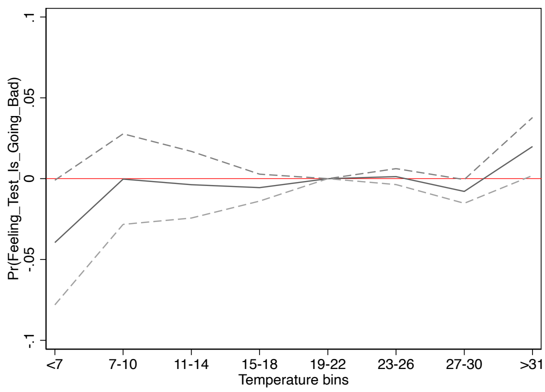
(a) Worried before test



(b) Anxiety



(c) Feeling confident



(d) Feeling test is going bad

Notes: Pooled sample of grade 10^{th} in school years 2011-12 to 2014-15 (student questionnaire was not administered for language test of grade 10^{th} from school years 2014-15 to 2016-17). Figures display non linear estimates of the effect of temperatures in the day of language test on students emotional outcomes. Dependent variables are dummies for each emotional perceptions retrieved from students' questionnaire. Estimates include controls for weather (10 bins of rainfall, windspeed and humidity) at the municipal level, the day of the test. Student-level controls include female, foreign, early enrolled, retained. We also control for class size. Fixed effects include: school year, school, day-of-week and region-by-school year. Max. temperature is in Celsius degree (°C), with reference category of 19-22°C. Confidence intervals are at 95%.

Tables

Table 1: Dates of the Tests

School Year	Month	Day	Math			Language		
			Grade 2nd	Grade 5th	Grade 10th	Grade 2nd	Grade 5th	Grade 10th
	5	8			x			x
2011-2012	5	9				x	x	
	5	11	x	x				
	5	7				x	x	
2012-2013	5	10	x	x				
	5	16			x			x
	5	6				x	x	
2013-2014	5	7	x	x				
	5	13			x			x
	5	6				x	x	
2014-2015	5	7	x	x				
	5	12			x			x
	5	4				x	x	
2015-2016	5	5	x	x				
	5	12			x			x
	5	3				x	x	
2016-2017	5	5	x	x				
	5	9			x			x

Notes: Dates of the test by grade from school year 2011-12 to 2016-17 in math and language tests.

Table 2: Summary Statistics

	Math		Language	
	Mean	S.d.	Mean	s.d.
Test variables:				
Correct answers (%)	57.068	21.650	65.008	20.358
Student cheating (%)	0.047	0.130	0.042	0.128
School monitoring (%)	0.065	0.246	0.064	0.244
Weather variables:				
Max. temperature	22.379	3.612	21.617	3.174
Wind speed	2.499	1.177	2.546	1.219
Tot. precipitation	2.069	7.806	1.812	7.413
Relative humidity	72.430	10.021	73.086	11.536
Students' and class characteristics:				
Female	0.490	0.499	0.490	0.451
Foreign	0.102	0.302	0.101	0.301
Early enrolled (%)	0.015	0.123	0.016	0.125
Retained (%)	0.068	0.251	0.063	0.243
Class size	18.791	4.641	18.814	4.588
Obs.	8,020,637		7,674,309	
# of schools	24,357		24,257	
# of municipalities	6,745		6,734	

Notes: Pooled sample of grades 2, 5 and 10 in school years from 2011-12 to 2016-17. School monitor variable is at the institute level. Test scores are standardized with 0 mean and unitary standard deviation within grade and academic year.

