I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

Lifestyle and Mental Health Disruptions during COVID-19

Osea Giuntella
Kelly Hyde
Silvia Saccardo
Sally Sadoff

DISCUSSION PAPER SERIES

Lifestyle and Mental Health Disruptions during COVID-19

Osea Giuntella
University of Pittsburgh and IZA

Silvia Saccardo
Carnegie Mellon University

Kelly Hyde
University of Pittsburgh

Sally Sadoff
University of California San Diego

ABSTRACT

Lifestyle and Mental Health Disruptions during COVID-19*

COVID-19 has affected daily life in unprecedented ways. Using a longitudinal dataset linking biometric and survey data from several cohorts of young adults before and during the pandemic (N=685), we document large disruptions to physical activity, sleep, time use, and mental health. At the onset of the pandemic, average steps decline from 9,400 to 4,600 steps per day, sleep increases by about 25-30 minutes per night, time spent socializing declines by over half to less than 30 minutes, and screen time more than doubles to over 5 hours per day. The proportion of participants at risk of clinical depression increases to 65%, over twice the rate in the same population prior to the pandemic. Our analyses suggest that disruption to physical activity is a leading risk factor for depression during the pandemic. However, restoration of those habits–either naturally or through policy intervention–has limited impact on restoring mental well-being.

JEL Classification:	I10
Keywords:	COVID-19, mental health, lifestyle disruptions

Corresponding author:
Osea Giuntella
University of Pittsburgh
230 South Bouquet St.
Pittsburgh, PA, 15260
USA

E-mail: osea.giuntella@pitt.edu

Introduction

A mental health crisis has emerged during the COVID-19 pandemic. As of May 2020, about a third of Americans were estimated to be at risk of clinical depression and anxiety [US Census, 2020]. These rates are even higher among young adults, a population that has already seen a significant increase in the prevalence of mental health disorders over the past decade [Twenge et al., 2019]. An estimated 42% percent of adults age 18 to 29 are suffering from depression during COVID-19 compared to about 11% of young adults in 2019 [National Center for Health Statistics, 2019] and about 25% of college students prior to the pandemic [American College Health Association, 2018]. The rise in depression has occurred at the same time that stay-at-home orders, campus closures, and social distancing measures have caused major disruptions to everyday life, altering the way people live, work, study and interact.

In this paper, we document disruptions in physical activity, sleep and time use among young adults at the onset of the pandemic, and examine the relationship between these disruptions and mental health. We take advantage of a wellness study that has enrolled multiple cohorts of U.S. college students from February 2019 through June 2020. Participants received wearable devices (Fitbits) and answered repeated surveys about their well-being and time use over the course of a semester. Participants in the 2020 cohort began the study in February and continued participating after the university moved all classes online in March and encouraged students not to return to campus.

These data allow us to make two primary contributions. First, we can conduct longitudinal analysis examining how physical and mental health have evolved during the pandemic compared both to baseline pre-pandemic levels as well as to prior cohorts. Second, we can link biometric measures of physical activity and sleep to survey measures of mental well-being and social distancing. This approach allows us to identify risk factors for depression during COVID-19 and compare those factors to predictors of depression prior to the pandemic.

We first document large changes to physical activity and sleep. Over the course of the three-month semester, average steps decline by over half from 9,400 to 4,600 steps per day, overall physical activity declines by about 30%, and sleep increases by about 25-30 minutes per night. We also find dramatic shifts in self-reported time use. Time spent socializing with others

declines by over half to less than 30 minutes per day, while screen time more than doubles to over 5 hours per day (excluding screen time for classes or work). These lifestyle disruptions stand alongside stark increases in anxiety and depression during the pandemic. We estimate that at the end of the Spring 2020 semester in April, 65% of our participants were at risk of clinical depression. This is more than double the rate in the same population just two months earlier prior to COVID-19. Using difference-in-differences and individual fixed effects regressions, we show that the changes in physical activity, sleep, social interactions, screen time and depression are all statistically significant compared to changes in prior cohorts ($p < 0.001$). The concurrent decline of both physical and mental health is particularly worrisome, as prior work suggests that the co-existence of mental health problems alongside poor physical health worsens overall health outcomes [Scott et al., 2016].

To link lifestyle and mental health, we exploit our rich longitudinal data and use tree-based classification methods to identify risk factors for depression during COVID-19. Taken together, the baseline (pre-pandemic) predictors of depression in the 2020 cohort differ significantly from prior cohorts ($p < 0.001$). When we examine specific risk factors, we find that that lifestyle behaviors are more closely linked to depression during the pandemic than in prior cohorts. In contrast, baseline (pre-pandemic) measures of depression become substantially less predictive of endline depression two months later. That is, low baseline levels of depression do not necessarily protect against depression during the pandemic as they normally would in pre-pandemic cohorts. We additionally find evidence that those most resilient to stress and least prone to anxiety may be especially protected against depression during the pandemic – with these traits less important in pre-pandemic cohorts.

Finally, some of the typical relationships between baseline habits and endline depression diverge during the pandemic. For example, in prior cohorts risk of endline depression is lower for those whose physical activity is near recommended levels and who report more time spent socializing. These findings are in line with research suggesting an association between exercise, social connectedness and mental health [Salmon, 2001, Ströhle, 2009, Byrne and Byrne, 1993, Chekroud et al., 2018, Cruwys et al., 2013]. During COVID-19, these relationships between baseline behaviors and endline depression do not hold: participants who prior to the pandemic walk about the recommended number of steps and socialize more are at *higher* risk

3

of depression. Similarly, while baseline sleep duration is an important predictor of depression in prior cohorts, the relationship between sleep and depression is largely flat in 2020. Our results suggest that having healthy habits prior to the pandemic does not necessarily protect against depression during the pandemic.

These atypical relationships suggest that mental health during the pandemic may be driven less by whether participants have healthy habits prior to the pandemic and more by whether they experience large disruptions to those habits. We investigate this hypothesis and find that large disruptions in physical activity, sleep and time use are generally associated with higher rates of depression in the 2020 cohort. In particular, large disruptions in physical activity emerge as a leading risk factor for depression during COVID-19. Together with anxiety and resilience, declines in physical activity are the most important predictor of depression during the pandemic, more so even than baseline depression. In contrast to pre-pandemic cohorts, in which there is little relationship between disruptions and mental health, those participants who sustain their baseline exercise habits during the pandemic are at significantly lower risk of depression.[1]

Building on this analysis, after the Spring 2020 semester ended in April we continued to track a subsample of our participants through mid-June. In May 2020 we find evidence of a "bounce back" in behaviors with physical activity–and to a lesser extent sleep and time use–moving towards baseline pre-pandemic levels. However, this bounce back in lifestyle habits is not predictive of depression. While declines in physical activity are a leading predictor of depression in April 2020, reductions in those declines towards baseline levels do not predict depression in May 2020. These findings suggests that while maintaining habits is strongly associated with protection from the mental health effects of the pandemic, restoring those habits after they have been disrupted may not have a symmetric effect on restoring mental health.

To further test this hypothesis, we implemented a randomized experiment in June 2020. In the experiment, we randomized half our participants to receive incentives for walking at least 10,000 steps per day for two weeks. Our intervention significantly increased daily steps and physical activity, with the treatment group close to their baseline pre-pandemic levels. However, the impact on exercise did not translate into an improvement in mental health.

Taken together, our evidence suggests that the relationship between physical activity and

---

[1]These results do not appear to be driven by students' location during the lockdown.

mental health may not be symmetric. Whereas disruptions in physical activity may accelerate symptoms of depression, improvements in physical activity after habits have been disrupted may not lead to rapid recovery. This could occur, for example, if it takes longer to restore mental health than it does to upend it. It could also be the case that the relationship between physical activity and depression is driven more by mental health than it is by lifestyle habits. For example, the strong association between maintenance of healthy habits and depression during COVID-19 could partially reflect individuals' ability to adapt to adversity and sustain their lifestyle despite the pandemic. Such resilience in the face of large disruptions may be critical for well-being during COVID-19.

Our study contributes to the growing literature examining the impact of the coronavirus pandemic on physical and mental well-being. Lifestyle disruptions during COVID-19 have been documented in studies focusing on a single type of behavior, such as exercise [Fitbit Inc., 2019], sleep [Ong et al., 2020], social distancing [Allcott et al., 2020, Biroli et al., 2020] or mental health [Adams-Prassl et al., 2020b, Brodeur et al., 2020a, Beland et al., 2020, Brooks et al., 2020, Fetzer et al., 2020b,a, Gunnell et al., 2020, Office for National Statistics, Banks and Xu, 2020, Proto et al., 2020]. Related work using cross-sectional data finds an association between self-reported changes in physical activity during the pandemic and measures of mental health [Mayer et al., 2020].

This paper also relates to the broader research on the determinants of mental health [Davy-dov et al., 2010, Ridley et al., 2020, De Quidt and Haushofer, 2016, Allen et al., 2014] as well as work on health behavior change. Prior studies demonstrate how changing circumstances or context can quickly disrupt healthy habits [Wood et al., 2005, Milkman et al., 2014]. In addition to documenting such disruptions as a consequence of the pandemic, our work is among the first to investigate the relationships between lifestyle disruptions and well-being.

Methods

Three cohorts of students from the University of Pittsburgh received a wearable tracker (a Fitbit Alta HR device) and installed a custom-made smartphone app on their phone, which allowed us to track their activity data throughout the semester: Spring 2019 (n=140), Fall 2019

(n=317) and Spring 2020 (n=331).[2] Upon receiving the device, each participant completed a baseline survey with their demographics, self-reported health habits and mental health (see Online Appendix for the full survey). At the end of the semester, participants filled out an endline survey repeating most of the measures collected at the beginning of the semester, including mental health. In the 2020 cohort we added a mid-semester survey. We administered the survey on March 20th, a little over a week after the University of Pittsburgh announced (on March 11th) that classes would be moved online. The university also strongly encouraged students not to return to campus after spring break, which was extended by a week to end on March 22nd. This occurred at the same time that schools and businesses were shutting down in western Pennsylvania. In the mid-semester survey, we collected measures of mental health and surveyed students about changes to their lifestyle as a consequence of, and worries about, the pandemic (see Online Appendix for the full survey). We also collected information on individual location after the University announced remote learning. By the end of the term, approximately 77% of the students were not in Pittsburgh.

We continuously collected Fitbit data throughout the study, including steps, minutes of physical activity, resting heart rate, bedtime, wake up time, sleep duration and sleep disruption. Physical activity is measured as minutes in which a person is non-sedentary for at least ten continuous minutes, where non-sedentary minutes are defined as activity that raises heart rate enough to burn at least three times as many calories as at rest. Resting heart rate is the number of heart beats per minutes while at rest (the typical resting heart rate is 60-80 beats per minute). Sleep disruption is the percent of minutes in bed that a person is not asleep. We used wake up time and bedtime to create a measure of midsleep, which is the time at which a person has completed half their sleep for the night. Prior work suggests that circadian rhythm disruptions, reflected in changes in midsleep, are a symptom of depression [Li et al., 2013, Saxvig et al., 2012, Giuntella et al., 2017].

Throughout the semester we collected weekly measures of time use through a diary survey following the structure of the American Time Use Survey [Abraham and Flood, 2009]. We cat-

---

[2]Of these, 131 participants enrolled in the study in Fall 2019 and continued their participation in Spring 2020. We are missing some outcomes measures for some participants. Our analysis includes all participants who have at least one observation for the relevant outcome measure. Limiting the analysis to participants who have observations for all outcome measures does not change the results (Table S.3, Online Appendix).

egorize time use into the following categories: social interactions, screen, work, study, eating or preparing food, personal care, sleep, exercise, errands and commuting (see Online Appendix for the full breakdown). In our analyses, we do not use the data on sleep, exercise and commuting by walking, which we collect from the Fitbit. Social interactions measure time spent hanging out with friends. Screen time includes time spent watching TV, playing video games or surfing the internet, and excludes study or work time on the screen.

Our primary measure of mental health is depression, which we assessed using the Center for Epidemiologic Studies Depression Scale (CES-D, [Devin et al., 1988]). The CES-D is a validated self-report instrument designed to assess depressive symptoms in the general population. This 20-item scale assesses the frequency of symptoms of depression, such as helplessness or loneliness, on a scale from 0 (Rarely or None of the time) to 3 (Most or All of the time) and has a total score between 0-60. Our primary benchmark for depression is a CES-D score of 16 or above, which is considered the cutoff for clinical concern, implying high levels of depressive symptoms [Radloff, 1977].

We additionally assessed anxiety, resilience and life satisfaction. We assessed anxiety using the Generalized Anxiety Disorder scale (GAD-7, [Spitzer et al., 2006]), which assesses the frequency of anxiety related symptoms using a scale from 0 (Not at all) to 3 (Nearly every day). The GAD-7 asks 7 questions for a total score between 0-21. Our primary measure from GAD-7 is the average score. We also examine whether a participants' score exceeds a threshold of 7 (GAD-7>7, see Online Appendix for analysis), which has been suggested as a benchmark [Plummer et al., 2016]. We assess resilience – the ability to bounce back from stress and adversity – using the Brief Resilience Scale [Smith et al., 2008]. Finally, we ask about general life satisfaction using a single item question ("In general, how satisfied are you with your life?") rated on a 4-point scale (1=Very satisfied to 4=Very dissatisfied).

We also collected biweekly measures of mood, stress, and resilience to stress, using experience sampling techniques [Larson and Csikszentmihalyi, 2014]. In the Spring 2020 cohort, we added weekly survey measures in late March for types of social interactions (face-to-face, by phone or by text) and loneliness (measured via experience sampling). For four weeks during the semester we randomly assigned participants to interventions aimed at improving sleep

habits (see Online Appendix for details).[3] We include controls for treatment assignment in our analysis and find no predictive power of treatment on our outcomes of interest (see Tables S.4-S.6 in Online Appendix).

In our main parametric and non-parametric analyses of physical activity, sleep and time use, we compare average outcome measures in the baseline period (from enrollment to February 16) to averages of those measures collected from the end of Spring Break, when students resumed their classes remotely (March 23rd), until the end of the semester (April 20th). We use the same time windows for the Spring 2019 cohort and select comparable dates for the Fall 2019 cohort (see Online Appendix). Since the pandemic may have begun to affect our participants' lifestyles well before the campus decided to close, in the Online Appendix we repeat all analyses using an alternative definition of pre- and post-pandemic in which we compare the baseline period (from enrollment to February 16) to the post-baseline period, February 17th to the end of the semester. The alternative data definition yields similar results.

Results

Lifestyle disruptions. Our biometric and time use measures reveal that the pandemic led to major disruptions in daily behavior. Figure 1 plots average daily physical activity and sleep across the semester for the Spring 2019 and Spring 2020 cohorts.[4]

In the Spring 2019 cohort, daily steps are fairly constant with an average of about 9,800 steps throughout the term (Panel A). At the beginning of the semester (February), the Spring 2020 cohort is statistically indistinguishable from the Spring 2019 cohort. In March, there is a sharp drop in the average number of steps from 9,400 to 4,600, an over 50 percent decline ($p < 0.001$, from a regression of difference-in-differences across cohorts, see Table S.2, Panel B column 1 in the Online Appendix). We observe a similar pattern for physical activity (Panel B). Time spent in active (non-sedentary) activities dropped by about 30% compared to the beginning of the semester. The decline in active hours throughout the term is 10 times larger than in 2019

---

[3]For the Spring 2020 cohort, the intervention period took place February 17 to March 6 and March 23 to March 28 with a pause for spring break.

[4]For ease of interpretation over time, we show results for the the Spring 2019 and Spring 2020 cohorts, excluding the Fall 2019 cohort. Pooling the prior cohorts (Spring 2019 and Fall 2019) does not change the results (See Online Appendix Figure S.1 and Table S.4).

($p < 0.001$, Table S.2, Panel B, col 2 in the Online Appendix).

We also find large disruptions in sleep (Panel C of Figure 1), as students started to sleep about 30 minutes more per night throughout the pandemic ($p < 0.001$, see Online Appendix Table S.2). As shown in Panel D, the increase in sleep is driven by later wake up times: while before the lockdown the average wake up time was around 9 am, by March 15 it shifted to around 10am. Bedtime also increased but by only 25 minutes (see Online Appendix Figure S.5 for bedtime). Previous studies document that misalignment of sleep timing with respect to the natural dark-light cycle may have detrimental effects on sleep quality, health, and depression [Li et al., 2013, Wittmann et al., 2006, Walker et al., 2020, Supartini et al., 2016]. Thus, the later timing of sleep during the pandemic may have contributed to mood disorders, or exacerbated depression symptoms in individuals predisposed for mental health disorders.

