Testing, Stress, and Performance: How Students Respond Physiologically to High-Stakes Testing

Jennifer A. Heissel[1]*, Emma K. Adam[2], Jennifer L. Doleac[3], David N. Figlio[2], Jonathan Meer[3]

[1]Naval Postgraduate School, Monterey, CA 93943

[2]Northwestern University, Evanston, IL 60208

[3]Texas A&M University, College Station, TX 77845

November 2018

*Corresponding author: jaheisse@nps.edu; 831-656-6394. 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 AEFP, APPAM, Northwestern University, and the 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.

Abstract

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 does affect cortisol responses, and those responses have consequences for test performance. Those who responded most strongly – with either a large increase or large decrease in cortisol – scored 0.40 standard deviations lower than expected on the on the high-stakes exam.

Keywords: stress, cortisol, cognition, high-stakes testing

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, & Adam, 2017).

This study makes clear the potential policy implications of high-stakes test-induced stress. We document how high-stakes testing affects low-income children's stress biology, and we show how changes in children's physiological responses to high-stakes tests affect performance on standardized tests. Knowing the answers to these questions affects our understanding of how high-stakes test results should be used and interpreted.

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 ("cortisol reactivity") in response to a high-stakes standardized test administered to students in grades 3-8, relative to a regular baseline school week. We then examine whether differences in cortisol reactivity are associated with performance on the test. We find that students have 15% higher cortisol levels in the homeroom period just before taking the high-stakes test, relative to that same timeframe during weeks without testing. These differ-

[1] The Center for American Progress found that 49% of parents thought that 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, Carlson, Goforth, Embse, & Barterian, 2013).

[2] A variety of studies have examined cognitive tests in lab settings (Lupien et al., 2002; Stroud, Salovey, & Epel, 2002) or with researcher-administered tests in schools that did not matter for student or school outcomes (Blair, Granger, & Razza, 2005; Lindahl, Theorell, & Lindblad, 2005). These studies do not include baseline, non-testing weeks in their analysis. Other studies have looked at adult responses in undergraduate and medical students (Malarkey, Pearl, Demers, Kiecolt-Glaser, & Glaser, 1995; Weekes et al., 2006).

ences are driven by boys, whose homeroom cortisol is 35% higher during testing weeks than regular weeks.[3] While our entire sample can be considered economically disadvantaged, we also find suggestive evidence of differences by level of disadvantage, with the largest cortisol effects for those living in high-poverty and high-crime neighborhoods. We also show that both large increases and large decreases in cortisol from the baseline week to the high-stakes testing week are associated with much lower test scores on the high-stakes test, relative to how we would expect students to perform based on other in-school academic performance (e.g., grades). High-stakes testing appears to be inducing large cortisol increases in some students, perhaps disrupting their ability to concentrate. For other students, their response to the stressor appears to be disengagement with their environment, as captured by large cortisol decreases and also resulting in worse test outcomes.

Descriptive studies show that children from low-SES and racial/ethnic minority groups have lower average scores on standardized academic tests relative to high-SES and white families (Bradbury, Corak, Waldfogel, & Washbrook, 2015; Reardon, 2011). Low-SES 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 & 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 having disparate effects on test performance. This, in turn, has implications for whether standardized tests are a fair means of evaluating student ability and school quality.

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. However, large increases in cortisol can make concentration difficult, while limited increases or reduced cortisol

[3] This is consistent with previous evidence that males show larger cortisol responses to achievement-related stressors (Stroud, Salovey, & Epel, 2002; Weekes et al., 2006).

may be a sign of disengagement with a task. In particular, those who experience prolonged stress exposure may get "burned out," in the sense that they are unable to respond to acute stressors (see review in McEwen, 1998).

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 sensitive period 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 chronic stress associated with poverty affects subsequent behavior. Third, and most importantly, we provide the first evidence on how differences in cortisol responses affect performance on 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.

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 concludes.

2. Background on biological stress responses and cortisol

Biological stress response includes 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 & 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 30-40 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, Hucklebridge, Stalder, Evans, & Thorn, 2010).

Real or perceived stressors can increase cortisol above typical diurnal levels.[4] For routine stressors (e.g., missing the bus), cortisol levels return to their usual daily pattern approximately an hour after the stressor has passed. According to the Adaptive Calibration Model, stress response is generally adaptive; for instance, the HPA axis may mobilize psychological and physiological responses when presented with a stressor (Del Giudice, Ellis, & Shirtcliff, 2011; Shirtcliff, Peres, Dismukes, Lee, & Phan, 2014). One at-home study had 24 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, Zeiders, Ehrlich, & Adam, 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

[4] This pattern has been consistently demonstrated in the psychology and endocrinology literature (see reviews in Adam, 2012; Miller, Chen, & Zhou, 2007; Sapolsky, Romero, & Munck, 2000).

[5] The Adaptive Calibration model attempts to build a model of the development of stress responsivity in general (Del Giudice, Ellis, & 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 & 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, Suay, González-Bono, & Serrano, 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, Sharkey, Torrats-Espinosa, Grant, and Adam (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.

that matter for them. Limited (or lowered) cortisol responses to stressors may be related to 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 gender. 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, while females may be more responsive to social rejection (Stroud et al., 2002). A meta-analysis of 28 studies similarly found larger cortisol responses to stressors in males than females (Sauro, Jorgensen, & 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 & 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. 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.

[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, & 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 restore 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).

Two previous studies compare a baseline week of normal activity against a stressful testing week. Weeks et al. (2006) found that male undergraduate students had an increase in examination-week cortisol levels, while females 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 test during the school day as stressful was correlated with higher same-day cortisol and lower test performance in Swedish adolescents (Lindahl, Theorell, & Lindblad, 2005). Conversely, among young, low-income children in Head Start, 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 et al., 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, Schulz, Hellhammer, Stone, & Hellhammer, 2006). Whether cortisol improves or detracts from performance may depend on anxiety about the task at hand (Mattarella-Micke, Mateo, Kozak, Foster, & Beilock, 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%), economically disadvantaged (97%)[8], and from high-poverty neighborhoods (with 40% of block group households in poverty, mean block group income of $27,000, and mean block group unemployment of 13%). The households were also in neighborhoods with a great deal of police activity, with a mean of 416 high-priority 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%, mean neighborhood incomes ranged from $9,000 to $58,000, neighborhood unemployment rates ranged from 0 to 74%, 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]

3.1 Cortisol data

We collected salivary cortisol samples from 93 pre-adolescent and adolescent volunteers in grades 3-8, across three schools from the charter school network. We 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 and were briefed on the protocol individually.

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 they followed protocol. Participants were instructed to avoid eating, drinking, and brushing their teeth 30 minutes prior to each sample collection. A kitchen timer pre-set to 30 minutes was provided

[8] Economic disadvantage is indicated by eligibility for free or reduced-price lunch.

[9] The median household income in the US was $57,000 in 2015, with 13.5% of households in poverty (Proctor, Semega, & Kollar, 2016). New Orleans had a median household income of $37,000, with 26% of households in poverty in this period; the mean of $27,000 income in our sample is similar to the $28,000 median black family income in New Orleans (US Census Bureau, 2017).

