Rise and Shine: The Effect of School Start Times on Academic Performance from Childhood through Puberty*

Jennifer Heissel[†] and Samuel Norris[‡]

April 13, 2017

Abstract

We analyze the effect of school start time on academic performance. Sleep patterns are determined in part by sunrise times, which vary across time zones. Because school start times do not fully reflect this difference, we instrument for the hours of sunlight before school with the time zone boundary in Florida. We find that moving start times one hour later relative to sunrise increases test scores by 0.08 and 0.06 standard deviations for adolescents in math and reading, respectively. In math, the effect is larger for older children and co-varies with entry into an important pubertal stage. School districts can improve performance while maintaining the current distribution of start times by moving classes earlier for younger children and later for older children. (JEL I21, I28)

*We thank David Figlio, Jonathan Guryan, Ian Hodgson, Seema Jayachandran, Elizabeth Krasner, Matthew Notowidigdo, Matthew Pecenco, Krishna Pendakur, Max Tabord-Meehan and Elie Tamer for their thoughts and comments on this work. We also thank seminar participants at the University of Chicago and Northwestern University, as well as conference participants at the Association for Education Finance and Policy and the Association for Public Policy Analysis and Management. We thank Abigail Durgan for her research assistance. An earlier draft of this paper included data from Tennessee; additional thanks is due to the hundreds of Florida and Tennessee high school principals and administrators who responded to our data requests. Norris is grateful for financial support from the Social Sciences and Humanities Research Council of Canada through its Doctoral Fellowship Awards.

[†]School of Education and Social Policy, Northwestern University. jheissel@u.northwestern.edu

[‡]Department of Economics, Northwestern University. norris@u.northwestern.edu

1 Introduction

American teenagers are chronically sleep-deprived (Eaton et al., 2010). As children enter puberty, physiological changes delay the onset of sleep and make it more difficult to wake up early in the morning. By the end of middle school there is a large disconnect between physiological sleep patterns and school schedules: Hansen et al. (2005) find that students lose as much as 120 minutes of sleep per night after they start school in September, compared to the summer months when they can better control their own sleep schedules.

Sleep matters for learning and cognition. Important memory formation and consolidation processes occur overnight, as the brain replays patterns of brain activity exhibited during learning (Fogel and Smith, 2011; Maquet et al., 2000). Restricting sleep also reduces alertness and attention levels (Lufi et al., 2011; Sadeh et al., 2003), which likely affects students' ability to learn or take tests the next day. In light of these findings, the American Academy of Pediatrics recommends that adolescents wake up no earlier than 8:00 a.m. (2014). As of 2011, the median *start time* for American high schools was 8:00 a.m., suggesting that current policy may have cognitive costs for students.

Relatively little research has directly examined the effect of K-12 start times on academic performance. We study this question with a novel identification strategy that takes advantage of the biological effect of light on sleep patterns. Sleep timing is partially regulated by sunlight exposure; holding hours of darkness constant, more sunlight in the morning (and less at night) naturally moves bedtimes earlier and increases alertness in the morning (Crowley et al., 2007). Sunlight before school — as opposed to clock start times — is therefore the correct measure of policy when comparing between schools.[1] We expect that students exposed to more sunlight will improve their academic performance, and that this effect will be stronger for pubertal children because of their delayed sleep schedules (Carskadon et al., 1997). Our empirical strategy leverages the discontinuous change in sunrise times at a time zone border, combined with the fact that school start times do not fully adjust for this difference. Using a rich administrative dataset of all public school students in Florida between 2000 and 2013, we track children as they move across the Central-Eastern time zone boundary. Treating time zone as an instrument for sunlight before school, we identify the effect of start time relative to sunrise on academic performance conditional on student fixed effects and school characteristics.

We observe children moving across the time zone boundary at all ages between eight and fifteen, which allows us to estimate the age-specific effect of school start times over a range of develop-

[1] For any given school, clock start time is colinear with sunlight before school.

mental stages. An additional hour of sunlight before school has almost no effect on math scores for pre-pubescent children, but a large and abrupt effect appears for girls at age 11 and boys at age 13. This pattern corresponds exactly to the gender-specific median age of an important pubertal transition (Campbell et al., 2012), which we take as evidence that the causal pathway is linked to the physiological changes that occur during puberty. Specifically, a one-hour delay in relative start times increases standardized math scores by 0.081 standard deviations for adolescents, but only 0.009 SDs for pre-pubertal children. In reading, an extra hour of sunlight before school increases scores by 0.057 SDs for adolescents and 0.061 SDs younger children. The difference between groups is not statistically significant in reading, though the adolescent estimate is more precise and can be tested as different from zero. As children move over the time zone boundary, the change in scores occurs within a year of the change in sunlight exposure and persists over time.

Later relative start times do not increase learning time for adolescents, as measured by absences. Absences are reduced by 0.869 percentage points for younger children. Differences in how absence is measured across school types (elementary, middle and high schools) may be part of the reason behind the differences in outcomes we find here. We do not observe tardiness that does not result in an absence and therefore cannot rule it out as a causal channel, but our results are consistent with improved alertness and learning capacity as a result of later start times for adolescents.

We build on the current literature in two other ways. First, we provide evidence on whether improved achievement in high-morning-sunlight areas is a result of better learning throughout the year, or merely improvements in testing performance. Using variation in test timing over the sample years, we show that testing effects are unlikely to account for the math results. They may make up a portion of the gains from later start times in reading.

Second, we address a potentially important educational policy. Although moving start times later for all students would increase academic performance at a relatively low monetary cost (Jacob and Rockoff, 2011), interference with transportation and parental work schedules is a major concern for many districts. An alternative policy is to keep the same distribution of start times, but to adjust the opening order for schools in a way that is consistent with the physiological evidence: elementary schools, middle schools, and finally high schools. We show that most districts in the Florida panhandle do not follow this optimal pattern, but that the policy would increase math and reading scores by 0.06 and 0.04 SDs for high school students, with little negative effect for younger students. Although there may be other costs — in particular, young children might have to wait for the school bus in the dark — our paper is the first to quantify the academic benefits of this policy.

2 Background

2.1 Previous research

There have been several recent studies investigating the effect of daily start times on academic achievement, though none have examined the role that pubertal changes play in the effects. Wahlstrom et al. (1998) find that delaying school start times in Minneapolis public schools from 7:15 to 8:40 improved student sleep by 39 minutes and significantly decreased tardiness rates. Their measure of academic performance was teacher-assigned grades, where they found a positive but statistically significant effect.[2] A later paper by Hinrichs (2011) exploiting the same policy change finds no effect on ACT scores. Another approach is from Edwards (2012), who uses changes to busing schedules as a source of potentially exogeneous variation in start times. He finds evidence that delayed start times increase achievement for middle school students. The effect seems to be smaller for elementary students, but he notes that this may be a result of start times being much later for younger children in his sample. The results are not available by gender, which makes inference on the importance of puberty difficult. Finally, Carrell et al. (2011) study freshmen cadets at the United States Air Force Academy who were randomly assigned different school schedules, and who belonged to cohorts with different first-period start times. Using this random variation, they find that having a start time of 7:00 a.m. (versus no class in first period) decreases achievement by about 0.15 SDs in that class, and by about 0.10 SDs in subsequent classes.

2.2 Sunlight, sleep, and puberty

The role of sunlight in determining sleep schedules is well known. Sleep patterns are partially controlled by the circadian rhythm, which synchronizes to a 24-hour cycle using the daily variation in light and darkness (Crowley et al., 2007). In the morning, light on the outside of the eyelids suppresses production of the hormone melatonin and stimulates brain processes to increase alertness; darkness at night increases melatonin levels and feelings of tiredness (Arendt, 2000).

One of the most drastic and well-documented changes during adolescence is to the timing of sleep. As children move through puberty, nocturnal melatonin secretion is delayed several hours relative to adults and younger children (Carskadon et al., 1997, 2004). The result is that adolescent sleep patterns become more owl-like, with later bedtimes and wake times, even holding the level of darkness fixed

[2] Teacher-assigned grades may understate the effect of school-level interventions if teachers curve assigned grades within a given class and year.

(Carskadon et al., 1993, 2004; Crowley et al., 2007). Schools in the United States tend to begin early to accommodate after-school activities and parental work schedules, preventing adolescents from waking at their preferred later times and leading to an increasing disconnect between weekday and weekend sleep schedules during the school year (Jenni and Carskadon, 2012; Laberge et al., 2001). The result is low wakefulness and attention levels on school days (Lufi et al., 2011). More directly, sleep levels have large effects on cognitive performance (Sadeh et al., 2003; Walker and Stickgold, 2006).

Although boys and girls undergo similar sleep-related changes during adolescence, the age profile of puberty varies significantly by gender. Marshall and Tanner (1970) show that pubic hair development begins 1.5 years earlier for girls than for boys; there is a similar gap for attainment of other developmental thresholds. This variation in age at entry into successive pubertal stages generates an important testable prediction: if physiological changes are driving the increasing importance of school start times during high school, then the size of the start time effect will co-vary with the gender-specific entry into puberty. In contrast, other changes that might make start times more relevant to achievement — e.g., the transition to a block schedule, middle-school social pressures, or changes to after-school activities — likely affect both genders at the same age.

3 Identification strategy

Our goal is to estimate the causal effect of school start times on academic achievement and behavioral outcomes. One approach would be to regress outcomes on start times, but because start times are chosen by the policy-maker, this approach would generate upwards-biased coefficients if better-managed schools tend to also start later in the day.[3]

Instead, our identification strategy exploits the relationship between sunlight and sleep, along with variation in sunrise time between locations. The intuition is that sleep patterns are linked partially to sunrise and sunset times, rather than clock time. This means that in terms of student sleep and alertness, the policy-relevant measure of school start time is start time relative to sunrise. For a given school, this is an unnecessary distinction: the choice of when to start classes according to the clock is equivalent to deciding when to start classes relative to sunrise. Between schools in different locations, however, a given clock start time corresponds to different relative start times. This contrast is particularly stark at a time zone boundary. Suppose that there are two schools close together but on opposite

[3]Better schools may also start earlier; for example, they may start earlier to accommodate after-school activities. This fundamental uncertainty about the direction of the bias from OLS underlines the importance of good instruments in this context.

sides of the boundary, where the sun rises at 6:00 a.m. in Central Time (CT) and 7:00 a.m. in Eastern Time (ET). If both schools begin classes at 8:00 a.m. local time, students attending the school in CT will have one more hour of sunlight before the morning bell.[4] To translate this insight into credible estimates, we track academic achievement as students move between schools on different sides of the time zone boundary. As students move from CT to ET, they are exposed to less sunlight before school, which we expect will decrease academic achievement. Conversely, a student moving from ET to CT gains sunlight before school and should see their test scores increase.

Formally, we use the time zone as an instrument for the amount of sunlight before school, which we refer to as the relative start time. We then regress academic and behavioral outcomes on instrumented relative start time to estimate the causal effect of relative start times.

The exclusion restriction in this setting is that time zone is uncorrelated with other school and student characteristics that might also affect achievement. This assumption might not be realistic in certain contexts. If, for example, we regressed achievement on instrumented time zone for the entire state of Florida, our identifying assumption would be that the only difference between schools in CT and ET relevant to student achievement is variation in relative sunrise times. Even conditional on a robust set of controls, this assumption is unlikely to hold. Instead, we include a set of student fixed effects and identify the coefficients of interest using only within-student variation. This means that variation in our instrument comes only from students who move between time zones.

We relate outcomes to start times using the following functional form:

$$y_{it} = \delta_1 hours_{it} + \delta_2 hours_{it} \times \mathbb{1}[\text{puberty}] + X_{it}\beta + \gamma_i + \varepsilon_{it} \tag{1}$$

where y_{it} is the outcome of interest, $hours_{it}$ is the number of hours between sunrise and school start, X_{it} is a vector of controls and γ_i is an individual fixed effect. The first stage instruments for relative start time with an indicator for time zone $timezone_{it}$:

$$hours_{it} = \alpha_{11} timezone_{it} + \alpha_{12} timezone_{it} \times \mathbb{1}[\text{puberty}] + X_{it}\theta_1 + \eta_{1i} + u_{1it} \tag{2}$$

$$hours_{it} \times \mathbb{1}[\text{puberty}] = \alpha_{21} timezone_{it} + \alpha_{22} timezone_{it} \times \mathbb{1}[\text{puberty}] + X_{it}\theta_2 + \eta_{2i} + u_{2it} \tag{3}$$

[4]Children in CT will also have one less hour of sunlight after school. It is possible that this has an effect on academic outcomes, for example if less sunlight after school decreased sports participation and led to more homework time. As a policy matter, moving school start times later will always increase sunlight before school at the expense of sunlight after school; because we are interested in the effect of school start times as a policy we consider this a feature of our approach.

where η_i are individual fixed effects. The vector X_{it} typically includes longitude, which directly affects sunrise times, as well as school-level demographic controls to proxy for school quality.

Crucially, we allow the effect of start time to vary by pubertal status. Based on the biological evidence discussed in Section 2.2, we expect that students' natural sleep patterns will become more out-of-sync with their school schedule as they enter puberty. We therefore expect that δ_1 in Equation 1 will be positive because later start times likely increase performance for children of all ages, and that δ_2 will be positive to reflect the greater benefits of later start times for adolescents.

One potential concern with this strategy is that the vast majority of cross-boundary moves are over a great distance. Long-distance moves may be inherently disruptive and therefore have an independent effect on academic outcomes. We address this concern by including in our sample students who move schools, but *not* across the time zone boundary. These students identify a set of dummies for 1, 2, and 3+ years after the move, disentangling the effect of moving from the effect of moving across a time zone boundary.

4 Data

4.1 Academic outcomes

Our data come from Florida Department of Education (FDOE) administrative records for the fifteen school years from 1998-1999 through 2012-2013 (henceforth, 1999 through 2013). We exclude alternative schools, adult education centers, and virtual academies that may have non-standard start times. Our primary outcome of interest is individual-level scores on the annual Florida Comprehensive Assessment Test (FCAT) in math and reading; this test is considered 'high stakes' for students and schools. Students took the FCAT in math in grades 5 and 8 in years 1999 through 2000, grades 3 to 10 in 2001 through 2010, and grades 3 to 8 in 2011 through 2013. They took the FCAT in reading in grades grades 4 and 8 in 1999 through 2000 and grades 3 through 10 in 2001 through 2013. Scores are standardized by year and grade at the state level for each test, with a mean of zero and a standard deviation of one. In addition to the FCAT, the data include individual-level characteristics such as race, ethnicity, gender, free- or reduced-price lunch (FRL) eligibility, and absentee rates. We use student birthdays to calculate age at the start of the school year in September.[5]

The longitudinally-linked data allow us to follow students over time, as long as they remain within

[5]The FDOE uses September 1 as the kindergarten admission cutoff.

the Florida public school system. About 90% of students are matched year-to-year by social security number; the remainder are matched by name and birthday. This matching process is conducted by the FDOE and appears to contain a small number of errors caused by multiple students with similar names or birthdays. To account for this, we exclude students who move backwards more than two grades, fail and then skip a grade, have a change in birthday, are older than 15, or change gender from year-to-year. In total, these deletions amount to about 7% of the original dataset. We lose few students in the longitudinal analysis; among students who took the third grade FCAT before 2009, we observe 93% taking an FCAT the following year and over 80% taking an FCAT five years later.

We restrict the sample in two main ways to address possible threats to identification. First, we focus on the area near the time zone boundary. This reduces the likelihood that there are different economic trends on either side of the boundary, which could mean that moves in one direction were disproportionately induced by job loss. Parental job loss is often a stressor for children and may itself have a negative impact on academic achievement; this could bias our results in either direction. The area near the time zone boundary is known as the Florida panhandle, and is generally seen as distinct from the rest of the state.[6]

Second, we limit the sample to students who make a substantial move, which we define as consecutive appearances at schools further than 25 miles apart. This restriction is largely targeted at the within-time zone movers; we want to ensure that these students are subjected to something comparable to the disruptive, long-distance cross-time zone moves. The exact choice of 25 miles as the cutoff is admittedly arbitrary; in the Online Appendix, we show that the main results are similar when using 15, 20, or 30 miles as the cutoff, or defining a move as a change in school district.

Table 1 displays summary statistics for third-graders in the panhandle. Note that this is a subset of our main estimation sample; we do not require that we observe a student in third grade to include them in our main analysis. However, because we intend to show that test scores are directly affected by time zone through the start time channel, observed differences in test scores for older children are not informative about baseline characteristics. The third grade summary statistics in Table 1 are therefore as close to baseline summary statistics as is possible with our data, although there may already be some effect of differing relative start times.

