TECHNICAL APPENDIX

Technical appendix for:
Hot test days, lower math scores: How heat affects student achievement

October 2025

Sofia Postell, Megan Kuhfeld, Susan Kowalski, Jazmin Isaacs

NWEA, a division of HMH, supports students and educators worldwide by providing assessment solutions, insightful reports, professional learning offerings, and research services. Visit NWEA.org to find out how NWEA can partner with you to help all kids learn.

Suggested citation: Postell et al. (2025). Technical appendix for: Hot test days, lower math scores: How heat affects student achievement.

Table of Contents

List of Tables

List of Figures

1. Introduction

The purpose of this technical appendix is to more fully describe the sample, methods, and results of the research brief "Hot test days, lower math scores: How heat affects student achievement." We investigated two research questions in this brief:

1) What is the impact of maximum test-day temperature on students' test scores?
2) How are students in high-poverty schools impacted differently by test-day temperature?

2. Data

2.1. Data Sources

NWEA MAP Growth

The MAP Growth data for this study are from the NWEA anonymized longitudinal student achievement database. School districts use NWEA® MAP® Growth™ assessments[1] to monitor elementary and secondary students' reading and math achievement and gains, with assessments typically administered in the fall (usually between August and November), winter (usually December to March), and spring (late March through June). The NWEA data also include demographic information, including student race/ethnicity and gender. For each test event, we also know the time and date that the test occurred. In this analysis, we focus on fall testing, since the most extreme hot temperatures in a school year typically occur in early fall. Additionally, we use spring scores from the same calendar year as control variables. The fall testing dates in our sample ranged from approximately mid-July to mid-November of each year.

Fall and spring RIT scores in each subject were standardized relative to the 2025 MAP Growth Norms, which situates test performance relative to students in the same grade, subject, and time in school prior to testing.

Common Core of Data (CCD)

We used the 2023-24 school data files from the CCD to use for school enrollment information and free and reduced lunch (FRPL) percentages. To calculate the percentage of FRPL-eligible students within each school, we used either the total free and reduced lunch number or the number of students eligible under direct certification, based on which was reported and whether a school participated in the community eligibility provision. We also downloaded the Education Demographic and Geographic Estimates (EDGE) geocodes dataset for the same year to obtain school latitude and longitude coordinates.

National Oceanic and Atmospheric Administration (NOAA) Climate Data

To obtain temperature data for the testing dates in our sample, we downloaded daily temperature records from the NOAA Climate Data Online database. We downloaded all available temperature data for each state during our study period and filtered to weather stations with consistently reported daily maximum temperatures.

We used daily maximum temperature (grouped into ten-degree bins) as the primary independent variable in our analyses. The decision to group temperature into bins was motivated by initial observations that showed a non-linear relationship between temperature and

[1] In the 2023-24 school year, NWEA began the phased implementation of an enhanced item-selection algorithm (EISA) for the MAP Growth assessment, which altered the test scale of the math assessment. To account for the differences in test version, we converted all legacy MAP math test scores to be on the new EISA scale. For more detail on the score conversion process, please see NWEA's EISA documentation.

scores. Figure 1A visualizes this relationship by fitting a linear model which included prior spring scores and controls for grade, state, and year tested, and plotting the predicted fall scores against test-day temperature using locally estimated smoothing. The figure shows slight increases in scores in both subjects as temperatures rise before reaching 90°F, where scores show sharper declines.

2.2. MAP Growth Samples

To conduct our analyses, we created a sample of 3rd to 8th grade Math and Reading test scores from U.S. public schools in six selected states located in variety of U.S. regions. The six states were chosen based on a combination of several factors, primarily high MAP coverage and geographic variability. We limited to schools that administered MAP Growth to over 70% of the enrolled students in a grade and were less than 20 miles away from their closest weather station reporting daily temperature data from NOAA.

In order to appear in our sample, students in each year (2022, 2023, and 2024) had to have both a fall term MAP score and a score in the same subject in the prior spring. We also ensured that a student's school that they attended in the spring was matched to the same weather station as their fall school. We excluded students testing from 30-50 °F and from 110-120 °F due to the relatively low number of test events at those temperatures. We also excluded students whose race was reported as Native Hawaiian/Pacific Islander or Not Specified/Other, also due to very low number of students in those categories.