We next examine shifts in self-reported time use. Figure 2 shows average daily social interactions and screen time in February compared to April for the Spring 2019 and Spring 2020 cohorts. In the 2020 cohort, screen time - the time spent on screens outside of work or studying- more than doubled after the announcement that classes would be moved remotely, reaching an average of 5 hours per day at the end of the term (p<.001, see Online Appendix Table S.2, Panel C, column 2). By contrast, screen time averaged around 2 hours per day throughout the semester in 2019 with only a moderate increase at the end of the term. Similarly, we observe a substantial drop in the number of hours spent interacting with friends, from approximately a hour per day at the beginning of 2020 to less than 30 minutes per day at the end of April, an over 50% decline ($p < 0.001$, Online Appendix Table S.2, Panel C, column 1). This drop is consistent with self-reported declines in face-to-face interactions (see Figure S.6 in the Online Appendix). In the Online Appendix, we also document a drop in the number of work hours– driven by a subset of our participants who lost their jobs as a result of the campus closure–and a significant drop in the number of hours spent studying in the second half of the semester (Table S.2, Panel C).

Mental Health. As discussed in the Methods section above, our primary measures of mental health are assessments of depression (CES-D) and anxiety (GAD-7). Figure 3 shows average

Figure 1: Physical Activity and Sleep
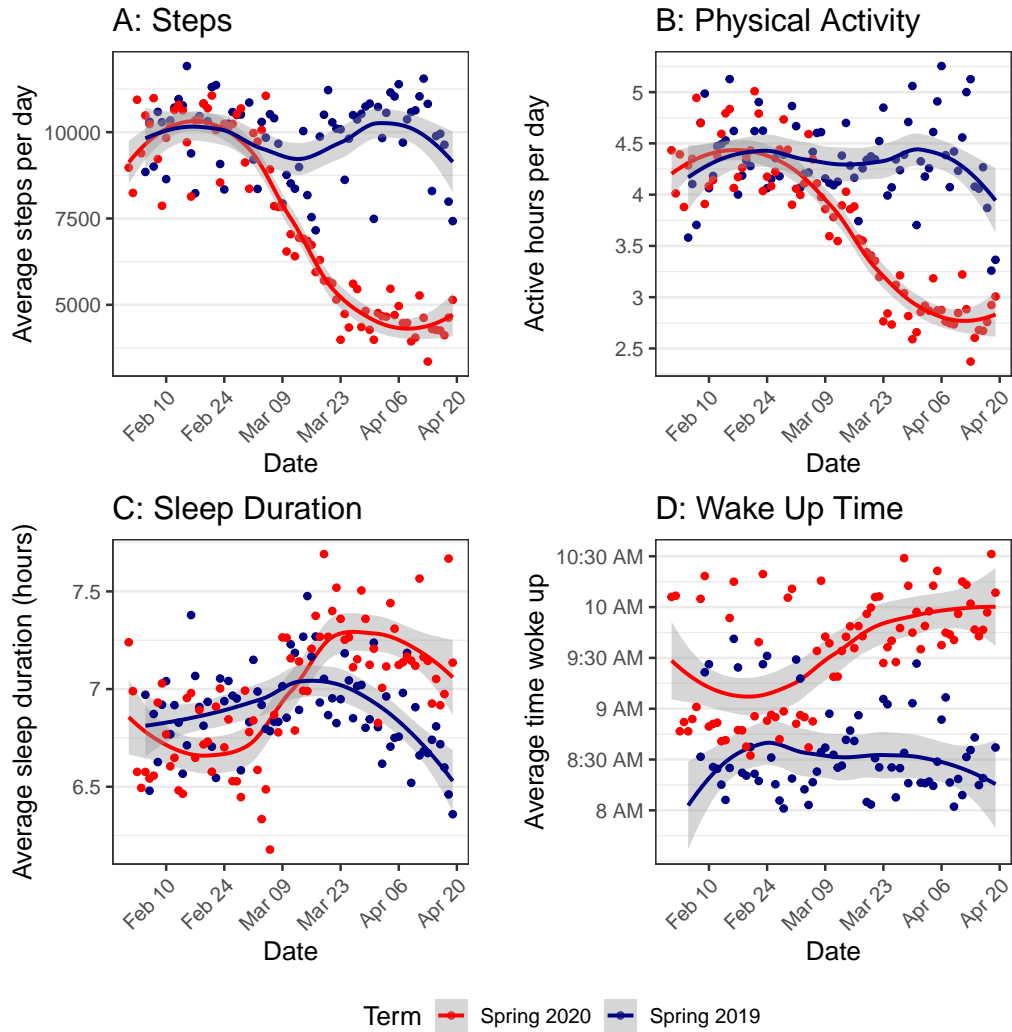
A: Steps

B: Physical Activity

C: Sleep Duration

D: Wake Up Time

Term — Spring 2020 — Spring 2019

Notes - The above figure plots the average outcomes by day for study participants in the Spring of 2019 (red) and Spring of 2020 (blue) cohorts. The curves fit a local polynomial regression (LOWESS).

Figure 2: Screen Time and Social Interactions
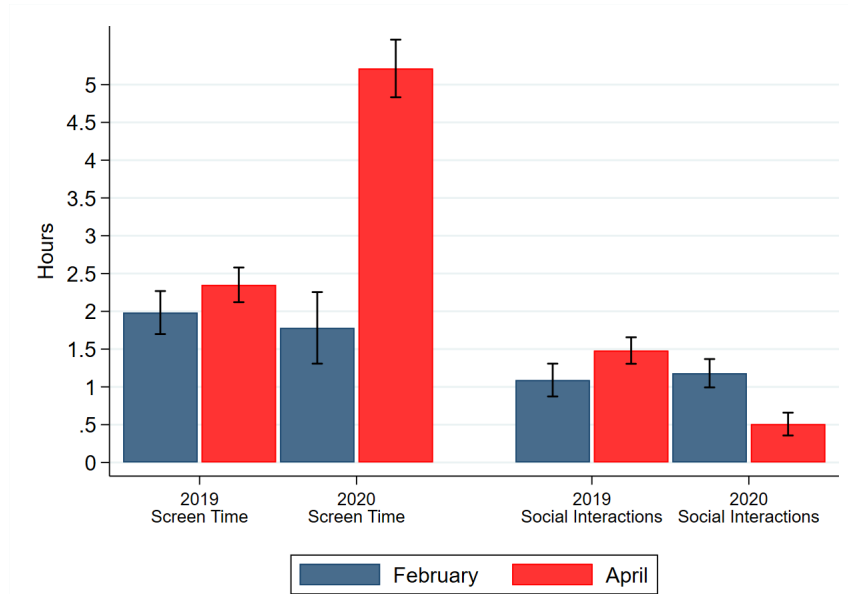
Notes - The figures show the average time spent with friends (social time) and the average screen time at the beginning (February) and end (April) of the semester during the Spring 2019 and Spring 2020 terms. Screen time includes time spent playing games, watching television, or surfing the Internet, and does not include time spent working or studying on a device. Bars indicate confidence intervals.

CES-D and GAD-7 scores for the Spring 2019 and Spring 2020 cohorts. We present the baseline measures taken at the beginning of the semester (February), the mid-semester measures taken in March (Spring 2020 cohort only) and the end of semester measures taken in April.

Figure 3: Depression and Anxiety
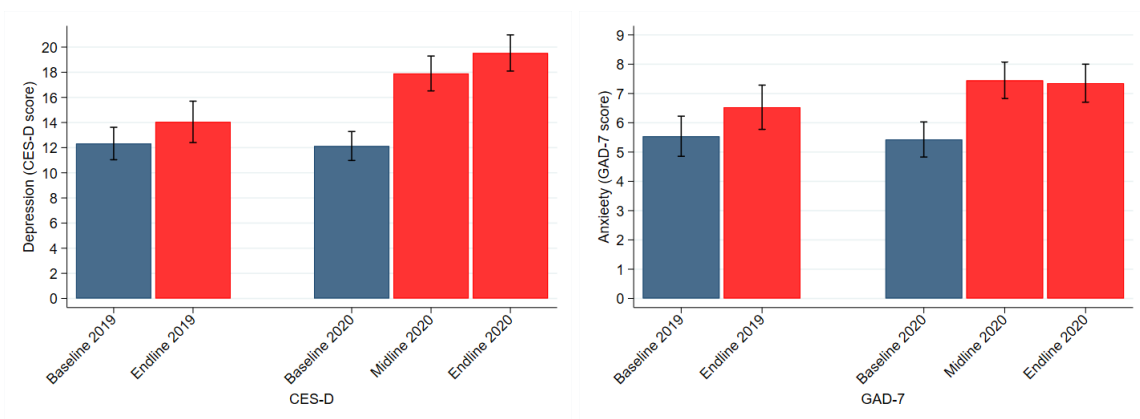
Notes - The figures show the average CES-D score and GAD-7 score at the beginning (February), middle (March, 2020 only) and end (April) of the semester during the Spring 2019 and Spring 2020 terms. Bars indicate confidence intervals.

Our results show large increases in anxiety and depression during the pandemic. In the Spring 2019 cohort there are only small increases in CES-D scores from the beginning to the end of the semester. By contrast, in the Spring 2020 cohort we see a substantial increase in the CES-D measure of depression, from an average of 12 to an average of 19 points ($p < 0.001$ from a a difference-in-differences regression across cohorts, see Online Appendix Table S.2, Panel A). We find a similar pattern when we look at anxiety using the GAD-7 scale. The anxiety score increases over 40 percent from 5.47 at baseline to 7.72 at midline, slightly decreasing to 7 at the end of the term ($p¡0.001$). By contrast, GAD-7 increases by about 9 percent from 5 to 5.5 in Spring 2019 ($p = 0.002$).

Figure 4 documents how the pandemic shifted the distribution of CES-D scores with a substantial increase in the share of subjects with a CES-D score above 15, the threshold commonly used to identify clinical depression. We estimate that as of April 2020, 65% of our participants were at risk of depression (a CES-D score of 16 or above). This is more than double the baseline rate of 31% just two months earlier prior to COVID-19 ($p < 0.001$). By comparison, these same rates increase by only 7 percentage points over the course of the Spring 2019 semester from 31% to 38% ($p = 0.105$).[5] Our baseline rates in both the 2019 and 2020 cohorts are similar to estimates from the broader college student population [American College Health Association, 2018].

The measures of anxiety and depression are highly correlated with self-reported life satisfaction ($r = -.57$, $p < 0.001$ for CES-D and $r = -.37$, $p < 0.001$ for GAD-7), which also decreases during the pandemic (Mean= 2.60 (sd=0.63) to 2.27 (sd=.73), $p < 0.01$). Further, these findings are corroborated by the experience sampling data on mood, which drops during the pandemic from 6.65 at baseline (in February) to 5.82 after the beginning of the lockdown in Pittsburgh (March 15) ($p < 0.01$), while experience sampling of stress remains similar ($p= 0.221$) and resilience only marginally decreased from 3.12 to 2.95 ($p = 0.532$).

---

[5]We note that that increase in the proportion of the population at risk for depression may overstate the true increase given that there are false positives using the CES-D scale.

Figure 4: Depression: Distribution of CES-D Score Across Cohorts
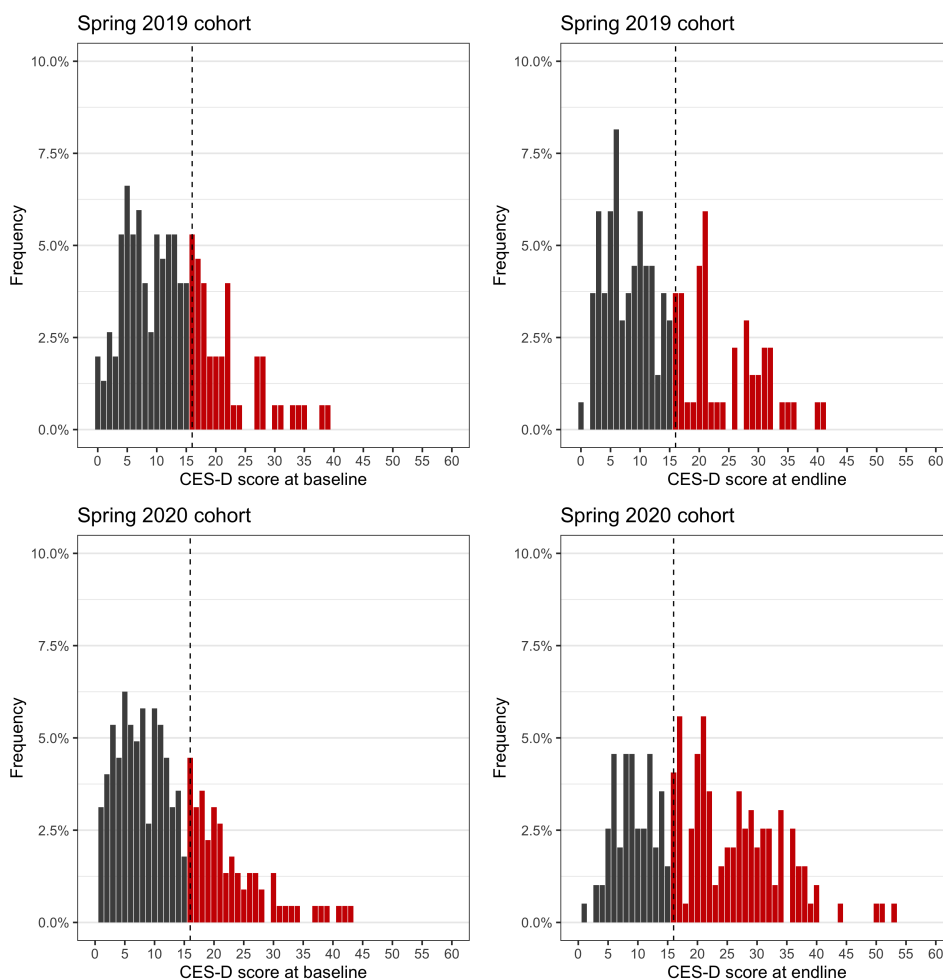
Notes - The figures show the distribution of CES-D score at baseline and endline during the Spring 2019 and Spring 2020 terms.

Predicting risk factors for depression. To better understand the dramatic rise in depression during COVID-19, we combine our rich data to identify risk factors for depression in the Spring 2020 cohort. We then compare those predictors to prior cohorts (pooling the Spring 2019 and Fall 2019 cohorts). We focus on risk factors for having an end of semester CES-D score that meets or exceeds 16, the threshold for clinical depression [Radloff, 1977]. Our baseline predictors capture measures collected over 14 days in February for the spring cohorts and 26 days in September for the fall cohort. We combine Fitbit data of physical activity and sleep with our survey measures of time use and demographics (see Online Appendix for the full list

13

of variables).

We feed the baseline variables as potential features into the XGBoost machine learning algorithm [Chen et al., 2019], a flexible and robust decision tree-based classification method. This methodology has two main advantages. First, as a decision-tree based algorithm, it is robust to multicollinearity among included features. This is relevant in our context because we have very detailed measures of interrelated individual behavior (e.g. time use categories, since more time spent on one type of activity necessarily implies less time spent on alternative activities). Second, it is nonparametric, and thus can flexibly account for relationships between features and depression without needing to specify a functional form ex-ante. This allows us to identify nonlinear relationships between the predictors and risk of depression, e.g. higher rates of depression associated with getting either too little or too much sleep.