Panel A presents school-level outcomes for all students in the panhandle (Column 1); for those

[6]The panhandle includes the following 19 counties: Bay, Calhoun, Escambia, Franklin, Gadsden, Gulf, Holmes, Jackson, Jefferson, Lafayette, Leon, Liberty, Madison, Okaloosa, Santa Rosa, Taylor, Wakulla, Walton, and Washington. The time zone boundary approximately bisects the area.

who move more than 25 miles (Column 2); and for those who move more than 25 miles between time zones, disaggregated by direction of move (Columns 3 and 4). Column 5 tests the difference between Columns 3 and 4. Movers come from nearly identical schools as non-movers on all dimensions. Comparing within cross-boundary movers, CT-ET movers come from fairly similar schools as ET-CT movers across most measures; two differences stand out as large and statistically significant. First, the schools in ET have a much larger percentage of black students. This occurs because most black students in our sample are from Tallahassee and its surrounding suburbs in ET. Second, the district-level third grade reading score of the cross-time zone movers' schools is 0.08 SDs higher in CT than in ET. This would be problematic for identification if it implied that underlying peer quality improves when students move from ET to CT. However, this pattern may actually be a *result* of later relative start times in CT, because these students have already been treated with four years of later relative start times in grades K-3. In contrast, peer covariates like FRL, which are less affected by sunlight levels, are more similar between time zones. As a precautionary measure, we control for some characteristics of the peer populations with demographic share controls in our main specifications. In the Online Appendix, we show that our results are robust to the inclusion of controls for peer mean test scores.

Panel B presents individual-level characteristics. The movers are quite similar to the overall panhandle population, which bodes well for external validity. Movers are 11 percentage points more likely to be FRL relative to the non-movers, but equally likely to be black. Their test scores are slightly lower than the non-movers (0.09 and 0.08 SDs lower in math and reading, respectively), possibly reflecting stress from the upcoming move or slightly higher poverty rates among movers.

The characteristics of cross-time zone movers who begin in CT and those who begin in ET are well-balanced in terms of demographic characteristics, although the third grade math score is an insignificant 0.06 SDs lower for the CT-ET movers. The CT-ET movers also have 1 percentage point lower absentee rates than ET-CT movers.

Overall, Table 1 tells us that the two different types of cross-time zone movers are similar but not identical in terms of third grade characteristics and those of the schools they attend. Equality of baseline outcomes is not strictly required for our identification strategy; we make only the difference-in-differences assumption that the unobserved changes in average achievement had the students moved at a different time (or moved but not been exposed to a different relative start time) be the same for both types of mover. There are two main ways that this could be violated: if the ET-CT movers are on a different trend than the CT-ET movers, or if there are different changes in school quality over the move for different mover types.

The patterns of achievement in the years before the move provide evidence on the similarity of the underlying trend for each of the mover groups. Figure 1 displays pre-move trends for four types of movers — two within a time zone (CT-CT and ET-ET) and two across (CT-ET and ET-CT) — estimated from a regression of test scores on the number of years until move interacted with mover type. We include a vector of controls[7] and a fixed effect for the period preceding a move for each student. The year before the move is the excluded category. The Figure shows that the trend for each mover group is similar: in both math and reading, the test scores for each group are statistically indistinguishable from each other during the pre-move period. Time until move is also not a very strong predictor of academic achievement; for all but two of the group-time combinations, we cannot reject that there is no difference in achievement between that year and the year immediately preceding the move. This suggests that the groups are on similar underlying trajectories, and that variation in post-move outcomes can be attributed to changes in sunlight before school, rather than differential trends.

One slightly surprising finding is that math scores trend upwards for all groups in the years before the move. Long-distance moves are often a result of parental divorce or job loss, which may occur several years before the move actually takes place. Because both of these events can increase stress levels for children, it might be expected that in the absence of controls, test scores would decline leading up to a move. In the Online Appendix we confirm this intuition; in a version of the same Figure without controls we show that both math and reading scores unconditionally decline in the years before a move. Although we prefer the version with controls to maintain comparability with our main results, the substantive conclusion in both cases remains the same: there are no large differential trends that would threaten our identification strategy.

Another violation of our exclusion restriction would arise if school or neighborhood characteristics changed dramatically over the move. In Appendix Table A4, we present evidence that changes in these characteristics are unlikely to drive our results. Taking the year before and after each move, we regress school characteristics on a set of student-move dummies and a dummy for each of the four types of move. Relative to the schools they started in, CT-ET movers move to schools with 4.5

[7] We include all controls from our baseline regressions, which we discuss more in Section 5. They include age-gender dummies, longitude, and school-level demographic means (male, FRL, black, Asian, and Hispanic). The longitude and demographic coefficients are identified from small deviations in school location and school demographics in the years before the move, but have no substantive effect on the coefficients of interest. We include them for comparability with our main regressions.

percentage points fewer FRL students, 14.0 percentage points more black students, and a median zip code income $5,700 higher (ET-CT movers see approximately the opposite changes). In the absence of any other intervention, this might actually raise achievement for CT-ET movers given the strong relationship between average income and school quality, when in fact we see the opposite.

4.2 Imputing puberty

We do not directly observe the onset of puberty, and instead use data from the National Health and Nutrition Examination Survey (NHANES) to impute developmental stage by age and gender. NHANES is a nationally representative sample of US children ages 8 to 19, and includes information on Tanner Stage, a 1-5 scale of pubertal development based on pubic hair. We use the median age of entry into Tanner Stage 3 as our cutoff for adolescence, as changes in sleep patterns occur after the acceleration of pubertal development during Tanner Stage 3 (Campbell et al., 2012).[8]

Figure A2 in the Online Appendix displays the cumulative share of children who have reached Tanner Stage 3 by gender and age; the median age of entry occurs at 11 for girls and 13 for boys. We use these ages as the start of puberty in our analysis.

4.3 School start times

We define school start time as the start of the first class where learning takes place; this excludes homeroom and breakfast. Data were mostly available on school websites, and we followed up by phone with all remaining schools.

We did not collect information on historical school start times, which change with some regularity according to the school principals we spoke with while conducting the survey.[9] Given the identification strategy, our estimates will be consistent if there has been no change in the average start time for each time zone over the study period.[10] We believe that this condition is likely met: although there has been some recent discussion of school start time policy in the popular press, most of our data is from before this conversation reached the mainstream. Furthermore, the debate has never touched on

[8] A second version of the Tanner Stage uses genital and breast development to demarcate stages. We use the pubic hair definition because the scale is more closely associated with pubertal changes in sleep patterns (Campbell et al., 2012), although using the alternate definition does not substantively change our main results. Using pubic hair Tanner Stage 2 or 4 changes the precision but not the direction of our results. Full results are available in the Online Appendix.

[9] This means that any attempt to estimate Equation 1 by OLS would result in attenuated coefficients due to measurement error on the right hand side.

[10] Under a more restrictive linear relationship between achievement and start times, we require only that there has been no change in the difference in start times between the two time zones.

whether early start times are more onerous for students with a later sunrise time.

School start times range from 7:00 a.m. to 9:30 a.m. local time. The average start time is 8:10 a.m., and the median is 8:00, which is similar to the national average (NCES, 2012). There is some heterogeneity with age: the median elementary school student starts school at 7:55, the median middle schooler at 8:25, and the median high schooler at 7:50. Nationwide, it is common to have high schools start earlier than the other schools in the district, so these broad patterns are not surprising.

We use NCES school location data to calculate sunrise times for each school. Combining these with our school start time data, we average the difference over the school year before the testing date to construct a measure of relative start time, measured as the number of hours between sunrise and school start times.

5 Results

5.1 First stage

Our first stage is predicated on the idea that although school start times may differ across the time zone boundary, they do not do so enough to erase the one-hour difference in sunrise times. Figure 2 plots the hours of sunlight before school, or relative start time, in the years before and after a move for each of the four groups of movers. We estimate each point from a regression of relative start times on time relative to move for each group as well as an individual-move fixed effect and controls for longitude and school demographics. The year before the move is normalized to be zero; we adjust the level of the coefficients with the group mean of relative start times for one year before the move.[11] There are three important takeaways. First, students in Central Time have more sunlight before school than those in Eastern Time, as expected. Second, the cross-time zone movers neatly switch places as they move across the time zone boundary: the cross-time zone movers are now 'treated' with the start time of the other time zone. This shift allows us to identify the effects of start time relative to sunrise using only within-student variation. Third, the lines generally overlap within time zones, indicating that those who switch time zones are likely not selecting into schools in a way that affects sunlight before school.

More formally, Panel A of Table 2 presents the first stage regression of relative start times on time zone.[12] The first row displays the main effect for all students, and the second row displays the

[11] A version of this graph with unconditional means for each group-time bin shows similar patterns.

[12] The Online Appendix includes robustness checks using additional controls including urbanicity, log income, school

interaction effect for pubescent students. The third row is the p-value from a test for the combined significance of the effect for pubescents. Each specification includes individual and age-gender fixed effects. Column 1 has no additional controls. Column 2 adds longitude.[13] Columns 3 and 4 add demographic means at the district and school level, respectively. These demographic means include the percentage of students who are male, FRL, black, Hispanic, and Asian. Columns 5 through 7 are identical to Columns 2 through 4, but with the addition of indicator variables for 1, 2, and 3+ years after the move to account for potential disruption.[14]

All specifications yield similar estimates. We prefer Column 7 because it includes controls that address both disruption and potential changes in peer characteristics over the move. Across the columns, younger children in ET have about 25 fewer minutes of sunlight before school than children in CT, while those who have gone through puberty have about a 41 minute difference. The difference is less than 60 minutes for each age group, which is what we would expect if schools opened at the same clock time on either side of the time zone boundary. We take this as evidence that policymakers faced with later sunrise times may shift start times later to compensate, and that they may differentially shift elementary start times to prevent younger students from waiting for the bus in the dark.[15] The F-statistics for the first stage range from 825 to 2004, with an F-statistic of 1105 for our preferred model.

5.2 Effect of start times on academic achievement

Panels B and C of Table 2 contain estimates for the effect of relative start times on math and reading test scores. Each specification includes individual fixed effects and age-gender dummies,[16] and the columns add additional controls in the same order as Panel A.

size, student/teacher ratio, and other levels of demographic aggregation. The results are similar to Table 2.

[13]We also consider adding latitude as a control. However, our study area has a relatively small north-south dimension — from the top to the bottom of the panhandle, the difference in average sunrise time over the school year is less than a minute. When we include latitude as a control, the main results are very similar but slightly smaller in magnitude. These robustness checks can be found in the Online Appendix.

[14]We consider specifications that control for the time until the move. This has almost no effect on the other coefficients in both the first and second stage, but we do not pursue this avenue to avoid controlling for information that the students may not have themselves.

[15]When we look at results by age, the difference in sunlight before school is 22-23 minutes for elementary school students (typically ages 8-10 in our data), 28-30 minutes for middle school students (ages 11-13), and 47-59 minutes for high school students (ages 14-15).

[16]Test scores are normalized at the year-grade level, so if we included the entire state population the age-gender dummies would reflect only the age-varying gender gap. Because our sample is restricted to movers in the Florida panhandle, there may be additional age-varying differences relative to non-panhandle and non-mover students that the age-gender fixed effects pick up. They are particularly important to include because they function as a set of saturated dummy variables for puberty, which we interact with start time as a explanatory variable of interest.

In Panel B, the estimated effect of relative start times on math scores is similar after we add a control for longitude in Column 2. In all subsequent specifications, moving start times one hour later increases math scores for prepubescents by 0.009-0.020 SDs; none of the coefficients are close to statistically significant. For adolescents, later start times increase math scores by 0.077-0.084 SDs. Across specifications, both the adolescent level and the difference between adolescent and prepubescent scores is significantly different from zero at the 1% level.[17]

Panel C repeats the exercise for reading. The results are again consistent across the columns; in our preferred specification moving start times one hour later increases reading scores by 0.061 SDs for prepubescent students and by 0.057 SDs for adolescents. The overall effect for adolescents is statistically significant at the 1% or 5% level for all specifications, while for prepubescents it is either significant at the 5% or 10% level depending on the level of aggregation for the demographic controls. There is no statistical difference between pubertal and prepubertal effects. For adolescents, the effect size is larger in math than in reading across specifications, corroborating previous research on middle schoolers (Edwards, 2012; Ng et al., 2009).

5.3 Mechanisms

There are (at least) two reasons why school start times might affect academic achievement. First, later start times relative to sunrise may make it easier to get to school on time, reducing absences and increasing time spent on instruction. Alternatively, more sunlight before school may improve cognitive function by increasing sleep levels and alertness.

Panel D of Table 2 explores the relationship between start times and absences. Conditional on school or district level demographic controls, there is no statistically significant relationship between start times and absence rates for adolescents, although there is an estimated 0.9% decrease in absences for the younger students in the preferred specification. For all ages, later relative start times decrease absences, although the relationship is weaker for adolescents than for prepubescents, which is difficult to reconcile with the larger effects of start times on achievement we observe in math and reading. Comparing between age groups is somewhat fraught; because record-keeping is not standardized across schools, an elementary-aged child might be marked absent for the entire day when she is late in the morning, but a high schooler who is similarly late could be marked absent only for the first

[17] The difference in effect size by pubertal stage is striking, and corresponds with increasing sensitivity to start times during puberty. In Online Appendix Table A6 we estimate a version of Table 2 without the interaction. The average effect of start times on achievement is close to the average of the adolescent and pre-pubertal measures; the reading estimates are statistically significant but the math estimates are only sometimes statistically significant.

class but not as absent in the larger tracking system. However, that caveat addresses only differences between the age groups; in light of the moderate and imprecisely estimated effects on absences for all age groups we think it is unlikely that reductions in absences are a major causal channel through which later relative start times translate into improved test scores.[18]

The evidence is somewhat stronger in favor of sleep and alertness as the causal channel. Our data do not contain information on sleep, so we use the Child Development Supplement (CDS) of the Panel Study of Income Dynamics (PSID) to estimate the effect of the time zone boundary on sleep. The CDS collected time use diaries for students in 1997, 2002, and 2007, along with geographic and demographic information. We regress hours of sleep on a dummy variable for residence in ET for children within 400 miles of the CT-ET boundary.[19]

Table A7 in the Online Appendix shows that prepubescent children in ET get 6 minutes less sleep per night during the week than children in CT.[20] The difference in sleep is reversed on the weekend as they attempt to correct the sleep deficit; students in ET sleep 4 minutes *more*. After the onset of puberty, both gaps widen: children in ET get 17 minutes less sleep per night during the week, and compensate with 13 minutes more sleep per night on the weekend.

These findings indicate that children in ET are more sleep-deprived than children in CT, and that this gap increases in adolescence. If school start times in our Florida sample are representative of start times elsewhere, this suggests a passthrough from relative school start times to sleep of 40-50%, which is comparable to the 46% found by Wahlstrom et al. (1998). Thus, moving from ET to CT increases both sleep and test scores (and increases them more for adolescents), suggesting that levels of sleep and alertness in the morning are important causal channels through which later school start times increase achievement. There may be other changes in time use — descriptive research indicates that later start times decrease time spent on extracurricular activities, as well as reduce leisure time for girls and computer use for boys (Groen and Pabilonia, 2015; Wahlstrom et al., 1998) — but it is difficult to reconcile the patterns of achievement by developmental status with an explanation *not* revolving around the transition to puberty. More importantly, from the perspective of a policymaker the distinction is moot: whether the causal channel is before-school time or after-school time, changing the school start time will affect both channels.

[18]We do not have data on tardiness, which could also be affected by start times.

[19]The publicly-available CDS does not geocode individuals at a sub-state level, so we exclude all observations from states with multiple time zones — including Florida. See the Online Appendix for more information on sample construction.

[20]All estimates reported here include demographic controls; see Column 2. We conservatively cluster by state. The difference in sleep between children in ET and CT is statistically different for adolescents but not for prepubescent children.

5.4 Heterogeneity by age and gender

Rather than allowing the effect of relative start times to vary by pubertal status as in Equation 1, it is possible to estimate each age-gender-start time interaction term separately. If the increasing importance of start times for math performance is a function of puberty, the effect sizes should grow in importance as a larger share of the gender enters puberty. This is precisely what we see.

Figure 3 presents coefficients from a version of Equation 1 estimated separately by gender, with start time fully interacted with age. Because ages range from 8 to 15, this amounts to estimating

$$y_{it} = \sum_{a=8}^{15} \delta_a h_{it} \times \mathbb{1}[\text{age}=a] + X_{it}\beta + \gamma_i + \varepsilon_{it} \tag{4}$$

where $h_{it}\mathbb{1}[\text{age}=a]$ is instrumented by time zone interacted with age, and X_{it} is the baseline vector of controls. Starting in the upper left corner of Figure 3, there is a sharp spike in the effect of school start times on math scores at age 11 for girls, precisely when the median girl enters Tanner Stage 3. The effect of later school start times is statistically significantly different from zero for girls 11-13, but not for girls 10 or younger. For boys, in the upper right corner, the effect of start times on math scores is statistically indistinguishable from zero at the 10% level for ages 8 to 12, then jumps from 0.049 to 0.096 at 13 as the median boy enters Tanner Stage 3. The effect of start times is significantly different from zero at the 1% level for ages 14 and 15. This is evidence that the increasing importance of start times with age is driven by pubertal entrance, rather than other academic or behavioral changes.