Both the math and reading final samples contained approximately 1.8 million unique students. Table 1 displays the characteristics of the students in our sample by year/subject. Descriptive information for the schools in our samples along with comparisons to the population of all U.S. public schools enrolling students in any of grades 3 to 8 is provided in Table 2. There are approximately 5,000 schools in our Reading and Math samples. Our samples reflect a diversity of schools from across various locales (urban, suburban, town, and rural). Compared to the population of U.S. schools, our samples include schools serving a higher average percentage of Hispanic students and overrepresents urban schools.

3. Methods

Matching test events to heat data

We used EDGE school geocodes data from the CCD to obtain schools' latitude and longitude coordinates and then used those to match to the closest weather station in a given state using the Haversine Distance formula[2]. We limited to schools whose nearest weather station was less than 20 miles away. Once we matched schools to nearby weather stations, we merged in the daily maximum temperature on each testing date in the fall term. Table 3 shows the distribution of test events by the daily maximum temperature bin (pooling across the three years).

RQ1. What is the impact of maximum test-day temperature on students' test scores?

To answer our first research question, we ran a series of models regressing standardized fall test scores on the test-day maximum temperature bins and a varying set of background characteristics. In the first model, we included indicators for a student's grade level, state, and year tested. In the second model, we added controls for students' prior spring RIT

[2] Specifically, we used the distHaversine function from the geosphere package in R.

score in the same subject as well as for whether the test occurred in the afternoon. In the third and final model, we added controls for the student's race and gender. In all three models, the nesting of students in schools was accounted for with a random school intercept.

The equation for the full model (Model 3) is shown below, where the outcome variable, $z.F_{ij}$, represents the standardized fall RIT score for student *i* in school *j*, r_{0j} represents the random intercept for school *j*, and ε_{ij} is the residual error for student *i* in school *j*.

$$z.F_{ij} = \beta_0 + \beta_1 temp\_bin_{ij} + \beta_2 z.S_{ij} + \beta_3 grade_{ij} + \beta_4 race_{ij} + \beta_5 gender_{ij} + \beta_6 afternoon_{ij} + \beta_7 state_j + \beta_8 year_j + r_{0j} + \varepsilon_{ij}$$

In a supplemental analysis, we ran an additional model to analyze how scores are affected by relatively hot days compared to a school's mean testing temperature during the fall. In this analysis, we centered each test-day maximum temperature using the school-level mean temperatures. We then ran the following model using the group-mean centered temperature $(temp\_gmc_{ij})$ and controlling for the mean at the school level $(temp\_m_{j)}$.

$$z.F_{ij} = \beta_0 + \beta_1 temp\_gmc_{ij} + \beta_2 temp\_m_j + \beta_3 z.S_{ij} + \beta_4 grade_{ij} + \beta_5 race_{ij} + \beta_6 gender_{ij} + \beta_7 afternoon_{ij} + \beta_8 state_j + \beta_9 year_j + r_{0j} + \varepsilon_{ij}$$

RQ2. How are students in high-poverty schools impacted differently by test-day temperature?

Secondly, we examined whether the effect of test-day temperature on fall achievement varied depending on the level of poverty at the school. Specifically, we estimated the interaction between FRPL percentage and test-day temperature using the model below, where $FRPL_j X\ temp\_bin_{ij}$ represents the interaction between FRPL-percentage for school j and the temperature of the test event for student *i* in school *j.* There were four levels of school FRPL eligibility: <25%, 25-50%, 50-75%, and >75%, and 25-50% was used as the reference group. To calculate the total effect for each school-lunch subgroup using the model results as shown in Figure 2 of the brief, we added the main effect of each temperature bin on each subgroup to the interaction between each subgroup and temperature bin (using 51-60°F as the reference group).