The algorithm produces three main outputs of interest. First, overall measures of predictive accuracy, sensitivity and specificity. Predictive accuracy is the overall percentage of observations that are correctly predicted by the model, sensitivity is the percentage of positive cases (i.e., CES-D > 15) that are correctly predicted by the model, and specificity is the percentage of negative cases (i.e., CES-D ≤ 15) that are correctly predicted by the model. Second, a measure of the relative importance of each feature on a scale from 0 to 1, which approximates the average gain in predictive accuracy from using that feature in the model. Third, a matrix of SHAP (Shapley Additive Explanations) scores, which approximate the marginal contribution of a feature to a particular observation's predicted risk. To protect against overfitting, we use 10-fold cross-validation to determine the optimal number of boosting iterations. To address the potential for parameter instability, we repeat the entire process of fitting each model 10,000 times and average the estimated feature importance and SHAP values across those iterations. We conduct this process for the Spring 2020 sample and repeat it for the pooled Spring 2019 and Fall 2019 comparison group. The pooled 2019 model achieves 88% predictive accuracy with 88% sensitivity and 88% specificity. The 2020 model achieves 89% predictive accuracy with 88% sensitivity and 91% specificity (see Online Appendix for full description).[6]

---

[6]This analysis only includes participants who have both baseline and endline measures of mental health: Spring 2019 (n=133), Fall 2019 (N=279), Spring 2020 (N=211). There are 53 subjects who participated both in Fall 2019 and Spring 2020 and who responded to the Spring endline survey. These participants are included in both the Fall 2019 and Spring 2020 analysis. For these participants, their fall baseline measures of mental health serve as their

Before pooling the 2019 cohorts, we first conducted this process separately for each cohort: Spring 2019, Fall 2019 and Spring 2020. We then examined the accuracy of the Spring 2019 model for predicting endline depression in Fall 2019 compared to predicting endline depression in Spring 2020. The Spring 2019 model achieves 78% accuracy predicting endline depression in Fall 2019, significantly exceeding the no information rate (the average prediction accuracy of randomly assigning endline depression) of 58% (p < 0.001). In contrast, it achieves 64% accuracy predicting endline depression in Spring 2020. This coincides with the no information rate, which is also 64% (p = 0.60). The difference in predictive accuracy for the Fall 2019 and Spring 2020 cohorts is statistically significant ($p < 0.001$). That the Spring 2019 model does a better job predicting depression in Fall 2019 than in Spring 2020 is particularly noteworthy because we might think that due to seasonality the two spring semesters should be more similar, or due to changes across academic years the Fall 2019 and Spring 2020 semesters should be more similar. These results suggest that the predictors of depression during the pandemic are exceptional compared to prior cohorts.

We find further support for this hypothesis when comparing the specific risk factors for depression across cohorts. Figure 5 reports the cumulative importance–adding up to 1–of the different features for the pooled 2019 and Spring 2020 cohorts, grouped by category: mental health, physical activity, sleep, time use, demographics and other, which includes baseline self-reported health and treatment assignment in our sleep intervention (see Online Appendix for individual features in each category). Below the figure we list the three most important features for each model along with their relative importance.

Panel A shows the results of the models that predict endline depression using baseline (pre-pandemic) measures. There are substantial shifts in the importance of each category across cohorts. For the 2019 cohorts, who participated prior to the pandemic, baseline measures of mental health at the beginning of the semester largely explain depression rates at the end of the semester. Baseline measures of depression, anxiety, resilience and life satisfaction account

---

baseline measures in Spring 2020. Excluding these participants from the Spring 2020 analysis does not affect the results.

Figure 5: Importance

Baseline Models / Difference Models

Top features and importance:

Pooled 2019	Spring 2020	Pooled 2019	Spring 2020
CES-D: 0.375	GAD-7: 0.169	CES-D: 0.358	GAD-7: 0.141
GAD-7: 0.089	CES-D: 0.108	GAD-7: 0.083	Phys. act. diff: 0.120
Life sat.: 0.076	Resilience: 0.096	Life sat.: 0.070	Resilience: 0.089
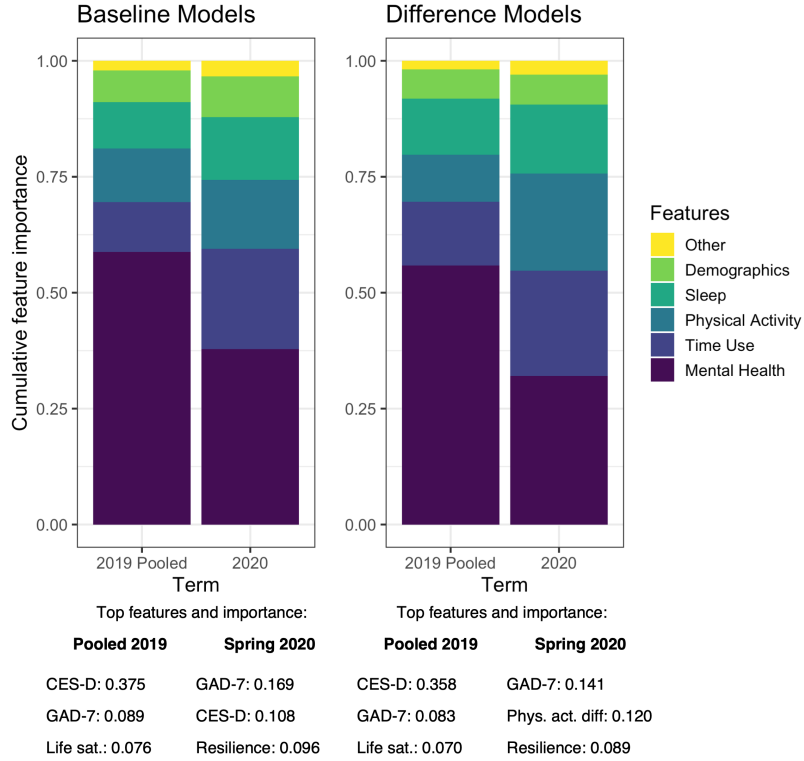
Notes - The figure displays the relative importance (out of 1) in the 2019 pooled cohorts and 2020 cohort of variables grouped by: mental health, time use, physical activity, sleep, demographics and other (baseline self-reported health and treatment assignment in our sleep intervention). The left panel models include baseline demographics, mental health, physical activity, sleep, and time use variables. The right panel models include baseline demographics and mental health; and, differences between endline and baseline levels of physical activity, sleep, and time use. See Online Appendix Tables S.11 and S.12 for the full breakdown.

for over half of the cumulative importance in the model, with baseline depression by far the leading factor, accounting for an estimated 38% of the predictive accuracy of the model on average. The cumulative importance of the baseline mental health measures declines during the pandemic from 0.55 in 2019 to to 0.375 2020. This shift is driven entirely by baseline depression: its relative importance declines from 0.38 in 2019 to 0.11 in 2020 and it is no longer the most important predictor of endline depression. That is, low baseline depression risk does not necessarily protect against depression during the pandemic. Instead, baseline anxiety emerges as the leading predictor of depression during the pandemic, increasing in relative importance from 0.09 in 2019 to 0.17 in 2020. We also see an increase in the relative importance of baseline measures of resilience from 0.05 in 2019 to 0.1 in 2020 (it is the 5th most important factor in 2019, see Online Appendix Table S.11).

The difference in the strength of baseline CES-D as an endline predictor during the pandemic is evident when comparing the SHAP values across cohorts. Figure 6 Panel 1A plots estimated SHAP values by baseline values of CES-D, where a higher SHAP value indicates a higher risk of endline depression. In the 2019 cohort, there is a sharp increase in risk for participants with baseline scores of 16 or above compared to those with lower scores. That is, being depressed at baseline largely predicts being depressed at endline two months later. The relationship has the same shape but is significantly flatter for the 2020 cohort – where prepandemic measures of depression do not map as strongly onto endline depression during the pandemic. Turning to baseline anxiety and resilience (Figure 6 Panels B and C), the strength of the relationship with endline depression is largely similar across cohorts, except in the tails.

In the 2020 cohorts we see sharp declines in depression for those with the very lowest levels of anxiety and very highest levels of resilience. Such levels of anxiety and resilience have a weaker relationship with depression in the 2019 cohort. These results suggest that those most resilient to stress and least susceptible to anxiety may be especially protected against depression during the pandemic. This is in line with work suggesting that resilience protects individuals against stressful events [Davydov et al., 2010], and helps them preserve their health in spite of adversity [Patel and Goodman, 2007].

Figure 6: Baseline Measures and Depression
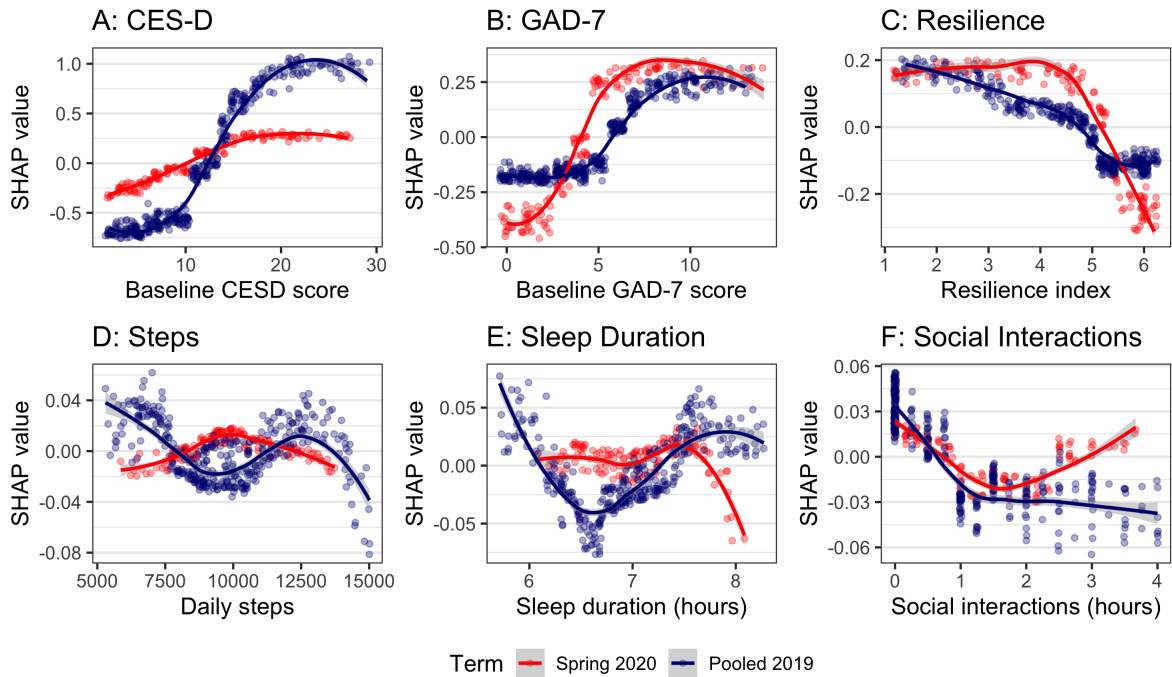
Notes - The figure displays the estimated additive feature contributions (SHAP values) for baseline CES-D scores (Panel A), baseline GAD-7 scores (Panel B), baseline Resilience (Panel C), baseline steps (Panel D), baseline Sleep duration in hours (E) and baseline social interactions (Panel F). A higher SHAP value indicates a higher risk of endline CESD>15.

18

As shown in Figure 5 Panel A, the relative importance of both biometric measures (physical activity and sleep) and self-reported time use measures increases in 2020 relative to 2019. In addition, unlike measures of baseline mental health–which largely differ across cohorts only in the strength of their relationship with endline depression–measures of physical activity, sleep and time use also differ across cohorts in the direction of their relationship with endline depression. For example, in 2019 risk of depression was highest for participants with lower baseline levels of physical activity (fewer than 7500 steps, Figure 6 Panel D; for the SHAP of physical activity see Online Appendix Figure S.9). This finding is consistent with prior research on the association between exercise and mental health [Salmon, 2001, Byrne and Byrne, 1993, Ströhle, 2009]. The relationship between depression and physical health is inverted in 2020 compared to prior cohorts, and is generally weaker. Participants whose baseline activity is about average (around 10,000 steps) are at highest risk for depression–the same activity levels that minimize risk reduction in prior cohorts.

We similarly find a divergence across cohorts for the relationship between baseline sleep habits and endline depression (Panel E). Risk reduction in the 2019 cohort is minimized for baseline sleep duration near the average of 6.5 to 7 hours of sleep per night, and increases sharply for those who sleep less than 6 hours at baseline. This is consistent with previous studies analyzing the relationship between sleep duration and depression [Zhai et al., 2015, Kaneita et al., 2006]. By contrast, in the 2020 cohort baseline sleep duration has little relationship with endline depression.[7]

Finally, as shown in Panel F, while the 2019 and 2020 cohorts show a similar increase in depression risk for those who report little or no time spent socializing, they diverge for those who socialize most at baseline. In the 2020 cohort, those who report the highest levels of socializing prior to the pandemic are at higher risk of depression during the pandemic. This runs counter to associations found in prior literature [Heinrich and Gullone, 2006, Kudielka et al., 2007, Adam et al., 2006], as well as the relationship in the 2019 cohorts in which those who socialize 1-4 hours per day are at the lowest risk for depression. We see a similar divergence in the relationship between screen time and depression across cohorts, as well as study and work

---

[7]There is a predicted decline in depression rates in the right tail of the data, but we interpret this with caution because it is based on fewer than 5% of our participants.

hours (see Online Appendix Figure S.8).

Lifestyle changes and depression. Why did many of the typical relationships between depression and baseline predictors differ in 2020 compared to past cohorts and the established literature? One potential reason is that the pandemic disrupted baseline habits in ways that also disrupted the associations between lifestyle and mental health. To investigate this hypothesis, we examine the extent to which the large disruptions to behaviors during the pandemic can help explain endline depression in 2020. We measure lifestyle changes as the difference between average behavior during the onset of the pandemic minus average baseline behavior. Figure 7 compares rates of depression among participants with smaller and larger disruptions using below- vs. above-median changes in: steps, physical activity, sleep duration, wake up time, social interactions and screen time. Depression rates are generally higher for participants who experience larger disruptions during the pandemic.[8] In particular, larger declines in physical activity, and especially active minutes, are associated with significantly higher rates of depression (p=.029, see regression in Online Appendix Table S.7). In the Online Appendix, we additionally examine the relationship between students' location during lockdown and depression (by the end of the term approximately 77% of students were not in Pittsburgh). We find no evidence of an association between depression and locations with higher COVID-19 cases or deaths (Table S.10).

Building on the above results suggesting the importance of lifestyle disruptions – rather than levels – we repeat our prediction exercise and focus on changes in lifestyle measures (i.e., physical activity, sleep and time use) instead of using those measures at baseline as predictors. We retain baseline measures of mental health and demographics. The 2019 differences model achieves 90% across the board (accuracy, sensitivity, specificity), while the 2020 differences model achieves 94% accuracy, 93% sensitivity and 96% specificity. In both 2019 and 2020, the differences model achieves higher levels of accuracy, sensitivity and specificity than the

---

[8]The exception to this pattern is social interactions, which potentially reflects the u-shaped relationship of depression with baseline social interactions discussed above – i.e., those with the lowest and highest baseline social interactions are at the highest risk for endline depression. In line with these results, there is also a u-shaped relationship with changes in social interactions and depression: those with the smallest changes (who were also those socializing least at baseline) and those with the largest changes (who were also those socializing most at baseline) have the highest rates of depression, with lower rates of depression for those just below and just above the median.

20

baseline model, but the differences are not statistically significant.

Figure 7: Habit Disruptions and Depression
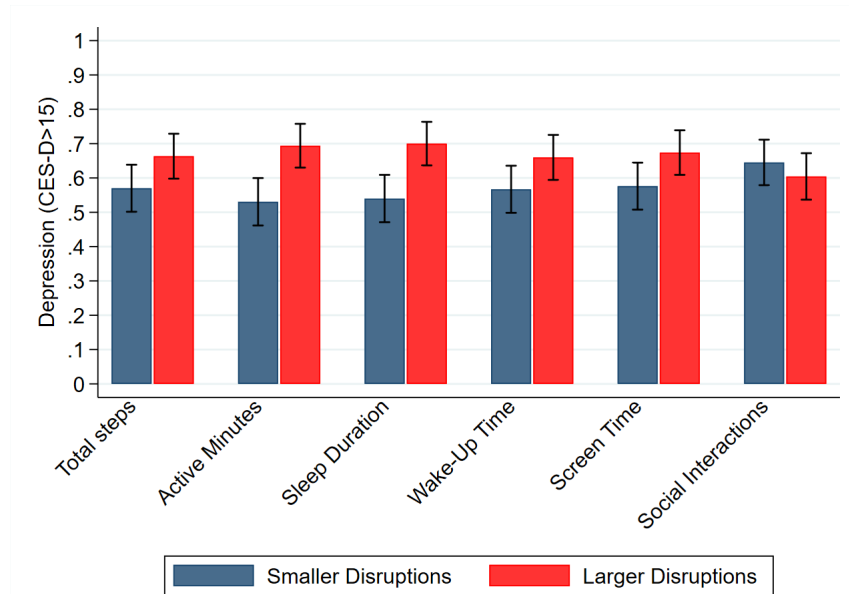
Notes - The figure reports the proportion of individuals reporting clinical depression (CES-D>15) below (Smaller Disruptions) and above (Larger Disruptions) median change in steps, sleep, social, and screen time.