The effect of start times on math scores is noticeably (though insignificantly) smaller for girls after age 13. One possible explanation is that certain stages of puberty are particularly important for sleep (Campbell et al., 2012), and girls have moved beyond this developmental stage by age 14. For example, Crowley et al. (2007) speculate that older adolescents may be less responsive to light than younger adolescents. However, there is no firm physiological evidence on sleep patterns or light sensitivity at a granular gender-age level, so resolution of this issue will have to wait for data which extends further into adolescence, especially for boys. There is persuasive evidence from Carrell et al. (2011) that start times have a large effect on achievement for college freshmen cohorts that include both boys and girls, so we think it is unlikely that the true effect is zero for 14 and 15 year old girls.

In reading, as one might expect from Table 2, there is no sharp change in the relationship between start time and achievement at the gender-specific puberty thresholds.

5.5 Heterogeneity by subgroup

Educational interventions often have a larger effect on disadvantaged students or students attending low-resource schools (see, e.g. Krueger et al. (2001)). In this case, however, there are more similarities than differences in effect sizes across racial, economic, and gender groups. The standard errors are large, but the results suggest that changes to start times will benefit all students, rather than certain demographic groups.

In Table 3, we apply our baseline regression of test scores and absence rates on start times for each of six demographic subgroups: whites and minorities;[21] FRL and non-FRL; and male and female. In math, the effect sizes are similar between white and minority students in Columns 1 and 2. For pubescents, a one-hour delay in relative start times increases math scores by 0.093 SDs for white students and 0.081 SDs for minority students. In reading, the effect sizes are 0.040 and 0.132, respectively, though this difference is not significant. None of the estimated effects for absences are statistically significant.

Columns 3 and 4 contrast FRL and non-FRL students. The effect size for math scores is significantly larger for the non-FRL adolescents at 0.147 SDs per hour, compared to 0.048 SD for FRL adolescents. There are no statistically significant differences for reading scores or absence rates.

Finally, Columns 5 and 6 indicate that the effect of relative start times on achievement and absences is similar for boys and girls. The difference is never statistically significant, and the effect sizes for both groups are similar to the overall estimates of Table 2.

5.6 Persistence of start times

To this point, we have not distinguished between a transitory and permanent effect of start times on academic achievement. This distinction could be important. If changing school start times from one year to another has an effect for (say) only one year while the student adjusts her sleep schedule, our estimates (which are essentially the average of achievement before and after the move) would overstate the long-term effect by averaging a positive effect in the first year with a zero effect in all other years. This would mean that our estimates would not correctly predict the long-term change in achievement as a result of changes in start time policy. We explore this possibility in Table 4, where we estimate a version of our baseline regression with relative start time by pubertal status interacted

[21] We count all non-white students as minorities. These results are not substantively affected by not counting Asians as minorities, or delineating the categories as black and non-black. In the latter case, however, the standard errors for the black sample are large.

with dummies for 1, 2 and 3+ years since move. Note that the pubescent effect is the total estimate for adolescents, rather than an interaction.

The results indicate that the short-term and long-term effects are quite similar; for prepubescent children the long-term math and reading coefficient is an insignificant 0.005-.011 SDs smaller. For adolescents, the math effect is 0.020 SDs lower in the long run; the difference is significant at the 1% level. The reading effect is 0.010 SDs higher in the long run; the difference between the short and long run is not statistically significant. In the long run, the effect is larger for adolescents than younger students in both subjects, although the difference is not statistically significant in reading. In both the short and long run, the adolescent effects differ from zero. We conclude that changes to start times improve math and reading achievement within a year of the change in sunlight exposure for adolescents, and the effects largely persist over time.

5.7 Learning versus testing

The positive effect of later relative start times on test scores has two potential causes: improved learning in the year leading up to the test, or better testing performance caused by increased alertness on the day of the test. Our approach so far has been to estimate the combined effect of learning and testing. Fully disentangling the two effects would require separate instruments for start times during the year and on the day of the test, which are unavailable in our data.

The data allow us to answer a related but less definitive question: does the relationship between sunlight and achievement vary with the amount of baseline test-day sunlight, holding sunlight during the school year constant? If so, this implies that changes to test-day relative start times matter for achievement. Estimates of the marginal effect of later relative start times at different levels of test-day sunlight can be combined with a mild assumption of diminishing returns to sleep to generate a lower bound on the size of the test-day start time effect.

This strategy is possible in our context because our data contain variation in test-day relative start time that is separate from the cross-time zone variation in start times. During the study period, testing dates moved from late February to mid-April. This changed levels of sunlight on the day of the test, but had only a small effect on average sunlight levels during the school year when learning occurred. Using these policy changes, we find that the lower bound on the test-day effect is relatively high for reading, but low for math. We interpret this as evidence in favor of potential testing effects in reading, but not as a definitive rejection of testing effects in math.

During the study period, the FDOE pushed the testing period later in two discrete steps. The first change was particularly useful for this research, because it moved the testing period from before to after the start of Daylight Saving Time. DST begins with a time change on the second Sunday of March in most of the United States.[22] Clocks "spring forward," moving sunrise one hour later and reducing the amount of sunlight before school. Figure 4 charts sunlight before school for 2000-2007, 2008-2009, and 2011-2013, corresponding to the three test-day policy eras.[23] In 2000-2007, testing took place just before the change to DST, meaning that there was a relatively large amount of sunlight before school; in ET, the average was 1 hour 20 minutes on the first day of testing. For 2008 and 2009, the test was moved two weeks later to directly after DST; the average amount of sunlight before school on the morning of the test in ET dropped to 28 minutes. In 2011, the test was moved one month later, increasing sunlight before school on the testing day to an average of 1 hour 9 minutes for 2011-2013.[24] Throughout the study period, the average sunlight before school in the school year leading up to the test barely changed, at 61, 56, and 59 minutes, respectively. Based on these differences, we group together 2000-2007 and 2011-2013 into a "late test time" treatment, and 2008-2009 into an "early test time" treatment.[25] As the testing date was moved back, preparation time increased for all students; however, because the early test time treatment occurred in the middle of the period (when the testing date was closest to the DST transition) the average preparation time is only five days longer for the late test time treatment group. Furthermore, neither of the changes in testing date correspond to any major changes in testing procedure or curriculum we could find, suggesting that any differences in performance between the policy eras can be attributed to test-day sunlight.

It is tempting to estimate the effect of earlier relative start times on the day of the test by regressing test scores on a dummy variable for the testing era. However, test scores are standardized by the mean statewide score in each grade-year, so direct comparisons between years are not possible. We instead test whether the effect of full-year relative start times on achievement changes depending on test-day sunlight. We estimate a second stage of:

[22] There have been changes in DST dates in the recent past; before 2007 DST started on the first Sunday of April. This change is not relevant for this research, because testing occurred before DST began in all years before the switch in DST dates.

[23] Specifically, the Figure shows 2007, 2008, and 2011, but all are archetypes of their eras.

[24] We exclude 2010 from analysis in this section because DST occurred during the testing period in this year, meaning that we cannot assign the test to either pre- or post-DST. We also exclude 1999 because testing occurred one month earlier, in the first week of February, where the sunrise time is between the early and late period.

[25] The main difference between 2000-2007 and 2011-2013 is that the average relative start time in the year preceding the test was slightly earlier in 2011-2013 because the extra month of class time was almost entirely after the DST transition. Excluding 2011-2013 from the regressions does not change our conclusions.

$$y_{it} = \phi_1 hours_{it} \mathbb{1}[\text{child} \cap \text{late test time}]_{it} + \phi_2 hours_{it} \mathbb{1}[\text{child} \cap \text{early test time}]_{it} +$$
$$\lambda_1 hours_{it} \mathbb{1}[\text{puberty} \cap \text{late test time}]_{it} + \lambda_2 hours_{it} \mathbb{1}[\text{puberty} \cap \text{early test time}]_{it} + X_{it}\beta + \gamma_i + \varepsilon_{it} \tag{5}$$

where X_{it} includes, in addition to the usual controls, dummies for the policy eras and their interaction with puberty.

Because sunlight before school during the year leading up to the test is nearly identical between eras, the difference in coefficients for a given age group represents the change in the effect of one extra hour of test-day morning sunlight on test scores between two different margins: 1 hour 17 minutes from sunrise (the average in the late testing years) and 28 minutes from sunrise (the average in the early testing years). If the coefficients are the same, that implies either that the effect of test-day sunlight is identical at the two margins, or that the effect of test-day sunlight is zero.[26] If they are different, that implies there is some effect of testing day sunlight on at least one of the margins. A smaller coefficient in the late-testing years is consistent with diminishing marginal returns to test-day sunlight.[27] Analogously to the main specification of Equation 1, we expect that $\lambda > \phi > 0$, since later start times should improve performance more for adolescents than for younger students.

Table 5 presents our findings for math and reading. Unlike the main table, the coefficients estimate the full effect for adolescents, rather than the difference between adolescents and younger children. We begin by verifying in Columns 1 and 3 that excluding 1999 and 2010 does not substantively affect our baseline results.

Columns 2 and 4 estimate Equation 5, allowing for a differential effect of start times on achievement as a function of baseline test-day start times. In Column 2, the math results are unchanged from our main specification: moving relative start times one hour later increases achievement at a similar rate in the two eras for adolescents (0.096 SDs per hour in the early versus 0.095 SDs in the late era), and the difference in estimates is statistically insignificant. Because we argue there should be diminishing marginal returns to more sunlight before school, we take the similarity in estimates be-

[26]The latter implication is technically a subset of the former, but the conceptual difference is important.

[27]Formally, this can be seen by modeling outcomes y as an additive function of full-year and test-day sunlight, $y = f_{year}(t_1) + f_{test}(t_2)$. We estimate $\beta_{early} = f'_{year}(56m) + f'_{test}(1h\ 17m)$ and $\beta_{late} = f'_{year}(56m) + f'_{test}(28m)$, where f' is the first derivative. Then, $\beta_{early} - \beta_{late} \approx f'_{test}(1h\ 17m) - f'_{test}(28m)$, so a positive difference is evidence for diminishing marginal returns. Any non-zero difference implies that the function relating performance and testing has a non-zero effect at (at least) one of the margins.

tween different test-day sunlight eras as evidence against test-day effects in math.[28] For children, the results are slightly more suggestive of testing effects, with larger effects for more sunlight on the test days with less sunlight before school (at 0.071 SDs per hour) than on the test days with more sunlight before school (at 0.022 SDs per hour). However, neither estimate statistically differs from zero, nor do they differ from each other.

In reading, the results are more strongly suggestive of testing effects. For younger children, one extra hour of morning sunlight increases test scores by 0.096 SDs in years with less sunlight before school (early years), while the effect is statistically insignificant and only 0.049 SDs in years with more sunlight before school (late years). For adolescents, the effect during the relatively earlier testing era is 0.104 SDs per hour of sunlight, compared to 0.045 SDs in the late era. The difference in estimates is statistically significant for adolescents, suggesting that test-day sunlight may be important for reading achievement. Under the assumption that changes to test-day relative start times do not change the effect of start times during the school year, and that there are decreasing marginal returns to later test-day start times, this indicates that the test-day effect is bounded at a minimum of 0.059 SDs per hour for adolescents (calculated as 0.104-0.045) and 0.047 for prepubescents (0.096-0.049) in the early start time years. This bounded effect implies that testing is a more important causal channel than learning for reading achievement.

There is, however, one important reason why the result in reading should be taken with some caution. In both of the early-testing years, the testing period began almost immediately after the switch to DST; one day after in 2008 and two days after in 2009. Because clocks move forward during the spring DST transition, students can lose up to an hour of sleep, depending on how much they adjust their sleep times. There is strong evidence that the DST transition negatively affects sleep levels and alertness: Smith (2016) finds an increase in the number of fatal car accidents in the six days following DST. We therefore interpret the difference in coefficients between the early- and late-baseline years as the difference in the gains from an hour of sunlight on test-day with a baseline of 1 hour 17 minutes sunlight before school and the gains from an hour of sunlight on test-day with a baseline of 28 minutes of sunlight before school *and* up to an hour of sleep deprivation. We have no information on the testing date for each student, so we cannot further stratify the start time effect as a function of number of days since the DST transition. However, since the testing period was longer than one week in both 2008 and

[28]Technically, the similarity between the early- and late-test time coefficients cannot be read as a failure to reject testing as an important input into math achievement. It is instead a rejection of a nonlinear relationship between achievement and test-day sunlight — it is consistent with an effect of test-day start times only if the relationship between achievement and start times is linear in the region between 28 minutes and 1 hour 17 minutes of sunlight before school.

2009, the test was likely taken a few days after the DST transition and perhaps as long as two weeks after, when transition-induced sleep loss has lessened. We therefore think that the safest interpretation is for moderate test-day effects in reading, of the same order as the full-year learning effects. At the very least, this result tells us that under an assumption of diminishing marginal returns to test-day sunlight, there are some situations (potentially including more sleep deprivation than is normal for this age group) where test-day sunlight has a large effect on academic achievement in reading. There is much more to be done to separately identify the effects of whole-year and test-day sunlight, but we leave this for future research.

5.8 Placebo time zone changes

The identification strategy in this paper leverages the discontinuity in sunrise times at the time zone boundary to estimate the effect of relative start times on academic performance. In a reduced form sense, we track students as they move east (west) over the time zone boundary in the Florida panhandle and find that scores decline (increase), as predicted by the earlier (later) relative start times.

Alternatively, perhaps moves to the east are score-decreasing for some reason unrelated to start times: schools are lower quality, or parents moving east get worse jobs and lower pay, which decrease investment in educational inputs. Our baseline specification includes controls for longitude and school demographics, which together control for any variation in underlying school or family characteristics that is linearly correlated with the demographic controls or varies linearly from east to west over the panhandle. If there are nonlinearities in this relationship, however, our method could misattribute variation in unobserved non-start time inputs to variation in start times, biasing our estimates.

In this section, we estimate placebo regressions that attempt to rule out a non-start time explanation. We generate placebo boundaries in ten mile increments from the true boundary; Figure 5 displays the estimated effect of moving over each placebo boundary, conditioning on true time zone, the regular vector of controls, and student fixed effects. We present estimates using cross-time zone movers, as well as restricting to only within-time zone movers. In Section 1.8 of the Online Appendix, we demonstrate that schools very close to the time zone boundary adopt start times similar to their cross-boundary counterparts; this means that there is a treatment effect of moving to or from the region directly adjacent to the boundary, even when the move is within time zone. We therefore exclude a 25 mile area around the true boundary (a version of the placebo test without this exclusion is available in the Online Appendix).[29]

Figure 5 displays the estimated coefficients for moving over placebo boundaries, placed in 10 mile increments from the true time zone boundary. In math, the placebo coefficients for the adolescent interaction are always smaller than the true coefficient, and usually significantly so. The true level coefficient is approximately zero, and the placebo coefficients bounce around that estimate, although we can sometimes reject they are zero. In reading, for both the within- and all-mover specifications, the placebo coefficients are almost always smaller than the true coefficients (and very imprecisely estimated when they are not). The true time zone-puberty interaction coefficient is approximately the same size as the placebos, although it is imprecise enough that we cannot differentiate it from zero in our main sample.

In summary, we estimate regressions of outcomes on placebo time zones, and find little evidence of changes in outcomes over the placebo boundaries, suggesting that the gains in achievement from westward moves are a function of crossing over the true time zone boundary and being exposed to later relative start times, rather than improvements in some other input.

5.9 Other effects of cross-time zone moves

A final threat to our identification strategy is the possibility that moving between time zones has a direct effect on family income or other characteristics. If these changes have an independent effect on academic performance, the exclusion restriction would be violated. Gibson and Shrader (2015) show that a one-hour delay in sunrise time reduces wages by between 0.5 and 4.5%. Given Dahl and Lochner's (2012) estimate of a 0.06 SD decrease in test scores per $1,000 decrease in EITC income, this could explain much of the test score effect. We do not observe parental income, and so cannot directly control for this possibility. However, there are three reasons to expect that a measure of income is not an important missing variable in our analysis. First, jobs are a primary reason for moving long distances and are chosen by the parents; wages are an important factor in job choice. It is therefore unlikely that movers are immediately treated with the average difference in wages given the change in sunrise times over the move. In fact, Gibson and Shrader (2015) argue that housing prices adjust to eliminate the incentive to move, and document that housing is indeed more expensive

[29] Excluding this region is not necessary in the main specification, as the IV estimate accounts for treatment bleed across time zones. However, our results are substantively the same even excluding this donut; we estimate that moving start times one hour later would improve math scores by 0.065 SDs for adolescents, and would have little effect on prepubescent math scores or reading scores for either age group. The number of students also decreases, resulting in larger standard errors on these estimates.

in early-sunrise cities. Disposable income would then be flat over the move, eliminating any effect on academic achievement. Second, in our sample zip-level income is higher in low-sunlight ET than in high-sunlight CT, which is the opposite of what is predicted by Gibson and Shrader (2015).[30] As we demonstrate in the Online Appendix, our results are unchanged by controls for zip-level income. Third, and most importantly, even if disposable income did increase as families moved over the time zone boundary, we would expect that children of all ages would benefit from the move. Instead, we observe larger increases in standardized test scores for pubertal children — and almost no increase for pre-pubertal children in math — suggesting that changes in sunlight before school are the most important causal factor.