$$z.F_{ij} = \beta_0 + \beta_1 temp\_bin_{ij} + \beta_2 z.S_{ij} + \beta_3 grade_{ij} + \beta_4 race_{ij} + \beta_5 gender_{ij} + \beta_6 afternoon_{ij} + \beta_7 state_j + \beta_8 year_j + \beta_9 FRPL_j X\ temp\_bin_{ij} + b_{0j} + \varepsilon_{ij}$$

4. Results

4.1. RQ1 Results

The first set of regression results for RQ1 are shown in Table 4 for math and Table 5 for reading. In math, results from Model 3 show consistent reductions in fall RIT score for all temperatures above the reference group, with the highest effect being for testing days above 100°F (-0.06 SD). The results of Model 3 in reading are largely insignificant, apart from a -0.01 SD effect for students testing on 61-70°F days.

The results for the supplemental analyses looking at the effect of a hotter test day relative to the school's average fall temperature are shown in Table 10 for math and Table 11 for reading. In math, small but statistically significant effects associated with a 1 degree increase in temperature were found for both higher group-mean centered testing temperatures (-0.001

SD) and for higher temperatures at the school-level (-.003 SD). In reading, higher test-day temperatures at the school-level were associated with lower scores (-0.003 SD), but higher temperatures within schools were found to have a positive association (0.001 SD).

4.2. RQ2 Results

The results for RQ2 are shown in Table 6 for math and Table 7 for reading. In math, significant decreases in scores were concentrated around schools with greater than 75% FRPL-eligible students, with the largest effects at 101–110°F testing days (−0.048 SD). Effects were largely insignificant for lower-poverty schools. In reading, significant but positive effects (0.03 SD) were found for the same group (days above 100°F, >75% FRPL), with similar insignificant results for other groups.

5. Sensitivity analyses

To understand how the findings varied across the three years of our study, we ran the same model below but separately by year to explore variation between years (see Tables 8 and 9). We also conducted a sensitivity analysis to test whether the calculated distance between schools and weather stations affected the results. In our main analysis, we included schools that were within 20 miles of a weather station, but we also re-ran the results with weather stations that were within 10 miles. This additional restriction did not have a significant effect on the analysis.

Table 1. Description of students in analytic sample

Sample	Year	N. Students	N. School	% White	% Asian	% Black	% Hispanic	% AIAN	% Female	% FRPL
Math	All	1,850,249	5,288	41.6	4.5	13.6	33.6	1	49	61.6
Math	2022	1,056,199	4,754	42.1	4.4	12.6	33.4	0.9	49	60.7
Math	2023	1,112,341	4,976	42.6	4.3	12.9	34.6	0.9	49	61.9
Math	2024	993,653	4,382	43.9	4.7	12.7	32.1	1.2	49	60.7
Reading	All	1,776,348	5,009	42.6	4.8	13.5	32.7	1	49	60.9
Reading	2022	1,014,897	4,477	43.5	4.8	12.9	31.7	0.8	48.9	60.3
Reading	2023	1,060,548	4,702	43.8	4.5	12.9	33.2	0.9	49	60.9
Reading	2024	973,724	4,200	44.3	4.9	12.4	32	1.2	48.9	60

Note. AIAN=American Indian or Alaska Native. Subject-specific samples for each year include students who tested in that subject in the fall of that year as well as in the spring of the previous school year. The total number of students reflects some students who appeared in the samples of multiple years.

Table 2. Sample school information relative to U.S. population of schools

Sample	N. Schools	% FRPL	% White	% Black	% Asian	% Hispanic	% City	% Suburb	% Rural	% Town
Math	5,288	61.6	41.6	13.6	4.5	33.6	38.8	35.1	16.3	9.9
Reading	5,009	60.9	42.6	13.5	4.8	32.7	38.9	34.1	16.8	10.3
Population of U.S. Public Schools Serving Grades 3-8	77,481	57.8	46.8	14.4	4.3	27.2	29	31.3	29.4	10.3

Note: % FRPL refers to school-level eligibility rates for free or reduced priced lunch. The source of the variables is the Common Core of Data (CCD) collected by the National Center for Educational Statistics for the 2023-24 school year. The U.S. public school population comparison was determined by limiting to the schools that offered any of grades 3-8.