The results of the models using differences in lifestyle behaviors are summarized in Panel B of Figure 5. First, this analysis reveals that baseline depression is no longer among the three top predictors of depression during the pandemic, as it was in 2019. Second, disruptions in physical activity emerge as a critical predictor of depression during the pandemic, along with baseline anxiety and resilience. Baseline anxiety and resilience are also associated with smaller disruptions to physical activity during the pandemic ($p = 0.015$ for GAD-7, $p = 0.062$ for resilience) with little relationship in the pre-pandemic cohorts ($p = 0.54$ for GAD-7 and $p = 0.631$ for resilience). These results reaffirm the findings discussed above for the Baseline model suggesting the importance of high levels of resilience and low levels of anxiety in protecting against depression during the pandemic; and link these traits to maintenance of physical habits in the face of the large disruptions caused by COVID-19.

The importance to well-being of maintaining habits is illustrated in Figure 8, which displays the SHAP for changes in physical activity and sleep. For physical activity (Panel A), we see that in the 2020 cohort, those with the largest decreases in active minutes are at significantly higher risk of endline depression. In 2019 there is no such relationship. The high-risk group in 2020 experiences a decline of about 1-3 fewer daily active hours, with disruptions of such magnitude largely absent in the 2019 cohort. Importantly for the 2020 cohort, those participants who maintain daily active hours similar to baseline (i.e., differences near zero) demonstrate strikingly lower risk of endline depression – a pattern that does not emerge for the 2019 cohort. These results suggest that sustaining healthy physical habits is strongly associated with well-being during the pandemic.

Figure 8: Changes in Activities and Depression
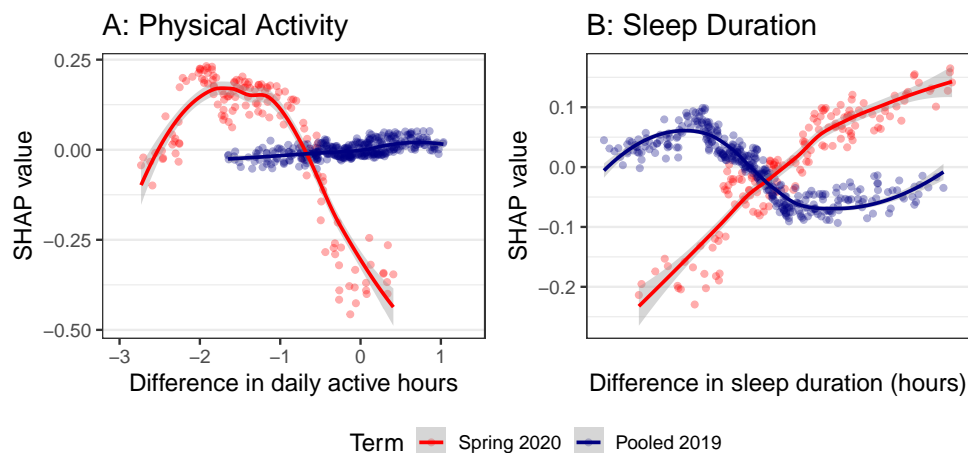
Notes - The figure displays the estimated additive feature contributions (SHAP values) for changes in physical activity, as measured in the number of active hours (left panel), and sleep duration (right panel). A higher SHAP value indicates a higher risk of endline CES-D>15.

We find a similar pattern for changes in sleep (Figure 8 Panel B). In 2020, risk of depression increases substantially for those with the largest increases in sleep during the pandemic while risk is lowest for those with no change or small decreases. In the 2019 cohort, the relationship is significantly flatter and moves in the opposite direction: decreases in sleep are associated with higher levels of depression while increases are associated with lower levels, consistent with

previous studies on the relationship between sleep duration and depression [Tsuno et al., 2005, Giuntella et al., 2017, Kaneita et al., 2006]. In the Online Appendix, we also show that limited changes in most lifestyle habits (e.g., steps, social interactions) are associated with lower risks of depression during the pandemic (Figure S.9).[9]

Physical Activity Bounce Back and Mental Health. In the 2020 cohort, we continued to track a subsample of our participants after the semester ended in April. These participants (n=205) agreed to continue wearing the Fitbit and complete weekly time use surveys as well as mental health surveys in May and June.[10] We find that in May, behaviors–and physical activity in particular–demonstrate a "bounce back" effect moving directionally towards baseline levels. As shown in Figure 9 Panels A, we find that physical activity increases in May to an average of 3 hours and 45 minutes compared to 3 hours and 10 minutes in April, closing 40% of the decline from baseline ($p < 0.01$, Online Appendix Table S.15). Daily steps similarly increase (Panel B) from 4400 in April to 5605 in May, which closes 27% of the decline from baseline ($p < 0.001$). We also observe small decreases in sleep duration (of about 10 minutes, $p < 0.05$) and a small increase in social interactions of about 10 minutes ($p < 0.1$), while screen screen time continues to increase ($p < 0.01$).

Our analysis of risk factors for depression during the pandemic predicts that such behavioral changes in the direction of baseline habits will be associated with lower levels of depression. To examine the role of bounce back, we test the predictions of our model (hereafter referred to as "the April model") out of sample. We use estimates from the April differences model that includes demographics, baseline measures of mental health and disruptions to the lifestyle behaviors collected during the semester (February – April) as predictors for end of semester depression. We apply those estimates to behaviors in May and examine how well they predict

---

[9]As shown in Figure 5, demographic features account for only a small share of the predictive accuracy for depression. In the Online Appendix we report regression analysis for demographic characteristics including gender, race/ethnicity, whether a student receives financial aid and whether a student is a first generation college student (Table S.8). Consistent with recent work [Adams-Prassl et al., 2020b], we find evidence that women experience larger increases in depression during the pandemic. We also explore the relationship between demographics and disruptions to physical activity and find that minority students demonstrate the largest declines in average daily steps (Online Appendix Table S.9).

[10]We examine differential attrition (Online Appendix Table S.19) and find no evidence that the subsample who continues differs significantly on demographics, baseline mental health, baseline physical activity or baseline sleep from those who attrit. The only statistically significant difference between the groups is that individuals who attrit from the sample report lower levels of baseline screen time.

depression out of sample at the end of May. Then to establish a benchmark, we repeat the XGBoost algorithm to generate a new model predicting depression in May directly with the same set of features (hereafter referred to as "the May model"). This allows us to compare predicted feature contributions (SHAP values) from the April model with the directly estimated contributions in the May model on a feature-by-feature basis.

We find suggestive evidence that the relationship between students' behaviors and depression are not symmetric—that is, rebounds in lifestyle behaviors are not as predictive of depression in May as the initial disruptions in those behaviors were of depression in April. This is particularly stark for differences in physical activity (see Panel C of Figure 9), which was a leading predictor of depression in the April model but has almost no discernible relationship with depression in the May model.

This finding suggests that, while the initial drop in physical activity at the onset of the pandemic was strongly associated with the resulting spike in depression among our sample, the degree to which participants resumed activity in the following months was not predictive of their recovery. Relative to physical activity, the April model performs better out-of-sample predicting the contributions of changes in sleep duration, social interactions, and screen time to risk of depression in May. However, the April model consistently overestimates the importance of these features relative to baseline CES-D score, which reemerges as a powerful predictor of depression in the May model, akin to the 2019 comparison models. This shift of predictive power back to the baseline mental health measures is the main reason why the April model underperforms predicting depression in May, achieving only 72% accuracy with 82% sensitivity and 62% specificity, while the May model achieves 98% accuracy, sensitivity, and specificity. The differences between the April and May model are statistically significant ($p < 0.001$).

The diminished importance of physical activity in the May model is consistent with the results of a randomized intervention to stimulate physical activity among our participants. In June 2020, we randomly assigned participants to receive incentives for walking a minimum of 10,000 steps a day. Participants in the treatment group received a monetary transfer of $5 every day they reached the minimum number of steps. The intervention began on June 1 and

24

Figure 9: Improved Physical Activity and Depression

Intervention Groups ● Control ● Treatment

Intervention Groups ● Control ● Treatment

Predicted (April model) ━━ Actual (May model)

Intervention Group ■ Control ■ Treatment
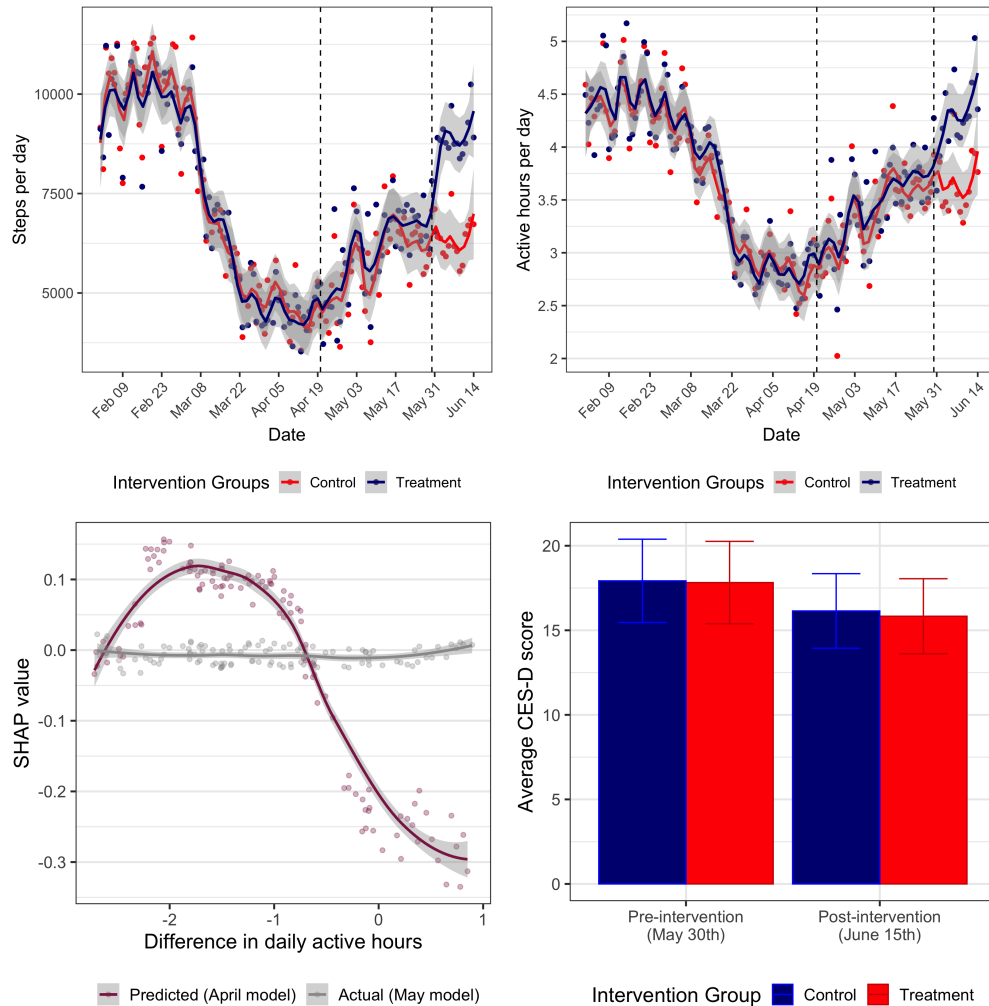
Notes - The top panels displays step (top-left) and active minutes (top-right) over time for the subset of participants who elected to participate in the second phase of the study. The dashed vertical lines indicate the end of the first phase of the study (April 20th) and the beginning of the intervention (June 1st). The bottom left panel indicates the predicted (based on the April model) and actual (based on the May model) relationship between CES-D and differences in the number of active hours between the end of the semester (April 20th) and the beginning of the intervention (June 1st). The bottom-right panel displays average CES-D scores before the intervention (May 30th) and after the intervention (June 15th) for the treatment and control group.

lasted 14 consecutive days. The control group received a similar distribution of payments (see Online Appendix for experimental procedures). After the end of the intervention, on June 16, we measured mental health again.

As shown on the right-hand sides of Figure 9 Panels A and B, the intervention had a large impact on physical activity, increasing average steps by about 2,000 steps ($p < 0.001$) and

active minutes by just over a half an hour, an 18% increase ($p < 0.001$). As a result, physical activity in the treatment group approached baseline pre-pandemic levels, averaging approximately 8,800 steps per day compared to about 6,400 in the control group. However, as shown in Panel D of Figure 9, we find no effect on CES-D scores measured at the end of the intervention period (see Online Appendix Table S.17 for regression analysis). We estimate a difference of -0.3 point (S.E. = 1.55, $p = 0.84$ compared to control). Our 95% confidence intervals rule out that treatment caused more than a 2 point change in the CES-D score, which is less than a third of the increase in average CES-D scores observed at the onset of the pandemic. We note that in both the control and treatment groups average CES-D scores improved in May and June compared to April (see Online Appendix Section 2.6). However, the evidence from both our bounce back analysis and randomized intervention suggests that increases in physical activity do not seem to accelerate these improvements in mental health.

Conclusion The global coronavirus pandemic has upended much of society in unprecedented ways. The measures adopted to mitigate the public health emergency, such as border closures, travel restrictions and lockdowns, have affected labor markets, consumption patterns, and economic activities all over the world [Coibion et al., 2020, Adams-Prassl et al., 2020b, Montenovo et al., 2020, Fetzer et al., 2020b, Brodeur et al., 2020b, Adams-Prassl et al., 2020a]. The impact of such disruptions on mental health is of critical policy concern. Over the last two decades mental health disorders have imposed a growing burden on society with estimated costs of over \$200 billion per year in the U.S. alone [Greenberg et al., 2015]. These costs may substantially increase as a result of the pandemic.

The consequences of COVID-19 for mental health have been dire, as highlighted in a May 2020 UN policy brief urging the international community to protect vulnerable populations [United Nations, 2020]. Among those identified as a specific population of concern were adolescents and young adults, who have faced large disruptions to their education and living situations and may suffer lifelong economic impacts from the pandemic. Our findings provide evidence for these disruptions and highlight the heavy toll of the pandemic on the well-being of college students.

We document several novel findings linking lifestyle disruptions to mental health. First, we show large disruptions to physical activity, sleep and time use, particularly at the onset of

the pandemic in March and April. Second, we document substantial declines in mental health with dramatic increases in depression. Third, we find that risk factors for depression diverge substantially during the pandemic compared to prior cohorts and the literature. In particular, having healthy habits and low risk of depression prior to the pandemic does not necessarily protect against depression during the pandemic, especially when those habits are not maintained while sheltering-in-place. Finally, while disruption of physical habits is a leading predictor of depression during COVID-19, the restoration of habits – either naturally or through policy intervention – has a limited impact on restoring well-being during the pandemic.

Our results suggest that the relationship between exercising habits and depression may not be symmetric: while disruptions may worsen mental health, the damage may not be easily undone once people are at significant risk for clinical depression. It is possible, for example, that our intervention to increase physical activity, which only lasted two weeks, was not long enough to lead to improvements in mental health. Future work could explore the effectiveness of longer interventions or attempt to restore physical health in conjunction with other important lifestyle habits (e.g., sleep, social interaction). Another possibility is that the relationship between physical activity and well-being is not driven primarily by lifestyle habits. For example, it may instead reflect that the kinds of people who are able to maintain their lifestyle during the pandemic are the kinds of people who are also better able to maintain their mental health in the face of major disruptions. We find evidence that these people may be those who, prior to the pandemic, are most resilient to stress and least prone to anxiety. Prior work has shown that it is possible to foster resilience [Chmitorz et al., 2018]. Future work could explore whether it is possible to do so in ways that help mitigate the large impact of the COVID-19 pandemic on physical and mental well-being.