6 Benefits of rearranging start times

Academic research and popular coverage of the potentially negative effects of early start times dates back at least as far as the late 1990s (New York Times, 1999; Wahlstrom et al., 1998). The evidence from the medical and physiological literature has grown so compelling that the American Academy of Pediatrics now recommends that middle and high schools delay start times to allow students to wake up no earlier than 8:00 am (2014). Despite the growing consensus, schools continue to open early; the median high school *opens* at 8:00 a.m. (NCES, 2012).

School districts, particularly those in large urban areas, often open different types of schools at different times. This structure is convenient for parents dropping off children at different schools, because it guarantees that a child in middle school will not need to be dropped off at the same time as a child in high school. It also allows school districts to use the same buses more intensively, saving on transportation costs. However, of the 19 school districts in the Florida panhandle, only 4 currently order their start times in the 'efficient' way. Inflexible parental schedules often preclude moving start times later for all students, since parents must be able to drop off their last child in time to get to work. In this section, we consider the academic effects of an alternative start time policy that better fits the physiological evidence but does not alter the overall distribution of start times: changing the opening order for different types of schools to elementary schools, middle schools, and finally high schools.

We operationalize this simple counterfactual by taking the average start time for each school type in each district, then assigning the earliest average start time to elementary schools, the next start time

[30]This does not seem to be a function of education, since literacy is actually marginally lower in ET (Authors' calculations from the NCES 2003 National Assessment of Adult Literacy).

to middle schools, and the latest time to high schools. We adjust the mean start time for each district so that it is the same in the counterfactual as in the real world.[31] We take the difference in relative start times for the counterfactual and real worlds for each school type and apply the coefficients from Table 3, weighting by the number of children in each district-school type. On average, this moves elementary start times 22 minutes earlier, middle schools 13 minutes earlier, and high schools 44 minutes later.

Figure 6 displays the effect on test scores, separated by gender and race. The counterfactual policy has been constructed so that if start times have an identical effect on children of all ages, the average increase in test scores will be zero. However, because the gains from later start times are smaller for younger children than for older children, our procedure has the effect of raising average academic achievement. In both math and reading, the effect is slightly (and usually insignificantly) negative for all groups of students in elementary and middle school. On average, elementary- and middle-school math and reading scores decline by 0.01 SDs. For high school students, the gains are large and statistically significant: in math, the proposed policy would increase minority student achievement in high school by 0.06 SDs in math and 0.08 SDs in reading. For white students, we expect that math scores would increase by 0.06 SDs and reading scores by 0.02 SDs. By gender, male high school students benefit slightly but insignificantly more compared to females. Using the coefficients from Table 2, the average effect is a 0.064 SD gain in math and a 0.044 SD gain in reading.

Furthermore, the high school results are good estimates for the overall change in achievement for each student by the end of high school. In Section 5.6 we show that increases in academic achievement occur immediately after the move and persist for years. That implies that back-loading the later start times will increase achievement as of the conclusion of high school by approximately the same amount as the single-year effect. Alternatively, taking the long-term estimates of Table 4 as given, the counterfactual would increase end-of-high school math scores by 0.05 SDs and reading scores by 0.037 SDs.[32]

One drawback of re-ordering start times would be that the youngest children may have to wait for the bus or walk to school in the dark. In December, the average sunrise would be only 53 minutes

[31] A clarifying example: if a district has 800 students in grade 9-12 schools with a start time of 7:00, 800 students in grades 6-8 schools with a start time of 7:30, and 1200 students in K-5 schools with a start time of 8:00, the mean district start time is 7:34. We would then set counterfactual start times to 7:08 in elementary school, 7:38 in middle school, and 8:08 in high school, with an average start time of 7:34. The procedure keeps the counterfactual mean start time the same as the status quo, and maintains the half hour spread in start times between school types.

[32] This math score is calculated by multiplying the long-term coefficent of $0.087 - 0.020 = 0.067$ by the average change in high school relative start times, 44 minutes.

before school starts, with 12% of elementary school students having less than half an hour between sunrise and school start in the darkest month. This would likely mean that a substantial number of very young students might need to travel to school in the dark, which presents a significant drawback to this proposal. Moving all school start times later, rather than re-ordering schools, would not have this problem.

In summary, we demonstrate that adjusting school start times so that high school students have the latest start time would significantly increase achievement for older children at a very low academic cost for younger children. Even when start times are reordered such that the average start time across the district remains the same, there are non-trivial gains in average academic performance that would benefit students in all demographic groups. These gains must be weighed against the costs of having younger children traveling to school in the dark.

7 Conclusion

We investigate the effect of daily school start times on academic performance. Adolescents in particular struggle with early start times; the onset of puberty shifts the sleep schedule back several hours, making any given start time more onerous for high schoolers than for students in other age groups. Our empirical strategy tracks academic performance in the same student before and after a cross-time zone move, which we use as an instrument for the amount of sunlight before school. Because the circadian rhythm is tied to variation in sunlight levels, this is a good approximation of a policy change in start times. Using a long individual panel from the state of Florida, we find that moving start times one hour later relative to sunrise would increase adolescent scores by 0.081 SDs in math and 0.057 SDs in reading. The increase in test scores can be observed immediately after the move, and persists for as long as we can measure it. Taking advantage of the fact that girls enter puberty two years earlier than boys, we document that the effect of relative start times on math performance spikes precisely at the gender-specific age of median entrance into an important pubertal stage. Previous research, which has mostly focused on a smaller age range of the population, has been unable to fully explore changes in the effect of start times over the pubertal transition.

These effects are cost-effective compared to other proposals to improve educational achievement, such as smaller classrooms or higher-skilled teachers. Specifically, reducing class size in elementary schools from 22 to 15 increases scores by 0.15-0.20 standard deviations (Schanzenbach, 2006), and a

1 standard deviation improvement in teacher quality increase scores by approximately 0.10 standard deviations (Chetty et al., 2011). Changes to school schedules would likely be much cheaper. Jacob and Rockoff (2011) suggest that the cost of moving start times one hour later is less than $150 per student per year and potentially as low as free. In contrast, reducing class sizes by a third costs approximately $6,200 per student per year.[33] The cost of such a large improvement in teacher quality is more difficult to evaluate, since the supply side of the teacher market is poorly understood. However, it is likely very large, if only because it would likely require hiring hundreds of thousands of new teachers.[34]

We simulate the effect of adjusting start times by school type to match students' developmental patterns while maintaining the same mean district start time. We estimate that this would increase math scores for high school students by 0.064 SDs and reading scores by 0.044 SDs, while having small and mostly statistically insignificant effects on scores for younger children. Alternatively, moving start times later across the board would increase achievement for all ages and demographics. In either case, adjustments on the start times margin seem to be significantly cheaper than adjustments to classroom size or teacher composition, suggesting that there may be large unrealized gains in this area.

There is one important caveat to our findings. Changes in school start times can increase achievement through either better learning in the year leading up to the test, or improved testing performance. We exploit a policy change in the testing date relative to Daylight Saving Time to learn whether test-day start times are important for achievement (but not by how much). We find suggestive evidence in favor of testing effects in reading, but not math. Our method is unable to precisely quantify the relative importance of testing and learning, but show that the magnitude is approximately the same for reading. We leave this as an important direction for future work.

Despite growing medical and physiological evidence that current school start times are too early for optimal adolescent cognitive functioning, there has been little policy response to move start times later. We add to this debate with direct evidence that more sunlight before school — or a later relative start time — increases academic achievement for children of all ages. The increase in scores is much larger for adolescents, implying that even when parental schedules preclude later start times for all children, districts can improve academic performance by adjusting the order in which school types open to correspond with students' changing sleep schedules. Specifically, high school students should

[33] These figures are from Schanzenbach (2006), inflated from 2002 to 2011 prices via the CPI.

[34] If teacher quality were distributed normally, then replacing the bottom half of teachers with average teachers would raise the average SD of teacher quality by only 0.4, and therefore test scores by 0.04 SD. According to the NCES, there were 3.7 million teachers in the United States in 2012. It is hard to imagine that finding 1.85 million new average-quality teachers could be done without significantly increasing wages.

begin school later in the day to compensate for pubertal changes that shift their circadian rhythm later, while elementary students should begin school the earliest. Despite the low costs of adopting this policy, the gains are quite large.

References

Adolescent Sleep Working Group and Committee On Adolescence, American Academy of Pediatrics (2014). School start times for adolescents. *Pediatrics*.

Arendt, J. (2000). Melatonin, circadian rhythms, and sleep. *New England Journal of Medicine*, 343(15):1114–1116.

Campbell, I. G., Grimm, K. J., de Bie, E., and Feinberg, I. (2012). Sex, puberty, and the timing of sleep EEG measured adolescent brain maturation. *Proceedings of the National Academy of Sciences*, 109(15):5740–5743.

Carrell, S. E., Maghakian, T., and West, J. E. (2011). A's from ZZZZ's? The causal effect of school start time on the academic achievement of adolescents. *American Economic Journal: Economic Policy*, 3(3):62–81.

Carskadon, M. A., Acebo, C., and Jenni, O. G. (2004). Regulation of adolescent sleep: Implications for behavior. *Annals of the New York Academy of Sciences*, 1021(1):276–291.

Carskadon, M. A., Acebo, C., Richardson, G. S., Tate, B. A., and Seifer, R. (1997). An approach to studying circadian rhythms of adolescent humans. *Journal of biological rhythms*, 12(3):278–289.

Carskadon, M. A., Vieira, C., and Acebo, C. (1993). Association between puberty and delayed phase preference. *Sleep*, 16:258–258.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2011). The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood. Technical report, National Bureau of Economic Research.

Crowley, S. J., Acebo, C., and Carskadon, M. A. (2007). Sleep, circadian rhythms, and delayed phase in adolescence. *Sleep Medicine*, 8(6):602–612.

Dahl, G. B. and Lochner, L. (2012). The impact of family income on child achievement: Evidence from the earned income tax credit. *The American Economic Review*, 102(5):1927–1956.

Eaton, D. K., McKnight-Eily, L. R., Lowry, R., Perry, G. S., Presley-Cantrell, L., and Croft, J. B. (2010). Prevalence of insufficient, borderline, and optimal hours of sleep among high school students–united states, 2007. *Journal of Adolescent Health*, 46(4):399–401.

Edwards, F. (2012). Early to rise? the effect of daily start times on academic performance. *Economics of Education Review*, 31(6):970–983.

Fogel, S. M. and Smith, C. T. (2011). The function of the sleep spindle: a physiological index of intelligence and a mechanism for sleep-dependent memory consolidation. *Neuroscience & Biobehavioral Reviews*, 35(5):1154–1165.

Gibson, M. and Shrader, J. (2015). Time use and productivity: The wage returns to sleep. *Working paper*.

Groen, J. A. and Pabilonia, S. W. (2015). Snooze or lose: High school start times and academic achievement. Technical report, Bureau of Labor Statistics, US Department of Labor.

Hansen, M., Janssen, I., Schiff, A., Zee, P. C., and Dubocovich, M. L. (2005). The impact of school daily schedule on adolescent sleep. *Pediatrics*, 115(6):1555–1561.

Hinrichs, P. (2011). When the bell tolls: The effects of school starting times on academic achievement. *Education*, 6(4):486–507.

Jacob, B. A. and Rockoff, J. E. (2011). Organizing schools to improve student achievement: Start times, grade configurations, and teacher assignments. *The Hamilton Project*, pages 1–28.

Jenni, O. G. and Carskadon, M. A. (2012). Sleep behavior and sleep regulation from infancy through adolescence: Normative aspects. *Sleep Medicine Clinics*, 7(3):529–538.

Krueger, A. B., Whitmore, D. M., et al. (2001). *Would smaller classes help close the black-white achievement gap?* Number 451. Industrial Relations Section, Princeton University.

Laberge, L., Petit, D., Simard, C., Vitaro, F., Tremblay, R., and Montplaisir, J. (2001). Development of sleep patterns in early adolescence. *Journal of Sleep Research*, 10(1):59–67.

Lufi, D., Tzischinsky, O., and Hadar, S. (2011). Delaying school starting time by one hour: Some effects on attention levels in adolescents. *Journal of Clinical Sleep Medicine: Official Publication of the American Academy of Sleep Medicine*, 7(2):137.

Maquet, P., Laureys, S., Peigneux, P., Fuchs, S., Petiau, C., Phillips, C., Aerts, J., Del Fiore, G., Degueldre, C., Meulemans, T., et al. (2000). Experience-dependent changes in cerebral activation during human rem sleep. *Nature neuroscience*, 3(8):831–836.

Marshall, W. A. and Tanner, J. M. (1970). Variations in the pattern of pubertal changes in boys. *Archives of disease in childhood*, 45(239):13–23.

NCES (2012). Average start time for public high schools and percentage distribution of start times in public high schools, by selected school characteristics. *Schools and Staffing Survey, National Center for Education Statistics*.

New York Times (August 1, 1999). Late to bed, early to rise makes a teen-ager ... tired.

Ng, E., Ng, D., and Chan, C. (2009). Sleep duration, wake/sleep symptoms, and academic performance in Hong Kong secondary school children. *Sleep and Breathing*, 13(4):357–367.

Sadeh, A., Gruber, R., and Raviv, A. (2003). The effects of sleep restriction and extension on school-age children: What a difference an hour makes. *Child development*, 74(2):444–455.

Schanzenbach, D. W. (2006). What have researchers learned from Project STAR? *Brookings papers on education policy*, pages 205–228.

Smith, A. C. (2016). Spring forward at your own risk: Daylight saving time and fatal vehicle crashes. *American Economic Journal: Applied Economics*, 8(2):65–91.

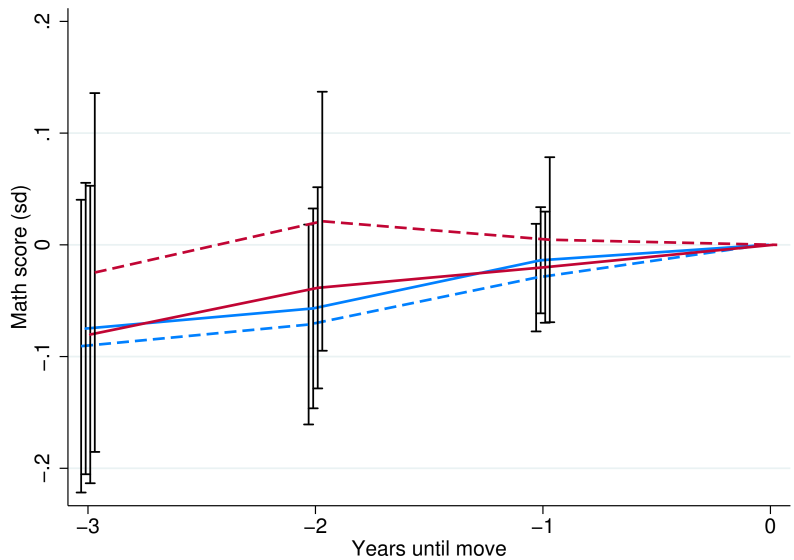
Wahlstrom, K., Wrobel, G., Kubow, P., et al. (1998). Minneapolis Public Schools start time study executive summary 1998.

Walker, M. P. and Stickgold, R. (2006). Sleep, memory, and plasticity. *Annual Review of Psycholoft*, 57:139–166.