Table 3. Number of tests per 10-degree bin of test-day maximum temperature

Temperature Bin	Math	Reading	% of Total (Math)	% of Total (Reading)
51-60	23,849	21,108	0.91	0.85
61-70	171,029	168,383	6.56	6.76
71-80	337,898	345,199	12.96	13.85
81-90	787,536	767,725	30.21	30.80
91-100	823,399	763,604	31.59	30.63
101-110	462,866	426,645	17.76	17.12

Table 4. RQ1 Results (Math)

	Dependent variable: Standardized fall RIT		
	(1)	(2)	(3)
61-70	0.003	-0.031***	-0.031***
	(0.007)	(0.003)	(0.003)
71-80	0.013*	-0.038***	-0.038***
	(0.007)	(0.003)	(0.003)
81-90	0.017**	-0.047***	-0.047***
	(0.007)	(0.003)	(0.003)
91-100	0.008	-0.048***	-0.047***
	(0.007)	(0.003)	(0.003)
101-110	0.025***	-0.061***	-0.060***
	(0.007)	(0.003)	(0.003)
Grade	X	X	X
Year	X	X	X
State	X	X	X
Pretest		X	X
Afternoon		X	X
Race			X
Gender			X
Constant	-0.061***	0.030***	0.033***
	(0.020)	(0.006)	(0.006)
Observations	3,162,193	3,162,193	3,162,193

Notes:
51-60°F was used as the reference group.

*p<0.1; **p<0.05; ***p<0.01

Table 5. RQ1 Results (Reading)

	Dependent variable: Standardized fall RIT		
	(1)	(2)	(3)
61-70	0.023***	-0.011***	-0.012***
	(0.007)	(0.004)	(0.004)
71-80	0.038***	-0.003	-0.004
	(0.007)	(0.004)	(0.004)
81-90	0.043***	-0.002	-0.002
	(0.007)	(0.004)	(0.004)
91-100	0.034***	-0.004	-0.004
	(0.008)	(0.004)	(0.004)
101-110	0.054***	0.006	0.007*
	(0.008)	(0.004)	(0.004)
Grade	X	X	X
Year	X	X	X
State	X	X	X
Pretest		X	X
Afternoon		X	X
Race			X
Gender			X
Constant	-0.071***	-0.022***	0.013**
	(0.019)	(0.006)	(0.006)
Observations	3,049,169	3,049,169	3,049,169

Notes:
51-60°F was used as the reference group.

*p<0.1; **p<0.05; ***p<0.01

Table 6. RQ2 Results (Math)

	Dependent variable: Standardized fall RIT
61-70 * <= 25% FRPL	-0.016*
	(0.010)
71-80 * <= 25% FRPL	0.007
	(0.010)
81-90 * <= 25% FRPL	-0.003
	(0.010)
91-100 * <= 25% FRPL	-0.016
	(0.010)
101-110 * < =25% FRPL	-0.002
	(0.010)
61-70 * 50-75% FRPL	0.001
	(0.009)
71-80 * 50-75% FRPL	0.009
	(0.009)
81-90* 50-75% FRPL	0.005
	(0.009)
91-100 * 50-75% FRPL	0.004
	(0.009)
101-110 * 50-75% FRPL	-0.002
	(0.009)
61-70 * > 75% FRPL	-0.037***
	(0.009)
71-80 * > 75% FRPL	-0.034***
	(0.009)
81-90 * > 75% FRPL	-0.039***
	(0.009)
91-100 * > 75% FRPL	-0.032***
	(0.009)
101-110 * > 75% FRPL	-0.048***
	(0.009)
Constant	0.038***
	(0.008)
Observations	3,117,327

Notes: *p<0.1; **p<0.05; ***p<0.01

51-60℉ is the reference group for temperature interactions. 25-50% FRPL is the reference group for FRPL interactions. Race, gender, state, year, afternoon, and pre-test controls were also included.