Table 3: Effect of Temperatures on Test Score and Manipulation

	Math		Language	
	Test score	Score manipulation	Test score	Score manipulation
	(1)	(2)	(3)	(4)
Temperature: <7°C	-0.031	0.002	0.022	0.007*
	(0.025)	(0.005)	(0.024)	(0.004)
Temperature: 7-10°C	-0.031*	-0.002	0.030	0.001
	(0.016)	(0.003)	(0.019)	(0.004)
Temperature: 11-14°C	0.005	0.002	0.016	0.002
	(0.010)	(0.002)	(0.011)	(0.002)
Temperature: 15-18°C	-0.003	0.000	0.005	0.001
	(0.005)	(0.001)	(0.005)	(0.001)
Temperature: 23-26°C	-0.007	0.003**	-0.000	0.001
	(0.005)	(0.001)	(0.004)	(0.001)
Temperature: 27-30°C	-0.022**	0.012***	0.007	0.007**
	(0.009)	(0.002)	(0.013)	(0.003)
Temperature: >31°C	-0.047***	0.002	-0.004	0.005
	(0.012)	(0.003)	(0.023)	(0.005)
Observations	8,020,637	8,020,637	7,674,309	7,674,309
F-stat.	3.544	3.544	0.507	1.325
P-val.	0.001	0.001	0.830	0.234

Notes: OLS estimates of standardized test scores and score manipulation in mathematics (Column 1 and 2) and language (Column 3 and 4) on bins of maximum temperatures observed the day of the test at the municipal level. Pooled sample of grades 2, 5 and 10 in school years from 2011-12 to 2016-17. Estimates include controls for weather (10 bins of rainfall, windspeed and humidity) at the municipal level, the day of the test. Student-level controls include female, foreign, early enrolled, retained. We also control for class size. Fixed effects include: school year, school-by-grade, day-of-week and region-by-school year. Max. temperature is in Celsius degree (°C), with reference category of 19-22°C. Standard errors, in parentheses, are clustered on municipalities. We also report F-statistics and p-values for the joint significance of temperature coefficients. Significance: *** p<0.01, ** p<0.05, * p<0.1.

Table 4: First-Stage Effect of Monitoring on Manipulation

	Score manipulation	
	Math	Language
	(1)	(2)
Monitoring	-0.041***	-0.036***
	(0.001)	(0.001)
F-stat.	1002.87	1053.40
Obs.	8,020,637	7,674,309

Notes: First-stage estimates of random monitoring on the fraction of score manipulation in mathematics (Column 1) and language (Column 2). Pooled sample of grades 2, 5 and 10 in school years from 2011-12 to 2016-17. Estimates include controls for weather (10 bins of rainfall, windspeed and humidity) at the municipal level, the day of the test. Student-level controls include female, foreign, early enrolled, retained. We also control for class size. Fixed effects include: school year, school-by-grade, day-of-week and region-by-school year. Max. temperature is in Celsius degree (°C), with reference category of 19-22°C. Standard errors, in parentheses, are clustered on municipalities. We also report F-statistics and p-values for the joing significance of temperature coefficients. Significance: *** p<0.01, ** p<0.05, * p<0.1.

Table 5: Effect of Temperatures on Test Score Net of Manipulation

	Test Score	
	Math	Language
	(1)	(2)
Temperature: <7°C	-0.041*	-0.008
	(0.023)	(0.028)
Temperature: 7-10°C	-0.022	0.028
	(0.018)	(0.018)
Temperature: 11-14°C	-0.003	0.009
	(0.010)	(0.011)
Temperature: 15-18°C	-0.004	-0.000
	(0.005)	(0.005)
Temperature: 23-26°C	-0.019***	-0.004
	(0.005)	(0.004)
Temperature: 27-30°C	-0.076***	-0.022*
	(0.009)	(0.012)
Temperature: >31°C	-0.055***	-0.026
	(0.015)	(0.018)
Manipulation	4.612***	4.006***
	(0.110)	(0.118)
Obs.	8,020,637	7,674,309
F-stat.	12.07	1.223
P-val.	0.001	0.286

Notes: 2SLS estimates of standardized test scores in mathematics (Column 1) and language (Column 2) on bins of maximum temperatures observed the day of the test at the municipal level. Manipulation is instrumented using random monitoring at the school level. Sanderson-Windmeijer F-statistics of excluded instrument: 1003.14, p =0.0000 for math test, and 1059.59, p =0.0000 for language test). Pooled sample of grades 2, 5 and 10 in school years from 2011-12 to 2016-17. Estimates include controls for weather (10 bins of rainfall, windspeed and humidity) at the municipal level, the day of the test. Student-level controls include female, foreign, early enrolled, retained. We also control for class size. Fixed effects include: school year, school-by-grade, day-of-week and region-by-school year. Max. temperature is in Celsius degree (°C), with reference category of 19-22°C. Standard errors, in parentheses, are clustered on municipalities. We also report F-statistics and p-values for the joint significance of temperature coefficients. Significance: *** p<0.01, ** p<0.05, * p<0.1.

