TESTING, STRESS, AND PERFORMANCE: HOW STUDENTS RESPOND PHYSIOLOGICALLY TO HIGH-STAKES TESTING

Jennifer A. Heissel
(corresponding author)
Graduate School of Defense Management
Naval Postgraduate School
Monterey, CA 93943
jaheisse@nps.edu

Emma K. Adam
School of Education and Social Policy
Northwestern University
Evanston, IL 60208
ek-adam@northwestern.edu

Jennifer L. Doleac
Department of Economics
Texas A&M University
College Station, TX 77845
jdoleac@tamu.edu

David N. Figlio
School of Education and Social Policy
Northwestern University
Evanston, IL 60208
figlio@northwestern.edu

Jonathan Meer
Department of Economics
Texas A&M University
College Station, TX 77845
jmeer@tamu.edu

Abstract

We examine how students' physiological stress differs between a regular school week and a high-stakes testing week, and we raise questions about how to interpret high-stakes test scores. A potential contributor to socioeconomic disparities in academic performance is the difference in the level of stress experienced by students outside of school. Chronic stress—due to neighborhood violence, poverty, or family instability—can affect how individuals' bodies respond to stressors in general, including the stress of standardized testing. This, in turn, can affect whether performance on standardized tests is a valid measure of students' actual ability. We collect data on students' stress responses using cortisol samples provided by low-income students in New Orleans. We measure how their cortisol patterns change during high-stakes testing weeks relative to baseline weeks. We find that high-stakes testing is related to cortisol responses, and those responses are related to test performance. Those who responded most strongly, with either increases or decreases in cortisol, scored 0.40 standard deviations lower than expected on the high-stakes exam.

https://doi.org/10.1162/edfp_a_00306

1. INTRODUCTION

The results of high-stakes standardized tests determine course placement, graduation, and college admission for students, result in sanctions or rewards for schools, and inform education policy. There is substantial resistance to testing regimes, often predicated on the notion that students are "stressed" by tests.[1] Yet, to our knowledge, no evidence exists on test-induced physiological stress among K–12 students in a real-world setting.[2] Understanding variation in test-induced stress responses and implications for performance is important for determining whether scores on high-stakes tests are reliable measures of ability and knowledge, or if they are biased by "stress disparities" between children (see review in Heissel, Levy, and Adam 2017).

This study raises important questions about the use of high-stakes testing. We document how high-stakes testing affects low-income children's stress biology in one charter school network, and we show how changes in children's physiological responses to high-stakes tests relate to performance on the standardized test. Our goal in this study is to identify these patterns in one setting, and to call for more research into this area in the field. Expanding our understanding of the relationship between stress and test performance will affect our understanding of how high-stakes test results should be used and interpreted. Throughout this paper, our footnotes include suggestions for future research in this area.

We use saliva-based measures of cortisol—a primary stress hormone that indicates how the biological stress system is functioning—among low-income students in New Orleans to document how cortisol levels change in response to a high-stakes standardized test administered to students in grades 3–8, relative to a regular baseline school week. We call this change "cortisol reactivity." We find that students have 18 percent higher cortisol levels in the homeroom period just before taking the high-stakes test, relative to that same timeframe during weeks without testing. These differences are driven by boys, whose homeroom cortisol is 35 percent higher during testing weeks than regular weeks.[3]

Everyone has a natural cortisol rhythm over the course of the day (described in more detail in section 2). Acute stressors are associated with increases in cortisol above these natural rhythms. An increase in cortisol is not necessarily bad—in the best case, it can provide the energetic boost one needs to respond to a challenge with attention and focus. How an individual responds to a given stressor is based on what is adaptive in that person's particular context (Del Giudice, Ellis, and Shirtcliff 2011; Shirtcliff et al. 2014).

1. The Center for American Progress found that 49 percent of parents thought there was too much testing in schools (Lazarín 2014), and the New York Association of School Psychologists provides an overview of many reported parent concerns (Heiser et al. 2015). These concerns are not unfounded: grade 3–5 students reported higher anxiety and stress symptoms following No Child Left Behind-required testing, relative to lower-stakes classroom testing (Segool et al. 2013).
2. A variety of studies have examined cognitive tests in lab settings (e.g., Lupien et al. 2002; Stroud, Salovey, and Epel 2002) or with researcher-administered tests in schools that did not matter for student or school outcomes (e.g., Blair, Granger, and Razza 2005; Lindahl, Theorell, and Lindblad 2005). Though these studies provide evidence of potential responses, they do not include baseline, non-testing weeks in their analysis. Other studies have looked at adult responses in undergraduate and medical students (Malarkey et al. 1995; Weekes et al. 2006).
3. This is consistent with previous evidence that males show larger cortisol responses to achievement-related stressors than females (Stroud, Salovey, and Epel 2002; Weekes et al. 2006).

What is adaptive to a given individual may differ by various background characteristics, and it is not necessarily adaptive in the context of an academic test. Although our entire sample can be considered economically disadvantaged, we find suggestive evidence of differences in cortisol changes by level of disadvantage, with the largest cortisol effects for those living in high-poverty and high-crime neighborhoods.

We next examine whether differences in cortisol reactivity are associated with test performance on a subsample of students for whom we have test score data. Large increases in cortisol can make concentration difficult, while reduced cortisol may be a sign of disengagement with a task. We show that both increases and decreases in cortisol from the baseline week to the high-stakes testing week are associated with lower test scores on the high-stakes test, relative to how we would expect students to perform based on other in-school academic performance (i.e., grades).

Descriptive studies show that children from low socioeconomic status and racial/ethnic minority groups have lower average scores on standardized academic tests relative to high socioeconomic status and white families (Reardon 2011; Bradbury et al. 2015). Low socioeconomic status and racial/ethnic minority individuals are also more likely to be exposed to stressful life events relative to higher income or white individuals (see review in Hatch and Dohrenwend 2007). These patterns are correlated, but the physiological stress response may provide a link between them. In particular, students who experience chronic stress may respond differently to new stressors, such as high-stakes tests. Persistent socioeconomic gaps in academic performance could be due in part to different responses to the stress of testing disparately affecting test performance. This disparate effect, in turn, has implications for whether standardized tests are a fair means of evaluating student ability and school quality.

This study makes several contributions. For one, we document cortisol patterns for a low-income 7-to-15-year-old student population about which there is limited evidence. This is the first study to take cortisol samples from such young students during the timeframe surrounding high-stakes testing, and our experience provides guidance for researchers interested in measuring cortisol levels in similar populations. Second, we document how cortisol patterns change for this population in response to a stressful event. This is relevant to understanding how students respond to tests in their actual school settings. Third, we provide the first evidence on how differences in cortisol responses are related to performance on real-world standardized tests. This is crucial for understanding the validity of those tests themselves and the interpretation of individual differences in test results, which can have important real-world consequences. Our analysis of test performance is necessarily correlational—who has larger cortisol responses is not randomly assigned. Changes to cortisol could be attributed to other outside shocks (e.g., changes to family income) that are also correlated with test scores. Still, after conditioning on in-school performance (grades), demographics, and neighborhood characteristics, we find that large changes to cortisol are associated with worse outcomes on the test. We use this evidence as a call for more research into the relationship between stress and test performance, across a variety of settings.

This paper proceeds as follows: In section 2 we provide more background on the science of biological stress responses and the cortisol hormone. Section 3 describes our data. Section 4 describes our analytic strategy. Section 5 presents our results. Section 6 discusses the results and conclusions.

2. BACKGROUND ON BIOLOGICAL STRESS RESPONSES AND CORTISOL

Biological stress responses include multiple systems, but this paper focuses on the hypothalamic-pituitary-adrenal (HPA) axis and its primary hormonal product, cortisol. Cortisol levels show a strong circadian rhythm across the day, known as the diurnal cortisol rhythm, with the highest cortisol levels occurring shortly after waking and the lowest levels occurring about thirty minutes after sleep begins (see Gunnar and Quevedo 2007 for more details). Two key measures in cortisol research are the waking cortisol level and the daily cortisol slope (i.e., the rate at which cortisol levels drop from wake to bedtime). The cortisol awakening response (CAR), a sharp increase in cortisol thirty to forty minutes after waking, is an additional measure. The CAR provides an energetic boost to help individuals meet the expected demands of the upcoming day (see review in Clow et al. 2010).

Real or perceived stressors can increase cortisol above typical diurnal levels.[4] For routine stressors (e.g., momentary loneliness [Doane and Adam 2010]), cortisol levels return to their usual daily pattern approximately an hour after the stressor has passed. According to the adaptive calibration model, the stress response is generally adaptive; for instance, the HPA axis may mobilize psychological and physiological responses when presented with a stressor (Del Giudice, Ellis, and Shirtcliff 2011; Shirtcliff et al. 2014). One at-home study had twenty-four participants (aged 21–42 years) recruited from a university community provide hourly cortisol samples over a 48-hour period. Rising cortisol was associated with subsequent-hour increases in positive emotions such as activeness, alertness, and relaxation, and marginally significant decreases in nervousness (Hoyt et al. 2016).

Broadly, high or rising cortisol occurs when individuals are in personally relevant situations, are engaged with their environment, and are facing a difficult (but not impossible) task. Low or diminishing cortisol occurs if an individual is disengaged from the environment, a task is impossible, or a task is no longer novel.[5] The HPA axis can also be anticipatory, with rising cortisol levels before an expected stressful event or changes to the CAR if the prior day was particularly stressful.[6] In the context of high-stakes testing, we may expect moderately increased cortisol before the test, particularly if the student expects the test to be difficult but manageable, with stakes that matter for them. Limited (or lowered) cortisol responses to stressors may be related to

4. This pattern has been consistently demonstrated in the psychology and endocrinology literature (see reviews in Sapolsky, Romero, and Munck 2000; Miller, Chen, and Zhou 2007; Adam 2012).
5. The adaptive calibration model attempts to build a model of the development of stress responsivity in general (Del Giudice, Ellis, and Shirtcliff 2011), and Shirtcliff et al. (2014) specifically focus on the cost/benefit of cortisol responsivity in individuals' particular contexts. This latter model specifically argues against the popular notion of cortisol as detrimental to health and well-being, and instead argues that cortisol responses can be beneficial in certain contexts. A large meta-analysis of 208 studies found that stressors that were uncontrollable or had a social-evaluative component (meaning that performance could be negatively judged by others) led to the largest increase in cortisol in laboratory settings (Dickerson and Kemeny 2004).
6. See Engert et al. (2013) for a summary of anticipatory cortisol in lab-based settings. The effect has also been demonstrated in the field: for instance, seventeen young men set to participate in a judo competition had higher cortisol on the day of the competition (but before the competition began) than at the same time on non-competition days (Salvador et al. 2003). For the CAR, Doane and Adam (2010) found that prior-day loneliness (a stressful experience) was associated with higher next-day cortisol in young adults; similarly, Heissel et al. (2018) demonstrated that nearby violent crime is associated with a larger CAR the following day in a sample of adolescents in a large Midwestern city, perhaps as the body anticipates a more stressful day ahead.

disengagement or "shutting down" in the face of the test; large increases in cortisol may reflect feeling threatened or overwhelmed in a way that is likely to prevent productive focus.