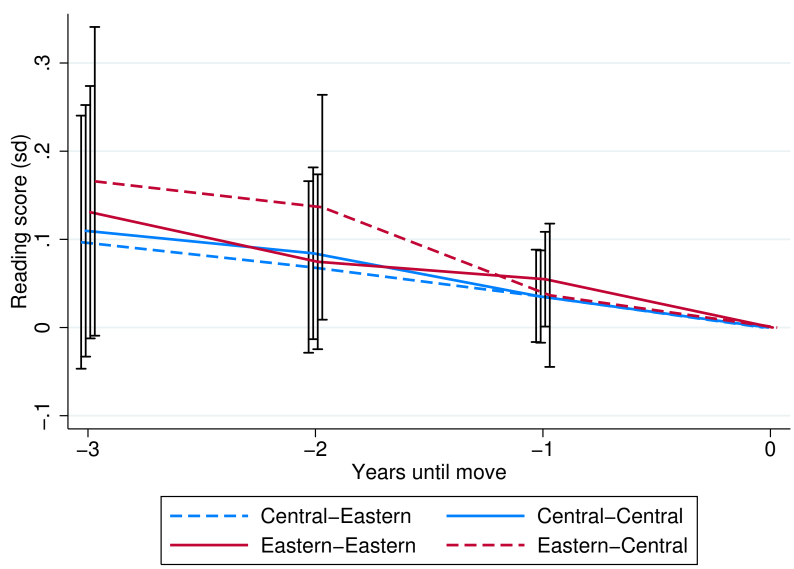
8 Figures

Figure 1: Pre-move trends in academic outcomes, by mover type

(a) Math scores

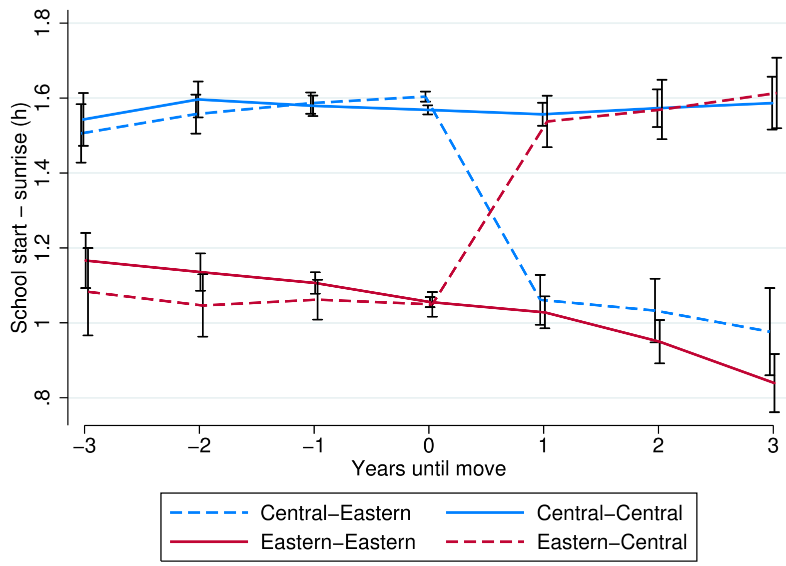


(b) Reading scores



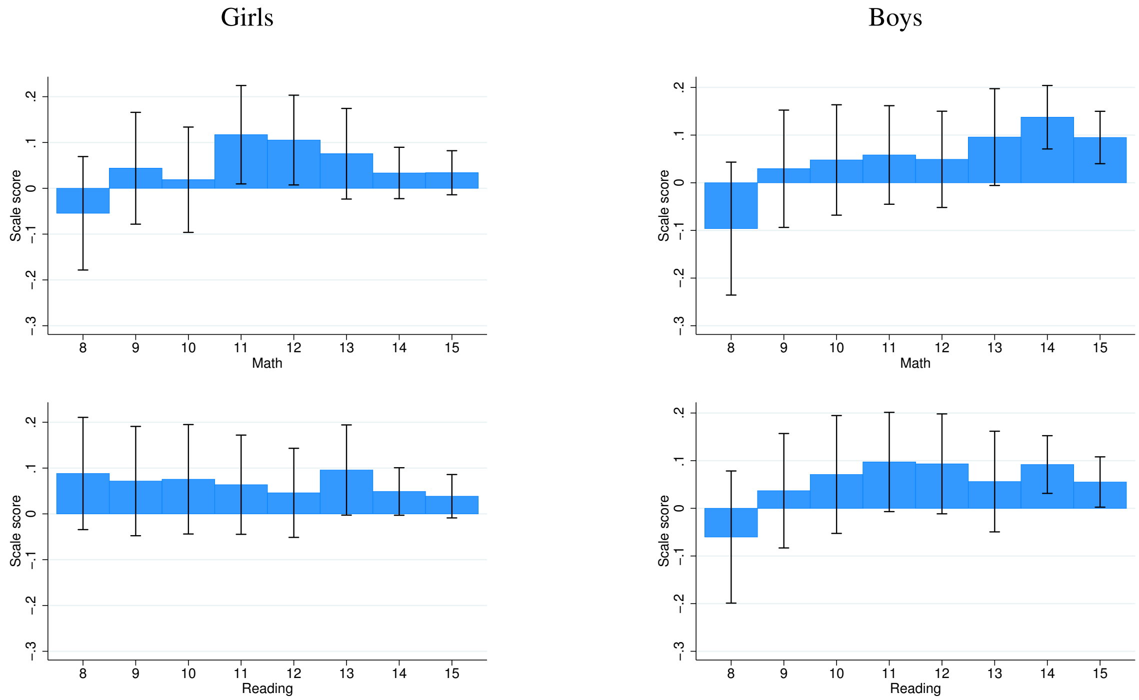
Displays the pre-move achievement trends for the four years leading up to a move of 25 miles or more. Results reported separately for four groups of movers: within CT, within ET, ET to CT, and CT to ET. Coefficients recovered from a regression of test scores on time-until-move dummies, a vector of controls (age-gender dummies, longitude, and school population shares for FRL, male, black, Asian, and Hispanic), and a fixed effect for the period before the move. Standard errors are clustered at the individual level, and included as bars representing 95% confidence intervals.

Figure 2: Hours of sunlight before school over move, by mover type



Displays the hours of sunlight before school for four groups: within CT, within ET, ET to CT, and CT to ET. Estimates are from a regression of relative school start time on time relative to move for each mover group, a vector of controls (age-gender dummies, longitude, and school population shares for FRL, male, black, Asian, and Hispanic), and a student-move fixed effect. The year before the move is normalized to be zero; we adjust the level of the coefficients with the group mean of relative start times for one year before the move. Standard errors are clustered at the individual level, and included as bars representing 95% confidence intervals.

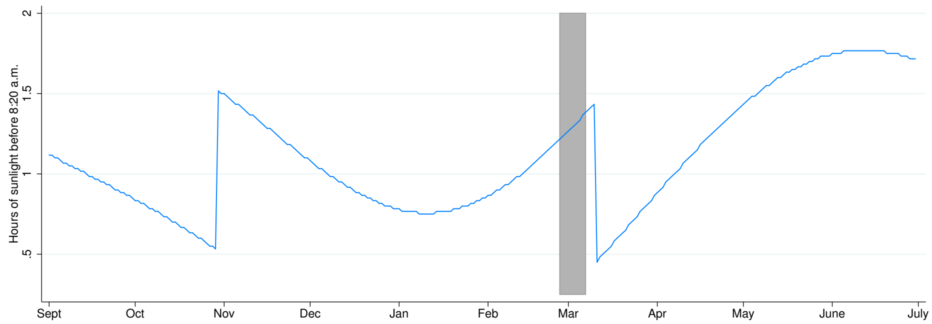
Figure 3: Effect of school start times on academic achievement, by age, gender, and subject



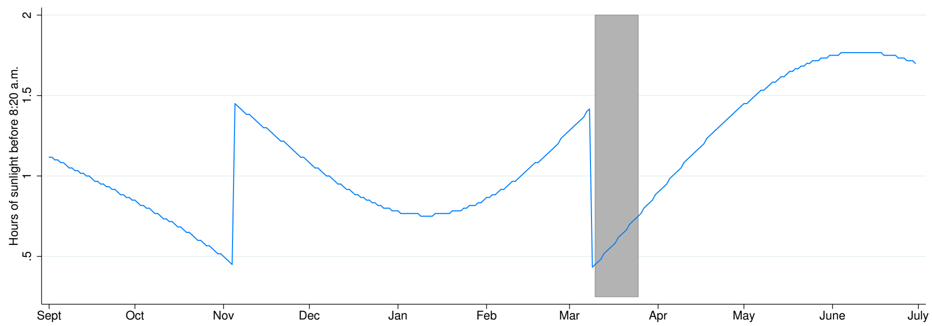
Each subfigure displays the age-gender specific effect of start times on academic achievement. Coefficients are from a regression of scale scores on school start time interacted with age, a vector of controls (age-gender dummies, longitude, and school population shares for FRL, male, black, Asian and Hispanic), and an individual fixed effect. Start time-age interactions are instrumented with time zone-age interactions. Sample is listed in the column headers, dependent variable is noted on the horizontal axis. Standard errors are clustered at the individual level, and included as bars representing 95% confidence intervals.

Figure 4: Hours of sunlight before 8:20 a.m. start time, by year with testing periods

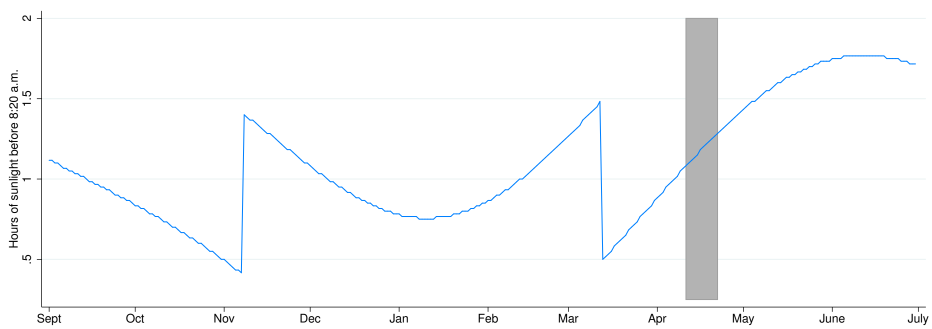
(a) 2000-2007



(b) 2008-2009

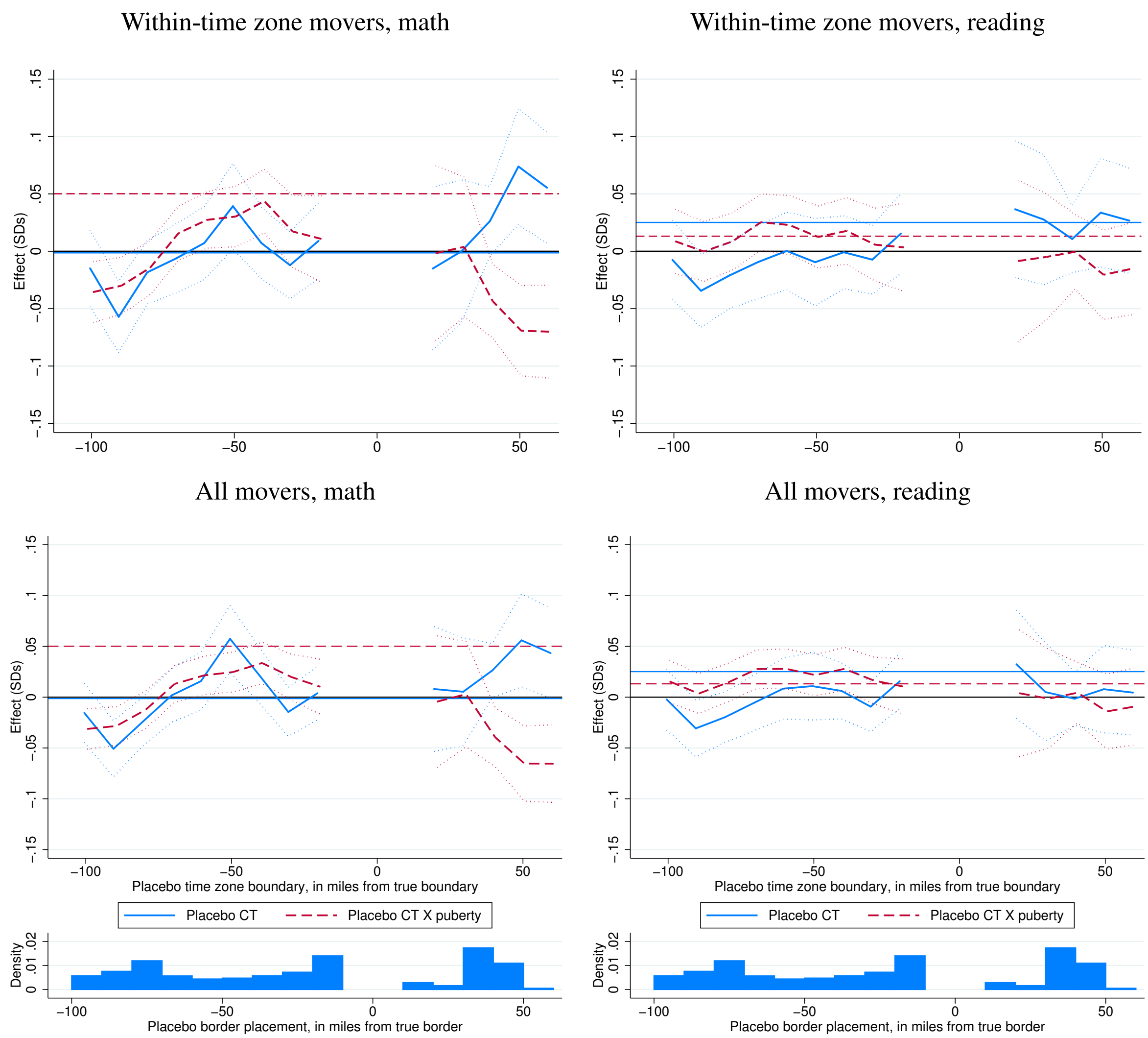


(c) 2011-2013



Amount of sunlight before school and testing dates for a hypothetical school for each of the three testing regimes. School location and opening time chosen to match the average test-day relative start time in ET in 2008. Grey areas represent testing periods. The figures display sunlight for 2007, 2008, and 2011, respectively, but all are archetypes of their era.

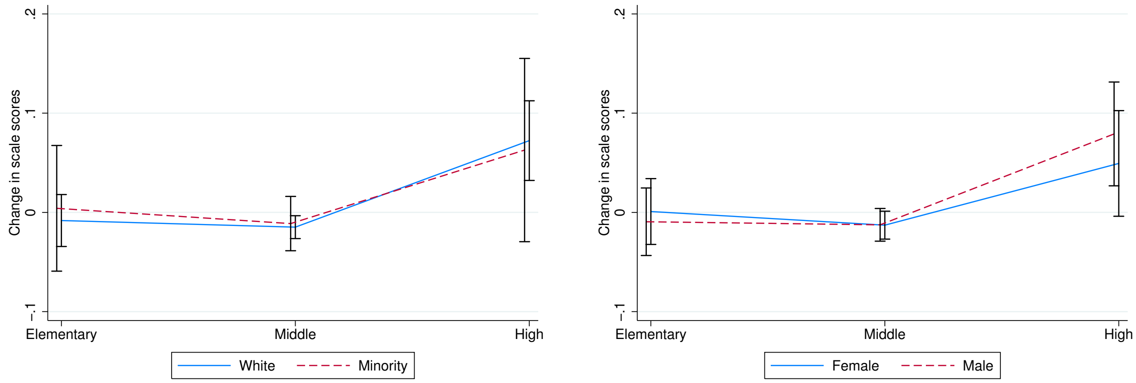
Figure 5: Effect of placebo time zones on academic achievement



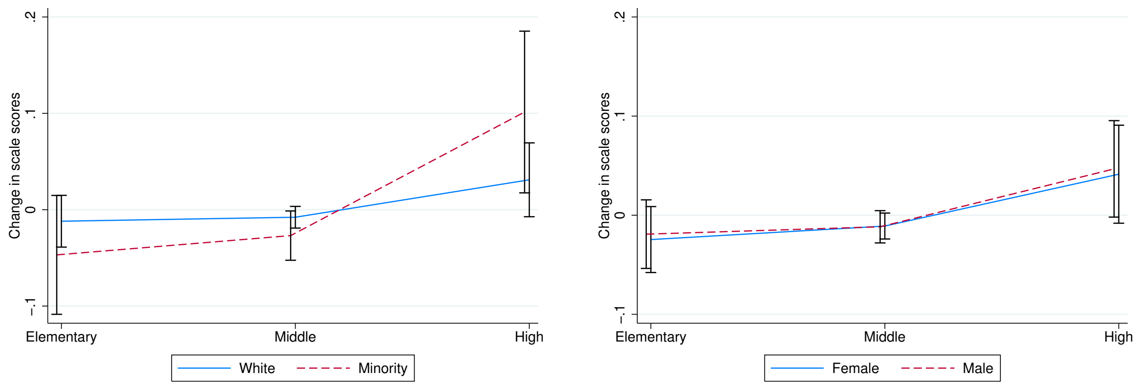
Dependent variable as noted in panel heading. Test scores measured in SDs normalized at the grade-year level for the entire state. Thin horizontal lines represent baseline coefficient estimates. We generate placebo time zones in ten mile increments from the true time zone boundary. Then, placebo coefficients are calculated from individual regressions of the outcome on the true time zone interacted with puberty, and the placebo time zone interacted with puberty. All specifications include age-gender dummies, longitude controls, school demographic means (FRL, male, black, Asian, and Hispanic) and individual fixed effects. Standard errors clustered at the individual level. We display results including and excluding cross-time zone movers. Sample excludes a 25 mile donut around the time zone boundary due to treatment bleed across the boundary.

Figure 6: Counterfactual change in test scores, reordered start times

(a) Math Scores



(b) Reading Scores



Estimated test score gains under a counterfactual policy where start times are adjusted to be later for older children. Adjustment is conducted by taking the average start time for each school type in each district (elementary, middle, and high), and swapping them between school types so that elementary schools open first, then middle schools, then high schools. We then adjust the level of all school times so that the mean counterfactual district start time is the same as the true mean start time. This results in bell times 22 minutes earlier for elementary schools, 13 minutes earlier for middle schools, and 44 minutes later for high schools. Gains are then calculated by multiplying the changes in start time for each child with the relevant coefficients from Table 3. Bars represent 95% confidence intervals.