Table 6: SUMMARY STATISTICS - EMOTIONAL PERCEPTIONS DURING THE TEST

	Mean	S.D	Obs.
	Panel A: Math – 5th grade		
$\mathbb{1}$(Worried before the test)	0.552	0.497	2,285,576
$\mathbb{1}$(Anxiety during the test)	0.174	0.379	2,283,346
$\mathbb{1}$(Feeling test was not going well)	0.453	0.497	2,278,711
$\mathbb{1}$(Feeling confident during the test)	0.531	0.499	2,277,589
	Panel B: Math – 10th grade		
$\mathbb{1}$(Worried before the test)	0.276	0.447	1,068,721
$\mathbb{1}$(Anxiety during the test)	0.111	0.315	1,068,755
$\mathbb{1}$(Feeling test is not going well)	0.328	0.461	1,066,898
$\mathbb{1}$(Feeling confident during the test)	0.694	0.460	664,602
	Panel C: Language – 5th grade		
$\mathbb{1}$(Worried before the test)	0.553	0.497	2,202,472
$\mathbb{1}$(Anxiety during the test)	0.172	0.378	2,200,503
$\mathbb{1}$(Feeling test was not going well)	0.452	0.497	2,195,066
$\mathbb{1}$(Feeling confident during the test)	0.531	0.499	2,196,075
	Panel D: Language – 10th grade		
$\mathbb{1}$(Worried before the test)	0.288	0.453	798,554
$\mathbb{1}$(Anxiety during the test)	0.114	0.318	798,620
$\mathbb{1}$(Feeling test was not going well)	0.331	0.471	393,099
$\mathbb{1}$(Feeling confident during the test)	0.694	0.460	797,462

Notes: Pooled sample school years 2011-12 to 2016-17 for grade 5th and 2011-12 to 2014-15 for grade 10th. The emotional perceptions are dummy indicators retrieved from self reported answers on a student questionnaire. These variables are available for grades 5th and 10th only.

Appendix A

In this appendix we present a simple generalization of the conceptual framework proposed in Section 2. We keep all the assumptions made so far, but we relax the hypothesis of equal contribution of true cognitive performance and manipulation to the observed score.

$$P^O = P + f(C) = g(T) + f(h(g(T)) \tag{8}$$

In this equation the part of score obtained through manipulation enters as $f(C)$. This means that answers obtained through manipulation are no more assumed to contribute similarly to those provided by students themselves. For instance, while cognitive performance is always effective and adds a positive amount to the observed score, manipulation could be both as effective as cognitive performance or noneffective (e.g. in case of bad manipulation).

Deriving equation (8) by T we obtain:

$$\frac{dP^O}{dT} = \frac{dg}{dT}\left(1 + \frac{df}{dh} \times \frac{dh}{dg}\right) \tag{9}$$

This expression states that the difference between the true effect of temperature on cognitive performance ($\frac{dg}{dT}$) and the observed effect ($\frac{dP^O}{dT}$) depends both on how students or teachers adjust the level of manipulation when cognitive performance varies ($\frac{dh}{dg}$) and on its effectiveness ($\frac{df}{dh}$). When there is no compensation at all ($\frac{dh}{dg} = 0$) the true effect and the observed one coincide. When students or teachers compensate because of temperature induced cognitive performance deterioration, the extent of the distortion depends on the effectiveness of manipulation. Unfortunately, our data do not allow to distinguish the bias coming from $\frac{dh}{dg}$ and that from $\frac{df}{dh}$, as we only observe their product.