Stress patterns also differ by sex. Females' CARs tend to peak later in the day than males' CARs (Stalder et al. 2016). Moreover, males may be more responsive to achievement-related stressors, whereas females may be more responsive to social rejection (Stroud, Salovey, and Epel 2002). A meta-analysis of twenty-eight studies similarly found larger cortisol responses to stressors in males than females (Sauro, Jorgensen, and Pedlow 2003). In the context of high-stakes testing, we may then expect larger cortisol responses to high-stakes testing from male students.

Of particular concern in this context, long-term stress exposure can lead to changes in the HPA axis that can be maladaptive in some contexts, including school. For instance, hypocortisolism is a condition that can follow a period of chronic stress, wherein the HPA axis shows low levels of cortisol and no longer responds to stressors (see summaries in McEwen 1998; McEwen and Gianaros 2010). This is one reason we might expect that children with high-stress backgrounds respond less-optimally (physiologically) to a high-stakes test. However, our results are more consistent with a story that chronic stress is associated with *high* cortisol reactivity in this population.

HPA axis activity may affect cognitive performance during test-taking by affecting memory recall. Associations between cortisol and memory recall generally displays an inverse-U pattern in laboratory-based studies.[7] In particular, inducing large increases *or* decreases in cortisol results in worse memory recall. If cortisol and memory recall are related, then differences in stress response may lead to different test outcomes even among students with equal ability who have learned the same amount of material. If the students most likely to be "stressed testers" come from already-disadvantaged backgrounds, this pattern may exacerbate the observed achievement gaps on high-stakes tests.

Two previous studies compared a baseline week of normal activity against a stressful testing week. Weekes et al. (2006) found that male undergraduate students had an increase in examination-week cortisol levels, while female undergraduates did not. The authors found no link between psychological (self-reported) stress and physiological stress as measured by cortisol. In contrast, Malarkey et al. (1995) collected cortisol and other measures on medical students one month before, during, and two weeks after examinations. They found increases in cortisol during the test week but only for those students who perceived the test as stressful. Neither set of authors examined performance on the tests and its relationship to cortisol.

Other research has not included baseline stress levels but instead examined same-day changes in cortisol in response to stressors. Perceiving a researcher-administered

7. When cortisol is administered synthetically before a lab-based memory assessment, humans generally have worse memory recall, relative to participants who did not receive a dose of synthetic cortisol (see review in Het, Ramlow, and Wolf 2005). However, randomly varying the levels of synthetically administered cortisol (from 0 to 24 mg) across participants was associated with an inverse-U shaped pattern, with the best memory recall at moderate elevations (Schilling et al. 2013). Another study pharmacologically decreased cortisol levels, then restored baseline cortisol levels with hydrocortisone replacement treatment, for treated participants. The researchers tested memory function after each manipulation, finding impaired recall after the induced cortisol decrease. Subsequent hydrocortisone replacement restored memory recall to the placebo level (Lupien et al. 2002).

test during the school day as stressful was correlated with higher same-day cortisol and lower test performance in Swedish adolescents (Lindahl, Theorell, and Lindblad 2005). Conversely, among young, low-income children in a Head Start program, having a larger same-day cortisol response to a stressor was correlated with better cognition and behavioral outcomes than those without a cortisol response (Blair, Granger, and Razza 2005). Adults with higher anxiety had larger increases in cortisol in response to performance tasks than those who did not (Malarkey et al. 1995; Schlotz et al. 2006). Whether cortisol improves or detracts from performance may depend on anxiety about the task at hand (Mattarella-Micke et al. 2011).

Overall, the relationships between perceived stress, stress hormones, and performance on a task are complicated and related to a wide variety of background characteristics. These relationships highlight the importance of accounting for baseline differences in cortisol patterns for individual students: Do students perform poorly because of elevated cortisol, or do the students who perform poorly in general also tend to have high cortisol levels in regular, non-tested weeks? In addition, it is not obvious that a real-world high-stakes test will lead to a physiological reaction in a group of young, low-income students. If reactions do occur, it is not obvious who would be most affected, or how such reactions might correspond to performance on the test. This study contributes to our understanding of these dynamics by measuring how cortisol changes in response to a high-stakes test for grade-school students from disadvantaged backgrounds.

3. DATA

Our data consist of cortisol measures, student diaries, and administrative data on student demographics and academic performance, for students from a charter school network in New Orleans. Descriptive statistics are in table 1. The participants were almost all black (95 percent), economically disadvantaged (97 percent),[8] and from high-poverty neighborhoods (with 40 percent of block group households in poverty, mean block group income of $27,000, and mean block group unemployment of 13 percent). The households were also in neighborhoods with a great deal of police activity, with a mean of 416 high-priority emergency (911) calls within a quarter-mile of their home in the prior year. However, these averages mask heterogeneity: The fraction of neighborhood households in poverty ranged from 14 to 91 percent, mean neighborhood incomes ranged from $9,000 to $58,000, neighborhood unemployment rates ranged from 0 to 74 percent, and the number of nearby high-priority 911 calls ranged from 0 to 1,380 in the prior year. On average, the participants are disadvantaged relative to the overall population, but there is significant variation within the sample.[9]

Cortisol Data

We collected salivary cortisol samples from ninety-three pre-adolescent and adolescent volunteers in grades 3–8, across three schools from the charter school network. We

8. Economic disadvantage is indicated by eligibility for free or reduced-price lunch.
9. The median household income in the United States was $57,000 in 2015, with 13.5 percent of households in poverty (Proctor, Semega, and Kollar 2016). New Orleans had a median household income of $39,000, with nearly 25 percent of households in poverty in this period; the mean of $27,000 income in our sample is similar to the $26,000 median black family income in New Orleans (Litten 2016).

Table 1. Descriptive Statistics

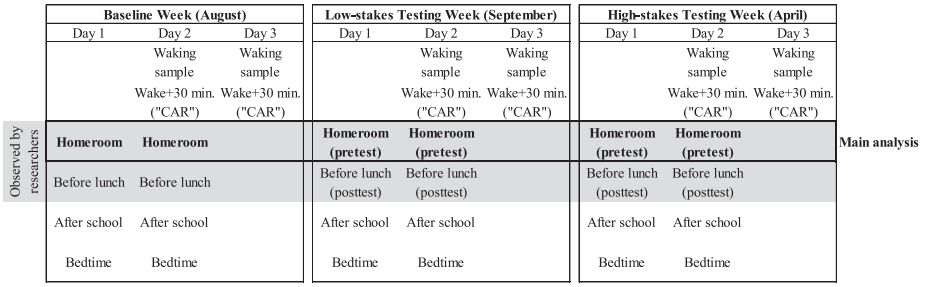
	Mean (1)	SD (2)	Min (3)	Max (4)	Count (5)
Grade	5.77	1.84	3.00	8.00	93
Age (fall 2015)	11.59	2.06	7.90	15.60	93
Female	0.55	0.50	0.00	1.00	93
Limited English proficiency	0.03	0.18	0.00	1.00	93
Exceptional child	0.13	0.34	0.00	1.00	92
Gifted	0.03	0.18	0.00	1.00	92
Black	0.95	0.23	0.00	1.00	93
Economically disadvantaged	0.97	0.16	0.00	1.00	84
Section 504 plan	0.29	0.45	0.00	1.00	84
McKinney-Vento Act	0.08	0.28	0.00	1.00	84
Priority 1 911 calls within 0.1 mi of home	83.94	73.99	0.00	351.00	85
Priority 1 911 calls within 0.25 mi of home	415.81	311.47	0.00	1,380.00	85
Neighborhood fraction houses in poverty	0.40	0.17	0.14	0.91	86
Neighborhood median income	26,830	11,246	9,327	58,194	80
Neighborhood fraction unemployed	0.13	0.11	0.00	0.74	86
N	93				

Notes: Section 504 is a civil rights law that prohibits discrimination against individuals with disabilities. Section 504 ensures that the child with a disability has equal access to an education. The child may receive accommodations and modifications. The McKinney-Vento Education of Homeless Children and Youth Assistance Act is a federal law that ensures immediate enrollment and educational stability for homeless children and youth. McKinney-Vento provides federal funding to states for the purpose of supporting district programs that serve homeless students. SD = standard deviation.

recruited participants through flyers distributed by their school, obtained parental consent and participant assent, and briefed participants on the protocol during homeroom on their first day of collection. Some participants joined the study late ($N = 13$ joined in week 2) and were briefed on the protocol individually. These students were mainly from school 2.[10]

To provide the samples, participants let saliva collect in their mouth, then used a small straw to drain the saliva into a small vial; this is called the passive drool technique. Participants watched a saliva sample demonstration at the first collection, had a video demonstration available, and received reminder texts from the research team during the data collection to ensure that they followed protocol. Participants were instructed to avoid eating, drinking, and brushing their teeth 30 minutes prior to each sample collection. A kitchen timer preset to 30 minutes was provided to aid in the timing of sample 2. Participants were instructed to refrigerate their home samples as soon as possible after collection and return their home samples to the research team in homeroom every day.

10. Based on a linear probability model regressing an indicator for joining in week 2 on the demographic characteristics of the ninety-three students, those with a 10 percentage point higher in-school science grade were 12.2 percentage points more likely to join in week 2 (relative to week 1), every one year increase in age was associated with a 10.8 percentage point decrease of joining in week 2, and being in school 2 was associated with a 58.8 percentage point increase in joining in week 2. Many students in school 2 mistook our real cash incentives as the points-based behavioral "dollars" the school used; they decided to join the study upon learning that the compensation was in fact real dollars.