Table 7. RQ2 Results (Reading)

	Dependent variable: Standardized fall RIT
61-70 * <= 25% FRPL	-0.015
	(0.012)
71-80 * <=25% FRPL	-0.001
	(0.012)
81-90 * <=25% FRPL	-0.005
	(0.012)
91-100 * <=25% FRPL	-0.016
	(0.012)
101-110 * < =25% FRPL	0.011
	(0.012)
61-70 * 50-75% FRPL	-0.011
	(0.010)
71-80 * 50-75% FRPL	-0.009
	(0.010)
81-90 * 50-75% FRPL	-0.015
	(0.010)
91-100 * 50-75% FRPL	-0.017
	(0.011)
101-110 * 50-75% FRPL	-0.021**
	(0.011)
61-70 * >75% FRPL	0.021*
	(0.011)
71-80 * >75% FRPL	0.019*
	(0.011)
81-90 * >75% FRPL	0.016
	(0.011)
91-100 * >75% FRPL	0.022*
	(0.011)
101-110 * >75% FRPL	0.027**
	(0.012)
Constant	0.026***
	(0.009)
Observations	2,998,501
Notes:	*p<0.1; **p<0.05; ***p<0.01

51-60℉ is the reference group for temperature interactions. 25-50% FRPL is the reference group for FRPL interactions. Race, gender, state, year, afternoon, and pre-test controls were also included.

Table 8. Year by year RQ1 results (Math)

	Dependent variable: Standardized fall RIT

	2022	2023	2024
61-70	0.010**	-0.057***	-0.078***
	(0.005)	(0.008)	(0.014)
71-80	0.009*	-0.057***	-0.090***
	(0.005)	(0.008)	(0.014)
81-90	-0.003	-0.055***	-0.091***
	(0.005)	(0.008)	(0.014)
91-100	-0.002	-0.056***	-0.101***
	(0.006)	(0.008)	(0.014)
101-110	-0.010	-0.052***	-0.121***
	(0.007)	(0.008)	(0.015)
Pretest	X	X	X
Grade	X	X	X
State	X	X	X
Afternoon	X	X	X
Race	X	X	X
Gender	X	X	X
Constant	-0.049***	0.038***	0.126***
	(0.008)	(0.009)	(0.015)
Observations	1,056,199	1,112,341	993,653
Note:		*p<0.1; **p<0.05; ***p<0.01	

Table 9. Year by year RQ1 results (Reading)

	Dependent variable: Standardized fall RIT		
	2022	2023	2024
61-70	-0.015**	0.020**	-0.017
	(0.006)	(0.010)	(0.019)
71-80	-0.015**	0.023**	-0.014
	(0.006)	(0.010)	(0.019)
81-90	-0.023***	0.038***	-0.022
	(0.007)	(0.010)	(0.019)
91-100	-0.029***	0.043***	-0.044**
	(0.007)	(0.010)	(0.019)
101-110	-0.031***	0.037***	-0.028
	(0.008)	(0.010)	(0.020)
Pretest	X	X	X
Grade	X	X	X
State	X	X	X
Afternoon	X	X	X
Race	X	X	X
Gender	X	X	X
Constant	0.038***	-0.038***	0.035*
	(0.009)	(0.011)	(0.020)
Observations	1,014,897	1,060,548	973,724
Note:		*p<0.1; **p<0.05; ***p<0.01	

Table 10. Results of within-school analysis (Math)

	Dependent variable: Standardized fall RIT
Group mean-centered temperature	-0.001***
	(0.00004)
School mean temperature	-0.003***
	(0.0003)
Constant	0.278***
	(0.026)
Observations	3,162,193

Notes: *p<0.1; **p<0.05; ***p<0.01

Model included controls for grade, year, state, prior spring score, and student race. School mean temperature refers to the mean temperature of all the test days at a given school during the fall.

Table 11. Results of within-school analysis (Reading)

	Dependent variable: Standardized fall RIT
Group-mean centered temperature	0.001***
	(0.0001)
School mean temperature	-0.003***
	(0.0003)
Constant	0.286***
	(0.026)
Observations	3,049,169

Notes: *p<0.1; **p<0.05; ***p<0.01

Model included controls for grade, year, state, prior spring score, and student race. School mean temperature refers to the mean temperature of all the test days at a given school during the fall.

Figure 1A. Relationship Between Test-Day Temperatures and RIT Scores