Appendix B

Table B1: Effect of Temperatures on Test Score – Between Subjects Specification

	Test Score
	(1)
Temperature: <7°C	-0.055***
	(0.020)
Temperature: 7-10°C	0.016
	(0.010)
Temperature: 11-14°C	0.000
	(0.005)
Temperature: 15-18°C	-0.003
	(0.002)
Temperature: 23-26°C	-0.001
	(0.001)
Temperature: 27-30°C	-0.043***
	(0.002)
Temperature: >31°C	-0.042***
	(0.004)
Manipulation	6.282***
	(0.072)
Obs.	11,347,227
F-stat.	99.47
P-val.	0.001

Notes: 2SLS estimates of test score on bins of maximum temperatures observed the day of the test at the municipal level. Pooled sample of subjects math and language, and grades 2, 5 in school years from 2011-12 to 2016-17. Estimates include controls for weather (10 bins of rainfall, windspeed and humidity) at the municipal level, the day of the test. Student-level controls include female, foreign, early enrolled, retained. We also control for class size. Fixed effects include: student-by-grade-year, subject, day-of-week and region-by-academic year. Max. temperature is in Celsius degree (°), with reference category of 19-22°. Standard errors, in parentheses, are clustered on municipalities. Significance: *** p<0.01, ** p<0.05, * p<0.1.

Table B2: Falsification Test - Effect of Temperatures on Test Score

	Test Score	
	(1)	(2)
	Math	Language
Temperature: <7°	-0.0053	-0.0032
	0.0264	0.0194
Temperature: 7-10°C	-0.0003	-0.0002
	0.0129	0.0135
Temperature: 11-14°C	0.0001	0.0018
	0.0075	0.0067
Temperature: 15-18°C	-0.0001	0.0010
	0.0035	0.0031
Temperature: 23-26°C	0.0006	0.0001
	0.0031	0.0027
Temperature: 27-30°C	-0.0003	-0.0004
	0.0041	0.0074
Temperature: >31°C	-0.0005	-0.0002
	0.0103	0.0137
Obs.	8,020,637	7,674,309

Notes: OLS estimates of standardized test scores in mathematics (Column 1) and language (Column 2) on bins of maximum temperatures observed the day of the test at the municipal level. Pooled sample of grades 2, 5 and 10 in school years from 2011-12 to 2016-17. Estimates are obtained by reshuffling dates across municipalities within the same school year and grade (50 iterations). Estimates include controls for weather (10 bins of rainfall, windspeed and humidity) at the municipal level, the day of the test. Student-level controls include female, foreign, early enrolled, retained. We also control for class size. Fixed effects include: school year, school-by-grade, day-of-week and region-by-school year. Max. temperature is in Celsius degree (°C), with reference category of 19-22°C. Standard errors, in parentheses, are clustered on municipalities. Significance: *** p<0.01, ** p<0.05, * p<0.1.

Table B3: Effect of Temperatures during Math Test on Emotional Outcomes

	Math							
	Grade 5th				Grade 10th			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Worried before test	Anxiety	Feeling confident	Feeling bad	Worried before test	Anxiety	Feeling confident	Feeling bad
Temperature: <7°C	-0.013	0.012	0.039***	-0.008	-0.050***	0.018**	-0.041**	0.009
	(0.016)	(0.017)	(0.012)	(0.020)	(0.015)	(0.009)	(0.019)	(0.029)
Temperature: 7-10°C	-0.005	0.008	-0.016	0.029***	-0.012	0.003	-0.002	0.011
	(0.011)	(0.009)	(0.010)	(0.010)	(0.011)	(0.005)	(0.014)	(0.007)
Temperature: 11-14°C	-0.005	-0.003	-0.016***	0.009*	-0.003	0.007**	-0.004	0.002
	(0.006)	(0.004)	(0.006)	(0.006)	(0.007)	(0.003)	(0.011)	(0.005)
Temperature: 15-18°C	-0.002	-0.001	-0.007***	0.008***	-0.003	0.006***	-0.007*	0.001
	(0.002)	(0.001)	(0.002)	(0.002)	(0.004)	(0.002)	(0.004)	(0.003)
Temperature: 23-26°C	0.004**	0.003**	0.009***	-0.013***	-0.004	0.001	0.001	0.003
	(0.002)	(0.001)	(0.001)	(0.003)	(0.003)	(0.001)	(0.003)	(0.003)
Temperature: 27-30°C	0.011***	0.010***	0.017***	-0.023***	-0.000	0.006**	-0.008*	0.004
	(0.002)	(0.002)	(0.003)	(0.004)	(0.005)	(0.003)	(0.004)	(0.004)
Temperature: >31°C	0.015***	0.007**	0.022***	-0.035***	0.013	0.009	0.019**	-0.019**
	(0.004)	(0.003)	(0.004)	(0.004)	(0.009)	(0.007)	(0.008)	(0.008)
Obs.	2,762,131	2,759,589	2,754,726	2,753,782	1,467,974	1,468,602	1,075,936	1,466,552
F-stat.	3.536	3.928	13.16	11.43	3.393	2.912	3.752	1.855
P-val.	0.001	0.001	0.001	0.001	0.001	0.005	0.001	0.073