9 Tables

Table 1: Sample characteristics, Florida panhandle movers

	Panhandle (1)	Movers (2)	CT-ET (3)	ET-CT (4)	Difference (3)-(4)
Panel A: School characteristics					
FRL (fraction)	0.54	0.55	0.56	0.56	0.000
	[0.27]	[0.24]	[0.21]	[0.30]	(0.038)
Male (fraction)	0.51	0.51	0.51	0.51	0.003
	[0.02]	[0.02]	[0.03]	[0.03]	(0.004)
Black (fraction)	0.25	0.26	0.20	0.37	-0.168***
	[0.27]	[0.28]	[0.22]	[0.47]	(0.057)
Hispanic (fraction)	0.04	0.03	0.03	0.03	-0.008
	[0.04]	[0.04]	[0.02]	[0.07]	(0.008)
Asian (fraction)	0.02	0.02	0.01	0.01	0.004
	[0.02]	[0.02]	[0.02]	[0.02]	(0.003)
District Grade 3 math scores (SD)	0.11	0.11	0.12	0.08	0.039
	[0.22]	[0.25]	[0.21]	[0.34]	(0.043)
District Grade 3 reading scores (SD)	0.15	0.15	0.17	0.09	0.084**
	[0.22]	[0.23]	[0.17]	[0.34]	(0.041)
District Grade 3 absentee rates	4.54	4.48	4.39	4.74	-0.358
	[0.77]	[1.13]	[1.70]	[1.40]	(0.227)
1999 median income by zip, logged	10.67	10.64	10.59	10.62	-0.036
	[0.27]	[0.26]	[0.26]	[0.36]	(0.051)
Student/teacher ratio	15.43	15.72	15.40	15.80	-0.400
	[1.20]	[1.40]	[2.15]	[1.71]	(0.273)
Charter school (fraction)	0.02	0.01	0.01	0.02	-0.015
	[0.12]	[0.07]	[0.05]	[0.14]	(0.017)
Urban (fraction)	0.27	0.24	0.18	0.27	-0.086
	[0.48]	[0.47]	[0.49]	[0.63]	(0.084)
Panel B: Individual characteristics					
FRL (=1)	0.55	0.66	0.67	0.69	-0.017
	[0.50]	[0.47]	[0.47]	[0.46]	(0.025)
Male (=1)	0.52	0.51	0.51	0.52	-0.003
	[0.50]	[0.50]	[0.50]	[0.50]	(0.026)
Black (=1)	0.26	0.26	0.25	0.26	-0.008
	[0.44]	[0.44]	[0.44]	[0.44]	(0.023)
Hispanic (=1)	0.04	0.04	0.02	0.03	-0.008
	[0.19]	[0.20]	[0.15]	[0.18]	(0.009)
Asian (=1)	0.02	0.01	0.01	0.01	0.002
	[0.13]	[0.11]	[0.10]	[0.09]	(0.005)
Math score (SD)	0.11	0.02	-0.06	0.00	-0.064
	[0.96]	[0.92]	[0.88]	[0.88]	(0.047)
Reading score (SD)	0.15	0.07	0.00	0.00	0.003
	[0.97]	[0.93]	[0.90]	[0.93]	(0.048)
Absentee rate	4.52	5.60	5.44	6.46	-1.026***
	[4.44]	[5.18]	[5.16]	[5.60]	(0.325)
Observations	186,278	13,788	713	726	

Sample is all third graders in the panhandle. Categorical variables are reported as 0-1. Absentee rate is reported as the percentage (0-100) of days missed in the school year to ease interpretation. Standard deviations in square brackets. Standard errors in parentheses and clustered at the school level in Panel A, unclustered in Panel B. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2: Academic and behavioral outcomes on start time, with student fixed effects

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: First stage, relative start time (hours)							
CT (=1)	0.471***	0.345***	0.424***	0.415***	0.346***	0.424***	0.415***
	(0.016)	(0.021)	(0.020)	(0.020)	(0.021)	(0.020)	(0.020)
CT X Puberty	0.264***	0.265***	0.306***	0.265***	0.265***	0.306***	0.265***
	(0.012)	(0.012)	(0.011)	(0.011)	(0.012)	(0.011)	(0.011)
P(CT+CT X puberty=0)	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Panel B: IV estimates, math test scores (SDs) on relative start time							
Start time - sunrise (h)	-0.063**	0.014	0.020	0.010	0.012	0.020	0.009
	(0.026)	(0.041)	(0.036)	(0.035)	(0.041)	(0.036)	(0.035)
Start time X puberty	0.099***	0.074***	0.058***	0.074***	0.073***	0.057***	0.073***
	(0.018)	(0.020)	(0.021)	(0.019)	(0.020)	(0.021)	(0.019)
P(Start+Start X puberty=0)	0.042	0.002	0.001	0.001	0.002	0.001	0.001
Cragg-Donald F-stat	1101.18	404.14	588.90	541.51	405.14	588.76	542.01
Panel C: IV estimates, reading test scores (SDs) on relative start times							
Start time - sunrise (h)	0.064**	0.088**	0.081**	0.061*	0.087**	0.081**	0.061*
	(0.027)	(0.041)	(0.037)	(0.036)	(0.041)	(0.037)	(0.036)
Start time X puberty	-0.005	-0.014	-0.023	-0.005	-0.013	-0.023	-0.004
	(0.018)	(0.021)	(0.022)	(0.020)	(0.021)	(0.022)	(0.020)
P(Start+Start X puberty=0)	0.000	0.004	0.008	0.014	0.004	0.008	0.014
Cragg-Donald F-stat	1230.00	485.69	637.13	618.88	486.65	637.22	619.26
Panel D: IV estimates, absence rate (%) on relative start times							
Start time - sunrise (h)	-0.937***	-1.885***	-0.696	-0.856*	-1.860***	-0.718	-0.869*
	(0.361)	(0.594)	(0.476)	(0.487)	(0.590)	(0.474)	(0.485)
Start time X puberty	0.481**	0.846***	0.365	0.443*	0.857***	0.395	0.469*
	(0.245)	(0.295)	(0.286)	(0.268)	(0.294)	(0.285)	(0.268)
P(Start+Start X puberty=0)	0.062	0.008	0.264	0.206	0.010	0.274	0.219
Cragg-Donald F-stat	689.75	273.69	425.19	383.57	274.18	425.38	383.62
Longitude	No	Yes	Yes	Yes	Yes	Yes	Yes
District quality	No	No	Yes	No	No	Yes	No
School quality	No	No	No	Yes	No	No	Yes
Time since move	No	No	No	No	Yes	Yes	Yes

Dependent variable as noted in panel heading. Test scores measured in SDs normalized at the grade-year level for the entire state. Absentee rate is the fraction of days the child missed school. Start time and its interaction with puberty are instrumented by time zone. Sample is all children who moved. All specifications include age-gender dummies and individual fixed effects. Sample size is fixed within panels: 34,018 students and 115,778 student-years in Panel A, 24,768 students and 99,835 student-years in Panel B, 25,191 students and 104,791 student-years in Panel C, and 15,906 students and 66,263 student-years in Panel D. Standard errors in parentheses and clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3: Academic and behavioral outcomes on start time, by group with student fixed effects

	White	Non-white	Non-FRL	FRL	Male	Female
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Math test scores (SDs)						
Start time - sunrise (h)	0.021	-0.017	0.045	-0.015	0.027	-0.008
	(0.039)	(0.095)	(0.055)	(0.046)	(0.050)	(0.050)
Start time X puberty	0.072***	0.098**	0.102***	0.063**	0.076***	0.072***
	(0.022)	(0.046)	(0.032)	(0.025)	(0.027)	(0.028)
P(Start+Start X puberty=0)	0.000	0.182	0.000	0.137	0.003	0.069
Cragg-Donald F-stat	459.66	84.63	177.22	373.97	263.79	277.79
Number of students	17013	7755	10052	14716	12380	12388
Observations	70535	29300	40140	59695	49436	50399
Panel B: Reading test scores (SDs)						
Start time - sunrise (h)	0.034	0.135	0.072	0.056	0.055	0.072
	(0.040)	(0.092)	(0.056)	(0.047)	(0.051)	(0.050)
Start time X puberty	0.006	-0.003	-0.028	0.006	0.006	-0.018
	(0.024)	(0.046)	(0.035)	(0.025)	(0.028)	(0.029)
P(Start+Start X puberty=0)	0.113	0.018	0.215	0.037	0.060	0.101
Cragg-Donald F-stat	516.36	100.07	221.60	407.29	289.00	333.87
Number of students	17264	7927	10284	14907	12560	12631
Observations	73872	30919	42458	62333	51752	53039
Panel C: Absence rate (%)						
Start time - sunrise (h)	-0.357	-2.012	-1.094	-0.619	-0.564	-1.277*
	(0.531)	(1.312)	(0.737)	(0.625)	(0.622)	(0.752)
Start time X puberty	-0.193	1.723***	0.298	0.533	0.201	0.794**
	(0.324)	(0.622)	(0.411)	(0.343)	(0.379)	(0.377)
P(Start+Start X puberty=0)	0.123	0.720	0.089	0.840	0.379	0.346
Cragg-Donald F-stat	320.62	58.76	116.36	270.00	193.18	190.14
Number of students	10613	5293	6383	9523	8019	7887
Observations	45654	20609	26483	39780	32994	33269

Dependent variable as noted in panel heading. Test scores measured in SDs normalized at the grade-year level for the entire state. Absentee rate is the percent of days the child missed school. Start time and its interaction with puberty are instrumented by time zone. Sample is all children who moved more than 25 miles. All specifications include age-gender dummies, longitude controls, school demographic means (FRL, male, black, Asian, and Hispanic) and individual fixed effects. Standard errors in parentheses and clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4: Persistence in effects of relative start time on student outcomes, with student fixed effects

	Math score (in SD)		Reading score (in SD)	
	(1)	(2)	(3)	(4)
Start time - sunrise (h) (prepubescent)	0.009	0.007	0.061*	0.052
	(0.035)	(0.036)	(0.036)	(0.036)
Start X moved two years ago (pre)		0.002		0.011
		(0.009)		(0.009)
Start X moved 3+ years ago (pre)		-0.011		-0.005
		(0.012)		(0.012)
Start time - sunrise (h) (pubescent)	0.082***	0.087***	0.057**	0.048**
	(0.025)	(0.026)	(0.023)	(0.024)
Start X moved two years ago (pub)		-0.016***		-0.004
		(0.006)		(0.006)
Start X moved 3+ years ago (pub)		-0.020***		0.010
		(0.007)		(0.007)
P[Start (pre) = Start (pub)]	0.000	0.000	0.826	0.861
P[Start (pre) = Start (pub), long run]		0.000		0.577
Cragg-Donald F-stat	542.01	107.47	619.26	124.19
Number of students	24,768	24,768	25,191	25,191
Observations	99,835	99,835	104,791	104,791

Dependent variable as noted in panel heading. Test scores measured in SDs normalized at the grade-year level for the entire state. Start time and its interaction with puberty are instrumented by time zone and the interaction of time zone and puberty. Sample is all children who moved more than 25 miles. All specifications include age-gender dummies, longitude controls, school demographic means (FRL, male, black, Asian, and Hispanic) and individual fixed effects. Standard errors in parentheses and clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 5: Academic outcomes, for testing before and after DST

	Math (SD)		Reading (SD)	
	(1)	(2)	(3)	(4)
Start time - sunrise (h) (prepubescent)	0.030		0.056	
	(0.038)		(0.038)	
Start time - sunrise (h) (pubescent)	0.096***		0.060**	
	(0.027)		(0.024)	
Start time X prepubescent, late test time		0.022		0.049
		(0.039)		(0.039)
Start time X prepubescent, early test time		0.071		0.096**
		(0.046)		(0.047)
Start time X adolescent, late test time		0.095***		0.045*
		(0.030)		(0.026)
Start time X adolescent, early test time		0.096***		0.104***
		(0.025)		(0.026)
Era X puberty controls	No	Yes	No	Yes
P[Early = late test (Prepub)]		0.165		0.192
P[Early = late test (Adol)]		0.967		0.001
Cragg-Donald F-stat	468.563	229.684	542.050	269.539
Number of students	23,618	23,618	24,152	24,152
Observations	89,707	89,707	94,515	94,515

Dependent variable as noted in panel heading. Test scores measured in SDs normalized at the grade-year level for the entire state. Start time and its interactions are instrumented by time zone and the interaction of time zone and interactions. Sample is all children who moved more than 25 miles. All specifications include age-gender dummies, longitude controls, school demographic means (FRL, male, black, Asian, and Hispanic), time since move dummies, and individual fixed effects. Sample includes years 2000-2013 excluding 2010, when testing took place over the DST time change. Standard errors in parentheses and clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

1 Online Appendix for *Rise and Shine: The Effect of School Start Times on Academic Performance from Childhood through Puberty*

1.1 Robustness checks for mover definition

Our identifying variation comes from students who move between schools in different time zones in the Florida panhandle. Most of these moves are quite long-distance; the median move is 83 miles. The disruption inherent in such a move may have an independent effect on achievement, which is important to control for in our context. To help identify the effect of moving, as well as the effect of other school-level covariates, we include in our sample students who move within a time zone. This requires defining what constitutes a move by setting a threshold distance between the schools the student attended. Otherwise, graduating from middle school to high school would constitute a move. A high threshold has the advantage of making the move more likely to match a cross-time zone move in terms of disruptiveness; a low threshold increases sample size and precision.

We settled on a threshold of 25 miles, but our results are robust to other threshold choices. Table A1 presents estimates for 15, 20, 25, and 30 mile thresholds for math and reading outcomes. We also consider defining a move as any move between different school districts, although this will include students who move less disruptive distances, such as when families move to a nearby suburb that happens to be in a different district. Across all definitions, the results are broadly consistent. In math, the effect for prepubescent children ranges from 0.009 to 0.037 SDs; the effect for adolescents ranges from 0.067 to 0.084 SDs. In reading, the range is 0.034 to 0.061 for younger children and 0.044 to 0.057 for adolescents. The effects statistically differ from zero for adolescents for both math and reading across all distances.

1.2 Specification robustness checks

We include two sets of control variable robustness checks. First, in Table A2, we consider different levels of aggregation for the demographic share controls (FRL, male, black, Asian, and Hispanic). Instead of aggregating at the school-year level, as we do in our main results, we consider district-year, district third graders-year,[1] school-year, and school-grade-year. All specifications include age-gender dummies and an individual fixed effect. For each level of aggregation, we present one specification with no other controls, one that adds urban dummies and log income controls, and a final model that

[1]District third graders-year is the demographic means for the third graders in the given district-year.

includes school size and student/teacher ratio.

Comparing across the rows of Table A2, the results are largely unchanged. In Panel A, all specifications show an effect size in math of 0.003-0.037 SDs for prepubescents, and 0.062-0.096 for adolescents. The effect is statistically significant at the the 1% level or better for adolescents but null for younger students. In reading, the estimates are also similar across specifications: 0.046-0.087 SDs for prepubescents, and 0.044-0.074 SDs for adolescents. The prepubescent effect is occasionally significant at the 5% level; the adolescent effect has a p-value of about 1%.

For absences, the inclusion of demographics (but not the level of aggregation) makes a substantive difference in the results. Comparing Columns 1-3 with Columns 4-15, the inclusion of demographic controls (at any level of aggregation) reduces the size of the suspension effect from about 1.5 percentage points and significant at the 1% level to about 0.8 percentage points and significant at the 10% level for prepubescents. The adolescent effects are generally null once we control for demographics. Since there may be significant between-school differences in policies for counting absences (and these may be correlated with school demographics), we think that the results with demographic controls are more trustworthy. It is therefore reassuring that they are the same regardless of the level of demographic aggregation.

Table A3 contains sur second control robustness check. Columns 1 and 3 restate our baseline results for math and reading. Columns 2 and 5 include controls for latitude; average sunrise times over the school year vary by about a minute over the north-south range of the panhandle[2] and this could conceivably have some affect on sleep (in contrast, the east-west variation in sunrise times from longitude is nearly 20 minutes, excluding the time zone change). The addition of latitude has a moderately sized but statistically insignificant effect on the prepubescent coefficients. The change in the adolescent coefficients is smaller.

In Columns 3 and 6 of Table A3 we test whether the inclusion of third grade district test scores as control variables affects the results. Third grade test scores are appealing as a summary measure of district quality, but may be endogeneous if start times affect performance for children in kindergarten to third grade. For this reason we do not include them in our main specification, but it is reassuring that they have little effect on the results.

[2]The average disguises some larger differences over the year; but it is never larger than three minutes.

1.3 Changes in school characteristics over the move

A potential threat to our identification strategy is changes in school and peer characteristics as students move between time zones. If students moving from CT to ET move to significantly worse schools, while ET-CT movers moved to better schools, it would not be surprising that student achievement declined upon entering ET and rose upon exiting. Because, on average, there is less sunlight before school in ET than in CT, this could generate a spurious positive relationship between relative school start times and academic achievement.

We consider this question directly in Table A4. We take the years directly before and after each move, and term these pairs of years a *moving episode*.[3] We then regress school- and zip-level characteristics on moving episode fixed effects and move indicators for the four different types of movers: Eastern-Eastern, Central-Central, Eastern-Central, and Central-Eastern. Each coefficient is a measure of the change in characteristics over the move. As outcomes, we consider the five school-level demographic share controls included in our preferred specification (percent FRL, male, black, Asian, and Hispanic), as well as school student/teacher ratio and zipcode-level median income as a measure of school and community resources.