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.

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[10] (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 1-2.[11] 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.[12] Sample 3 had the most consistent timing and the highest completion rate across days and is the focus of the majority of our analysis.

3.2 Student diaries

Participants filled out diaries at each data 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.

---

[10] Homeroom started at 7:00am in School 1 and 8:00am in Schools 2 and 3.

[11] Seventy-eight percent of individual-week-sample number combinations had at least one sample, though completion rates varied by sample (see Figure A2). Samples were stored at -20°C before shipment to Trier, Germany, where they were assayed in duplicate using time-resolved fluorescent-detection immunoassay (Dressendörfer, Kirschbaum, Rohde, Stahl, & Strasburger, 1992).

[12] 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%; out-of-school compliance was 72%. Changes in school scheduling meant that Sample 4 timing had a wider variance than Sample 3. Figure A1 displays the distribution of sample timing by sample and school.

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.

3.3 Administrative data

The charter network provided administrative data including participants' high-stakes math, science, and English/language arts (ELA) tests and low-stakes math, science, ELA, and social studies tests. 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.[13] The administrative data also included in-school grades (on a 0-100 scale) for each academic quarter in math, science, ELA, and social studies.

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 mimicked the structure of the year-end high-stakes test.[14] 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 our attention on Sample 3.[15] Our first analysis examines whether the level of cortisol

[13] 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.

[14] This amount of testing is not atypical; for instance, Chicago Public Schools had a testing schedule comparable to our charter school network in the 2015-16 school year (Chicago Public Schools, 2015).

[15] See appendix for details on sampling timing by week, sample, and school (Figure A1) and by percent of samples provided by week and sample (Figure A2).

in the homeroom period (just before the test was administered) changed from baseline to the low-stakes and high-stakes testing weeks. This time period is particularly important given that it reflects the level of cortisol that students bring into the test setting. This analysis examines how homeroom cortisol changed from baseline to the testing weeks with the following specification:

$$\ln(\text{[illegible]}_{iwd}) = \beta_0 + \beta_1 LowStakes_w + \beta_2 HighStakes_w + \beta_3 Time_{iwd} + \beta_4 Time_{iwd}^2 + \beta_5 Waketime_{iwd} + \beta_6 CAR_{iwd} + \gamma_i + \varepsilon_{iwd}$$

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, and CAR_{iwd} is an indicator for whether the homeroom sample was 15-60 minutes after the individual's wake time that day. 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. 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. The individual fixed effects γ_i account for any observed and unobserved factors that are constant across an individual over time (e.g., gender, 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.

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 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:

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

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 ***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% or lower, -10 to -30%, -10 to 10% (the reference bin), 10% to 30%, 30% to 50%, 50% to 70%, and 70% 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 ***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 under- or over-performance 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.[16] Some participants were missing either requisite cortisol or testing data, and the N in

[16] There is no statistical difference in our estimates if we do not include the demographic controls in this analysis; we include them for completeness.

the final analysis is 68.[17] Given this smaller sample, we interpret the academic performance results as suggestive and do not conduct subgroup analyses.

5. Results

5.1 Changes in cortisol daily rhythms

Figure 1 displays the cortisol patterns from wake to eight hours post-wake for baseline, low-stakes, and high-stakes weeks using locally-weighted scatterplot 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 during 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.

5.2 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. The coefficient on low-stakes (high-stakes) testing estimates 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. The homeroom estimates were similar whether controlling for wake time or not, and going forward we prefer the more conservative estimate that controls for wake time and CAR. On average, cortisol was 15% 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

[17] 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. (Estimates available upon request.)

relative to the baseline week, though as expected the coefficients are positive but smaller than the coefficient for the high-stakes week.

The final three columns 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. Males had large average increases in homeroom cortisol in the low-stakes testing week (30%) and high-stakes testing week (35%), relative to the baseline week. The female effect sizes statistically differed from the male estimates; the difference relative to baseline was -4% in the low-stakes week (calculated as 0.299-0.338) and 0% in the high stakes week. Neither of the female estimates statistically differed from zero. Turning to neighborhood characteristics, those from higher-poverty neighborhoods had larger average increases in homeroom cortisol than those from lower-poverty neighborhoods in the high-stakes week (26% versus 4%), relative to baseline, though the difference between groups was not statistically significant. Similarly, those from neighborhoods with above-median number of high-priority 911 calls had larger average increases in homeroom cortisol than those from below-median neighborhoods in the high-stakes week (24% versus 6%), though the difference between groups was not statistically significant. While those facing chronic stress do respond to the stress of testing differently, 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.

There was higher cortisol before the test, on average, relative to the baseline week, but there was also substantial variation in reactivity. Figure 2 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 performance on the test.

5.3 Differences in academic outcomes

Figure 3 examines how cortisol reactivity 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 24 to 67% relative to baseline during high-stakes testing. Quintile 2 includes those with little change, ranging from -22% to +12%. Quintile 3 participants have moderate increases, from 13% to 52%. The final two quintiles cover those with large increases, from 53 to 90% in Quintile 4 and over 100% increases in Quintile 5.

Quintile 2 differs significantly from Quintile 1, with test scores 0.45 standard deviations higher (conditional on demographic controls, concurrent cortisol, and in-school grades; *p*-value=0.015). The final three quintiles do not differ significantly from Quintile 1. We reject that the five categories are the same with an F-test (F(4, 45)=2.90; *p*-value=0.032), but we cannot reject that Quintiles 1, 3, 4, and 5 are statistically significantly different from each other (F(3, 45)=0.76; *p*-value=0.521). 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.

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, & 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.[18] The pattern is an inverse-U, but contrary to prior work, we find no evidence of an improvement in outcomes for moderate increases in cortisol.[19]

---

[18] The model also includes a quadratic control for concurrent cortisol, but we find no relationship between concurrent cortisol and outcomes on the test.

[19] The scatterplot of these data does not clearly indicate a U-shaped pattern. The quadratic term for responsivity is statistically significant (β=-5.166; *p*-value=0.008), indicating that we find an inverse-U type pattern for increases in cortisol, relative to baseline. The linear estimate is null in this model, and it is actually slightly negative (β=-0.144; *p*-value=0.790). We also tested a cubic function, finding a relationship with test scores for the quadratic term (β=-4.992; *p*-value=0.016) but no relationship for the linear (β=-0.122; *p*-value=0.879) or cubic (β=-1.430; *p*-value=0.844) terms. The quadratic term was also the only statistically significant coefficient in a quartic function.

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% to +10% 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, large decreases *and* large increases in cortisol were associated with underperformance on the high-stakes test. This is in line with the lab-based evidence described in Section 2 that inducing large decreases or increases in cortisol reduced memory recall. Grouping the "change" bins together, an increase of more than 10% or a decrease of more than 10% was associated with a 0.461 standard deviation decrease in the test score (*p*-value=0.007), relative to those with little cortisol responsivity (-10% to +10%), holding school-year academic grades, demographic characteristics, and concurrent cortisol constant. The estimates are fairly similar when broken up by those who increase more than 10% (0.446 standard deviation lower scores relative to those with -10% to +10% change, *p*-value=0.012) and those who decrease more than 10% (0.500 standard deviation lower scores, *p*-value=0.014).