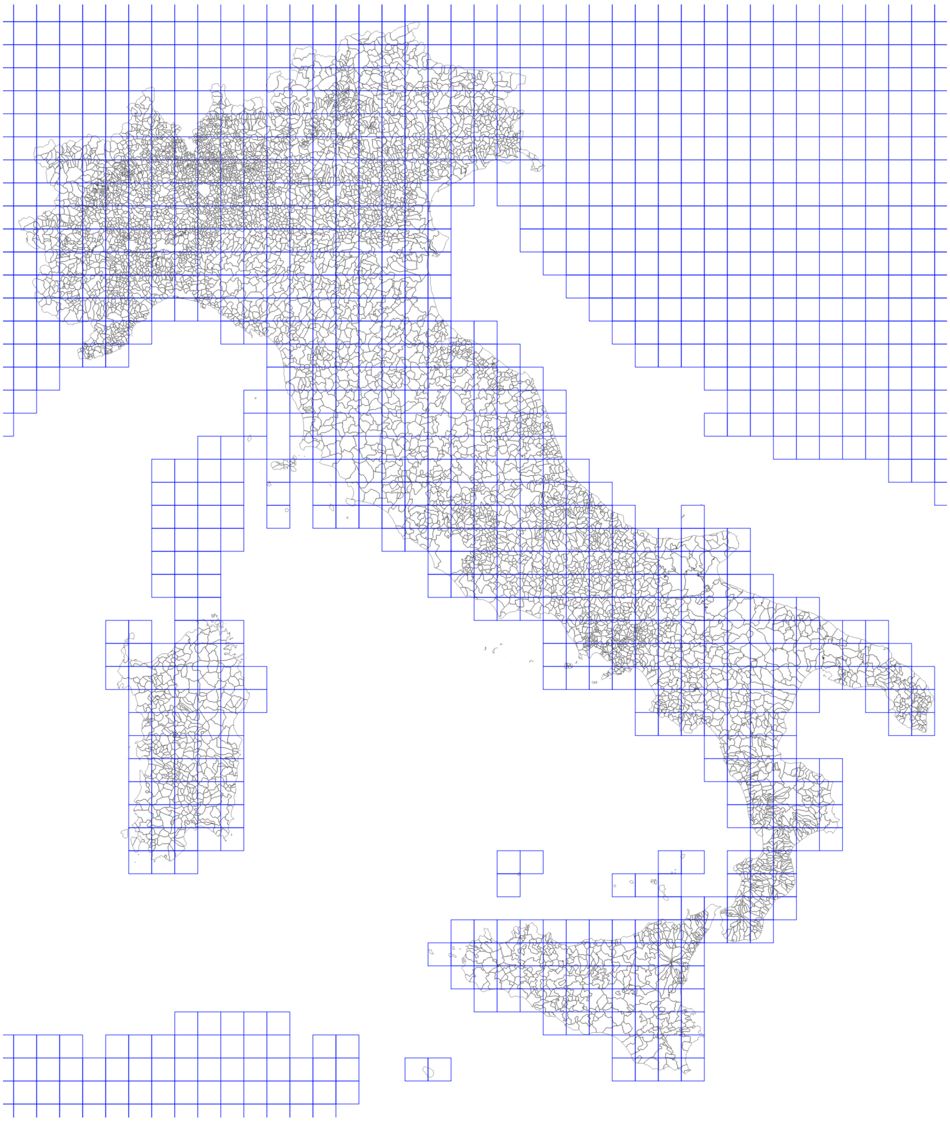
Notes: OLS estimates in a pooled sample of school years 2011-12 to 2016-17 for grade 5th and 2011-12 to 2014-15 for grade 10th in math test. Dependent variables are dummies for each emotional perceptions retrieved from students' questionnaire. Estimates include controls for weather (10 bins of rainfall, windspeed and humidity) at the municipal level, the day of the test. Student-level controls include female, foreign, early enrolled, retained. We also control for class size. Fixed effects include: school year, school, day-of-week and region-by-school year. Max. temperature is in Celsius degree (°C), with reference category of 19-22°C. Standard errors, in parentheses, are clustered on municipalities. We also report F-statistics and p-values for the joint significance of temperature coefficients. Significance: *** p<0.01, ** p<0.05, * p<0.1.