Note: Testing occurred between homeroom and lunch in weeks 2 and 3. CAR = cortisol awakening response.

Figure 1. Research Design by Sample, Day, and Week

Figure 1 displays the experimental design. Saliva sample collection occurred during three weeks of the 2015–16 academic year: a baseline week (no testing; late August), a low-stakes testing week (internal school testing; early September); and a high-stakes testing week (statewide testing; late April). During each week, participants provided saliva samples at six points over a 24-hour period: at wake (sample 1), 30 minutes after wake (to capture the CAR; sample 2), during homeroom[11] (sample 3), before lunch (sample 4), after school (sample 5), and at bedtime (sample 6). Data were collected over a 48-hour period each week, beginning in homeroom on the first day, such that day 1 included samples 3–6, day 2 included samples 1–6, and day 3 included samples 1 and 2.[12] Low- and high-stakes testing (during testing weeks) occurred just after homeroom and ended before lunch. Homeroom (before the test, sample 3) and before-lunch (after the test, sample 4) saliva samples were collected under the supervision of the research team, and timing was verified by the team.[13] Sample 3 had the most consistent timing and the highest completion rate across days and is the focus of the majority of our analysis.

We dropped week 1 samples for two students who had extremely high cortisol levels (mean = 15.43 μg/dL and 3.50 μg/dL, relative to the overall mean of 0.15 μg/dL), likely indicating that they were taking cortisol-containing medication. Per typical protocol for cortisol, we top-code cortisol levels to 1.80 μg/dL.[14] Online appendix table A.1 contains descriptive statistics for the homeroom cortisol samples by school.

11. Homeroom started at 7:00 a.m. in school 1 and 8:00 a.m. in schools 2 and 3.
12. Seventy-eight percent of individual-week-sample number combinations had at least one sample, although completion rates varied by sample (see figure A.1, which is available in a separate online appendix that can be accessed on *Education Finance and Policy*'s Web site at https://doi.org/10.1162/edfp_a_00306). Samples were stored at −20°C before shipment to Trier, Germany, where they were assayed in duplicate using time-resolved fluorescent-detection immunoassay (Dressendörfer et al. 1992).
13. Samples 1, 2, 5, and 6 were taken out of school, and timing was reported by students and verified against diary entries. The in-school compliance rate was 89 percent; out-of-school compliance was 72 percent. Future researchers working in this area could consider focusing only on in-school sampling in order to ensure compliance and reduce costs. Changes in school lunch scheduling meant that sample 4 timing had a wider variance than sample 3. Figure A.2 in the online appendix displays the distribution of sample timing by sample and school.
14. Online appendix figure A.3 displays an exercise using alternative means of limiting potentially contaminated samples, from including all available cortisol to dropping anything above the 90th percentile.

Student Diaries

Participants filled out diaries at each saliva collection. The sample 1 diary included questions about the prior night's sleep and that morning's wake time. We coded wake time for each day as the minimum reported timing across the sample 1 cortisol sample, the sample 1 diary entry, and diary-reported daily waking time. Students took sample 1 on days 2 and 3; by design day 1 did not have a reported wake time. If students were missing the wake time measure, we imputed it using the mean wake time by individual by week, then (if still missing) the mean wake time by individual, then (if still missing) the mean wake time by school by week.

We calculated time since wake for each sample as the length of time between that day's wake time and the reported timing of the cortisol sample. If missing sample timing, we imputed it using that sample's diary time, then (if still missing) the mean of the sample timing by individual by sample number, then (if still missing) the mean of the sampling timing by school by sample number.

Administrative Data

The charter network provided administrative data including participants' scores on low-stakes math, science, English Language Arts (ELA), and social studies tests and high-stakes math, science, and ELA tests. The administrative data also included in-school grades (on a 0–100 scale) for each academic quarter in math, science, ELA, and social studies. In the test score analysis, we dropped students' missing test score data or missing cortisol data in the baseline week or high-stakes testing week, leaving us with $N = 68$ students in the test score subsample. We lose fifteen students who did not have baseline week cortisol (thirteen who joined in week 2 and two who did not have homeroom-specific data), eight students who appear to have moved out of the charter network,[15] and two with missing test score data (but with cortisol data, meaning they were in school at least part of the week). We used a linear probability model to regress an indicator for being in the final test score sample on the demographic characteristics and in-school grades for the ninety-three participants. Those with 10-point higher science grades had a 13.1-percentage point higher probability of being in the final sample; all other demographic characteristics were not statistically related to being in the final test score sample.

We converted each test score into standardized Z-score units by grade; the resulting scores should thus be interpreted as the distance from the average score in standard deviations.[16]

The results of the high-stakes test in our study mattered for the school, as they contributed to the letter grade (A–F) rating given to the school by Louisiana's Department of Education. However, the test had no direct repercussions for individual students. In addition, the students took a variety of other tests in their school system throughout the year, including a series of tests that were only used for internal assessment but

15. The students who appeared to move out of the school system did not have cortisol data in the high-stakes testing week, test score data, or quarter 4 grades. In robustness testing, we examine how changes in cortisol relate to performance on the low-stakes test, which was only two weeks after the baseline week and thus had less time for students to switch schools.

16. Z-scores are calculated by subtracting the mean score on a given test in a given grade from the individual's score on that test, then dividing by the standard deviation of that test in that grade.

mimicked the structure of the year-end high-stakes test.[17] Given how often these students were tested, we might expect them to be so accustomed to the process that even high-stakes tests would not be perceived as stressful. This will reduce the likelihood of finding any effect of testing on cortisol responses.

4. ANALYTIC STRATEGY

Given the greater control the research team had over the before-test homeroom sample collection, and because our main objective is understanding the high-stakes testing period, we focus most of our attention on sample 3, taken in the homeroom period just before the test was administered. This time period is particularly important given that it reflects the level of cortisol that students bring into the test setting. Our first analysis examines whether the level of cortisol in the homeroom period changed from baseline to the low-stakes and high-stakes testing weeks with the following specification:

$$\ln(cortisol_{iwd}) = \beta_0 + \beta_1 LowStakes_w + \beta_2 HighStakes_w + \beta_3 Time_{iwd} + \beta_4 Time^2_{iwd} + \beta_5 Waketime_{iwd} + \beta_6 CAR_{iwd} + \gamma_i + \varepsilon_{iwd}, \quad (1)$$

where $LowStakes_w$ is equal to 1 in the low-stakes testing week and zero otherwise, $HighStakes_w$ is equal to 1 in the high-stakes PARCC testing week and zero otherwise, $Time_{iwd}$ is time of the sample relative to the end of homeroom for individual i in week w on day d, $Waketime_{iwd}$ is that day's approximate wake time for the individual (measured in hours relative to midnight), and CAR_{iwd} is an indicator for whether the homeroom sample was 15–60 minutes after the individual's wake time that day. A control for CAR may be necessary if a student woke up late relative to school start and took their homeroom sample 15–60 minutes post-waking. We control for a quadratic of $Time_{iwd}$ because the level of cortisol falls at a decreasing rate throughout the day; not including the quadratic does not change the results. The individual fixed effects γ_i account for any observed and unobserved factors that are constant across an individual over time (e.g., sex, intelligence, personality, constant health) and allows us to isolate within-student changes in cortisol from week to week. Standard errors are clustered at the individual level. The analysis indicates whether, holding other individual-specific factors constant, homeroom cortisol levels change from baseline to the testing weeks.[18]

Supplementary analyses test for variation based on proxies for chronic stress—specifically, poverty rates and crime rates in students' neighborhoods. We might expect students' responsiveness to the stress of the test to differ if they are chronically stressed. We also tested for differences by gender.

Finally, we examined whether cortisol reactivity to high-stakes testing was associated with performance on the high-stakes test. We controlled for participant

17. This amount of testing is not atypical; for instance, Chicago Public Schools had a testing schedule comparable to our charter school network (Chicago Public Schools 2020).

18. One concern could be that seasonality in cortisol levels could lead us to falsely attribute changes in cortisol to the test. Alternatively, cortisol sampling itself could be stressful, but habituation could occur as individuals take more cortisol samples. If anything, prior research suggests that both habituation and seasonality in cortisol would work against finding increased cortisol, as both more sampling exposure and springtime are associated with lower cortisol levels than less sampling exposure and fall, respectively (King et al. 2000). We could not test this theory under a schedule that worked for our charter school network, though future studies should consider adding an additional "control" sample just before or after the high-stakes testing period.

demographics, academic grades in the school in the first three quarters of the year, sample timing, and school characteristics. Estimated effects on high-stakes test performance can be interpreted as differences relative to how we would expect participants to perform based on their academic performance in daily school settings. We estimate the following model:

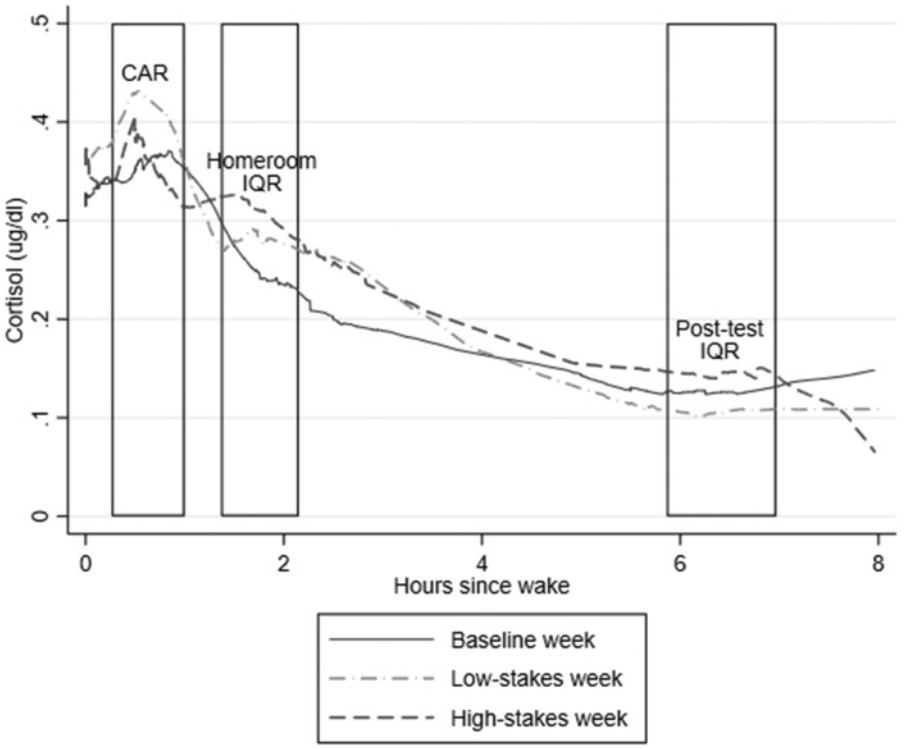
$$TestZScore_i = \beta_0 + \boldsymbol{Responsivity}_i\gamma + \beta_1 CurrentCortisol_i + \beta_2 CurrentCortisol_i^2 + \boldsymbol{T}_i\alpha + \boldsymbol{X}_i\delta + \varepsilon_i, \quad (2)$$

where $TestZScore_i$ is the average Z-score of the math, science, and ELA high-stakes tests; $CurrentCortisol_i$ is the mean individual homeroom cortisol in the high-stakes testing week; $CurrentCortisol_i^2$ allows the marginal effect of cortisol to change as the cortisol level increases; $\boldsymbol{T}_i$ is a vector of grades in school (on a 0–100 scale) in academic quarters 1–3 in math, science, ELA, and social studies; and $\boldsymbol{X}_i$ is a vector of other individual characteristics from school administrative data (age, gender, exceptional child status, whether the student had a Section 504 plan, homelessness, and school controls).

The primary variable of interest is $\boldsymbol{Responsivity}_i$, which is a vector of indicator variables representing 20 percentage-point bins for the change in homeroom cortisol levels from baseline to the high-stakes testing week. Bins are grouped as follows: −30 percent or lower, −10 to −30 percent, −10 to 10 percent (the reference bin), 10 percent to 30 percent, 30 percent to 50 percent, 50 percent to 70 percent, and 70 percent or higher. We will show that alternative bin cutoffs lead to qualitatively similar conclusions. Statistically significant coefficients for $CurrentCortisol_i$ would indicate that the same-day level of cortisol is related to test performance; statistically significant coefficients for $\boldsymbol{Responsivity}_i$ would indicate that the change in cortisol level from baseline to the test week is related to performance.

The vector of in-school grades accounts for regular, non-high-stakes student performance, and any observed effects of responsivity or current cortisol would indicate underperformance or overperformance on the test beyond what is predicted by those test scores and demographic factors. To the extent that some demographics (e.g., homelessness) themselves cause chronic stress that in turn could affect cortisol responses, our main estimates could underestimate the effect of cortisol responsivity.[19] Still, other unobserved shocks to individuals (e.g., parental job loss) could conceivably be related to both changes in cortisol and test performance.[20] Moreover, some participants were missing either requisite cortisol or testing data, and the *N* in the final analysis is 68.[21] Given potential omitted variable bias and this smaller sample, we interpret the academic performance results as suggestive and do not conduct subgroup analyses.

19. There is no statistical difference in our estimates if we do not include the demographic controls in this analysis; we include them for completeness.
20. In future research efforts, an additional cortisol data collection closer to the time of the high-stakes test could reduce this concern. Practically, one challenge researchers may face is that a lot of testing (both high- and low-stakes) occurs in the spring, so it may be difficult to find a "control" week near the high-stakes testing week. Moreover, each sample collection is a burden on the school and students, and schools may be hesitant to allow too much access during the spring testing period.
21. Results were similar when we imputed responsivity for those missing baseline cortisol measures, using the change in cortisol from the low-stakes to the high-stakes testing week.



Notes: We use locally weighted scatter plot smoothing to display the data, which does not impose parameters on the pattern. Boxes include the cortisol awakening response (CAR, 15–60 minutes post-waking) and the interquartile range (IQR) of timing for the before-test (homeroom) and post-test (before lunch) samples. $N = 93$ individuals included over multiple days.

Figure 2. Cortisol Patterns from Wake to Eight Hours Post-Wake for Baseline, Low-Stakes, and High-Stakes Weeks

5. RESULTS

Changes in Cortisol Daily Rhythms

Figure 2 displays the cortisol patterns from wake to eight hours post-wake for baseline, low-stakes, and high-stakes weeks using locally weighted scatter plot smoothing, which does not impose parameters on the pattern. Cortisol followed the expected diurnal pattern in the baseline week. We see the sharp rise in the cortisol awakening response (15–60 minutes after waking), following by falling cortisol as time passes.

The pattern visibly differs in the high-stakes testing week, with a less-pronounced CAR and much higher levels of cortisol during the homeroom period. In the baseline week, cortisol levels were not elevated above the expected slope during homeroom, which provides an important test on our hypothesis: We would not expect elevated cortisol during homeroom in a regular school week. Cortisol elevations during the low-stakes test week were in between the baseline and high-stakes test weeks in the homeroom period.

Changes in Before-Test Cortisol

The estimates in table 2 show whether, within individuals, homeroom cortisol levels differed from baseline to the testing weeks. All columns include individual fixed effects

Table 2. Changes in Level of Before-Testing Homeroom Period Cortisol by Week

	All (1)	All (2)	By Gender (3)	By Poverty (4)	By Local 911 Calls (5)	By Ability (6)
Low-stakes testing	0.123	0.101	0.310**	0.069	0.267*	0.269**
	(0.087)	(0.087)	(0.113)	(0.156)	(0.118)	(0.091)
High-stakes testing	0.204**	0.176*	0.327**	0.295*	0.320**	0.273**
	(0.075)	(0.076)	(0.119)	(0.123)	(0.121)	(0.092)
Low-stakes × female			−0.353*			
			(0.166)			
High-stakes × female			−0.259+			
			(0.150)			
Low-stakes × lower poverty				0.018		
				(0.188)		
High-stakes × lower poverty				−0.240		
				(0.164)		
Low-stakes × lower crime					−0.369+	
					(0.196)	
High-stakes × lower crime					−0.266	
					(0.171)	
Low-stakes × higher ability						−0.329*
						(0.163)
High-stakes × higher ability						−0.190
						(0.142)
Time of day	−0.149	0.034	0.038	0.133	0.261	0.047
	(0.612)	(0.644)	(0.646)	(0.713)	(0.721)	(0.642)
Time of day-squared	0.150	0.349	0.345	0.409	0.445	0.345
	(0.619)	(0.682)	(0.689)	(0.749)	(0.750)	(0.678)
Wake time		−0.183	−0.192	−0.184	−0.191	−0.178
		(0.139)	(0.136)	(0.138)	(0.136)	(0.141)
CAR timeframe		−0.101	−0.085	−0.151	−0.118	−0.077
		(0.175)	(0.175)	(0.194)	(0.191)	(0.175)
p(sum low-stakes testing = 0)			0.721	0.404	0.487	0.668
p(sum high-stakes testing sum = 0)			0.469	0.610	0.644	0.467
Observations	489	489	489	454	448	489
Participants	93	93	93	86	85	93

Notes: Robust standard errors clustered by student identification. Analysis conducted at the student-day level. Outcome is the natural log of cortisol. Data come from saliva collected in homeroom. Each column represents a different regression estimate. Model limits the comparison to within-individuals, accounting for any constant observed and unobserved characteristics. Wake time is the approximate wakeup time for the day, measured with error. Column 2 is the preferred overall model. Columns 3–6 conduct the analysis by interacting the test with indicator variables for the given group. Column 4 is separated by median neighborhood poverty level (40 percent), column 5 is separated by median number of 911 calls within 0.25 of home address within a year (median = 240 calls), and column 6 is separated by median first quarter grades expressed in Z-scores (median = −0.15 standard deviations). Table includes *p*-values of the estimated difference in these groups for the change in cortisol for the low- and high-stakes weeks.

$^{+}p < 0.10$; $^{*}p < 0.05$; $^{**}p < 0.01$.

and a quadratic of time relative to waking for a given day. The coefficient on low-stakes (high-stakes) testing approximates whether the level of cortisol differs from the baseline week to the low-stakes (high-stakes) testing week. Column 1 does not include wake time or controls for whether the sample was taken during the CAR period (15-60 minutes post-wake), as these were necessarily imputed on day 1 of each week. However, later wake times were associated with higher waking cortisol, so we added controls for wake time and CAR in column 2.[22] The homeroom estimates were similar whether

22. For reference, online appendix table A.2 displays the analysis without fixed effects but with demographics controls, as well as post-double-selection LASSO methods (Belloni, Chernozhukov, and Hansen 2014) to select

controlling for these variables or not, and going forward we prefer the more conservative estimate that controls for wake time and CAR. On average, student levels of cortisol were 18 percent higher in homeroom in the high-stakes week relative to the same students' homeroom cortisol at baseline. There was no statistical difference in cortisol in the low-stakes week relative to the baseline week, though as expected the coefficients are positive. The high-stakes and low-stakes weeks do not statistically differ from one another; future analyses should explore this relationship further.[23]

Columns 3 through 6 of table 2 examine subgroups. All estimates within a column come from the same regression, and the bottom rows of the table test whether the sum of the main effect and the interaction for a given test differs from zero. Male students had large average increases in homeroom cortisol in the low-stakes testing week (31 percent) and high-stakes testing week (33 percent), relative to the baseline week. The female effect sizes statistically differed from the male estimates; the difference relative to baseline was −4 percent in the low-stakes week (calculated as 0.310–0.353) and +7 percent in the high-stakes week. Neither of the female estimates statistically differed from zero, as indicated by the *p*-values at the bottom of the table. Turning to neighborhood characteristics, we first divide individuals by the median neighborhood poverty rate observed in our sample (40 percent); lower-poverty neighborhood ranged from 14 percent to 40 percent (with a mean of 28 percent) and higher-poverty neighborhoods ranged from 41 percent to 91 percent (with a mean of 53 percent). Those from higher-poverty neighborhoods had larger average increases in homeroom cortisol than those from lower-poverty neighborhoods in the high-stakes week (30 percent versus 6 percent), relative to baseline, though the difference between groups was not statistically significant. Similarly, those from neighborhoods with an above-median number of high-priority 911 calls within 0.25 miles (median = about 340 calls) had larger average increases in homeroom cortisol levels than those from below-median neighborhoods in the high-stakes week (32 percent versus 5 percent).[24] The difference between groups was again not statistically significant. While the stress responses of those facing chronic stress do differ from their less-stressed peers, we do not find evidence of pervasive hypocortisolism (the lack of ability to respond to a stressor), per se; indeed, those participants have moderately larger increases in cortisol. Note that our sample size is a bit smaller in the neighborhood analysis due to missing or difficult-to-geocode addresses.