Results were similar using the low-stakes test scores instead of quarter 1-3 grades to control for baseline ability (-0.560 standard deviations lower for large reactivity participants, relative to the +/-10% group, *p*-value=0.002; N=63), when using the low-stakes test as an additional outcome (-0.291 standard deviations, *p*-value=0.014; N=138 tests for 75 participants), and without controlling for concurrent cortisol (-0.453 standard deviations, *p*-value=0.007; N=68). The estimates are based on the average score across the math, ELA, and science tests to decrease variability in scores; post hoc analyses demonstrated that the effects were negative for all three individual tests, with the largest estimate in science.[20] There was no relationship between concurrent level of homeroom cortisol during the testing week itself and outcomes on the test. Feeling nervous about the test was not associated with any difference in outcomes (with a coefficient on an indicator variable for nervousness equal to -0.020 standard deviations, *p*-value=0.879), nor was

[20] The high-reactivity scores were lower than the low reactivity (+/-10%) scores in science (-0.649, *p*-value=0.002), reading (-0.495, *p*-value=0.066), and math (-0.240, *p*-value=0.110). Hausman tests indicated that these coefficients sizes did not statistically differ across the three models *p*-value=0.295) and that they jointly differed from zero (*p*-value=0.000).

there a significant interaction between responsivity and nervousness (the coefficient on the interaction was 0.011 standard deviations, *p*-value=0.974).[21]

We do not conduct the full binning exercise by subgroups due to small sample size. However, when we grouped the estimates into three groups (decreases in cortisol of more than 10%, a low reactivity group between -10% and 10%, and increases over more than 10%), we found no statistically significant differences in the patterns by gender, neighborhood poverty, or neighborhood crime.[22] 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.

While we prefer a bin-based specification for flexibility, the choice of -10% to +10% 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% confidence intervals. The X-axis starts at 10% to match the estimate above, showing that a change of more than 10% above or 10% below baseline cortisol levels is associated with statistically significantly lower test scores, relative to those with cortisol responsivity between -10% and 10%. If, instead, we set the reference range to be +/-20%, those whose cortisol dropped 20% or more had 0.315 standard deviations lower test scores (*p*-value=0.063) and those whose cortisol increased 20% or more had 0.330 standard deviations lower test scores (*p*-value=0.016), relative to those in the -20% to +20% range. Neither of the differences are statistically significant at the 5% level when we reach the +/-25% range; neither are statistically significant at the 10% level once we reach the +/-29% range.

Figures 4 and 5 show that the estimates are noisy, with a lot of 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

[21] Their diary questionnaire included a question about nervousness on various topics; this was a 0-3 scale for older middle school students and 0/1 for elementary school students. We created an indicator variable with any reported nervousness about tests equal to one, and 68% of respondents reported feeling nervous about tests during home during the testing week.

[22] When interacting demographics with an indicator for reactivity, the coefficient was -0.518 standard deviations for high-reactivity male participants; the effect was -0.611 standard deviations for high-reactivity females, relative to non-reactors in the +/- 10% range (*p*-value of male-female difference=0.780). The coefficient was -0.412 for participants from lower-poverty neighborhoods and -0.259 for higher-poverty participants (*p*-value of difference=0.648). The coefficient was -0.548 for participants from lower-crime neighborhoods and -0.650 for higher-crime participants (*p*-value of difference=0.756).

tests are associated with worse performance on the test, but there is much more to be done in this area.

5.4 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.[23] Throughout the year, the average student got into at least some trouble on 38% 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 34 percentage points *less* likely to get in trouble than on regular school days (*p*-value=0.000).[24] We do not take these estimates as a measure of acting out, necessarily, given the discretion that teachers have in assigning points to students.[25] 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 that those who had decreased cortisol or those who did not have a strong cortisol response.[26] We found no evidence of a difference in the probability of getting in trouble by those

[23] 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).

[24] 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 0.4 percentage points every 10 calendar days. Tuesdays were the most likely day to get in trouble, followed by Wednesday, Monday, Thursday, and (much less likely) Friday.

[25] 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. 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 scored – 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.

[26] We interacted test type with an indicator for a responsivity greater than 10% and an indicator for a responsivity of less than -10%.

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 naturalistic settings, and how any responses were associated with performance on the high-stakes test. Children displayed a statistically significant increase in cortisol level in anticipation of high-stakes testing; this pattern was driven by males. 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.

Large decreases and large 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 (15%) was associated with much lower test scores, relative to those with little change in cortisol. An increase of more than 10% or a decrease of more than 10% was associated with a 0.4 SD decrease in test scores, relative to those with little change. This is equivalent to approximately 80 points on the 1600-point SAT scale. Concurrent cortisol during the test was not a statistically significant predictor of performance: it was the cortisol change relative to baseline that predicted outcomes.

Future analyses 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. That said, our findings that students respond physiologically to high-stakes tests are relevant to schools that emphasize high-stakes testing – and particularly those that serve disadvantaged populations. Large cortisol responses – either positive or negative – were associated with worse test performance, perhaps introducing a “stress bias” and making tests a less reliable indicator of student learning.

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, 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 there is 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.

The same can be said of school policies that use test scores. 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. Overall, we find that students do physiologically respond 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 first 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 employed in education policy and practice.

References

Adam, E. 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. https://doi.org/10.1111/j.1540-5834.2012.00657.x

Blair, C., Granger, D., & Razza, R. P. (2005). Cortisol reactivity is positively related to executive function in preschool children attending Head Start. *Child Development*, *76*(3), 554–567. https://doi.org/10.1111/j.1467-8624.2005.00863.x

Bradbury, B., Corak, M., Waldfogel, J., & Washbrook, E. (2015). *Too Many Children Left Behind: The U.S. Achievement Gap in Comparative Perspective*. New York: Russell Sage Foundation.