References

Katherine G Abraham and Sarah M Flood. American time use survey data extract builder (atus-x). *International Journal for Time Use Research*, 6:167–168, 2009.

Emma K Adam, Louise C Hawkley, Brigitte M Kudielka, and John T Cacioppo. Day-to-day

dynamics of experience–cortisol associations in a population-based sample of older adults. *Proceedings of the National Academy of Sciences*, 103(45):17058–17063, 2006.

Abi Adams-Prassl, Teodora Boneva, Marta Golin, and Christopher Rauh. Inequality in the impact of the coronavirus shock: Evidence from real time surveys. *CEPR Discussion Paper No. DP14665*, 2020a.

Abi Adams-Prassl, Teodora Boneva, Marta Golin, Christopher Rauh, et al. The impact of the coronavirus lockdown on mental health: Evidence from the US. *HCEO Working Paper 2020-030*, 2020b.

Hunt Allcott, Levi Boxell, Jacob Conway, Matthew Gentzkow, Michael Thaler, and David Y Yang. Polarization and public health: Partisan differences in social distancing during the coronavirus pandemic. *NBER Working Paper N.26946*, (w26946), 2020.

Jessica Allen, Reuben Balfour, Ruth Bell, and Michael Marmot. Social determinants of mental health. *International Review of Psychiatry*, 26(4):392–407, 2014.

American College Health Association. National College Health Assessment. Technical report, 2018. `https://www.acha.org/NCHA/ACHA-NCHA_Data/Publications_and_Reports/NCHA/Data/Reports_ACHA-NCHAIIc.aspx`.

James Banks and Xiaowei Xu. The mental health effects of the first two months of lockdown and social distancing during the covid-19 pandemic in the uk. *IFS Working Paper W20/16*, 20(16), 2020.

Zachary Beattie, Y Oyang, A Statan, A Ghoreyshi, A Pantelopoulos, A Russell, and CJPM Heneghan. Estimation of sleep stages in a healthy adult population from optical plethysmography and accelerometer signals. *Physiological Measurement*, 38(11):1968, 2017.

Louis-Philippe Beland, Abel Brodeur, Derek Mikola, and Taylor Wright. The short-term economic consequences of covid-19: Occupation tasks and mental health in canada. 2020.

Pietro Biroli, Steven Bosworth, Marina Della Giusta, Amalia Di Girolamo, Sylvia Jaworska, and Jeremy Vollen. Family life in lockdown. 2020.

Abel Brodeur, Andrew Clark, Sarah Fleche, and Nattavudh Powdthavee. Covid-19, lockdowns and well-being: Evidence from google trends. 2020a.

Abel Brodeur, David M Gray, Anik Islam, and Suraiya Bhuiyan. A literature review of the economics of covid-19. *IZA Discussion Paper N.13411*, 2020b.

Samantha K Brooks, Rebecca K Webster, Louise E Smith, Lisa Woodland, Simon Wessely, Neil Greenberg, and Gideon James Rubin. The psychological impact of quarantine and how to reduce it: rapid review of the evidence. *The Lancet*, 2020.

A Byrne and DG Byrne. The effect of exercise on depression, anxiety and other mood states: a review. *Journal of Psychosomatic Research*, 37(6):565–574, 1993.

Sammi R Chekroud, Ralitza Gueorguieva, Amanda B Zheutlin, Martin Paulus, Harlan M Krumholz, John H Krystal, and Adam M Chekroud. Association between physical exercise and mental health in 1· 2 million individuals in the usa between 2011 and 2015: a cross-sectional study. *The Lancet Psychiatry*, 5(9):739–746, 2018.

Tianqi Chen, Tong He, Michael Benesty, and Vadim Khotilovich. Package 'xgboost'. *R version*, 90, 2019.

A Chmitorz, A Kunzler, I Helmreich, O Tüscher, R Kalisch, T Kubiak, M Wessa, and K Lieb. Intervention studies to foster resilience–a systematic review and proposal for a resilience framework in future intervention studies. *Clinical Psychology Review*, 59:78–100, 2018.

Olivier Coibion, Yuriy Gorodnichenko, and Michael Weber. Labor markets during the covid-19 crisis: A preliminary view. *NBER Working Paper N. 27017*, 2020.

Tegan Cruwys, Genevieve A Dingle, Catherine Haslam, S Alexander Haslam, Jolanda Jetten, and Thomas A Morton. Social group memberships protect against future depression, alleviate depression symptoms and prevent depression relapse. *Social Science & Medicine*, 98:179–186, 2013.

Dmitry M Davydov, Robert Stewart, Karen Ritchie, and Isabelle Chaudieu. Resilience and mental health. *Clinical Psychology Review*, 30(5):479–495, 2010.

Jonathan De Quidt and Johannes Haushofer. Depression for economists. *NBER Working Paper N. 22973*, 2016.

G Devin, C Orme, C Costello, et al. Measuring depressive symptoms in illness populations: psychometric properties of the ces-d scale. *Psychol Health*, 2(139):56, 1988.

Thiemo Fetzer, Lukas Hensel, Johannes Hermle, and Christopher Roth. Coronavirus perceptions and economic anxiety. *Review of Economics and Statistics*, pages 1–36, 2020a.

Thiemo Fetzer, Marc Witte, Lukas Hensel, Jon Jachimowicz, Johannes Haushofer, Andriy Ivchenko, Stefano Caria, Elena Reutskaja, Christopher Roth, Stefano Fiorin, et al. Global behaviors and perceptions in the covid-19 pandemic. *PsyArXiv*, 2020b.

Fitbit Inc. The impact of coronavirus on global activity. Technical report, 2019. `https://blog.fitbit.com/covid-19-global-activity/`.

Osea Giuntella, Wei Han, and Fabrizio Mazzonna. Circadian rhythms, sleep, and cognitive skills: evidence from an unsleeping giant. *Demography*, 54(5):1715–1742, 2017.

Paul E Greenberg, Andree-Anne Fournier, Tammy Sisitsky, Crystal T Pike, and Ronald C Kessler. The economic burden of adults with major depressive disorder in the United States (2005 and 2010). *The Journal of Clinical Psychiatry*, 2015.

David Gunnell, Louis Appleby, Ella Arensman, Keith Hawton, Ann John, Nav Kapur, Murad Khan, Rory C O'Connor, Jane Pirkis, Eric D Caine, et al. Suicide risk and prevention during the covid-19 pandemic. *The Lancet Psychiatry*, 7(6):468–471, 2020.

Benjamin Handel and Jonathan Kolstad. Wearable technologies and health behaviors: new data and new methods to understand population health. *American Economic Review*, 107(5): 481–85, 2017.

Ben B. Hansen and Stephanie Olsen Klopfer. Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics*, 15(3):609–627, 2006.

Liesl M Heinrich and Eleonora Gullone. The clinical significance of loneliness: A literature review. *Clinical Psychology Review*, 26(6):695–718, 2006.

Yoshitaka Kaneita, Takashi Ohida, Makoto Uchiyama, Shinji Takemura, Kazuo Kawahara, Eise Yokoyama, Takeo Miyake, Satoru Harano, Kenshu Suzuki, and Toshiharu Fujita. The relationship between depression and sleep disturbances: a Japanese nationwide general population survey. *The Journal of Clinical Psychiatry*, 2006.

Kurt Kroenke, Robert L Spitzer, Janet BW Williams, Patrick O Monahan, and Bernd Löwe. Anxiety disorders in primary care: prevalence, impairment, comorbidity, and detection. *Annals of Internal Medicine*, 146(5):317–325, 2007.

Brigitte M Kudielka, Louise C Hawkley, Emma K Adam, and John T Cacioppo. Compliance with ambulatory saliva sampling in the chicago health, aging, and social relations study and associations with social support. *Annals of Behavioral Medicine*, 34(2):209–216, 2007.

Reed Larson and Mihaly Csikszentmihalyi. The experience sampling method. In *Flow and the foundations of positive psychology*, pages 21–34. Springer, 2014.

Hyun-Ah Lee, Heon-Jeong Lee, Joung-Ho Moon, Taek Lee, Min-Gwan Kim, Hoh In, Chul-Hyun Cho, and Leen Kim. Comparison of wearable activity tracker with actigraphy for sleep evaluation and circadian rest-activity rhythm measurement in healthy young adults. *Psychiatry Investigation*, 14(2):179, 2017.

Jun Z Li, Blynn G Bunney, Fan Meng, Megan H Hagenauer, David M Walsh, Marquis P Vawter, Simon J Evans, Prabhakara V Choudary, Preston Cartagena, Jack D Barchas, et al. Circadian patterns of gene expression in the human brain and disruption in major depressive disorder. *Proceedings of the National Academy of Sciences*, 110(24):9950–9955, 2013.

S.M. Lundberg, G. Erion, and H. Chen. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2:56–67. doi: 10.1038/s42256-019-0138-9. URL https://doi.org/10.1038/s42256-019-0138-9.

Yael Mayer, Shir Etgar, Noga Shiffman, and Ido Lurie. The fear of covid-19 familial infection scale: Initial psychometric examination. *PsyArXiv*, 2020.

Katherine L Milkman, Julia A Minson, and Kevin GM Volpp. Holding the hunger games hostage at the gym: An evaluation of temptation bundling. *Management Science*, 60(2): 283–299, 2014.

Laura Montenovo, Xuan Jiang, Felipe Lozano Rojas, Ian M Schmutte, Kosali I Simon, Bruce A Weinberg, and Coady Wing. Determinants of disparities in covid-19 job losses. *NBER Working Paper N.27132*, 2020.

National Center for Health Statistics. Early release of selected mental health estimates. Technical report, 2019. `https://www.cdc.gov/nchs/data/nhis/earlyrelease/ERmentalhealth-508.pdf`.

Office for National Statistics. United Coronavirus and anxiety, Great Britain: 3 April 2020 to 10 May 2020, year=2020, note=`https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/articles/coronavirusandanxietygreatbritain/3april2020to10may2020`. Technical report.

JL Ong, TY Lau, SAA Massar, ZT Chong, BKL Ng, D Koek, W Zhao, BTT Yeo, K Cheong, and MWL Chee. Covid-19 related mobility reduction: Heterogenous effects on sleep and physical activity rhythms. *arXiv preprint arXiv:2006.02100*, 2020.

Vikram Patel and Anna Goodman. Researching protective and promotive factors in mental health. *Oxford University Press*, 2007.

Faye Plummer, Laura Manea, Dominic Trepel, and Dean McMillan. Screening for anxiety disorders with the gad-7 and gad-2: a systematic review and diagnostic metaanalysis. *General Hospital Psychiatry*, 39:24–31, 2016.

Eugenio Proto, Climent Quintana-Domeque, et al. Covid-19 and mental health deterioration among bame groups in the uk. Technical report, 2020.

Lenore Sawyer Radloff. The ces-d scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, 1(3):385–401, 1977.

Matthew W Ridley, Gautam Rao, Frank Schilbach, and Vikram H Patel. Poverty, depression, and anxiety: Causal evidence and mechanisms. *NBER Working Paper N.27157*, 2020.

Peter Salmon. Effects of physical exercise on anxiety, depression, and sensitivity to stress: a unifying theory. *Clinical Psychology Review*, 21(1):33–61, 2001.

Ingvild W Saxvig, Ståle Pallesen, Ane Wilhelmsen-Langeland, Helge Molde, and Bjørn Bjorvatn. Prevalence and correlates of delayed sleep phase in high school students. *Sleep Medicine*, 13(2):193–199, 2012.

Kate M Scott, Carmen Lim, Ali Al-Hamzawi, Jordi Alonso, Ronny Bruffaerts, José Miguel Caldas-de Almeida, Silvia Florescu, Giovanni De Girolamo, Chiyi Hu, Peter De Jonge, et al. Association of mental disorders with subsequent chronic physical conditions: world mental health surveys from 17 countries. *JAMA Psychiatry*, 73(2):150–158, 2016.

Bruce W Smith, Jeanne Dalen, Kathryn Wiggins, Erin Tooley, Paulette Christopher, and Jennifer Bernard. The brief resilience scale: assessing the ability to bounce back. *International Journal of Behavioral Medicine*, 15(3):194–200, 2008.

Bruce W Smith, Erin M Tooley, PJ Christopher, and VS Kay. Resilience as the ability to bounce back from stress: A neglected personal resource? *The Journal of Positive Psychology*, 5(3): 166–176, 2010.

Robert L Spitzer, Kurt Kroenke, Janet BW Williams, and Bernd Löwe. A brief measure for assessing generalized anxiety disorder: the gad-7. *Archives of Internal Medicine*, 166(10): 1092–1097, 2006.

Andreas Ströhle. Physical activity, exercise, depression and anxiety disorders. *Journal of Neural Transmission*, 116(6):777, 2009.

Atin Supartini, Takanori Honda, Nadzirah A Basri, Yuka Haeuchi, Sanmei Chen, Atsushi Ichimiya, and Shuzo Kumagai. The impact of sleep timing, sleep duration, and sleep quality on depressive symptoms and suicidal ideation amongst Japanese freshmen: the equsite study. *Sleep Disorders*, 2016, 2016.

Norifumi Tsuno, Alain Besset, and Karen Ritchie. Sleep and depression. *The Journal of Clinical Psychiatry*, 2005.

Jean M Twenge, A Bell Cooper, Thomas E Joiner, Mary E Duffy, and Sarah G Binau. Age, period, and cohort trends in mood disorder indicators and suicide-related outcomes in a nationally representative dataset, 2005–2017. *Journal of Abnormal Psychology*, 128(3):185, 2019.

United Nations. United Nations Report. Technical report, 2020. `https://unsdg.un.org/sites/default/files/2020-05/UN-Policy-Brief-COVID-19-and-mental-health.pdf`.

US Census. Household Pulse Survey. Technical report, 2020. `https://www.census.gov/householdpulsedata`.

William H Walker, James C Walton, A Courtney DeVries, and Randy J Nelson. Circadian rhythm disruption and mental health. *Translational Psychiatry*, 10(1):1–13, 2020.

Marc Wittmann, Jenny Dinich, Martha Merrow, and Till Roenneberg. Social jetlag: misalignment of biological and social time. *Chronobiology International*, 23(1-2):497–509, 2006.

Wendy Wood, Leona Tam, and Melissa Guerrero Witt. Changing circumstances, disrupting habits. *Journal of Personality and Social Psychology*, 88(6):918, 2005.

Long Zhai, Hua Zhang, and Dongfeng Zhang. Sleep duration and depression among adults: A meta-analysis of prospective studies. *Depression and Anxiety*, 32(9):664–670, 2015.

Online Appendix

1.1 Methods

1.1.1 Data, Setting and Recruitment

In January 2019, we began to run a randomized experiment on wellness and, as part of the study, we collected a rich longitudinal data set of measures of students' physical health (physical activity, sleep, heart rate), time use, and mental health. The study was approved by the University of Pittsburgh Institutional Review Board and was pre-registered in the AEA RCT Registry (RCT ID #AEARCTR-0003235). We have collected data from three cohorts of students in Spring 2019, Fall 2019, and Spring 2020. At the beginning of each semester, we invited college students at the University of Pittsburgh to enroll in a semester long experiment on wellness. We recruited participants from the Pittsburgh Experimental Economics Lab using the SONA online management system. To be eligible for the study, participants had to have a smartphone and be willing to wear and routinely synchronize a wearable device (Fitbit) throughout the semester. Students interested in participating in the study received $6 for taking part in an initial 30-minute session in the laboratory and received an additional payment at study completion. Participants were guaranteed a payment of $30 for the study, and the opportunity to receive additional earnings based on luck and their decisions in the study.

During the initial session of the study, each participant received a Fitbit Alta HR device, registered for a Fitbit account, and installed a custom-made smartphone app on their smartphone, which allowed us to track their Fitbit data and to interact with the subjects throughout the study. During the laboratory session, we instructed participants to wear their device as much as possible, synchronize it daily, answer weekly surveys, and return the device at the end of the semester. All participants filled out an enrollment survey where we collected baseline information on demographics, self-reported health, screen time use, GPA, and mental health.