Our final subgroup analysis examines changes by student ability, based on median first-observed-quarter grades expressed in Z-scores (median = −0.15 standard deviation).[25] Increases in cortisol were driven by lower-achieving students; there is no average change for above-median students.

a set of controls to avoid over-fitting but minimize omitted variable bias. All models indicate broadly similar results.

23. In a larger sample, it would be useful to know if students acclimate to low-stakes testing, high-stakes testing, both, or neither. We only had one school with students in grades 3–4 and two with students in grades 6–8, which means we cannot separate out school-specific effects from age- or grade-specific effects.
24. Lower-crime addresses ranged from 0 to 338 high-priority calls within 0.25 miles in a year (with a mean of 191 calls) and higher-crime addresses ranged from 343 to 1,380 annual calls (with a mean of 645 calls).
25. Scores in the lower-achieving group ranged from −1.55 to −0.15 standard deviations from the mean (mean of this group = ₋0.80 standard deviation), while the higher-achieving group ranged from −0.11 to 2.14 standard deviation (with a mean of 0.82 standard deviation).

There was higher cortisol before the test, on average, relative to the baseline week, but there was also substantial variation in reactivity. Figure 3 displays the density of the change ("responsivity") from baseline to the low-stakes testing week and from baseline to the high-stakes testing week. Although, on average, cortisol was higher in testing weeks, some individuals had little change and others actually had lower cortisol in the testing weeks—either due to the noisiness of the cortisol sampling or perhaps due to disengagement from the stressful situation. We next test whether these different responses were associated with different performances on the test.

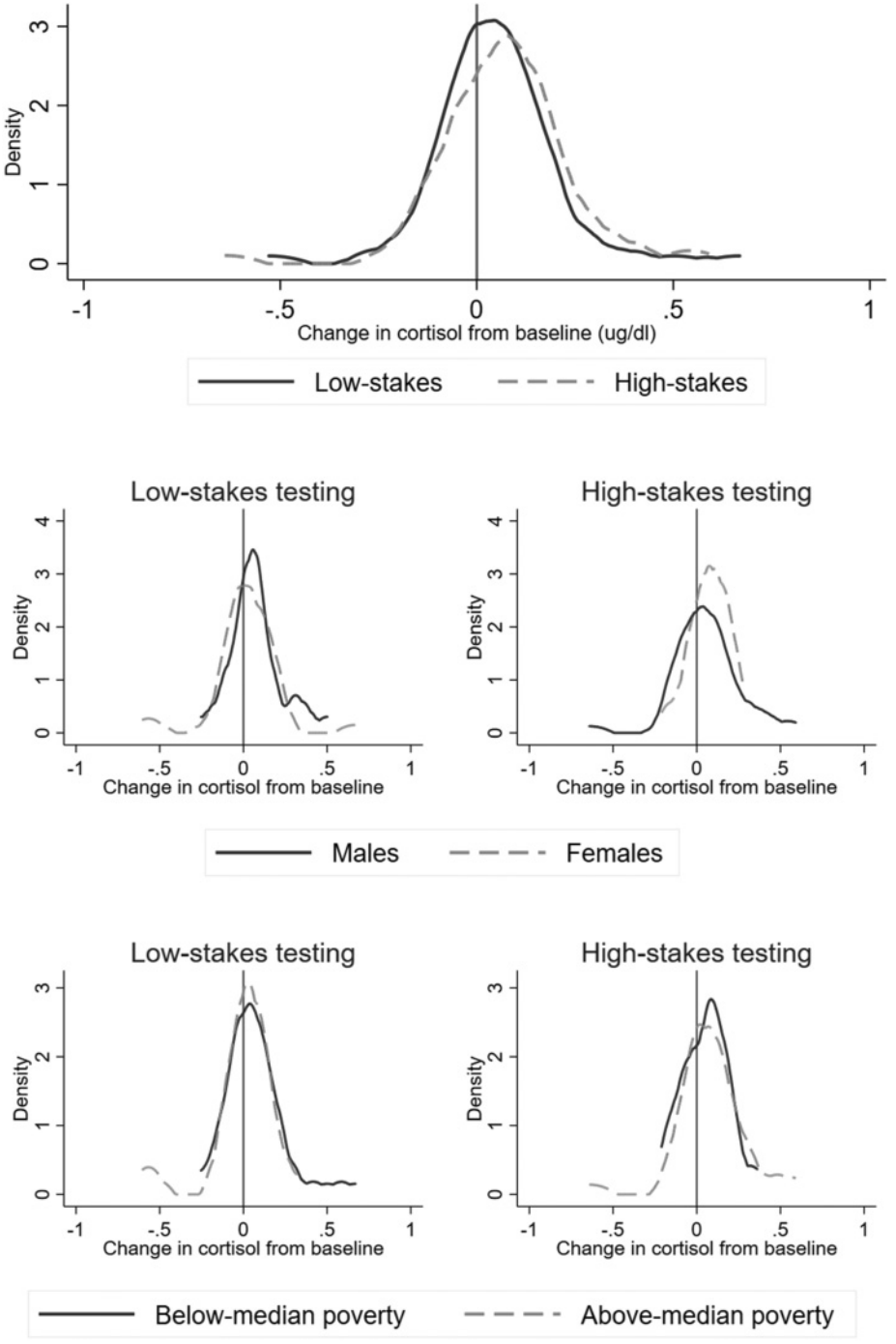
Differences in Academic Outcomes

Figure 4 examines how cortisol reactivity was related to test outcomes. It is unclear how best to measure this relationship, and we include several approaches for transparency. First, panel A breaks the subgroup of participants with the requisite data into quintiles based on the percentage change in their homeroom cortisol from baseline to the high-stakes testing week. Quintile 1, the reference group, includes those whose cortisol fell 22 to 78 percent relative to baseline during high-stakes testing. Quintile 2 includes those with little change, ranging from −21 percent to +12 percent. Quintile 3 participants had moderate increases, from 13 percent to 52 percent. The final two quintiles cover those with large increases, from 53 to 119 percent in quintile 4 and over 119 percent increases in quintile 5.

Quintile 2 differs significantly from quintile 1, with test scores 0.38 standard deviation higher (conditional on demographic controls, concurrent cortisol, and in-school grades; p-value = 0.027). The final three quintiles do not differ significantly from quintile 1. We reject that the five quintiles are the same at the 10 percent level with an F-test ($F(4, 42) = 2.15$; p-value = 0.091), but we cannot reject that quintiles 1, 3, 4, and 5 are statistically the same ($F(3, 42) = 0.87$; p-value = 0.465). In other words, it appears that participants in quintile 2, who have the least amount of change from baseline to the high-stakes test, outperform the other quintiles, conditional on the other control variables (although it is only marginally statistically significant).

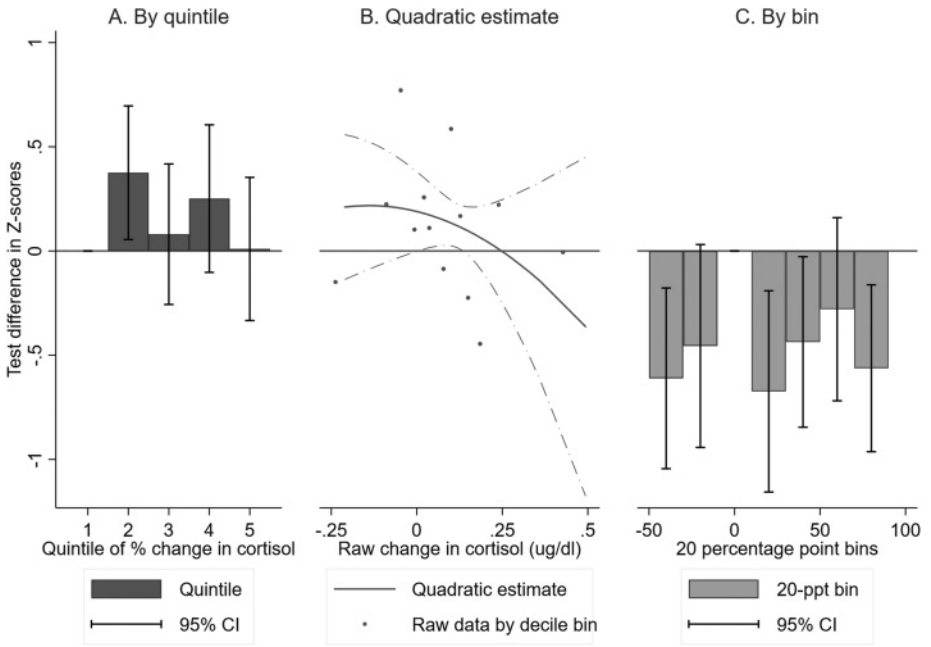
An alternative, parametric approach to the estimate is displayed in panel B. Prior research has found an inverse-U shape in the relationship between cortisol and outcomes (Het, Ramlow, and Wolf 2005; Schilling et al. 2013). Here, conditional on demographic controls and in-school grades, we model a quadratic estimate of the relationship between test score (the outcome) and the raw level of change in cortisol in micrograms per deciliter.[26] Panel B plots this estimate and its 95 percent confidence interval, as well as binned scatter plots for test scores and change in cortisol (with five to six observations per bin).[27] The pattern appears as an inverse-U, but contrary to prior work, we find no evidence of an improvement in outcomes for moderate increases in cortisol. Moreover, the quadratic term is not statistically significant, at least at our level of power.[28]

26. The model also includes a quadratic control for concurrent cortisol, but we find no relationship between concurrent cortisol and outcomes on the test.
27. Predicted line and 95 percent confidence interval estimated using Stata's *adjust* command, based on the mean of all other control variables.
28. The quadratic term for responsivity is $\beta = -1.446$; p-value = 0.132. The linear estimate is null in this model, and it is actually slightly negative ($\beta = -0.405$; p-value = 0.504). We also tested cubic and quartic functions but they were also not statistically significant. Future tests with more power may confirm the inverse-U pattern.