Table B4: Effect of Temperatures during Language Test on Emotional Outcomes

	Language							
	Grade 5th				Grade 10th			
	(1) Worried before test	(2) Anxiety	(3) Feeling confident	(4) Feeling bad	(5) Worried before test	(6) Anxiety	(7) Feeling confident	(8) Feeling bad
Temperature: <7°C	0.010	0.035***	0.010	0.036***	-0.050***	0.017**	-0.040**	0.003
	(0.015)	(0.010)	(0.015)	(0.013)	(0.014)	(0.009)	(0.020)	(0.029)
Temperature: 7-10°C	-0.003	0.013*	-0.013	0.034***	-0.012	0.002	-0.000	0.008
	(0.011)	(0.008)	(0.009)	(0.010)	(0.011)	(0.006)	(0.014)	(0.007)
Temperature: 11-14°C	0.003	0.009***	-0.006	0.027***	-0.004	0.005*	-0.004	0.002
	(0.005)	(0.003)	(0.005)	(0.005)	(0.007)	(0.003)	(0.011)	(0.005)
Temperature: 15-18°C	0.002	0.003**	-0.006***	0.015***	-0.001	0.006***	-0.006	0.000
	(0.002)	(0.001)	(0.002)	(0.002)	(0.004)	(0.002)	(0.004)	(0.003)
Temperature: 23-26°C	0.000	-0.001	-0.001	0.000	-0.005*	-0.001	0.001	0.003
	(0.002)	(0.001)	(0.002)	(0.002)	(0.003)	(0.001)	(0.003)	(0.003)
Temperature: 27-30°C	-0.001	0.021***	0.015***	0.006	-0.002	0.005*	-0.008**	0.006
	(0.004)	(0.003)	(0.004)	(0.004)	(0.005)	(0.003)	(0.004)	(0.004)
Temperature: >31°C	-0.003	0.007	0.002	0.014**	0.008	0.005	0.020**	-0.017**
	(0.006)	(0.004)	(0.005)	(0.006)	(0.008)	(0.007)	(0.009)	(0.007)
Obs.	2,664,898	2,662,633	2,657,994	2,657,175	1,461,578	1,462,193	1,070,560	1,460,135
F-stat.	0.370	9.480	4.764	10.14	3.977	2.711	3.645	1.906
P-val.	0.920	0.001	0.001	0.001	0.001	0.008	0.001	0.065

Notes: OLS estimates in a pooled sample of school years 2011-12 to 2016-17 for grade 5th and 2011-12 to 2014-15 for grade 10th in language test. Dependent variables are dummies for each emotional perceptions retrieved from students' questionnaire. Estimates include controls for weather (10 bins of rainfall, windspeed and humidity) at the municipal level, the day of the test. Student-level controls include female, foreign, early enrolled, retained. We also control for class size. Fixed effects include: school year, school, day-of-week and region-by-school year. Max. temperature is in Celsius degree (°C), with reference category of 19-22°C. Standard errors, in parentheses, are clustered on municipalities. We also report F-statistics and p-values for the joint significance of temperature coefficients. Significance: *** p<0.01, ** p<0.05, * p<0.1.

Figure B1: Weather Data Grid and Municipalities



Notes: Source: Agri-4-Cast data. Available at `https://agri4cast.jrc.ec.europa.eu/DataPortal/Resource_Files/SupportFiles/grid25.zip`