Chicago Public Schools. (2015). CPS Assessment Framework, SY2015-16. Retrieved September 18, 2018, from https://www.cps.edu/SchoolData/Documents/assessmentFramework.pdf

Clow, A., Hucklebridge, F., Stalder, T., Evans, P., & Thorn, L. (2010). The cortisol awakening response: More than a measure of HPA axis function. *Neuroscience & Biobehavioral Reviews*, *35*(1), 97–103. https://doi.org/10.1016/j.neubiorev.2009.12.011

Del Giudice, M., Ellis, B. J., & Shirtcliff, E. A. (2011). The Adaptive Calibration Model of stress responsivity. *Neuroscience & Biobehavioral Reviews*, *35*(7), 1562–1592. https://doi.org/10.1016/j.neubiorev.2010.11.007

Dickerson, S. S., & Kemeny, M. E. (2004). Acute stressors and cortisol responses: A theoretical integration and synthesis of laboratory research. *Psychological Bulletin*, *130*(3), 355–391. https://doi.org/10.1037/0033-2909.130.3.355

Doane, L. D., & Adam, E. K. (2010). Loneliness and cortisol: Momentary, day-to-day, and trait associations. *Psychoneuroendocrinology*, *35*(3), 430–441. https://doi.org/10.1016/j.psyneuen.2009.08.005

Dressendörfer, R. A., Kirschbaum, C., Rohde, W., Stahl, F., & Strasburger, C. J. (1992). Synthesis of a cortisol-biotin conjugate and evaluation as a tracer in an immunoassay for salivary cortisol measurement. *The Journal of Steroid Biochemistry and Molecular Biology*, *43*(7), 683–692. https://doi.org/10.1016/0960-0760(92)90294-S

Engert, V., Efanov, S. I., Duchesne, A., Vogel, S., Corbo, V., & Pruessner, J. C. (2013). Differentiating anticipatory from reactive cortisol responses to psychosocial stress. *Psychoneuroendocrinology*, *38*(8), 1328–1337. https://doi.org/10.1016/j.psyneuen.2012.11.018

Gunnar, M. R., & Quevedo, K. (2007). The neurobiology of stress and development. *Annual Review of Psychology*, *58*(1), 145–173. https://doi.org/10.1146/annurev.psych.58.110405.085605

Hatch, S. L., & Dohrenwend, B. P. (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. https://doi.org/10.1007/s10464-007-9134-z

Heiser, P., Simidian, G., Albert, D., Garruto, J., Catucci, D., Faustino, P., … Caci, K. (2015). *Anxious for success: High anxiety in New York's schools* (p. 10). New York: New York Association of School Psychologists. Retrieved from http://www.nyssba.org/clientuploads/nyssba_pdf/Test_Anxiety_Report.pdf

Heissel, J. A., Levy, D. J., & Adam, E. K. (2017). Stress, sleep, and performance on standardized tests: Understudied pathways to the achievement gap. *AERA Open*, *3*(3), 2332858417713488. https://doi.org/10.1177/2332858417713488

Heissel, J. A., Sharkey, P. T., Torrats-Espinosa, G., Grant, K., & Adam, E. K. (2018). Violence and vigilance: The acute effects of community violent crime on sleep and cortisol. *Child Development*, *89*(4), e323–e331. https://doi.org/10.1111/cdev.12889

Het, S., Ramlow, G., & Wolf, O. T. (2005). A meta-analytic review of the effects of acute cortisol administration on human memory. *Psychoneuroendocrinology*, *30*(8), 771–784. https://doi.org/10.1016/j.psyneuen.2005.03.005

Hoyt, L. T., Zeiders, K. H., Ehrlich, K. B., & Adam, E. K. (2016). Positive upshots of cortisol in everyday life. *Emotion*, *16*(4), 431–435. https://doi.org/10.1037/emo0000174

Lazarín, M. (2014). *Testing Overload in America's Schools* (p. 34). Center for American Progress. Retrieved from https://cdn.americanprogress.org/wp-content/uploads/2014/10/LazarinOvertestingReport.pdf

Lindahl, M., Theorell, T., & Lindblad, F. (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. https://doi.org/10.1111/j.1651-2227.2005.tb01922.x

Lupien, S. J., Wilkinson, C. W., Brière, S., Ménard, C., Ng Ying Kin, N. M. K., & Nair, N. P. V. (2002). The modulatory effects of corticosteroids on cognition: Studies in young human populations. *Psychoneuroendocrinology*, *27*(3), 401–416. https://doi.org/10.1016/S0306-4530(01)00061-0

Malarkey, W. B., Pearl, D. K., Demers, L. M., Kiecolt-Glaser, J. K., & Glaser, R. (1995). Influence of academic stress and season on 24-hour mean concentrations of ACTH, cortisol, and β-endorphin. *Psychoneuroendocrinology*, *20*(5), 499–508. https://doi.org/10.1016/0306-4530(94)00077-N

Mattarella-Micke, A., Mateo, J., Kozak, M. N., Foster, K., & Beilock, S. L. (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. https://doi.org/10.1037/a0023224

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

McEwen, B. S., & Gianaros, P. J. (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. https://doi.org/10.1111/j.1749-6632.2009.05331.x

Miller, G. E., Chen, E., & Zhou, E. S. (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. https://doi.org/10.1037/0033-2909.133.1.25

Proctor, B. D., Semega, J. L., & Kollar, M. A. (2016). *Income and Poverty in the United States: 2015* (No. P60-256). Washington, D.C.: US Census Bureau. Retrieved from https://www.census.gov/library/publications/2016/demo/p60-256.html

Reardon, S. F. (2011). The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. In G. J. Duncan & R. J. Murnane (Eds.), *Whither Opportunity? Rising Inequality, Schools, and Children's Life Chances* (pp. 91–116). New York, NY: Russell Sage Foundation.

Salvador, A., Suay, F., González-Bono, E., & Serrano, M. A. (2003). Anticipatory cortisol, testosterone and psychological responses to judo competition in young men. *Psychoneuroendocrinology*, *28*(3), 364–375. https://doi.org/10.1016/S0306-4530(02)00028-8

Sapolsky, R. M., Romero, L. M., & Munck, A. U. (2000). How do glucocorticoids influence stress responses? Integrating permissive, suppressive, stimulatory, and preparative actions. *Endocrine Reviews*, *21*(1), 55–89. https://doi.org/10.1210/er.21.1.55

Sauro, M. D., Jorgensen, R. S., & Pedlow, T. (2003). Stress, glucocorticoids, and memory: A meta-analytic review. *Stress*, *6*(4), 235–245. https://doi.org/10.1080/10253890310001616482

Schilling, T. M., Kölsch, M., Larra, M. F., Zech, C. M., Blumenthal, T. D., Frings, C., & Schächinger, H. (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. https://doi.org/10.1016/j.psyneuen.2013.01.001

Schlotz, W., Schulz, P., Hellhammer, J., Stone, A. A., & Hellhammer, D. H. (2006). Trait anxiety moderates the impact of performance pressure on salivary cortisol in everyday life. *Psychoneuroendocrinology*, *31*(4), 459–472. https://doi.org/10.1016/j.psyneuen.2005.11.003

Segool, N. K., Carlson, J. S., Goforth, A. N., Embse, N. von der, & Barterian, J. A. (2013). Heightened test anxiety among young children: Elementary school students' anxious responses to high-stakes testing. *Psychology in the Schools*, *50*(5), 489–499. https://doi.org/10.1002/pits.21689

Shirtcliff, E. A., Peres, J. C., Dismukes, A. R., Lee, Y., & Phan, J. M. (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. https://doi.org/10.1521/pedi.2014.28.1.40

Stalder, T., Kirschbaum, C., Kudielka, B. M., Adam, E. K., Pruessner, J. C., Wüst, S., … Clow, A. (2016). Assessment of the cortisol awakening response: Expert consensus guidelines. *Psychoneuroendocrinology*, *63*, 414–432. https://doi.org/10.1016/j.psyneuen.2015.10.010