Note: Figure includes estimates by gender and poverty, with above-median poverty indicating more poverty.

Figure 3. Distribution of the Change ("Responsivity") from Baseline to the Low-Stakes Testing Week and from Baseline to the High-Stakes Testing Week



Notes: Models regressed mean Z-score on different ways to measure of change in cortisol. All models control for quarters 1–3 grades for math, English Language Arts, science, and social studies; time of day; time-squared; age; indicators for female, exceptional child status, Section 504 status, and homelessness; and school indicator variables. $N = 68$ individuals. Analysis conducted at the student level. Results are similar when imputing cortisol changes for those missing baseline data. Panel A groups participants by quintile based on their percentage change in cortisol, in quintile 1 (−78% to −22%, $N = 13$), quintile 2 (−21% to +12%, $N = 13$), quintile 3 (+13 to +52%, $N = 14$), quintile 4 (+53% to +119%, $N = 14$), and quintile 5 (>119%, $N = 13$). The model also controls for five quintiles of concurrent (in the high-stakes week) cortisol. Panel B does not group participants into categories, but instead includes variables for responsivity and responsivity-squared term to measure whether an inverse-U pattern occurs. Displayed solid line maps this predicted *Z*-score; dashed lines provide 95% confidence intervals, based on predicted values and standard errors from Stata's *adjust* command. The model also controls for concurrent cortisol and concurrent cortisol-squared, as well as the typical demographics. Display includes a binned scatter plot of the raw data, with 5–6 observations per bin. Panel C groups participants by percentage-change in cortisol. Bins grouped by decreases greater than 30% ($N = 10$), −30% to −10% ($N = 5$), reference group at −10% to +10% ($N = 8$), +10% to +30% ($N = 8$), +30% to +50% ($N = 7$), 50% to 70% ($N = 8$), and increases greater than 70% ($N = 21$).

Figure 4. Change in Predicted Mean Z-Score (across Math, Science, and English Language Arts Tests) on the High-Stakes Test by Change in Cortisol from the Baseline to the High-Stakes Testing Week

Finally, our preferred model groups the estimates into 20-percentage point bins. The estimates are somewhat noisy, but relative to those in the low reactivity group (from −10 percent to +10 percent homeroom cortisol change from baseline to the high-stakes week), those with either large increases or decreases in cortisol from the baseline week performed worse on the standardized test. In other words, decreases *and* increases in cortisol were associated with underperformance on the high-stakes test. Grouping the "change" bins together, an increase of more than 10 percent or a decrease of more than 10 percent was associated with a 0.443 standard deviation decrease in the test score (p-value = 0.009), relative to those with little cortisol responsivity (−10 percent to +10 percent), holding school-year academic grades, demographic characteristics, and concurrent cortisol constant. Table 3 contains these results. The estimates are fairly similar when broken up by those who increase more than 10 percent (0.437 standard deviation lower scores relative to those with −10 percent to +10 percent change, p-value = 0.015)

Table 3. Changes in Test Scores by Cortisol Responsivity to the Test

	Preferred (1)	Low-stakes Scores as a Control (2)	No Concurrent Cortisol (3)	Adding Low-stakes Week (4)	Placebo: Within-baseline Change (5)	Placebo: Low-stakes Change (6)	Placebo: High-stakes Change (7)
	Panel A: Change in Test Scores for 10% Above/Below Baseline Cortisol in Testing Week						
±10% from baseline	−0.443**	−0.558*	−0.439**	−0.261*	−0.114	−0.129	−0.159
	(0.164)	(0.214)	(0.157)	(0.109)	(0.196)	(0.193)	(0.204)
	Panel B: Change in Test Scores for 10% Above or 10% Below Baseline Cortisol in Testing Week						
10% above baseline	−0.437*	−0.550*	−0.431*	−0.263*	−0.125	−0.077	−0.163
	(0.172)	(0.222)	(0.163)	(0.115)	(0.198)	(0.194)	(0.208)
10% below baseline	−0.458*	−0.589+	−0.458*	−0.258*	−0.046	−0.238	−0.154
	(0.192)	(0.292)	(0.176)	(0.127)	(0.218)	(0.207)	(0.217)
Controls:							
Q1–3 grades	Y	N	Y	Q1 only	Y	Y	Y
Low-stakes test scores	N	Y	N	N	N	N	N
Concurrent cortisol	Y	Y	N	Y	Y	Y	Y
Cortisol change from baseline	To high-stakes	To high-stakes	To high-stakes	To low- OR high-stakes	Within baseline	To low-stakes	To high-stakes
Test outcome	High-stakes	High-stakes	High-stakes	Low- OR high-stakes	High-stakes	High-stakes	Low-stakes
N	67	62	67	136	82	63	63

Notes: Robust standard errors. Analysis conducted at the student level. Outcome is the *Z*-score of the indicated test. Cortisol data comes from saliva collected in homeroom. Each column represents a different regression estimate. All models also control for school fixed effects; a quadratic of time relative to wake; an indicator for whether the cortisol sample occurred within the CAR timeframe; wake time; age; female; and indicators for economically disadvantaged, Section 504, and McKinney-Vento Act (see Table 1 details). Column 1 is the preferred overall model. Columns 2–4 test alternate specifications by changing the measure of baseline ability (Column 2), removing controls for concurrent cortisol (Column 3), or adding the low-stakes week as an additional observation (Column 4). Column 4 uses quarter 1 grades as the control for baseline ability. Columns 5–7 test placebos of whether general cortisol variability is associated with test scores. Column 5 assesses within-baseline week changes of cortisol predict scores on the high-stakes test. Column 6 (7) tests whether cortisol changes from baseline to the low-stakes (high-stakes) test week affect high-stakes (low-stakes) test outcomes.

$^{+}p < 0.10$; $^{*}p < 0.05$; $^{**}p < 0.01$.

and those who decrease more than 10 percent (0.458 standard deviation lower scores, p-value = 0.021).

A potential concern is that grades do not adequately account for student testing ability. Thus, column 2 of table 3 uses low-stakes test scores instead of quarters 1–3 grades to control for baseline ability. Results were similar (−0.558 standard deviation lower for large reactivity participants, relative to the ±10 percent group, p-value = 0.005; N = 62), though the N was also slightly lower because some students were missing low-stakes test results. Results were also similar without controlling for concurrent cortisol (−0.439 standard deviation, p-value = 0.007; N = 67) and when adding the low-stakes test as an additional outcome (−0.261 standard deviations, p-value = 0.019; N = 136 tests for 73 participants).[29] The estimates are based on the average score across the math, ELA, and science high-stakes tests to decrease variability in scores; post hoc analyses demonstrated that the effects were negative for all three individual tests, with the

29. Here, we add the low-stakes week as an additional observation, and we define reactivity based on the change from baseline to the given testing week.

largest estimate in science.[30] There was no relationship between a quadratic of concurrent level of homeroom cortisol during the testing week itself and outcomes on the test, with or without including reactivity from baseline.

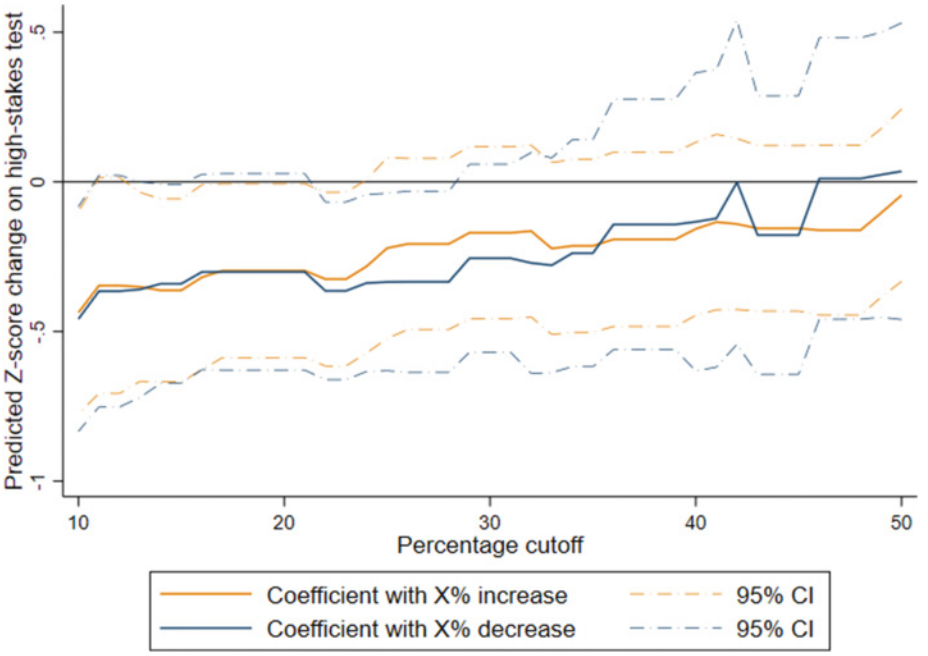
One concern with the analysis is that cortisol variability, rather than responsivity to the high-stakes test, is associated with worse outcomes on the test. We test this with three placebo measures in columns 5–7 of table 3. These all test whether changes between days unrelated to a given test predict performance on the test. Column 5 uses whether changes from day 1 to day 2 (or day 2 to day 3 for three individuals who joined on day 2) in the baseline week predict performance on the high-stakes test. It does not. As a second placebo in column 6, we test whether cortisol responsivity to the *low-stakes* test predicted performance months later on the *high-stakes* test. It did not, nor did responsivity to the high-stakes test predict performance on the low-stakes test in column 7. Thus, it does not appear that cortisol variability in general predicts performance on the high-stakes test. Instead, it is specifically changes from baseline to the high-stakes test week that are associated with performance on the high-stakes test itself.