After collecting about 1 to 4 weeks of baseline data, we randomized participants into a control group and different treatments aimed at improving their sleep habits via reminders to go to sleep, feedback on sleep behavior and monetary incentives for sleeping 7 hours by 9am every weekday. In the Feedback treatment, participants received feedback every morning, informing them whether they met the goal of sleeping 7 hours per night by 9am. In three additional

treatments, individuals received \$4.75 for every weeknight in which they slept at least 7 hours per night by 9am. In the CashNow treatment, participants received feedback and their rewards every morning. In the CashDelayed treatment, participants received feedback every morning but only received their payment at the end of the semester. In the CashDelayedNoFeedback treatment, participants did not receive any feedback and received their payments at the end of the semester. The interventions lasted 4 weeks. We continued following participants for an additional 2 to 4 weeks until the end of the semester.

Throughout the semester, students filled out weekly surveys about time use, and received weekly text messages aimed at measuring mood and resilience (alternated every week) using experience sampling techniques. At the end of the study, they filled out an endline survey which contained most of the questions we collected at baseline, including the mental health questions.

For the Spring 2020 cohort, we also administered a midline survey on March 15th, right after the university had announced that classes would be moved online, and students were encouraged not to come back to campus. The survey assessed the frequency of social interactions (face to face and via text-messages, calls and video calls), changes in day-to-day behaviors as a result of the pandemic, changes in employment status and living conditions, participants' beliefs over the pandemic, and mental health.

At the end of the study, participants received any outstanding payments they had earned through surveys or other decisions they made in the study. If their cumulative payments over the course of the study were less than \$30, we added to their payment the amount required to make their total payment equal \$30, as that was the guaranteed minimum cumulative earnings indicated in recruitment materials.

Table S.1 reports the main demographic characteristics of our sample in each wave. Our sample over-represents women, Asians, and freshmen. In our analysis we show the sensitivity of our findings to the inclusion of demographic controls (Table S.2) or individual fixed effects (Table S.6).

36

Table S.1: Summary Statistics

	Overall sample			Wave 1			Wave 2			Wave 3		
	Obs	Mean	Std.Dev.	Obs	Mean	Std.Dev.	Obs	Mean	Std.Dev.	Obs	Mean	Std.Dev.
Demographics												
Age	669	19.74	2.97	151	20.56	3.77	314	19.70	3.16	204	19.19	1.54
Female	670	0.68	0.47	151	0.78	0.41	315	0.58	0.49	204	0.76	0.43
Race												
Black	669	0.10	0.30	151	0.11	0.32	314	0.05	0.23	204	0.15	0.36
Hispanic	669	0.04	0.19	151	0.01	0.11	314	0.05	0.23	204	0.03	0.18
Asian	669	0.27	0.44	151	0.21	0.41	314	0.29	0.46	204	0.27	0.44
White	669	0.56	0.50	151	0.62	0.49	314	0.56	0.50	204	0.50	0.50
American Indian	669	0.00	0.04	151	0.00	0.00	314	0.00	0.06	204	0.00	0.00
Arab	669	0.02	0.13	151	0.05	0.21	314	0.01	0.11	204	0.00	0.07
Health												
% reporting poor health	670	0.08	0.27	151	0.07	0.25	315	0.08	0.28	204	0.08	0.28
College year												
Freshmen	669	0.44	0.50	151	0.38	0.49	314	0.40	0.49	204	0.56	0.50
Sophomore	669	0.23	0.42	151	0.18	0.38	314	0.27	0.45	204	0.22	0.41
Junior	669	0.15	0.35	151	0.19	0.40	314	0.14	0.35	204	0.11	0.32
Senior	669	0.18	0.38	151	0.25	0.44	314	0.19	0.39	204	0.11	0.32
Major												
Humanities	670	0.08	0.27	151	0.11	0.32	315	0.07	0.25	204	0.08	0.28
Business	670	0.13	0.34	151	0.09	0.29	315	0.19	0.39	204	0.20	0.40
Social science	670	0.16	0.37	151	0.17	0.37	315	0.14	0.35	204	0.00	0.00
No major	670	0.51	0.50	151	0.48	0.50	315	0.50	0.50	204	0.52	0.50
Medical	670	0.11	0.32	151	0.15	0.35	315	0.09	0.29	204	0.12	0.33
Financial aid	604	0.45	0.50	133	0.46	0.50	284	0.44	0.50	187	0.46	0.50

Notes - The table above reports the demographic characteristics of our sample in the three waves of our study.

1.1.2 Main Variables

1.1.3 Mental Health

We measured mental health in the enrollment survey, in the endline survey and, for the Spring 2020 cohort, in the midline survey, by administering the following scales.

Depression. To measure depression, we administer the Center for Epidemiologic Studies Depression Scale-D scale (CES-D, [Devin et al., 1988]). The CES-D scale is a widely used validated self-report instrument designed to assess depressive symptomatology in the general population. The scale is comprised of 20 items that assess the frequency of symptoms associated with depression, such as feelings of helplessness, worthlessness, loss of appetite, or loneliness, using a scale from 0 (Rarely or None of the Time) to 3 (Most or all of the times), see Figure S.7 for the full scale. An overall depression score is obtained by summing answers to all 20 question, with higher scores indicating greater depressive symptoms. Individuals with scores at or above the threshold of 16 points are classified to be at risk of clinical depression.

Anxiety. We measure anxiety using the GAD-7 scale ([Spitzer et al., 2006]), a 7 items scale designed to assess the presence and severity of generalized anxiety disorder. The instrument assesses the frequency of anxiety related symptoms over the past week using a Likert scale that ranges from 0 (Not at all) to 7 (Nearly every day), with total scores ranging form 0 to 21. Higher scores in this scale reflect greater anxiety severity. Individuals with scores above 10 are considered to be at risk of generalized anxiety disorder [Spitzer et al., 2006], though recent studies have identified 8 as a reasonable cut-point for identifying anxiety disorder [Kroenke et al., 2007, Plummer et al., 2016].

Resilience. We measure resilience, defined as "the ability to bounce back from stress" [Smith et al., 2010] using 5 of the 6 items of the Brief Resilience Scale developed and validated by [Smith et al., 2008]. Each item in the scale is rated on a 5-point scale, ranging from 1 (strongly disagree) to 5 (strongly agree). Items are averaged, and higher average scores are indicative of higher resilience.

Life Satisfaction We measure life satisfaction using a 4-point scale.

1.1.4 Physical Health

Personal wearable activity trackers, such as Fitbits, have been used in past work to study health behavior [e.g., Handel and Kolstad, 2017]. Through the wearable trackers we collected data on physical activity and sleep. In particular, we collected data on daily steps and heart rate. We also collected data on sleep duration, wake up time, bedtime, and sleep disruptions. Because the devices measure heart rate, they are able to distinguish between time spent in bed not sleeping, such as watching TV, and time spent sleeping. Previous work [e.g., Lee et al., 2017] finds that wearable activity trackers that detect heart rate perform fairly well in terms of tracking sleep compared to actigraphy, the more sophisticated method used in medical studies [Beattie et al., 2017].

Time-Diaries and Experience Sampling

Time use. Throughout the study, participants filled out time-use diaries every 2 weeks. They indicated how they spent the previous 24 hours by choosing from different categories, following the structure of the American Time Use Survey (https://www.bls.gov/tus). We categorize time use into the following categories: social interactions, screen, work, study, eating or preparing food, personal care, sleep, exercise, errands and commuting (see Online Appendix for the full breakdown). In our analyses, we do not use the data on sleep, exercise and commuting by walking, which we collect from the Fitbit. Social interactions measure time spent hanging out with friends. Screen time includes time spent watching TV, playing video games or surfing the internet, and excludes study or work time on the screen.

Experience Sampling. Every week, we elicited their mood or resilience via text message, alternating between the two questions every week. For mood, participant indicated, on a 10-point Likert scale, how happy they felt in that moment. For resilience, participants were asked to indicate, using a 5-point Likert scale, a) the extent to which they were facing stress in their life and b) how able they felt to deal with the stress they were facing. After March 15th, we also started to elicit loneliness by asking participants to rate, using a 5-point scale, how lonely they felt that day. We varied the time at which we texted participants and randomly assigned subjects to receive a text message with one of the questions at either 11am, 4pm, or 9pm on Fridays.

2 SUPPLEMENTARY RESULTS

2.1 Lifestyle and Mental Health Disruptions

2.2 Data Definition

Fitbit. Participants were given devices (Fitbit^{TM} Alta HR, Fitbit^{TM} Inc, San Francisco, CA). Data on sleep and physical activity were extracted from the Fitbit API. The data on activity included daily total steps, active hours (which is the sum of lightly active, fairly active and very active minutes) and resting heart rate levels. Sleep data included bedtime, wake up time, total sleep time, time in bed, and time in bed awake. Eighty three percent of participants synced their devices for 80% of the possible dates or more in Spring 2019, whereas 72% did so in Spring 2020. Ninety two percent of participants in the Fall 2019 Cohort synced their device for 80% of the possible dates. For robustness, we also conduct analyses that exclude 5% of participants who synced the least, see Figure S.2. The remaining 95% of the sample synced, on average, 75.8% of the maximum possible number of days they could have synced within each respective study term. The excluded 5% of participants synced only 8.6% of the days on average.

For each individual, we compute the average for all the variables considered (total steps, active hours, sleep duration, bedtime and wake up time). For the analysis of changes between beginning and end of the semester discussed in Figure 7, Tables S.2-S.6, as well as in the prediction exercise, we calculated the average for all the variables considered for all individuals who had synced at least twice (once during baseline and once at the onset of the pandemic). The baseline period spans study enrollment through February 16th in 2020 and February 24th in 2019. We compare variables at baseline to the period after the onset of the pandemic (from March 23rd to April 20th in both Spring 2019 and Spring 2020). For Fall 2019, the baseline period spans from study enrollment to September 30th. In the analyses that include Fall 2019, we compare baseline data to data collected from November 2nd through the end of the fall semester. Baseline periods span the period before we began the sleep intervention in each semester. For Fall 2019, we picked November 2nd as the start of the endline period to match the timing of the remote learning period in Spring 2020 (i.e., a comparable number of days since the beginning of each respective semester). Choosing alternative cutoffs yields similar results. In section 2.5, as a robustness check, we replicate all analyses using a different cutoff,

comparing data at the beginning of the semester (before we began the intervention in each semester) to all data collected after the beginning of the intervention.

Time Use. As explained in the main text, we collected data on time use using weekly surveys, following the structure of the American Time Use survey [Abraham and Flood, 2009]. We characterize time use into the following categories: social interactions (i.e., hanging out with friends), screen (which includes watching TV, surfing the internet and playing games), work, study, eating or preparing food, personal care, house-care and errands (which includes cleaning, laundry, paying bills, grocery shopping), sleep, exercise, and commuting. Average response rate to the weekly survey in 2020 was 75.9% and ranged between 60.6% (week 10 of the study) and 88.6% (week 1 of the study). Average response rate was 72% in Spring 2019 (ranging from 82.4% (week 2) to 54.1% in week 7) and was 67% in Fall 2019. For the analysis of changes between beginning and end of the semester we did the following. In Figure 2 and Tables S.2 and S.18, we show average time use at the beginning of the semester and average time use later in the semester using all responses. In Tables S.2-S.6, as well as in the prediction exercise, we include all participants who filled out the survey at least twice, at least once during our baseline data collection and at least once during the pandemic (After March 23). For each individual we consider the average in each period for all of the variables considered (e.g., social interaction, screen time).

Mental Health. All mental health questions were collected in the Baseline (upon enrollment) and endline survey. A total of 627 participants filled out the baseline survey. Of these 91.9% filled out the endline survey (100% in Spring 2019, 86.6% in Fall 2019 and 83% in Spring 2020). In Figure 3 we report average responses to the mental health assessments (CES-D and GAD-7) at the beginning and end of the semester. For Spring 2020, we also have a midline survey, which was filled out by 83.1% of participants. In our prediction exercise we focus on the 91.9% of participants for whom we have both baseline and endline measures of mental health.

2.3   Descriptive Figures including Fall 2019

Figure S.1 is analogous to Figure 1 from the main text. It plots steps over time pooling the data from the 2019 Spring (N=140) and 2019 Fall cohort (317) and compares them to Spring

2020 (N=331). Panel A shows steps, panel B shows physical activity measured by the number of active hours every day. Panel C displays sleep duration (in hours) and Panel D shows Wake up Time. Note that people in the 2020 cohort wake up 30 minutes later than those in the 2019 cohort, though this difference is not significant. Wake up time increases by over an hour at the onset of the pandemic.

While Figure S.1 includes any participant for whom we have at least one day of syncing, Figure S.2 reproduces the same figure but excludes participants who synced the least (bottom 5%). The figure uses 643 individuals across all terms.

Similarly, Figures S.3 and S.4 replicate the results presented in the main text in Figures 3 and 4 including data from the Fall term of 2019. These figures include 414 observations for 2019 and 204 observations for the 2020 term.

Figure S.1: Physical Activity and Sleep
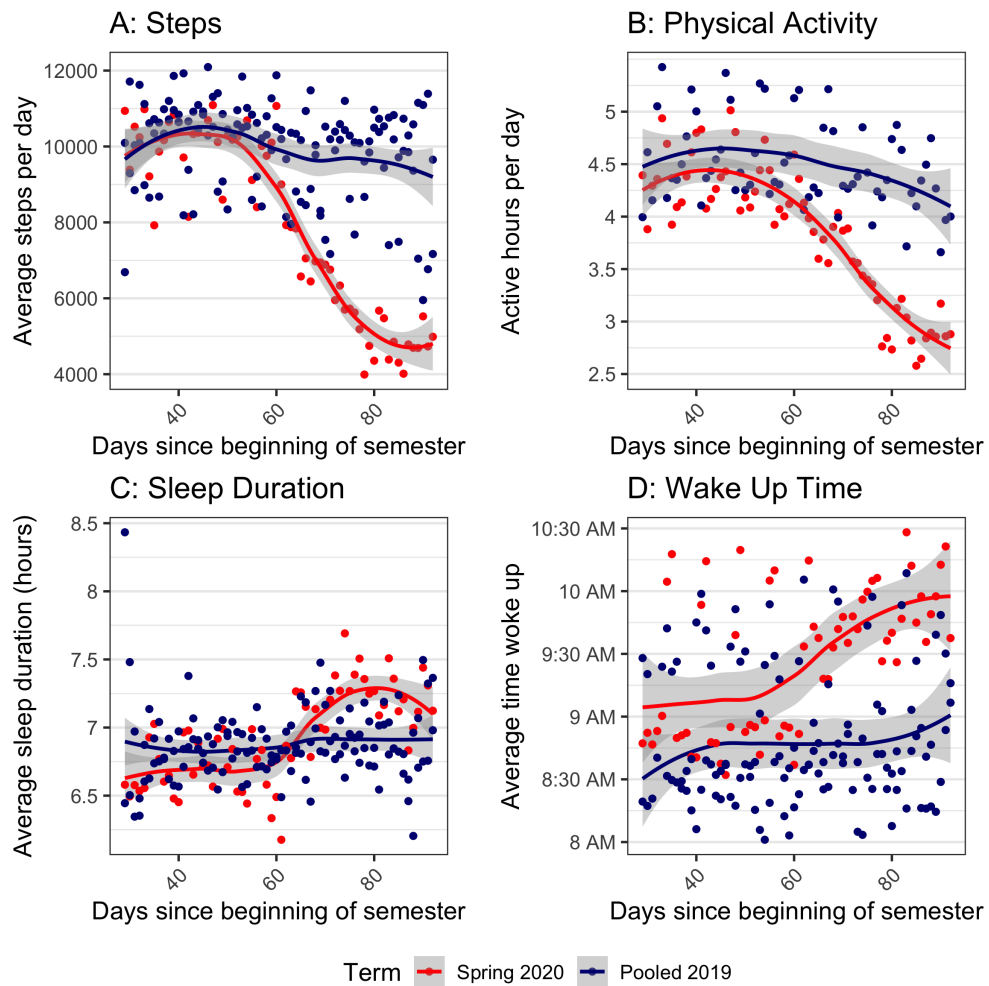
Notes - The above figure plots the average outcomes by day for study participants in the Spring and Fall of 2019 (red) and Spring of 2020 (blue). The lines report a kernel-weighted local polynomial regression.