Stroud, L. R., Salovey, P., & Epel, E. S. (2002). Sex differences in stress responses: Social rejection versus achievement stress. *Biological Psychiatry*, *52*(4), 318–327. https://doi.org/10.1016/S0006-3223(02)01333-1

US Census Bureau. (2017). U.S. Census Bureau QuickFacts: New Orleans city, Louisiana. Retrieved September 17, 2018, from https://www.census.gov/quickfacts/fact/table/neworleanscitylouisiana/INC110216#viewtop

Weekes, N., Lewis, R., Patel, F., Garrison-Jakel, J., Berger, D. E., & Lupien, S. J. (2006). Examination stress as an ecological inducer of cortisol and psychological responses to stress in undergraduate students. *Stress*, *9*(4), 199–206. https://doi.org/10.1080/10253890601029751

Figures

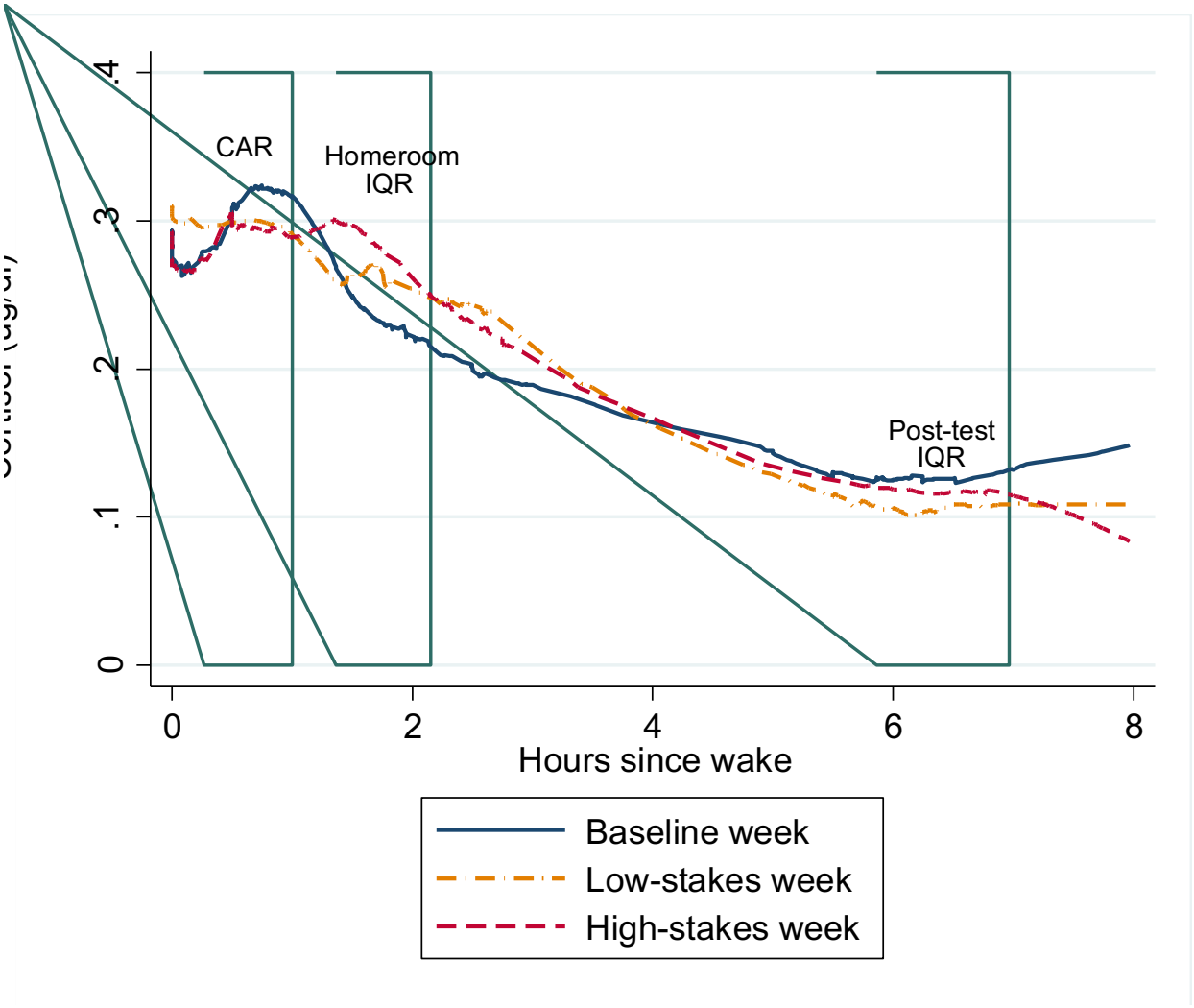

Fig. 1. Cortisol patterns from wake to eight hours post-wake for baseline, low-stakes, and high-stakes weeks using locally-weighted scatterplot smoothing to display the data, which does not impose parameters on the pattern. Boxes include the cortisol awakening response (CAR, 15-60 minutes post-waking) 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.