We do not conduct the full binning exercise by subgroups due to small sample size. However, when we compared lower-reactivity participants (±10 percent cortisol change) to higher-reactivity participants (greater than ±10 percent change), we found no statistically significant differences in the patterns by gender, neighborhood poverty, neighborhood crime, or prior grades.[31] So, although some groups are more likely to be high-reactivity than others, the relationship with test scores is similar among all high-reactivity participants, at least to the extent that we can test it in this setting.

Although we prefer a bin-based specification for flexibility, the choice of −10 percent to +10 percent as a reference group range is arbitrary. Thus, figure 5 displays the estimated effect for being above and below different cut points. The graph includes 95 percent confidence intervals. The *x*-axis starts at 10 percent to match the estimates above, showing that a change of more than 10 percent above or 10 percent below baseline cortisol levels is associated with statistically significantly lower test scores, relative to those with cortisol responsivity between −10 percent and 10 percent. If, instead, we set the reference range to be ±15 percent, those whose cortisol dropped 15 percent or more had 0.340 standard deviation lower test scores (p-value = 0.051) and those whose cortisol increased 15 percent or more had 0.362 standard deviation lower test scores (p-value = 0.025), relative to those in the −15 percent to +15 percent range. Neither of the differences is statistically significant at the 5 percent level when we reach the ±17 percent range; neither is statistically significant at the 10 percent level once we reach the ±29 percent range.

30. The high-reactivity scores were lower than the low-reactivity (±10 percent) scores in science (−0.625, p-value = 0.009), reading (−0.493, p-value = 0.121), and math (−0.212, p-value = 0.240). Hausman tests indicated these coefficients sizes did not statistically differ across the three models (p-value = 0.237) and they jointly differed from zero (p-value = 0.001).

31. When interacting demographic indicator variables with an indicator for reactivity, the coefficient was −0.639 standard deviation for high-reactivity male participants and −0.965 standard deviation for high-reactivity female participants, relative to non-reactors in the ±10 percent range (p-value of male–female difference = 0.317). The coefficient was −0.372 for participants from lower-poverty neighborhoods and −0.145 for higher-poverty participants (p-value of difference = 0.492). The coefficient was −0.557 for participants from lower-crime neighborhoods and −0.707 for higher-crime participants (p-value of difference = 0.638). The coefficient was −0.600 for participants who had below-median course grades and −1.000 for participants who had above-median course grades (p-value of difference = 0.182)



Notes: Each distance is a separate regression; two coefficients per regression displayed. Coefficients displayed are for a variable that is equal to 1 if the change from baseline to the high-stakes testing week is greater than the indicated level. $N = 67$.

Figure 5. Estimated Effect Size by Different Bounding Distances (±10% to ±50%)

Figures 4 and 5 show that the estimates are noisy, with considerable unexplained fluctuation in test scores, and that the best outcomes appear around where there is little cortisol change. Overall, we take this as suggestive evidence that large changes in cortisol in response to high-stakes tests are associated with worse performance on the test, but there is much more to be done in this area.

Misbehavior as a Potential Mechanism

One hypothesis is that a cortisol spike could be associated with "acting out" and misbehavior during the test, which could inhibit performance. We can assess this hypothesis because the charter network tracked behavior using a daily points-based system.[32] Throughout the year, the average student got into at least some trouble on 35 percent of school days.

Relative to a regular day, there were no differences in the probability of getting in trouble on a low-stakes test day. However, for the most important week of high-stakes testing, students were 26 percentage points *less* likely to get in trouble than on

32. Observed values for behavior infractions and rewards ranged from −30 to +10, with positive outcomes in areas such as "scholarship" (+5 points, 775 observed instances over the academic year across the 83 students with observed data) and being a "reading rockstar" (+10 points, 60 instances observed), and negative outcomes in areas such as "instigating and/or fighting/fronting (including play fighting)" (−20 points, 65 instances), a category called "bathroom" (−10 points, 833 instances), "major violations" (−10 points, 828 instances), "talking out of turn" (−5 points, 1,847 instances), and "line" (−2 points, 973 instances).

regular school days (p-value = 0.000).[33] We do not take these estimates as a measure of acting out, necessarily, given the discretion that teachers have in assigning points to students.[34] Perhaps teachers were more lenient in general on test days, or perhaps students had fewer opportunities to get in trouble. However, we did test whether those with large increases in cortisol had different drops in infractions than those who had decreased cortisol or those who did not have a strong cortisol response.[35] We found no evidence of a difference in the probability of getting in trouble by those with very large increases (or decreases) in cortisol level. As best as we can measure, then, we find no evidence that misbehavior is driving the results on the tests. Instead, we hypothesize that the ability to focus and recall information relevant to the test is affected.

6. DISCUSSION

This study examined whether children responded physiologically to high-stakes testing in a naturalistic setting, and how any responses were associated with performance on a high-stakes test. Children in one charter school network displayed a statistically significant increase in cortisol level in anticipation of high-stakes testing; this pattern was driven by male students. We also find some evidence that, among a sample of disadvantaged students, the most-disadvantaged students had the largest increase in cortisol in anticipation of the high-stakes test. These changes were driven by the occurrence of a test that mattered for schools but had limited consequence for individual students.

Moderate decreases and increases in cortisol were associated with underperformance on the high-stakes test, relative to what we would have expected from students given their in-school academic performance and other characteristics. Even the average increase in cortisol shown in table 2 (18 percent) was associated with lower test scores, relative to those with little change in cortisol. An increase of more than 10 percent or a decrease of more than 10 percent was associated with a 0.4 standard deviation decrease in test scores, relative to those with little change. This is equivalent to approximately 80 points on the 1,600-point SAT scale. Concurrent cortisol measured as linear, quadratic, or bins during the test was not a statistically significant predictor of performance; it was cortisol change relative to baseline that predicted outcomes.

Of course, one study on a small, nonrandom sample of students may not give us the true population-level effect of high-stakes tests on cortisol or how cortisol relates to test

33. We identified every low- and high-stakes test day during the academic year. Using student fixed effects, we regressed an indicator for these day types, indicators for day of the week, and a continuous variable measuring the day of the year on an indicator for the probability that a student got in trouble on a given day. Students were less likely to get in trouble as the year went on, with the daily probability of getting in trouble dropping about 0.47 percentage points every ten calendar days. Tuesdays were the most likely day to get in trouble, followed by Wednesday, Monday, Thursday, and (much less likely) Friday.

34. Anecdotally, during our data collection we observed multiple instances of students acting out or acting differently during the high-stakes testing period than during the other data collection weeks. For example, a student was throwing up in the back of the room after the test; we were told protocol was to allow the students to leave their seats if they had to throw up. Another student "made a run for it" and led the staff on a chase through the school when we brought him to the hallway for his saliva sampling; they found him hiding in the kitchen. The behavior of students—and how that behavior might affect test scores—is an area in need of further systematic study for those who want to use test scores to make high-stakes decisions about students and school.

35. We interacted test type with an indicator for a responsivity greater than 10 percent and an indicator for a responsivity of less than −10 percent.

performance. We view this study as a call for additional work in this area. Specifically, we identify four non-exclusive themes in need of future research. First, future analyses should replicate that students do indeed have an increase in cortisol and that it differs by various attributes. Such research should examine a more diverse population of students, rather than the largely low-income, mostly black population we examined here. A larger sample size would permit a greater degree of heterogeneity analysis than is possible in the present study in order to more robustly test whether different groups respond differentially to high-stakes testing. One possibility is collecting cortisol samples around the time of SAT or ACT testing, as these tests have real-world implications for students.

Second, research must confirm whether moderate changes in cortisol are associated with worse performance in other settings. Researchers rely on high-stakes tests as a measure of academic performance to evaluate various education and social policies. Such research may accept that high-stakes tests are noisy measures of ability or knowledge, but it generally assumes that the noise is evenly distributed across the socioeconomic spectrum. If, however, certain groups are systematically "stressed testers"—that is, they have large physiological reactions to the high-stakes testing setting—the policies recommended by such research may be suboptimal. As an extreme example, consider a world where all children learn the same amount of material during the year but group A has a bigger physiological reaction, and subsequently lower scores, than group B. Examining test scores would lead researchers to conclude there is an achievement gap between these groups and that group A needs intervention. But in reality, both groups learn the same amount of material and can perhaps even apply that material similarly in the real world. The policy solution in this case would be much different than if learning differed between groups. Such test-day stress deficits are not the only cause of achievement gaps, but they may explain part of existing disparities. Future research should examine how large a role they play. A key consideration in any such research is causality. Researchers cannot assign stress responses to students in real-world tests, but real-world tests may be more stressful to students than lab-based tests. Carefully designed research, which includes measures of performance outside of the high-stakes tests, will be necessary to move understanding forward.

Third, researchers should consider how school policies that use test scores may exacerbate or alleviate disparities among groups. If certain groups are more likely to be stressed testers, then, holding baseline knowledge constant, those stressed testers will be disadvantaged by admission or graduation policies based on high-stakes tests. Researchers should carefully consider how policy decisions interact with biological responses to testing.

Finally, if new work confirms that testing causes stress for students in ways that impact their performance, a logical question is what schools can do to mitigate the stress response—or at least the effects of the stress response on performance. Potential options include mindfulness programs (Zenner, Herrnleben-Kurz, and Walach 2014), integrated mental health interventions (Fazel et al. 2014), or yoga (Ehud, An, and Avshalom 2010), among other interventions. Though such programs may have benefits well beyond test scores, researchers could investigate whether they are associated with changes in biological stress reactions to high-stakes testing.

Given the prevalence of high-stakes testing in U.S. education policy, much more work is needed in this area. If the patterns of test-induced stress that we find in this study continue to hold up, it might suggest that high-stakes testing results should be used and interpreted differently than the way they are currently implemented in education policy and practice.

ACKNOWLEDGMENTS

We thank the anonymous school district and its staff for their invaluable cooperation, as well as Kaho Arakawa, Chernjen Lee, Royette Tavernier, members of the COAST Lab at Northwestern University, and seminar participants at Northwestern University and the AEFP, APPAM, and Western Economic Association meetings. Laura Scaramella at the University of New Orleans provided access to laboratory space. We are grateful for funding from the Spencer Foundation (grant no. 2015000117) and the Institute for Policy Research at Northwestern University.

REFERENCES

Adam, Emma K. 2012. Emotion-cortisol transactions occur over multiple time scales in development: Implications for research on emotion and the development of emotional disorders. *Monographs of the Society for Research in Child Development* 77(2): 17–27.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies* 81(2): 608–650.