43

Figure S.2: Physical Activity and Sleep, Alternative Sample

A: Steps

B: Physical Activity

C: Sleep Duration

D: Wake Up Time

Term —●— Spring 2020 —●— Spring 2019
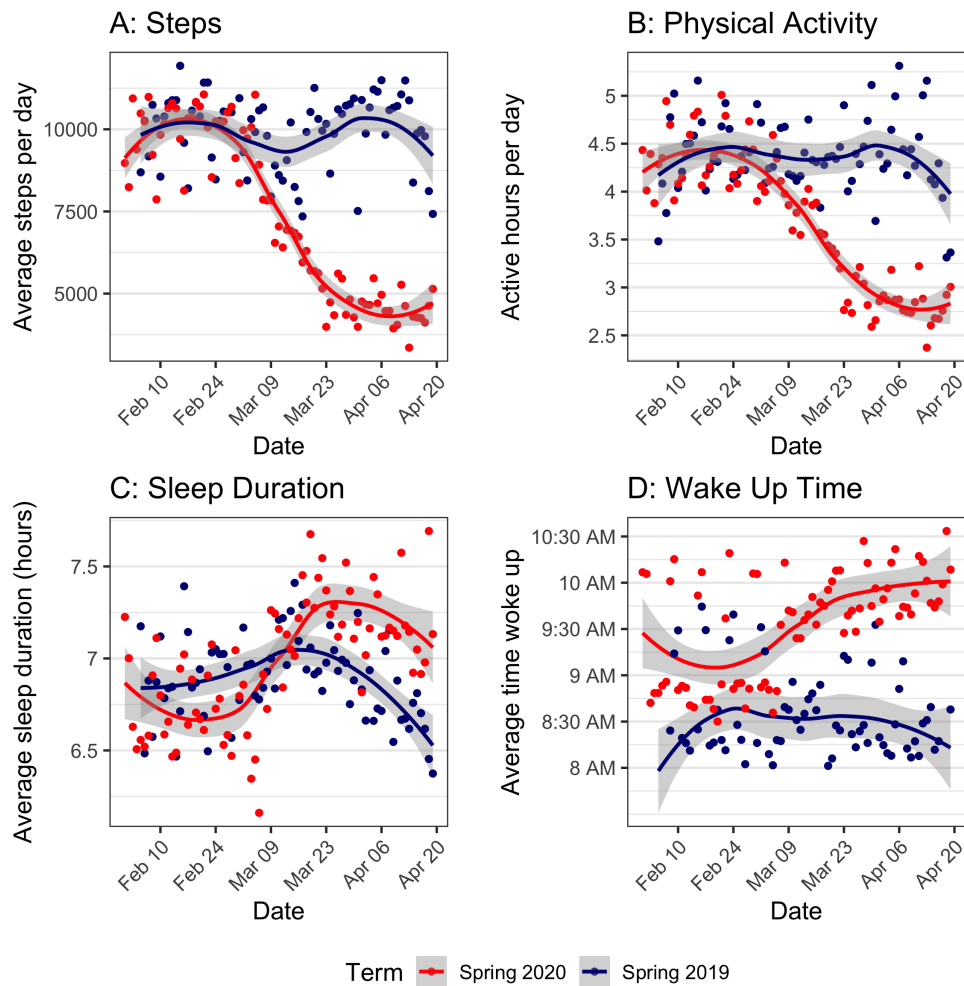
Notes - The sample excludes individuals who synced least (bottom 5%). The above figure plots the average outcomes by day for study participants in the Spring of 2019 (red) and Spring of 2020 (blue). The lines report a kernel-weighted local polynomial regression.

Figure S.3: Depression and Anxiety, Alternative Sample
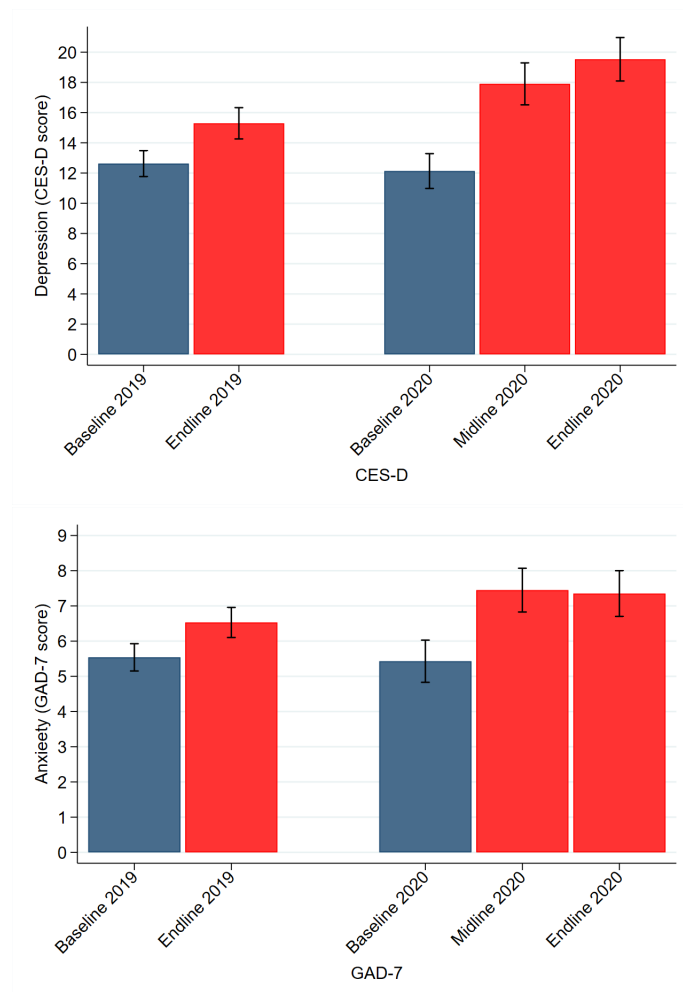
Notes - The figures show the average CES-D score and GAD-7 score during Spring and Fall 2019 and Spring 2020 terms. Bars indicate confidence intervals.

2.3.1 Additional Figures

Figure S.5 shows bedtime at the beginning and end of the semester for the Spring 2019 and 2020 cohort (Panel A) and for the pooled 2019 cohorts and 2020 cohort (Panel B). Bedtime differed among the two cohorts and this difference did increase, but less dramatically than the difference observed for wake-up time. Since the relative increase in wake up time is larger than the relative increase in bedtime, we also see that mid-sleep (the time at which participants complete half their sleep) becomes later in the 2020 cohort at the onset of the pandemic. The

Figure S.4: Depression: Distribution of CES-D Score Across Cohorts, Alternative Sample

increase in the mid-sleep measure captures misalignment of sleep timing with respect to the natural dark-light cycle.

Figure S.5: Bedtime
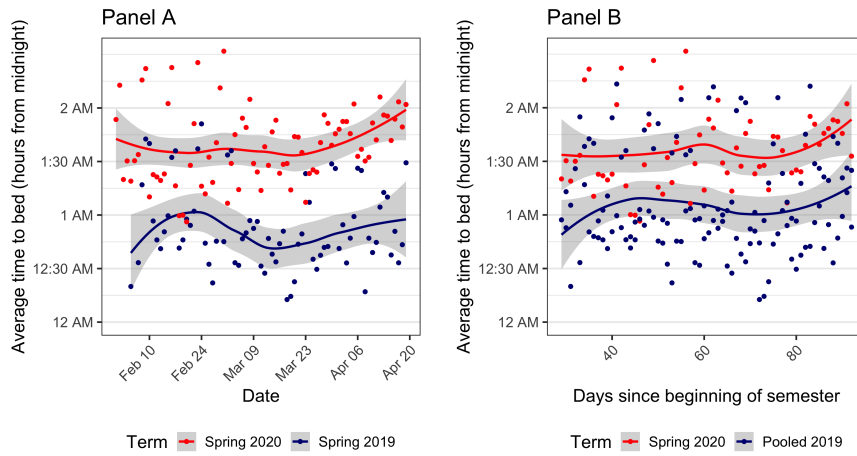
Notes - Panel A of the above figure plots the average bedtime by day for study participants in the Spring 2019 (red) and Spring of 2020 (blue). The lines report a kernel-weighted local polynomial regression. Panel B reproduces Panel A using the pooled Spring 2019 and Fall 2019 cohort as a comparison group.

Mid-Semester Data

Starting on March 20th and until the end of the semester, we added questions aboout the nature of social interactions to the time use surveys. Figure S.6 shows the average number of unique interactions that occurred face to face, via calls or text message. On the April 14th survey we separated calls into phone calls and video calls.

47

Figure S.6: Social Interactions

CES-D scale questions
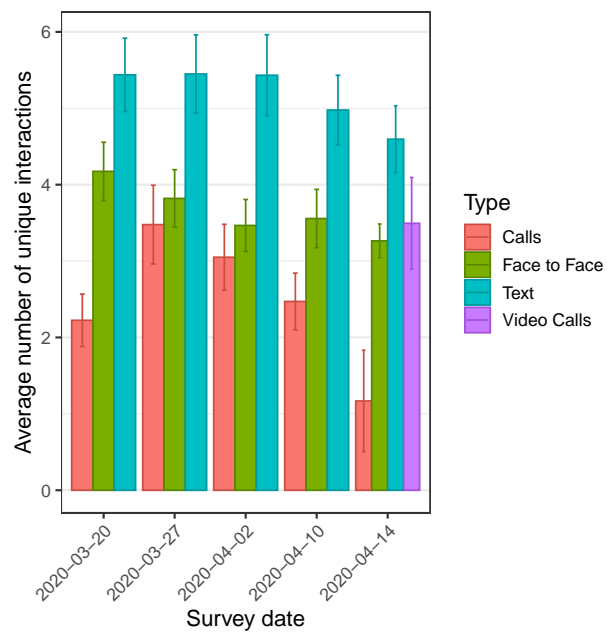
Notes - The figure above reports item-by item CES-D differences at baseline, midline and endline inspring 2020.

z

Table S.2: Difference-in-difference Analysis: Spring 2019 & Spring 2020

	(1)	(2)	(3)	(4)
Panel A: Mental Health	CES-D	CES-D>15	GAD-7	GAD-7>7
end of term * 2020	4.940***	0.234***	0.639	0.055
	(1.041)	(0.058)	(0.478)	(0.058)
end of term	1.932***	0.064	1.072***	0.113***
	(0.698)	(0.043)	(0.343)	(0.042)
Observations	779	779	779	779
R-squared	0.152	0.109	0.074	0.060
Number of individuals	460	460	460	460
Mean of Dep. Var.	14.59	0.408	6.241	0.331
std. dev.	9.640	0.492	4.449	0.471
Panel B: Activity and Sleep	Total Steps	Active Hours	Sleep Hrs	Wake Up Time
end of term * 2020	-4,037.415***	-0.916***	0.467***	0.522***
	(248.585)	(0.091)	(0.094)	(0.124)
end of term	-74.740	-0.039	-0.036	-0.103
	(154.025)	(0.050)	(0.073)	(0.083)
Observations	928	913	875	875
R-squared	0.401	0.234	0.116	0.099
Number of individuals	491	486	474	474
Mean of Dep. Var.	8079	4.125	7.115	7.803
std. dev.	3544	1.132	0.825	1.252
Panel C: Time Use	Social Interactions	Screen Hours	Work Hours	Study Hours
end of term * 2020	-0.959***	3.333***	-0.265	-1.545***
	(0.153)	(0.219)	(0.240)	(0.313)
end of term	0.096	-0.236	0.185	-0.683***
	(0.127)	(0.157)	(0.198)	(0.252)
Observations	719	719	719	719
R-squared	0.144	0.429	0.196	0.239
Number of individuals	460	460	460	460
Mean of Dep. Var.	1.100	3.212	1.854	4.408
std. dev.	1.188	2.488	2.113	2.749

Notes - All estimates include dummies for year and wave, and controls for gender, age, college year, parental highest education, financial aid, race, ethnicity, a dummy for whether the individual stayed in the Pittsburgh area after the beginning of remote learning, and dummies controlling for treatment assignment.Standard errors are clustered at the individual level.

*** p<0.01, ** p<0.05, * p<0.1

Table S.3: Difference in difference Analysis: Spring 2019 & Spring 2020, Restricted Sample

	(1)	(2)	(3)	(4)
Panel A: Mental Health	CES-D	CES-D>15	GAD-7	GAD-7>7
end of term * 2020	5.755***	0.233***	1.072**	0.090
	(0.000)	(0.000)	(0.031)	(0.142)
end of term	1.182	0.060	0.709*	0.077*
	(0.109)	(0.187)	(0.054)	(0.083)
Observations	660	660	660	660
R-squared	0.157	0.112	0.077	0.068
Individuals	377	377	377	377
Mean of Dep. Var.	14.40	0.409	6.139	0.326
std. dev.	9.275	0.492	4.305	0.469
Panel B: Activity and Sleep	Total Steps	Active Hours	Sleep Hrs	Wake Up Time
end of term * 2020	-4,609.210***	-1.148***	0.402***	0.589***
	(0.000)	(0.000)	(0.000)	(0.000)
end of term	-164.511	-0.009	0.033	-0.057
	(0.304)	(0.878)	(0.597)	(0.399)
Observations	660	660	660	660
R-squared	0.418	0.273	0.155	0.138
Individuals	377	377	377	377
Mean of Dep. Var.	8623	4.253	7.150	7.873
std. dev.	3366	1.055	0.749	1.103
Panel C: Time Use	Social Interactions	Screen Hours	Work Hours	Study Hours
end of term * 2020	-1.085***	3.314***	-0.266	-1.417***
	(0.000)	(0.000)	(0.273)	(0.000)
end of term	0.147	-0.242	0.259	-0.690***
	(0.241)	(0.125)	(0.193)	(0.006)
Observations	660	660	660	660
R-squared	0.145	0.441	0.209	0.249
Individuals	377	377	377	377
Mean of Dep. Var.	1.141	3.124	1.890	4.520
std. dev.	1.196	2.427	2.122	2.683

Notes - All estimates include dummies for year and wave, and controls for gender, age, college year, parental highest education, financial aid, race, ethnicity, a dummy for whether the individual stayed in the Pittsburgh area after the beginning of remote learning, and dummies controlling for treatment assignment.Standard errors are clustered at the individual level.

*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Table S.4: Difference in difference Analysis: Spring 2019/ Fall 2019 & Spring 2020

	(1)	(2)	(3)	(4)
Panel A: Mental Health	CES-D	CES-D>15	GAD-7	GAD-7>7
end of term * 2020	4.093***	0.202***	0.622*	0.078*
	(0.858)	(0.044)	(0.366)	(0.044)
end of term	2.898***	0.101***	1.113***	0.087***
	(0.419)	(0.023)	(0.191)	(0.023)
Observations	1,379	1,379	1,379	1,379
R-squared	0.096	0.069	0.048	0.033
Number of individuals	672	672	672	672
Mean of Dep. Var.	14.41	0.386	6.175	0.331
std. dev.	10.02	0.487	4.468	0.471
Panel B: Activity and Sleep	Total Steps	Active Hours	Sleep Hrs	Wake Up Time
end of term * 2020	-3,241.600***	-0.840***	0.283***	0.511***
	(221.588)	(0.086)	(0.078)	(0.125)
end of term	-931.017***	-0.129***	0.145***	-0.067
	(118.867)	(0.044)	(0.055)	(0.086)
Observations	1,550	1,530	1,424	1,424
R-squared	0.298	0.167	0.103	0.163
Number of individuals	685	679	666	666
Mean of Dep. Var.	8516	4.249	7.050	7.405
std. dev.	3438	1.152	0.882	1.461
Panel C: Time Use	Social Interactions	Screen Hours	Work Hours	Study Hours
end of term * 2020	-0.749***	3.073***	-0.169	-2.050***
	(0.111)	(0.172)	(0.156)	(0.223)
end of term	-0.125*	0.031	0.087	-0.201
	(0.070)	(0.087)	(0.091)	(0.140)
Observations	1,284	1,284	1,284	1,284
R-squared	0.120	0.371	0.154	0.214
Number of individuals	633	633	633	633
Mean of Dep. Var.	1.227	2.712	1.758	4.993
std. dev.	1.320	2.235	2.062	2.840

Notes - All estimates include dummies for year and wave, and controls for gender, age, college year, parental highest education, financial aid, race, ethnicity, a dummy for whether the individual stayed in the Pittsburgh area after the beginning of remote learning, and dummies controlling for treatment assignment.Standard errors are clustered at the individual level.