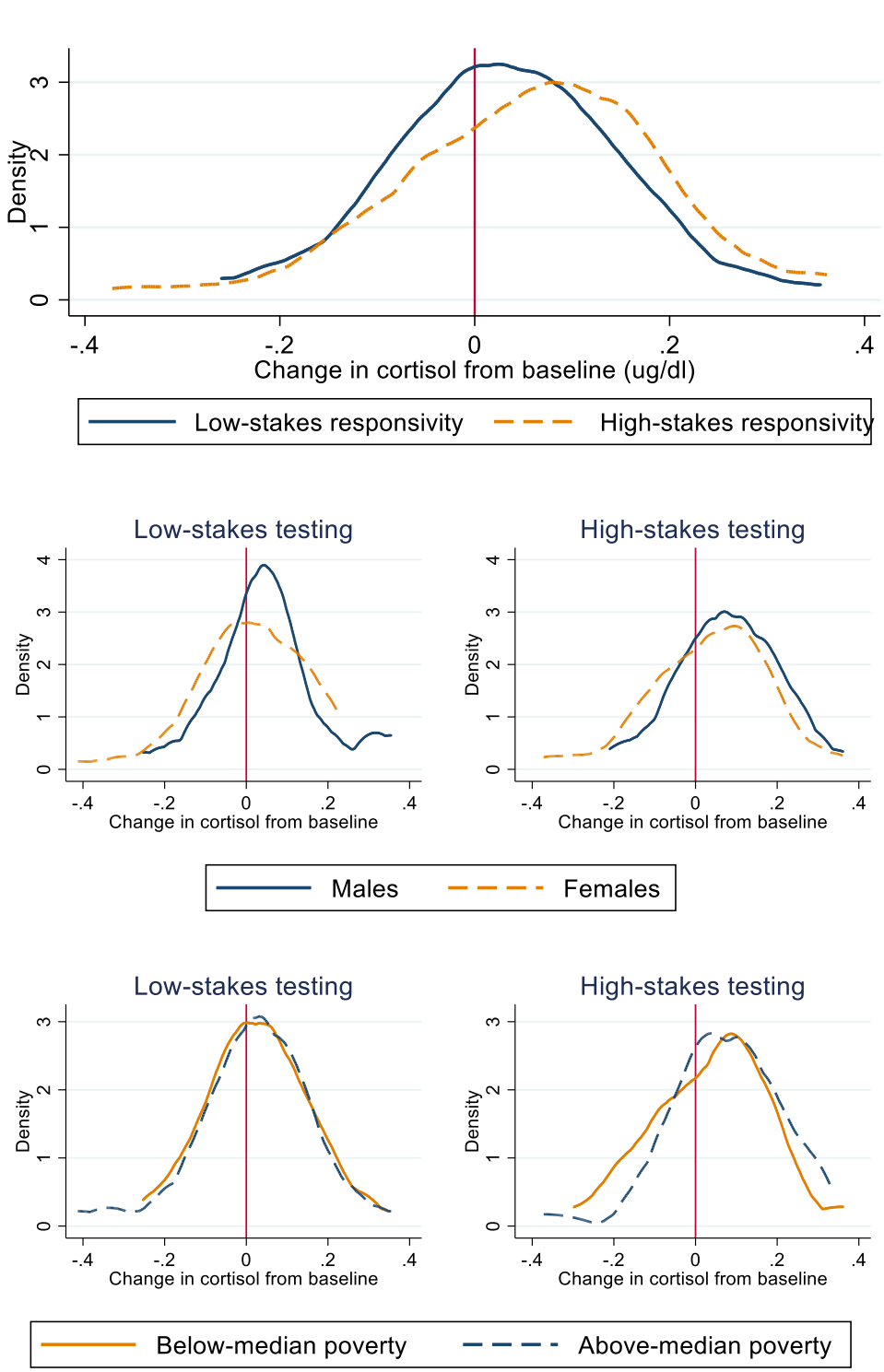

Fig 2. Distribution of the change ("responsivity") from baseline to the low-stakes testing week and from baseline to the high-stakes testing week. Includes estimates by gender and poverty.

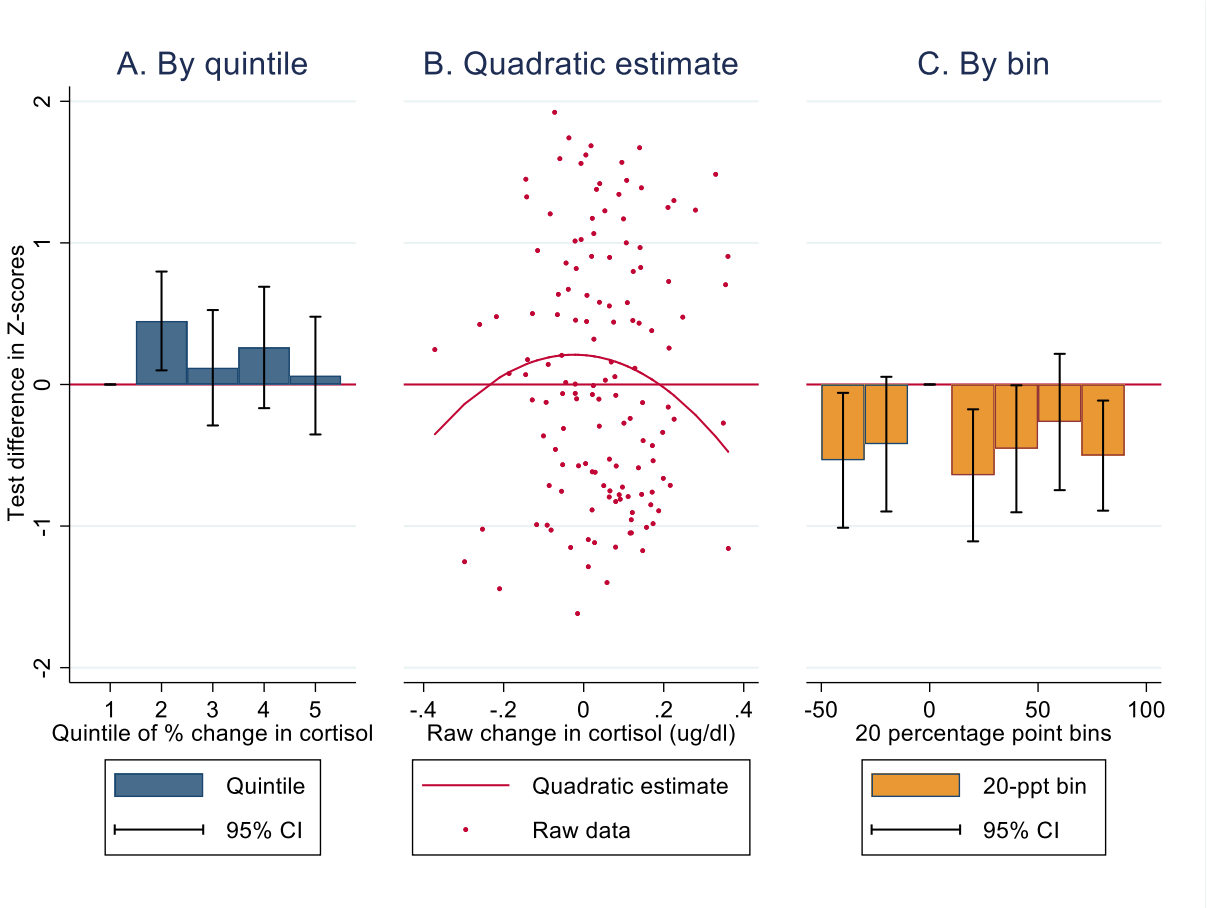

Fig. 3. Change in predicted mean Z-score (across math, science, and English language arts tests) on the high-stakes test by change in cortisol from the baseline to the high-stakes testing week. Models regressed mean Z-score on different ways to measure of change in cortisol. All models control for quarters 1-3 grades for math, ELA, 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 (-67.2 to -24.0%, N=13), Quintile 2 (-21.6% to +12.2%, N=14), Quintile 3 (+13.4 to +52.8%, N=14), Quintile 4 (+52.9% to +90.3%, N=14), and Quintile 5 (+106.4% to 455.1%). 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. Model also controls for concurrent cortisol and concurrent cortisol-squared. **Panel C** groups participants by percentage-change in cortisol. Bins grouped by decreases greater than 30% (N=11), -30 to -10% (N=5), reference group at -10 to +10% (N=8), +10 to +30% (N=8), +30 to +50% (N=6), 50 to 70% (N=12), and increases greater than 70% (N=18).