Blair, Clancy, Douglas Granger, and Rachel Peters Razza. 2005. Cortisol reactivity is positively related to executive function in preschool children attending Head Start. *Child Development* 76(3): 554–567.

Bradbury, Bruce, Miles Corak, Jane Waldfogel, and Elizabeth Washbrook. 2015. *Too many children left behind: The U.S. achievement gap in comparative perspective*. New York: Russell Sage Foundation.

Chicago Public Schools. 2020. *Chicago Public Schools student assessments*. Available https://www.cps.edu/academics/student-assessments/. Accessed 13 October 2020.

Clow, Angela, Frank Hucklebridge, Tobias Stalder, Phil Evans, and Lisa Thorn. 2010. The cortisol awakening response: More than a measure of HPA axis function. *Neuroscience & Biobehavioral Reviews* 35(1): 97–103.

Del Giudice, Marco, Bruce J. Ellis, and Elizabeth A. Shirtcliff. 2011. The adaptive calibration model of stress responsivity. *Neuroscience & Biobehavioral Reviews* 35(7): 1562–1592.

Dickerson, Sally S., and Margaret E. Kemeny. 2004. Acute stressors and cortisol responses: A theoretical integration and synthesis of laboratory research. *Psychological Bulletin* 130(3): 355–391.

Doane, Leah D., and Emma K. Adam. 2010. Loneliness and cortisol: Momentary, day-to-day, and trait associations. *Psychoneuroendocrinology* 35(3): 430–441.

Dressendörfer, R. A., C. Kirschbaum, W. Rohde, F. Stahl, and C. J. Strasburger. 1992. Synthesis of a cortisol-biotin conjugate and evaluation as a tracer in an immunoassay for salivary cortisol measurement. *Journal of Steroid Biochemistry and Molecular Biology* 43(7): 683–692.

Ehud, Miron, Bar-Dov An, and Strulov Avshalom. 2010. Here and now: Yoga in Israeli schools. *International Journal of Yoga* 3(2): 42–47.

Engert, Veronika, Simona I. Efanov, Annie Duchesne, Susanne Vogel, Vincent Corbo, and Jens C. Pruessner. 2013. Differentiating anticipatory from reactive cortisol responses to psychosocial stress. *Psychoneuroendocrinology* 38(8): 1328–1337.

Fazel, Mina, Kimberly Hoagwood, Sharon Stephan, and Tamsin Ford. 2014. Mental health interventions in schools in high-income countries. *Lancet Psychiatry* 1(5): 377–387.

Gunnar, Megan R., and Karina Quevedo. 2007. The neurobiology of stress and development. *Annual Review of Psychology* 58(1): 145–173.

Hatch, Stephani L., and Bruce P. Dohrenwend. 2007. Distribution of traumatic and other stressful life events by race/ethnicity, gender, SES and age: A review of the research. *American Journal of Community Psychology* 40(3–4): 313–332.

Heiser, Paul, Gayle Simidian, David Albert, John Garruto, Dawn Catucci, Peter Faustino, Kara McCarten May, and Kelly Caci. 2015. Anxious for success: High anxiety in New York's schools. Available www.nyssba.org/clientuploads/nyssba_pdf/Test_Anxiety_Report.pdf. Accessed 8 October 2020.

Heissel, Jennifer A., Dorainne J. Levy, and Emma K. Adam. 2017. Stress, sleep, and performance on standardized tests: Understudied pathways to the achievement gap. *AERA Open* 3(3): 1–17.

Heissel, Jennifer A., Patrick T. Sharkey, Gerard Torrats-Espinosa, Kathryn Grant, and Emma K. Adam. 2018. Violence and vigilance: The acute effects of community violent crime on sleep and cortisol. *Child Development* 89(4): e323–e331.

Het, Serkan, G. Ramlow, and Oliver T. Wolf. 2005. A meta-analytic review of the effects of acute cortisol administration on human memory. *Psychoneuroendocrinology* 30(8): 771–784.

Hoyt, Lindsay T., Katharine H. Zeiders, Katherine B. Ehrlich, and Emma K. Adam. 2016. Positive upshots of cortisol in everyday life. *Emotion* 16(4): 431–435.

King, Jean A., Milagros C. Rosal, Yunsheng Ma, George Reed, Terri-Ann Kelly, and Ira S. Ockene. 2000. Sequence and seasonal effects of salivary cortisol. *Behavioral Medicine* 26(2): 67–73.

Lazarín, Melissa. 2014. *Testing overload in America's schools*. Available https://cdn.americanprogress.org/wp-content/uploads/2014/10/LazarinOvertestingReport.pdf. Accessed 8 October 2020.

Lindahl, Mats, Töres Theorell, and Frank Lindblad. 2005. Test performance and self-esteem in relation to experienced stress in Swedish sixth and ninth graders—Saliva cortisol levels and psychological reactions to demands. *Acta Pædiatrica* 94(4): 489–495.

Litten, Kevin. 2016. *New Orleans poverty rates fall in 2015, still higher than state average*. Available https://www.nola.com/news/politics/article_8b5169be-4ab4-5609-b65c-d8fd215e0808.html. Accessed 13 October 2020.

Lupien, Sonia J., Charles W. Wilkinson, Sophie Brière, Catherine Ménard, N. M. K. Ng Ying Kin, and N. P. V. Nair. 2002. The modulatory effects of corticosteroids on cognition: Studies in young human populations. *Psychoneuroendocrinology* 27(3): 401–416.

Malarkey, William B., Dennis K. Pearl, Laurence M. Demers, Janice K. Kiecolt-Glaser, and Ronald Glaser. 1995. Influence of academic stress and season on 24-hour mean concentrations of ACTH, cortisol, and β-endorphin. *Psychoneuroendocrinology* 20(5): 499–508.

Mattarella-Micke, Andrew, Jill Mateo, Megan N. Kozak, Katherine Foster, and Sian L. Beilock. 2011. Choke or thrive? The relation between salivary cortisol and math performance depends on individual differences in working memory and math-anxiety. *Emotion* 11(4): 1000–1005.

McEwen, Bruce S. 1998. Stress, adaptation and disease: Allostasis and allostatic load. *Annals of the New York Academy of Sciences* 840: 34–44.

McEwen, Bruce S., and Peter J. Gianaros. 2010. Central role of the brain in stress and adaptation: Links to socioeconomic status, health, and disease. *Annals of the New York Academy of Sciences* 1186: 190–222.

Miller, Gregory E., Edith Chen, and Eric S. Zhou. 2007. If it goes up, must it come down? Chronic stress and the hypothalamic-pituitary-adrenocortical axis in humans. *Psychological Bulletin* 133(1): 25–45.

Proctor, Bernadette D., Jessica L. Semega, and Melissa A. Kollar. 2016. *Income and poverty in the United States: 2015*. Available https://www.census.gov/library/publications/2016/demo/p60-256.html. Accessed 8 October 2020.

Reardon, Sean F. 2011. The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. In *Whither opportunity? Rising inequality, schools, and children's life chances*, edited by Greg J. Duncan and Richard J. Murnane, pp. 91–116. New York: Russell Sage Foundation.

Salvador, A., F. Suay, E. González-Bono, and M. A. Serrano. 2003. Anticipatory cortisol, testosterone and psychological responses to judo competition in young men. *Psychoneuroendocrinology* 28(3): 364–375.

Sapolsky, Robert M., L. Michael Romero, and Allan U. Munck. 2000. How do glucocorticoids influence stress responses? Integrating permissive, suppressive, stimulatory, and preparative actions. *Endocrine Reviews* 21(1): 55–89.

Sauro, Marie D., Randall S. Jorgensen, and Teal Pedlow. 2003. Stress, glucocorticoids, and memory: A meta-analytic review. *Stress* 6(4): 235–245.

Schilling, Thomas M., Monika Kölsch, Mauro F. Larra, Carina M. Zech, Terry D. Blumenthal, Christian Frings, and Hartmut Schächinger. 2013. For whom the bell (curve) tolls: Cortisol rapidly affects memory retrieval by an inverted U-shaped dose–response relationship. *Psychoneuroendocrinology* 38(9): 1565–1572.

Schlotz, Wolff, Peter Schulz, Juliane Hellhammer, Arthur A. Stone, and Dirk H. Hellhammer. 2006. Trait anxiety moderates the impact of performance pressure on salivary cortisol in everyday life. *Psychoneuroendocrinology* 31(4): 459–472.

Segool, Natasha K., John S. Carlson, Anisa N. Goforth, Nathan von der Embse, and Justin A. Barterian. 2013. Heightened test anxiety among young children: Elementary school students' anxious responses to high-stakes testing. *Psychology in the Schools* 50(5): 489–499.

Shirtcliff, Elizabeth A., Jeremy C. Peres, Andrew R. Dismukes, Yoojin Lee, and Jenny M. Phan. 2014. Hormones: Commentary. Riding the physiological roller coaster: Adaptive significance of cortisol stress reactivity to social contexts. *Journal of Personality Disorders* 28(1): 40–51.

Stalder, Tobias, Clemens Kirschbaum, Brigitte M. Kudielka, Emma K. Adam, Jens C. Pruessner, Stefan Wüst, Samantha Dockray, Nina Smyth, Phil Evans, Dirk H. Hellhammer, Robert Miller, Mark A. Wetherell, Sonia J. Lupien, and Angela Clow. 2016. Assessment of the cortisol awakening response: Expert consensus guidelines. *Psychoneuroendocrinology* 63:414–432.

Stroud, Laura R., Peter Salovey, and Elissa S. Epel. 2002. Sex differences in stress responses: Social rejection versus achievement stress. *Biological Psychiatry* 52(4): 318–327.

Weekes, Nicole, Richard Lewis, Falgooni Patel, Jared Garrison-Jakel, Dale E. Berger, and Sonia J. Lupien. 2006. Examination stress as an ecological inducer of cortisol and psychological responses to stress in undergraduate students. *Stress* 9(4): 199–206.

Zenner, Charlotte, Solveig Herrnleben-Kurz, and Harald Walach. 2014. Mindfulness-based interventions in schools—A systematic review and meta-analysis. *Frontiers in Psychology* 5: Article 603.