The first two rows of Table A4 show that peer quality changed slightly over the move for within-time zone movers. ET-ET movers had 4.5 percentage points fewer FRL classmates; CT-CT movers had 1.7 percentage points fewer. School quality as measured by the student/teacher ratio increased slightly for both groups. Median income rose by $1,000 for within-ET movers and fell by $430 for within-CT movers. These differences are statistically significant, but none are particularly large or striking.

The cross-time zone movers tell a slightly different story. Eastward movers generally ended up in a richer area — 4.5 percentage points fewer FRL classmates and $5,700 higher median income — and had 14.0 percentage points more black classmates and 0.5 percentage points more Hispanic classmates. School quality as measured by the student/teacher ratio was unchanged. ET-CT movers saw approximately the opposite changes in medium income and percent of black students. The economic and peer changes may work in opposite directions in this case, making it unclear in which direction the overall bias goes. However, neither the inclusion of demographic controls (in Table 2) or income controls (in Table A2) substantively changes our results, suggesting that changes in peer characteristics have only a moderate effect on outcomes over the move, and do not significantly affect our results.

[3] Since occasionally a student will move in consecutive years, a small number of observations are repeated.

1.4 Performance trend before move

In Section 4.1, we show that test score trends are similar for all groups of movers in the years before the move. However, math scores trend *up*, which is somewhat surprising since the disruption of the upcoming move would be expected to reduce scores. Figure A1 show results from a regression of scale scores on time-until-move dummies and a fixed effect for the period until the move. This is identical to the regression displayed in Figure 1, but without controls. The Figure confirms that unconditionally, test scores trend down in both math and reading before a move. This is largely a result of removing the age-gender fixed effects, which soak up any time trend. Comparing across different groups of movers, the trends are slightly further apart than in the version with controls, but are still generally statistically indistinguishable.

1.5 Robustness checks for puberty definition

One of our main interests in this paper is how the effect of relative school start times varies with pubertal status. This requires a working definition of puberty, and there are several defensible alternatives. Pubertal development is typically measured with the Tanner Scale. There are two versions; one that uses levels of pubic hair to define the stages and another that uses breast and genital development. We rely on the pubic hair version of the Scale, which Campbell et al. (2012) indicate is more closely associated with pubertal changes in sleep patterns. They also note that changes in sleep patterns begin during Stage 3, so we use the age of median attainment (by gender) of Stage 3 as the definition of puberty.

Table A5 shows our main results with three alternative definitions of puberty: pubic hair Stage 2, pubic hair Stage 4, and breast/genital Stage 3. These changes typically shift the age of puberty by at most a year, and not necessarily for both genders. The results are largely unchanged, although slightly attenuated in some specifications. Because this definition of puberty is a worse fit for the underlying biological processes, this is unsurprising.

1.6 Estimates without interactions

Table A6 displays a version of our baseline model without an interaction between relative start time and pubertal status. Allowing for heterogeneity by pubertal status is important, but for completeness we have included this specification.

Across the rows, the change in sunlight is about 30 minutes over the time zone border. For both math and reading, the effect of moving start times one hour later is about the average of the child and adolescent effects from Table 2. In math, the estimated effect is 0.043 SD per hour by the final column, and the estimates are only occasionally statistically significant. In reading, the effect is 0.059 SD per hour by the final column, and the effect sizes are all significant to at least the 5% level in all estimates. The attendance results vary, with a decrease of 0.7 percentage points in absence per hour of sunlight by the final column.

1.7 PSID data definitions

In this paper, we demonstrate that students treated with later relative start times have higher academic achievement. However, we do not directly observe sleep levels in the academic outcomes dataset. To more concretely link changes in start times to changes in sleep, we use the Child Development Supplement of the Panel Study of Income Dynamics (PSID) to estimate the effect of the time zone boundary on sleep. The survey collected time use diaries for students on a weekend day and a weekday in the years 1997, 2002, and 2007. We include all states with a single time zone,[4] and all children who were 6-19 during the survey and within 400 miles of the ET-CT time zone boundary. Our aim is descriptive, so we regress daily hours of sleep on a fully interacted set of dummies for puberty, CT, and whether the night was a weekend. In our preferred specification, we also include controls for gender, black/non-black, and FRL status. We expect that children in CT will have more sleep on weekdays when they face earlier relative start times, and those in ET will compensate with more sleep on weekends.

Table A7 contains the results. As discussed in Section 5.3, children in CT get 6 minutes more sleep per night during the week than children in ET; during puberty they get 17 minutes more. On the weekend, children in ET compensate for low levels of sleep during the week by sleeping 10 minutes more per night in the years before puberty and 19 minutes more while in puberty. We conservatively cluster at the state level. The coefficient for the difference in sleep between adolescents in CT and ET is significant at the 10% level; most others are not. Including student fixed effects suggests a slightly larger difference between the time zones: the decrease in sleep during puberty is 15 minutes smaller for adolescents in CT than in ET. This set of results corresponds to a pass-through rate of about 40-50% from school start times to sleep if Florida panhandle school start times are representative of the

[4]The CDS does not geocode individuals at a sub-state level in the publicly available version, which precludes analysis using observations in states with multiple time zones — including Florida.

rest of the US near the ET-CT time zone boundary. This number is close to the 46% pass-through reported by Wahlstrom (1998).

1.8 Treatment bleed for schools near the time zone boundary

In the placebo analysis, we study how test scores change when students move east-west or west-east but *not* across the true time zone boundary. Ideally, we would examine within-time zone moves to and from the region directly adjacent to the boundary, to help test whether there are unobservable changes in the school or community environment that occur nearby, but not exactly at, the time zone boundary.

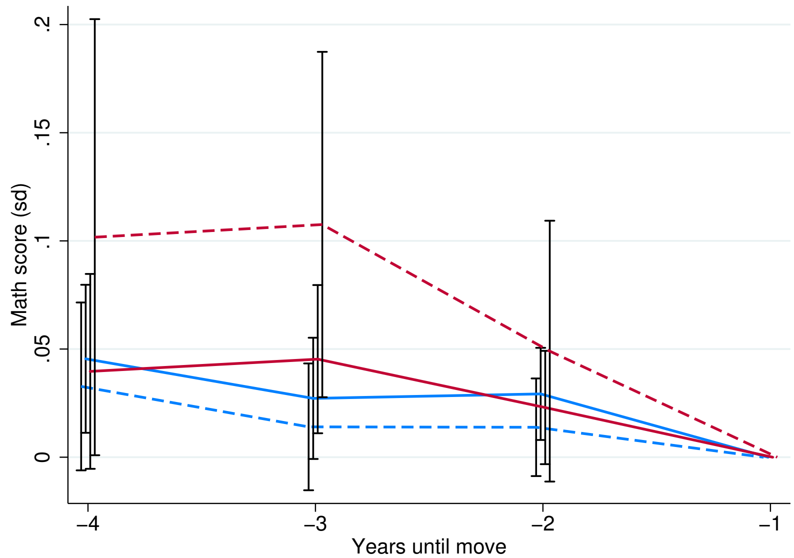
This approach will be problematic if there is an effect of being near the time zone boundary on school start times —- then, moving from directly beside the boundary in CT to a city fifty miles west could increase relative start times, directly increasing test scores. Figure A3 displays a nonparametric regression of relative start times on distance to the time zone boundary, estimated separately for each time zone. In the region directly adjacent to the boundary, start times veer towards the other time zone's norm, particularly for adolescents. We interpret this as the synchronization of start times across time zones, which allows parents to help their children prepare for school before going to work, whether or not they are commuting across time zones. This also means that start times are later for students moving west either from the region directly beside the boundary in CT, or *to* the region directly beside the boundary in ET.

In the main placebo results, we account for the treatment bleed across time zones by taking out a 25 mile "donut" around the time zone boundary. However, in the interest of completeness we include the unexcised version in Figure A4. The difference with Figure 5 is most stark in the puberty-time zone coefficient for math, where there is a consistent effect above the size of the true coefficient. Comparing between figures, removing the donut around the time zone boundary reduces the size of *all* placebo coefficients. The placebo effect is coming largely from individuals moving between the area close to the true time zone boundary and the rest of the study area, not individuals moving between areas far from the time zone boundary.

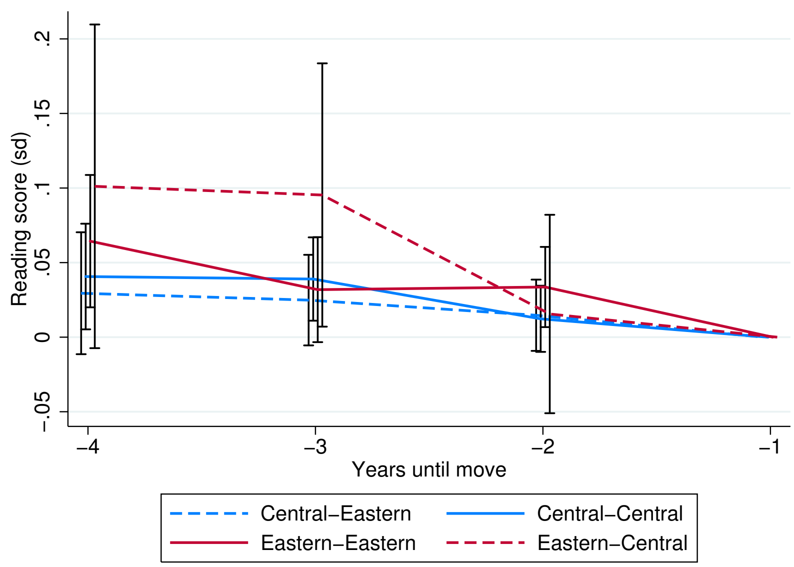
1.9 Online Appendix Figures

Figure A1: Pre-move trends in academic outcomes, by mover type without additional controls

(a) Math scores

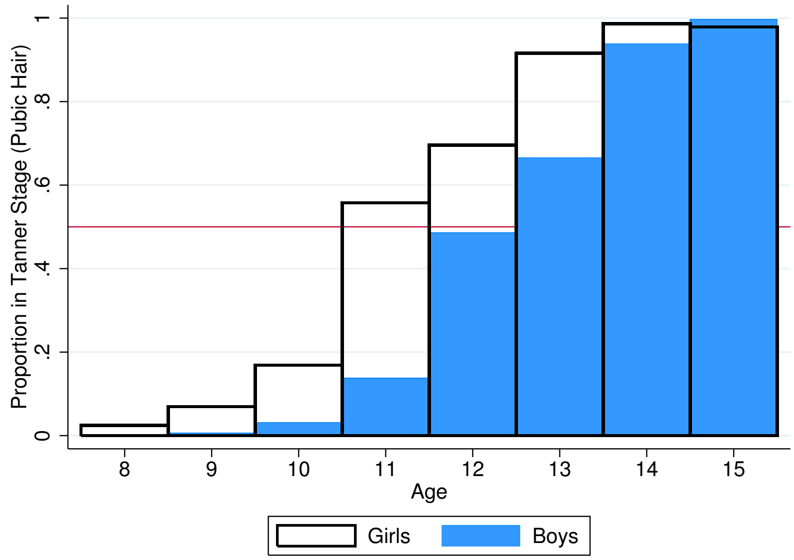


(b) Reading scores



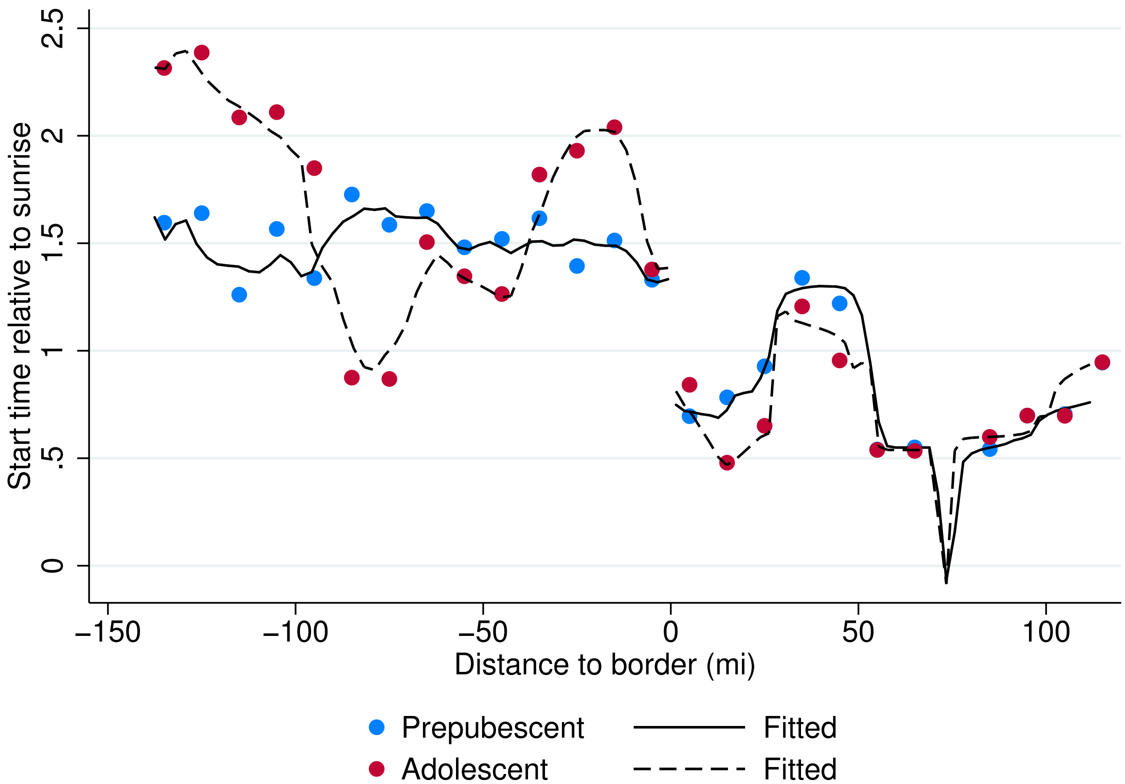
Displays the pre-move achievement trends for the four years leading up to a move of 25 miles. Results reported separately for four groups of movers: within CT, within ET, ET to CT, and CT to ET. Coefficients recovered from a regression of test scores on time-until-move dummies and a fixed effect for the period before the move. Standard errors are clustered at the individual level, and included as bars representing 95% confidence intervals.

Figure A2: Tanner stage 3 proportions by age and sex



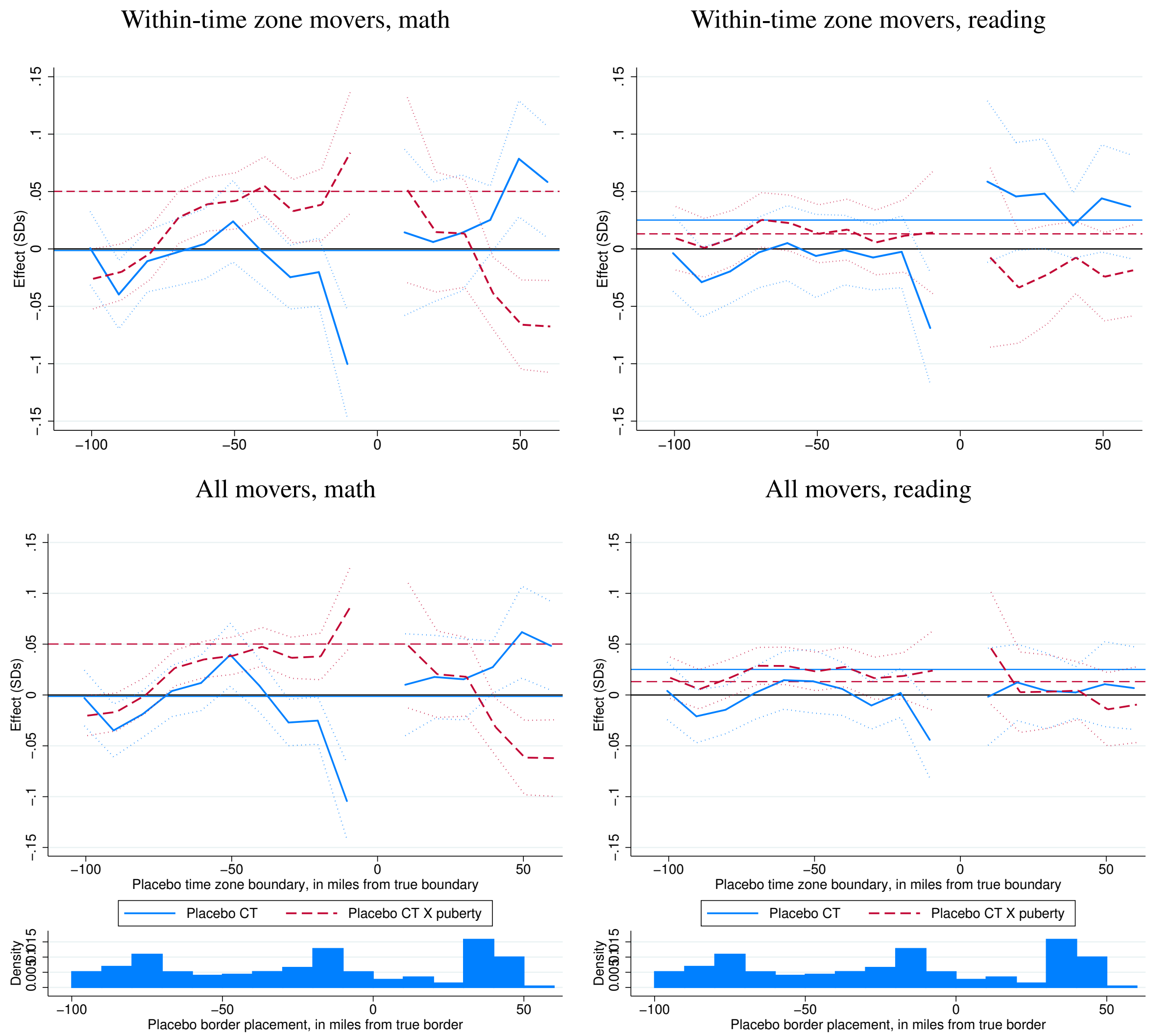
Displays proportion of children who had entered the Tanner Stage for pubic hair development at a given age for males and females. Horizontal line represents median child entering the stage.