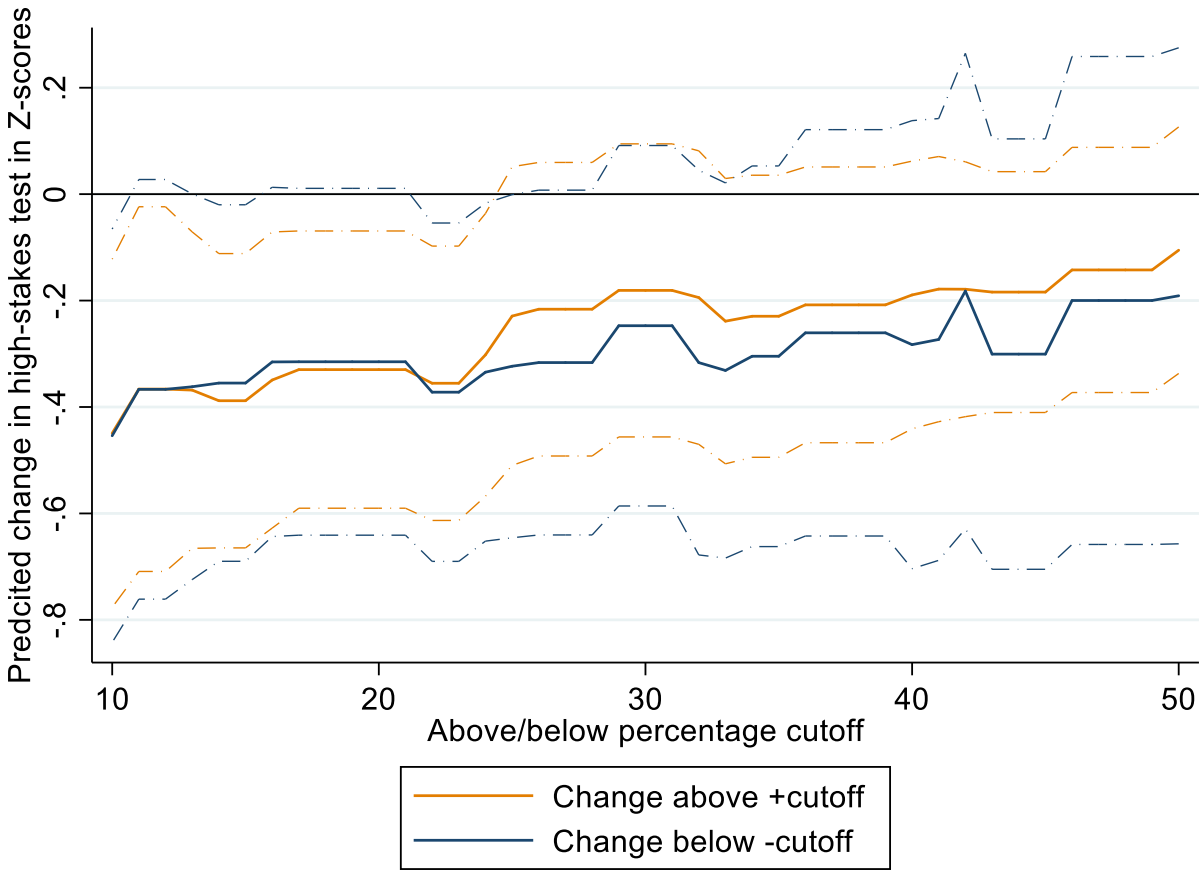

Fig. 4. Estimated effect size by different bounding distances (+/-10% to +/-50%). Each distance is a separate regression; two coefficients per regression displayed. Coefficients displayed are for a variable that is equal to 1 if the change from baseline to the high-stakes testing week is greater than the indicated level. N=68.

Tables

Table 1: Descriptive Statistics

	(1) Mean	(2) SD	(3) Min	(4) Max	(5) Count
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	136.22	121.47	0.00	531.00	85
Priority 1 911 calls within 0.25 mi of home	813.92	728.62	0.00	4291.00	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.

Table 2. Changes in level of before-testing homeroom period cortisol by week

	(1) All	(2) All	(3) By gender	(4) By poverty	(5) By local 911 calls
Low-stakes testing	0.112	0.095	0.299*	0.126	0.212+
	(0.077)	(0.076)	(0.121)	(0.136)	(0.111)
High-stakes testing	0.167*	0.147*	0.352**	0.263*	0.236*
	(0.073)	(0.073)	(0.123)	(0.112)	(0.114)
Low-stakes X female			-0.338*		
			(0.158)		
High-stakes X female			-0.349*		
			(0.150)		
Low-stakes X lower poverty				-0.078	
				(0.178)	
High-stakes X lower poverty				-0.219	
				(0.162)	
Low-stakes X lower crime					-0.257
					(0.178)
High-stakes X lower crime					-0.176
					(0.170)
Time of day	-0.018	0.151	0.150	0.303	0.332
	(0.565)	(0.572)	(0.565)	(0.620)	(0.633)
Time of day-squared	0.315	0.509	0.493	0.620	0.598
	(0.577)	(0.609)	(0.608)	(0.651)	(0.662)
Wake time		-0.131	-0.135	-0.141	-0.151
		(0.128)	(0.124)	(0.130)	(0.126)
CAR timeframe		-0.135	-0.127	-0.184	-0.154
		(0.167)	(0.158)	(0.183)	(0.182)
p(sum low-stakes testing=0)			0.685	0.642	0.730
p(sum high-stakes testing sum=0)			0.965	0.695	0.616
Observations	478	478	478	445	440
Participants	93	93	93	86	85

Robust standard errors clustered by student ID. Analysis conducted at the student-day level. + p < 0.10, * p < 0.05, ** p < 0.01, *** p < 0.001. Outcome is the natural log of cortisol. Data comes 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 3-5 conduct the analysis by interacting the test with indicator variables for the given group. Table includes p-values of the estimated difference in these groups for the change in cortisol for the low- and high-stakes weeks.

Appendix

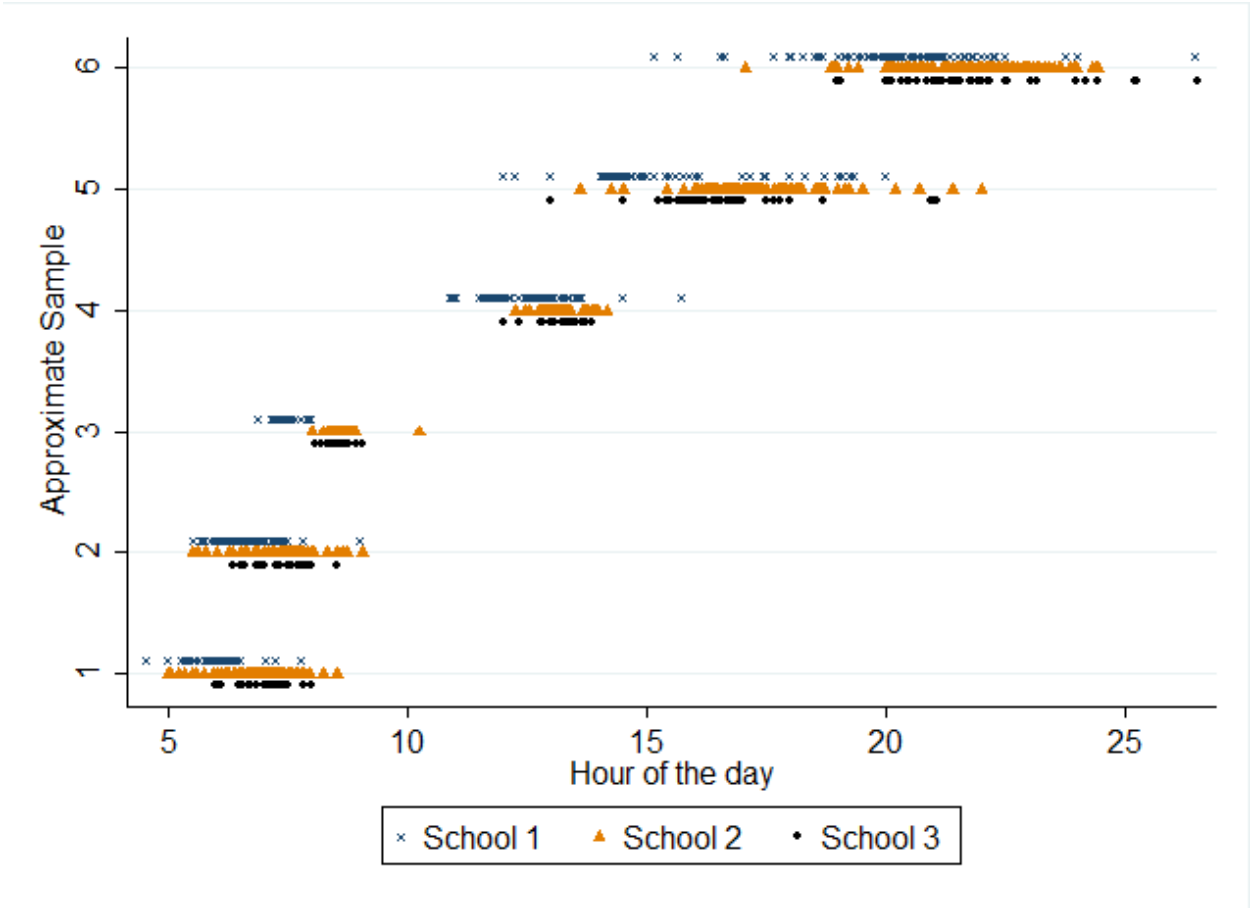

Fig. A1. Distribution of the timing of samples by collection period and school. Study protocol specified Sample 1 as wake; Sample 2 as wake+30 minutes, Sample 3 as before-test (homeroom), Sample 4 as after-test (before-lunch), Sample 5 as after school, and Sample 6 as bedtime. Research team supervised and verified timing for collection in Sample 3 and Sample 4; on-the-ground school needs meant that the timing of Sample 4 changed week-to-week. Sample 3 is the most consistently-timed sample. Times greater than 24 indicates a bedtime after midnight.

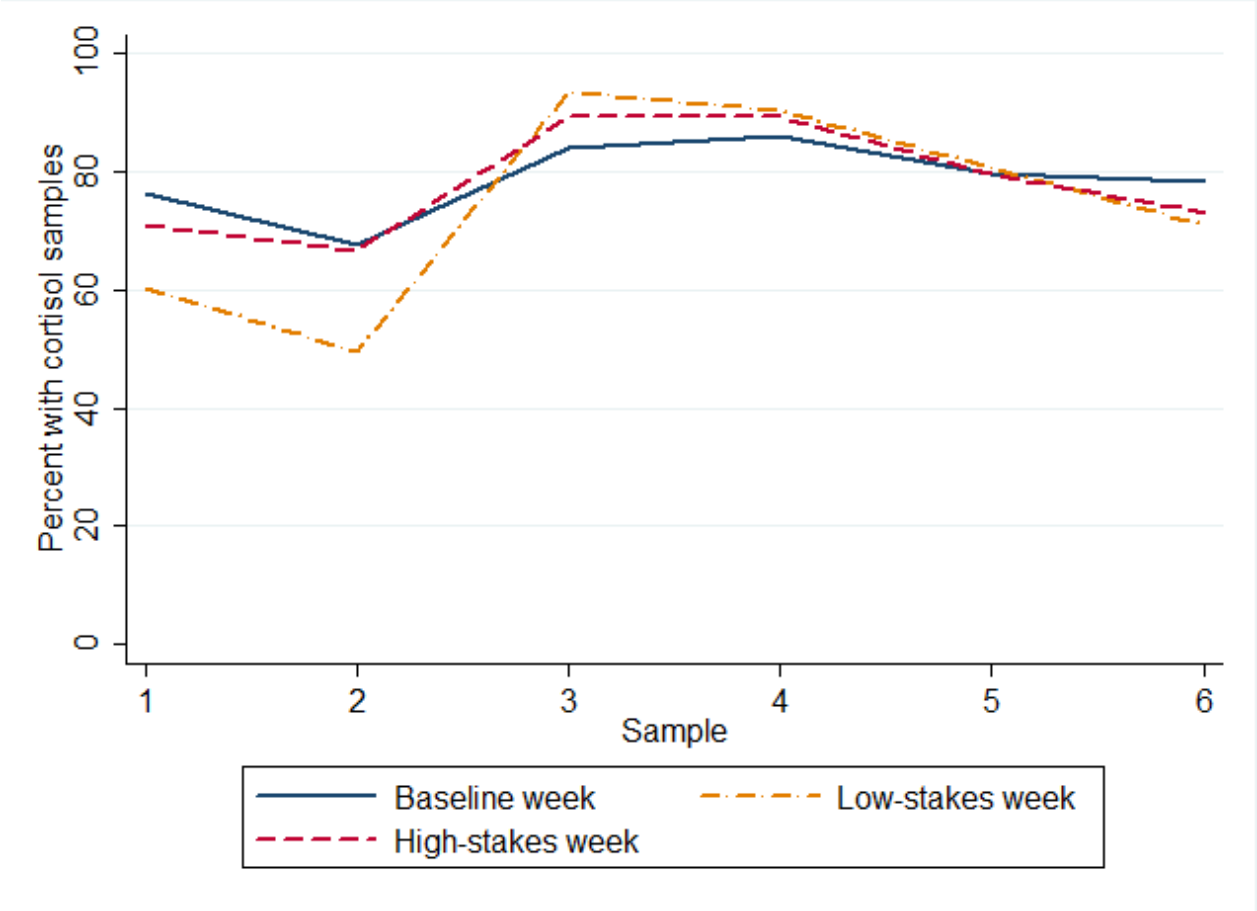

Fig. A2. Percent of participants with at least one sample by sample number and week. Study protocol specified Sample 1 as wake; Sample 2 as wake+30 minutes, Sample 3 as before-test (homeroom), Sample 4 as after-test (before-lunch), Sample 5 as after school, and Sample 6 as bedtime. Research team supervised and verified timing for collection in Sample 3 and Sample 4.