Figure A3: Relative start time near the time zone boundary



Displays a nonparametric regression of relative start time (start time minus sunrise) on distance to the time zone boundary, estimated separately for each time zone. Scatter points are ten mile bin averages.

Figure A4: Effect of placebo time zones on academic achievement, no sample exclusion near true time zone boundary



Dependent variable as noted in panel heading. Test scores measured in SDs normalized at the grade-year level for the entire state. Thin horizontal lines represent baseline coefficient estimates. We generate placebo time zones in ten mile increments from the true time zone boundary. Then, placebo coefficients are calculated from individual regressions of the outcome on the true time zone interacted with puberty, and the placebo time zone interacted with puberty. All specifications include age-gender dummies, longitude controls, school demographic means (FRL, male, black, Asian, and Hispanic) and individual fixed effects. Standard errors clustered at the individual level. We display results including and excluding cross-time zone movers.

1.10 Online Appendix Tables

Table A1: Academic outcomes on school start time for varying mover definitions, with student fixed effects

	Math (SDs)					Reading (SDs)				
	(1) dist	(2) 15 mi	(3) 20 mi	(4) 25 mi	(5) 30 mi	(6) dist	(7) 15 mi	(8) 20 mi	(9) 25 mi	(10) 30 mi
Start time - sunrise (h)	0.037	0.029	0.014	0.009	0.009	0.037	0.034	0.026	0.061*	0.053
	(0.034)	(0.033)	(0.031)	(0.035)	(0.037)	(0.036)	(0.034)	(0.032)	(0.036)	(0.038)
Start time X puberty	0.036**	0.038**	0.070***	0.073***	0.060***	0.007	0.011	0.018	-0.004	-0.008
	(0.018)	(0.017)	(0.018)	(0.019)	(0.022)	(0.019)	(0.018)	(0.018)	(0.020)	(0.023)
P(Start+Start X puberty=0)	0.001	0.002	0.000	0.001	0.004	0.029	0.025	0.033	0.014	0.049
Cragg-Donald F-stat	610.14	611.40	677.49	542.01	542.98	684.27	701.42	766.47	619.26	612.31
Number of students	33712	35744	28969	24768	21557	34144	36197	29393	25191	21957
Observations	143921	153462	120233	99835	84165	150800	160997	126110	104791	88408

Dependent variable as noted in panel heading. Test scores measured in SDs normalized at the grade-year level for the entire state. Start time and its interaction with puberty are instrumented by time zone and the interaction of time zone and puberty. All specifications include age-gender dummies, longitude controls, school demographic means (FRL, male, black, Asian, and Hispanic), and individual fixed effects. Standard errors in parentheses and clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A2: Academic and behavioral outcomes on start time, with student fixed effects

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
Panel A: Math Test Scores (SDs)															
Start time - sunrise (h)	0.012	0.011	0.009	0.020	0.028	0.028	0.031	0.036	0.037	0.009	0.014	0.014	0.003	0.012	0.012
	(0.041)	(0.035)	(0.035)	(0.036)	(0.037)	(0.036)	(0.035)	(0.036)	(0.036)	(0.035)	(0.034)	(0.034)	(0.036)	(0.034)	(0.034)
Start time X puberty	0.073***	0.051***	0.054***	0.057***	0.037*	0.039*	0.065***	0.041**	0.043**	0.073***	0.050***	0.053***	0.076***	0.050***	0.053***
	(0.020)	(0.019)	(0.019)	(0.021)	(0.021)	(0.021)	(0.020)	(0.020)	(0.020)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)
P(Start+Start X puberty=0)	0.002	0.005	0.005	0.001	0.003	0.002	0.000	0.001	0.000	0.001	0.003	0.002	0.001	0.004	0.003
Cragg-Donald F-stat	405.14	593.76	604.37	588.76	593.74	606.81	580.48	599.33	612.93	542.01	640.62	655.01	534.48	638.53	655.08
Number of students	24768	23516	23516	24768	23516	23516	24545	23294	23294	24768	23516	23516	24765	23514	23514
Observations	99835	91853	91853	99835	91853	91853	98751	90852	90852	99835	91853	91853	99823	91846	91846
Panel B: Reading Test Scores (SDs)															
Start time - sunrise (h)	0.087**	0.061*	0.061*	0.081**	0.075**	0.074**	0.071**	0.065*	0.065*	0.061*	0.049	0.048	0.051	0.046	0.046
	(0.041)	(0.035)	(0.035)	(0.037)	(0.037)	(0.036)	(0.035)	(0.036)	(0.035)	(0.036)	(0.034)	(0.034)	(0.036)	(0.034)	(0.034)
Start time X puberty	-0.013	-0.009	-0.008	-0.023	-0.022	-0.022	-0.011	-0.013	-0.013	-0.004	-0.003	-0.003	0.000	-0.002	-0.002
	(0.021)	(0.020)	(0.020)	(0.022)	(0.021)	(0.021)	(0.021)	(0.020)	(0.020)	(0.020)	(0.019)	(0.019)	(0.020)	(0.019)	(0.019)
P(Start+Start X puberty=0)	0.004	0.015	0.015	0.008	0.014	0.014	0.008	0.015	0.014	0.014	0.027	0.027	0.025	0.030	0.030
Cragg-Donald F-stat	486.65	679.86	687.26	637.22	648.26	671.04	656.76	675.89	697.05	619.26	729.44	746.01	616.60	725.65	742.75
Number of students	25191	24048	24048	25191	24048	24048	24963	23823	23823	25191	24048	24048	25189	24045	24045
Observations	104791	96788	96788	104791	96788	96788	103547	95641	95641	104791	96788	96788	104776	96776	96776
Panel C: Absence Rates															
Start time - sunrise (h)	-1.860***	-1.463***	-1.431***	-0.718	-0.709	-0.695	-0.848*	-0.789*	-0.772*	-0.869*	-0.874*	-0.859*	-0.965**	-0.904*	-0.880*
	(0.590)	(0.505)	(0.502)	(0.474)	(0.483)	(0.479)	(0.460)	(0.471)	(0.467)	(0.485)	(0.467)	(0.464)	(0.492)	(0.470)	(0.466)
Start time X puberty	0.857***	0.677**	0.637**	0.395	0.330	0.304	0.439	0.353	0.320	0.469*	0.384	0.365	0.491*	0.396	0.367
	(0.294)	(0.278)	(0.275)	(0.285)	(0.286)	(0.283)	(0.274)	(0.278)	(0.275)	(0.268)	(0.268)	(0.265)	(0.269)	(0.270)	(0.266)
P(Start+Start X puberty=0)	0.010	0.012	0.011	0.274	0.182	0.166	0.156	0.117	0.103	0.219	0.091	0.087	0.151	0.081	0.077
Cragg-Donald F-stat	274.18	413.70	416.25	425.38	431.70	439.86	453.02	458.47	467.24	383.62	451.74	458.44	373.38	447.12	454.86
Number of students	15906	15130	15130	15906	15130	15130	15906	15130	15130	15906	15130	15130	15903	15128	15128
Observations	66263	61128	61128	66263	61128	61128	66263	61128	61128	66263	61128	61128	66252	61122	61122
Urban and log income	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Size and S/T ratio	No	No	Yes	No	No	Yes	No	No	Yes	No	No	Yes	No	No	Yes
District controls	No	No	No	Yes	Yes	Yes	No	No	No	No	No	No	No	No	No
District grade 3 controls	No	No	No	No	No	No	Yes	Yes	Yes	No	No	No	No	No	No
School controls	No	No	No	No	No	No	No	No	No	Yes	Yes	Yes	No	No	No
School-grade controls	No	No	No	No	No	No	No	No	No	No	No	No	Yes	Yes	Yes

Dependent variable as noted in panel heading. Test scores measured in SDs normalized at the grade-year level for the entire state. Absentee rate is the fraction of days the child missed school. Start time and its interaction with puberty are instrumented by time zone. Sample is all children who moved more than 25 miles. All specifications include age-gender dummies, longitude, and individual fixed effects. Standard errors in parentheses and clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A3: Outcomes on school start time, with latitude and school test grade scores

	Math			Reading		
	(1)	(2)	(3)	(4)	(5)	(6)
Start time - sunrise (h)	0.009	-0.035	0.015	0.061*	0.035	0.051
	(0.035)	(0.033)	(0.037)	(0.036)	(0.034)	(0.037)
Start time X puberty	0.073***	0.085***	0.073***	-0.004	0.004	-0.001
	(0.019)	(0.019)	(0.020)	(0.020)	(0.020)	(0.020)
Latitude controls	No	Yes	No	No	Yes	No
Third grade district scores	No	No	Yes	No	No	Yes
P(Start+Start X puberty=0)	0.001	0.029	0.001	0.014	0.069	0.035
Cragg-Donald F-stat	542.01	631.95	508.46	619.26	715.55	589.27
Number of students	24768	24768	24288	25191	25191	24730
Observations	99835	99835	97483	104791	104791	102276

Dependent variable as noted in panel heading. Test scores measured in SDs normalized at the grade-year level for the entire state. Start time and its interaction with puberty are instrumented by time zone and the interaction of time zone and puberty. Sample is all children who moved more than 25 miles. All specifications include age-gender dummies, longitude controls, school demographic means (FRL, male, black, Asian, and Hispanic), and individual fixed effects. Standard errors in parentheses and clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A4: Florida school and peer characteristics on move

	% FRL (1)	% male (2)	% black (3)	% Hispanic (4)	% Asian (5)	S/T (6)	Med income (7)
Move, ET-ET	-4.494***	-0.452***	0.186	-0.100	0.263***	0.258***	1010.277*
	(0.726)	(0.118)	(0.801)	(0.224)	(0.059)	(0.081)	(601.359)
Move, CT-CT	-1.681***	-0.316***	-0.582**	0.110***	-0.011	0.190***	-429.606***
	(0.280)	(0.054)	(0.227)	(0.037)	(0.025)	(0.038)	(162.849)
Move, ET-CT	0.115	-0.009	-15.350***	0.025	0.426***	0.124	-4778.338***
	(0.923)	(0.162)	(1.015)	(0.183)	(0.084)	(0.103)	(731.901)
Move, CT-ET	-4.513***	-0.557***	13.965***	0.495***	0.023	0.113	5729.001***
	(0.939)	(0.163)	(1.010)	(0.166)	(0.088)	(0.101)	(752.117)
P(ET-CT=CT-ET)	0.002	0.029	0.000	0.105	0.003	0.944	0.000
Observations	31763	31763	31763	31763	31763	31763	27747

Dependent variable as noted in panel heading. Regression is of school/zip summary stat on move, with student X moving event FE. Standard errors in parentheses and clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A5: Alternative definitions of puberty

	Math (SDs)				Reading (SDs)			
	(1) Preferred	(2) Stage 2	(3) Stage 4	(4) BG	(5) Preferred	(6) Stage 2	(7) Stage 4	(8) BG
Start time - sunrise (h)	0.009	0.011	0.032	0.025	0.061*	0.057	0.056	0.058
	(0.035)	(0.036)	(0.035)	(0.035)	(0.036)	(0.036)	(0.036)	(0.036)
Start time X puberty	0.073***	0.064***	0.029	0.040**	-0.004	0.003	0.006	0.002
	(0.019)	(0.019)	(0.020)	(0.019)	(0.020)	(0.020)	(0.021)	(0.020)
P(Start+Start X puberty=0)	0.001	0.003	0.005	0.008	0.014	0.012	0.002	0.010
Cragg-Donald F-stat	542.01	566.32	444.15	542.35	619.26	655.35	487.58	615.52
Number of students	24768	24768	24768	24768	25191	25191	25191	25191
Observations	99835	99835	99835	99835	104791	104791	104791	104791

Dependent variable as noted in panel heading. Test scores measured in SDs normalized at the grade-year level for the entire state. Absentee rate is the fraction of days the child missed school. Start time and its interaction with puberty are instrumented by time zone and the interaction of time zone and puberty. Sample is all children who moved more than 25 miles. All specifications include age-gender dummies, longitude controls, school demographic means (FRL, male, black, Asian, and Hispanic) and individual fixed effects. Standard errors in parentheses and clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A6: Academic and behavioral outcomes on start time, with student fixed effects

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Panel A: First stage, relative start time (hours)						
CT (=1)	0.598***	0.475***	0.585***	0.547***	0.475***	0.584***	0.547***
	(0.015)	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)
Observations	115778	115778	115778	115778	115778	115778	115778
	Panel B: IV estimates, math test scores (SDs) on relative start time						
Start time - sunrise (h)	-0.005	0.047	0.048*	0.044	0.045	0.048*	0.043
	(0.019)	(0.034)	(0.028)	(0.029)	(0.034)	(0.028)	(0.029)
Cragg-Donald F-stat	2254.173	744.796	1120.532	1002.330	746.364	1120.434	1003.020
	Panel C: IV estimates, reading test scores (SDs) on relative start times						
Start time - sunrise (h)	0.061***	0.081**	0.069**	0.059**	0.080**	0.069**	0.059**
	(0.019)	(0.032)	(0.028)	(0.028)	(0.032)	(0.028)	(0.028)
Cragg-Donald F-stat	2587.05	911.72	1209.23	1151.57	913.31	1209.80	1152.03
	Panel D: IV estimates, absence rate (%) on relative start times						
Start time - sunrise (h)	-0.664**	-1.539***	-0.549	-0.670*	-1.510***	-0.559	-0.672*
	(0.275)	(0.501)	(0.391)	(0.407)	(0.499)	(0.389)	(0.405)
Longitude	No	Yes	Yes	Yes	Yes	Yes	Yes
District quality	No	No	Yes	No	No	Yes	No
School quality	No	No	No	Yes	No	No	Yes
Time since move	No	No	No	No	Yes	Yes	Yes
Cragg-Donald F-stat	1394.52	475.67	721.91	669.77	476.44	722.82	669.98

Dependent variable as noted in panel heading. Test scores measured in SDs normalized at the grade-year level for the entire state. Absentee rate is the fraction of days the child missed school. Relative start time instrumented by time zone. Sample is all children who moved more than 25 miles. All specifications include age-gender dummies and individual fixed effects. Sample size is fixed within panels: 34018 students and 115778 student-years in Panel A, 24768 students and 99835 student-years in Panel b, 25191 students and 104791 student-years in Panel C, and 15906 students and 66263 student-years in Panel D. Standard errors in parentheses and clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A7: Hours of sleep by time zone

	(1)	(2)	(3)
Central	0.081	0.103	
	(0.088)	(0.131)	
Puberty	-0.451***	-0.804***	-0.676***
	(0.055)	(0.122)	(0.134)
Weekend	1.421***	1.192***	1.229***
	(0.102)	(0.158)	(0.158)
Central X weekend	-0.107	-0.166	-0.102
	(0.156)	(0.194)	(0.188)
Central X puberty	0.218	0.183	0.257
	(0.139)	(0.185)	(0.195)
Weekend X puberty	0.384***	0.616***	0.586***
	(0.087)	(0.161)	(0.150)
Central X wkend X puberty	-0.215	-0.149	-0.229
	(0.168)	(0.239)	(0.224)
P(Central + Central X weekend = 0)	0.830	0.566	
P(Central + Central X puberty = 0)	0.074	0.085	
Demographic controls	No	Yes	No
Student fixed effects	No	No	Yes
Observations	6,084	3,737	6,084

Dependent variable is hours of sleep per night. Sample is all children 6-19 in the Child Development Supplement of the Panel Study of Income Dynamics within 400 miles of the ET-CT time zone boundary in a state with a single time zone. Demographic controls in Column 2 include gender, race, and FRL status. Standard errors in parentheses and clustered at the state level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.