*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Table S.5: Fixed effects analysis: Spring 2019 & Spring 2020

	(1)	(2)	(3)	(4)
Panel A: Mental Health	CES-D	CES-D>15	GAD-7	GAD-7>7
end of term * 2020	5.682***	0.238***	1.079**	0.074
	(0.969)	(0.059)	(0.440)	(0.057)
end of term	1.895***	0.075*	1.052***	0.113***
	(0.639)	(0.042)	(0.319)	(0.042)
Observations	779	779	779	779
R-squared	0.303	0.177	0.167	0.090
Number of individuals	460	460	460	460
Mean of Dep. Var.	14.59	0.408	6.241	0.331
std. dev.	9.640	0.492	4.449	0.471
Panel B: Activity and Sleep	Total Steps	Active Hours	Sleep Hrs	Wake Up Time
end of term * 2020	-4,580.660***	-1.112***	0.385***	0.545***
	(199.454)	(0.075)	(0.089)	(0.119)
end of term	-139.423	-0.039	-0.037	-0.098
	(136.178)	(0.046)	(0.073)	(0.083)
Observations	928	913	875	875
R-squared	0.750	0.537	0.105	0.080
Number of individuals	491	486	474	474
Mean of Dep. Var.	8079	4.125	7.115	7.803
std. dev.	3544	1.132	0.825	1.252
Panel C: Time Use	Social Interactions	Screen Hours	Work Hours	Study Hours
end of term * 2020	-1.008***	3.324***	-0.341	-1.411***
	(0.154)	(0.205)	(0.232)	(0.291)
end of term	0.136	-0.348**	0.254	-0.760***
	(0.129)	(0.147)	(0.192)	(0.244)
Observations	719	719	719	719
R-squared	0.232	0.614	0.007	0.359
Number of individuals	391	391	391	391
Mean of Dep. Var.	1.100	3.212	1.854	4.408
std. dev.	1.188	2.488	2.113	2.749

Notes - All estimates include a year dummy and individual fixed effects. The sample is restricted to observations for which all the outcomes are available. Standard errors are clustered at the individual level.

*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Table S.6: Fixed-Effects Analysis: Spring/Fall 2019 & Spring 2020

	(1)	(2)	(3)	(4)
Panel A: Mental Health	CES-D	CES-D>15	GAD-7	GAD-7>7
end of term * 2020	4.415***	0.199***	0.959***	0.092**
	(0.773)	(0.044)	(0.333)	(0.043)
end of term	3.090***	0.113***	1.151***	0.091***
	(0.393)	(0.022)	(0.183)	(0.023)
Observations	1,379	1,379	1,379	1,379
R-squared	0.240	0.136	0.129	0.063
Number of individuals	672	672	672	672
Mean of Dep. Var.	14.41	0.386	6.175	0.331
std. dev.	10.02	0.487	4.468	0.471
Panel B: Activity and Sleep	Total Steps	Active Hours	Sleep Hrs	Wake Up Time
end of term * 2020	-3,572.531***	-0.940***	0.250***	0.539***
	(183.750)	(0.073)	(0.074)	(0.122)
end of term	-1,050.772***	-0.168***	0.094*	-0.117
	(109.877)	(0.040)	(0.052)	(0.085)
Observations	1,550	1,530	1,424	1,424
R-squared	0.607	0.379	0.058	0.068
Number of individuals	685	679	666	666
Mean of Dep. Var.	8516	4.249	7.050	7.405
std. dev.	3438	1.152	0.882	1.461
Panel C: Time Use	Social Interactions	Screen Hours	Work Hours	Study Hours
end of term * 2020	-0.749***	3.073***	-0.169	-2.050***
	(0.111)	(0.172)	(0.156)	(0.223)
end of term	-0.125*	0.031	0.087	-0.201
	(0.070)	(0.087)	(0.091)	(0.140)
Observations	1,284	1,284	1,284	1,284
R-squared	0.135	0.497	0.006	0.237
Number of individuals	633	633	633	633
Mean of Dep. Var.	1.227	2.712	1.758	4.993
std. dev.	1.320	2.235	2.062	2.840

Notes - All estimates include a year dummy and individual fixed effects. Standard errors are clustered at the individual level.
*** p<0.01, ** p<0.05, * p<0.1

Table S.7: Habit Disruptions and Depression during the Pandemic

	(1)	(2)	(3)	(4)	(5)	(6)
Total steps	-0.093					
	(0.069)					
Active minutes		-0.163**				
		(0.069)				
Sleep duration			0.160**			
			(0.068)			
Wake up time				0.093		
				(0.070)		
Screen Time					0.098	
					(0.071)	
Social Interactions						-0.041
						(0.072)
Observations	201	196	200	194	184	184
R-squared	0.009	0.028	0.027	0.009	0.010	0.002
Mean of Dep. Var.	0.323	0.316	0.320	0.314	0.321	0.321
std. dev.	0.469	0.466	0.468	0.465	0.468	0.468

Notes - The sample is restricted to the individuals observed in Spring 2020. The table above reports the coefficients of univariate regression of an indicator for $CES-D > 15$ on dummy variables identifying larger disruptions in total number of steps, active minutes, sleep duration, wake up time, screen time, and social interactions.

*** p<0.01, ** p<0.05, * p<0.1

Table S.8: Gender, Race, SES and Mental Health During Covid-19

	CESD>15 at endline							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female	0.130*							0.131*
	(0.075)							(0.078)
White		0.014						
		(0.066)						
Black			0.027					0.012
			(0.095)					(0.103)
Asian				-0.063				-0.057
				(0.070)				(0.078)
Hispanic					0.039			0.045
					(0.159)			(0.144)
Some financial aid						0.008		0.012
						(0.067)		(0.073)
First generation student							-0.089	-0.064
							(0.104)	(0.115)
CESD>15 at baseline	0.347***	0.353***	0.349***	0.357***	0.351***	0.353***	0.348***	0.354***
	(0.061)	(0.063)	(0.061)	(0.062)	(0.061)	(0.062)	(0.061)	(0.064)
Observations	205	205	205	205	205	205	205	205
R-squared	0.128	0.114	0.115	0.118	0.115	0.114	0.118	0.120
Mean of Dep. Var.	0.615	0.615	0.615	0.327	0.615	0.615	0.615	0.615
std. dev.	0.488	0.488	0.488	0.470	0.488	0.488	0.488	0.488

Notes - We report the correlation between gender, race, and SES and mental health at the end of the 2020 term.
*** $p<0.01$, ** $p<0.05$, * $p<0.1$

2.4 Predicting clinical depression

2.4.1 Methods and Variables

To complement the descriptive statistics, difference-in-differences and individual fixed-effects regressions relating lifestyle disruptions and mental health during COVID-19, we use the XGBoost (eXtreme Gradient Boosting) algorithm [Chen et al., 2019] to predict risk of clinical depression (CES-D score > 15) for each participant in the Spring 2019, Fall 2019, and Spring 2020 cohorts.

To protect against overfitting, we use 10-fold cross-validation to determine the optimal number of boosting iterations (i.e., number of base learner decision trees to build). This sets apart a portion of the data as a testing sample, fits the model on the remaining training sample, and

Table S.9: Gender, Race, SES and Physical Activity During Covid-19

	steps							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female	-115.165							-80.797
	(354.219)							(336.717)
White		1,338.485***						
		(322.643)						
Black			-1,657.492***					-1,785.109***
			(355.625)					(395.671)
Asian				-594.600*				-944.029**
				(355.169)				(372.315)
Hispanic					-213.141			-487.808
					(729.013)			(816.443)
Some financial aid						825.633**		510.050
						(339.458)		(358.213)
First generation student							-956.978**	-547.545
							(384.732)	(382.830)
total steps	0.708***	0.659***	0.704***	0.685***	0.710***	0.708***	0.707***	0.665***
at baseline	(0.078)	(0.079)	(0.075)	(0.081)	(0.078)	(0.079)	(0.078)	(0.080)
Observations	205	205	205	205	205	205	205	205
R-squared	0.128	0.114	0.115	0.118	0.115	0.114	0.118	0.120
Mean of Dep. Var.	0.615	0.615	0.615	0.327	0.615	0.615	0.615	0.615
std. dev.	0.488	0.488	0.488	0.470	0.488	0.488	0.488	0.488

Notes - We report the correlation between gender, race, and SES and physical activity at the end of the 2020 term.

*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Table S.10: Location during Lockdown and Depression

	Dependent variable:			
	CES-D > 15 at endline			
	(1)	(2)	(3)	(4)
High case exposure	0.005 (0.064)			
High death exposure		0.009 (0.064)		
High case-fatality ratio			−0.007 (0.064)	
Stayed in Pittsburgh				0.019 (0.078)
Baseline CES-D	0.021*** (0.004)	0.021*** (0.004)	0.021*** (0.004)	0.021*** (0.004)
Observations	202	202	202	202
R-squared	0.126	0.126	0.126	0.126
Mean of Dep. Var.	0.62	0.62	0.62	0.62
std. dev.	0.49	0.49	0.49	0.49

Notes - "High case exposure," "High death exposure," and "High case-fatality ratio" mean the number of confirmed COVID-19 cases per million, the number of confirmed COVID-19 deaths per million, and the case-fatality ratio (confirmed cases divided by confirmed deaths) as of April 17, 2020 (the date of the endline survey in Spring 2020) in the participant's county of residence were greater than the sample median, respectively. County of residence was determined from the geolocation coordinates recorded at the time of their response to the endline survey in Spring 2020. "Stayed in Pittsburgh" is a dummy variable that equals 1 if the recorded geolocation of participant's responses to the time use surveys after the transition to online classes on March 11, 2020 were within 25 kilometers of campus. The sample for this table excludes individuals that were not in the United States at the time of the endline survey.

*p<0.1, **p<0.05, ***p<0.01

reports the number of boosting iterations after which the testing sample error began to increase. We then run the model using this number of iterations. The maximum tree depth, number of parallel trees, and learning rate were selected via grid search to minimize prediction error.[11]

For each cohort, we repeat the algorithm 10,000 times and take the average predictions, feature importance measures, and feature contributions (SHAP values), respectively. Because the XGBoost algorithm randomly subsamples features and observations in each training instance, even if the predictions are stable across instances, there is some randomness in selected features and their relative importance. Since we are interested not only in predictive accuracy but also the relative importance of features (see Figure 5) and their contributions to depression risk, we average across a large number of iterations to achieve stable measures of relative importance and estimated feature contributions.

The measure we use to determine relative feature importance is *gain*, which is defined as the average increase in predictive accuracy that is achieved from splitting a decision tree using that feature [Chen et al., 2019]. The XGBoost algorithm produces three options for feature importance: gain, *cover*, which is based on the number of training instances in which the feature is used to split a tree, including branches; and *frequency*, which is based on the number of times the feature is used to split a tree. Each of these measures is normalized so that their values across all features in the model sum to 1. We use gain because it is the most direct measure of the predictive power of a feature, which captures how much of the observed risk of depression can be explained by a particular feature. For example, in the 2019 cohorts, the feature with the highest gain by far is baseline CES-D score, which suggests that preexisting risk of depression can explain a greater share of the observed depression at endline than any particular time use activity, demographic characteristic, or biometric observation.

We also use the algorithm output to calculate SHAP (Shapley Additive Explanation) values for each observed value of each feature [Lundberg et al.]. SHAP values estimate the increase or decrease in predicted depression risk that is attributable to a particular observation of a particular feature. For example, Figure 6 shows that in the pooled Spring and Fall 2019 cohort model,

---

[11]The selected hyperparameter values were `eta` (the learning rate) = 0.3, `subsample` = 0.5 (as an additional protection against overfitting, this randomly samples half of the observations as the training sample), `colsample_bytree` = 0.5 (this reduces dimensionality by taking a random subsample of half the features when building each tree), `num_parallel_trees` = 10, and `max_depth` = 4.

a value of 25 or higher in the CES-D score at baseline was assigned a SHAP value of approximately 1.0, which means that the predicted probability of endline depression is increased by approximately 1.[12] The absolute magnitude of these SHAP values provides an additional measure of feature importance to complement gain, and their sign represents the directionality of the relationship between ranges of feature values and endline depression. Negative SHAP values suggest that a particular value of a feature is associated with reduced risk of depression at endline, while positive SHAP values suggest increased risk. SHAP values close to zero suggest that a particular value of a feature is not significantly associated with risk of endline depression in either direction.

To generate average predictions from the 10,000 repeated iterations of the algorithm, we take the mean of the raw output for each observation. We then dichotomize the prediction by assigning 1 to observations above the mean prediction and 0 to those at or below the mean prediction. These dichotomous predictions are used to measure the predictive accuracy of each model, as reported in the results section above. The same process of averaging the raw output scores across 10,000 iterations and dichotomizing the result is used to determine the accuracy when using the Spring 2019 model to predict depression in Fall 2019 and Spring 2020 out-of-sample respectively.

The variables included as features in the baseline (without disruptions) model are described below:

Demographics: Highest education level completed by parents, self-reported health status, gender, age, race/ethnicity, category of college major (social science, medical, STEM, business, humanities), current year in school, assigned treatment group in prior wellness study from which participants were recruited

Mental Health: Baseline CES-D score, baseline GAD-7 score, baseline self-reported life satisfaction, baseline resilience index

Time Use: Daily hours spent studying, daily hours spent working, daily hours spent eating or preparing meals and snacks, daily hours of screen time (playing games, watching TV, or surfing the internet), and daily hours of social interactions

---

[12]If CES-D score at baseline was the only feature in the model, this would suggest near-certainty of depression at endline; however, since SHAP values are additive, there may be negative SHAP values associated with other features for the same individual.

Physical Activity: Daily active (non-sedentary) hours, daily steps, resting heart rate

Sleep: Average mid-sleep time (time at which half of the total sleep duration was completed), sleep disruptions (percentage of minutes in bed during which individual is not asleep), daily hours of sleep

In the models including lifestyle disruptions, we use the difference from baseline to midline (for Spring 2020) or endline (for Spring 2019 and Fall 2019, which did not have a midline survey) for the features in the sleep, physical activity, and time use categories above.

2.4.2 Results

Table S.11 displays the relative importance of all predictiors included in the Baseline models. Figure S.8 displays additional SHAPs for the main lifestyle and time use measures: Sleep Disruption, Mid-Sleep, Studying, Working and Screen Time.

Table S.12 displays the relative importance of all predictiors included in the diffrences models. Figure S.9 displays additional SHAPs for the main lifestyle and time use measures: Sleep Disruption, Mid-Sleep, Studying, Working and Screen Time.

Table S.11: Relative Importance of Predictors: Baseline Model

Baseline Model	2019 Pooled	2020
Mental Health - Baseline	**0.59**	**0.38**
CES-D	0.38	0.11
GAD-7	0.09	0.17
Life Satisfaction	0.08	0.01
Resilience	0.05	0.10
Time Use - Baseline	**0.11**	**0.22**
Studying	0.03	0.06
Eating/Cooking	0.02	0.03
Social Interactions	0.02	0.05
Screen	0.02	0.04
Working	0.01	0.05
Physical Activity - Baseline	**0.10**	**0.14**
Steps	0.04	0.04
Active Hours	0.03	0.05
Resting Heart Rate	0.03	0.05
Sleep - Baseline	**0.12**	**0.15**
Mid-Sleep	0.05	0.05
Sleep Duration	0.04	0.04
Sleep Disruption	0.03	0.06
Demographics	**0.07**	**0.09**
Female	0.00	0.01
Age	0.01	0.01
Financial Aid	0.01	0.01
Parents' Highest Education	0.01	0.02
Ethnicity: Black	0.00	0.00
Ethnicity: Hispanic or Latino	0.00	0.00
Ethnicity: Asian or Pacific Islander	0.01	0.01
Ethnicity: Caucasian/White	0.01	0.00
Year in College: 1	0.00	0.01
Year in College: 2	0.00	0.01
Year in College: 3	0.00	0.00
Year in College: 4	0.00	0.00
Field of Study: Social Science	0.00	0.00
Field of Study: Humanities	0.01	0.00
Field of Study: Business	0.00	0.00
Field of Study: Medical	0.00	0.00
Other	**0.02**	**0.03**
Self-reported Health	0.01	0.02
Incentivized to Sleep	0.01	0.02

Figure S.8: Additional SHAP Figures from Baseline Model
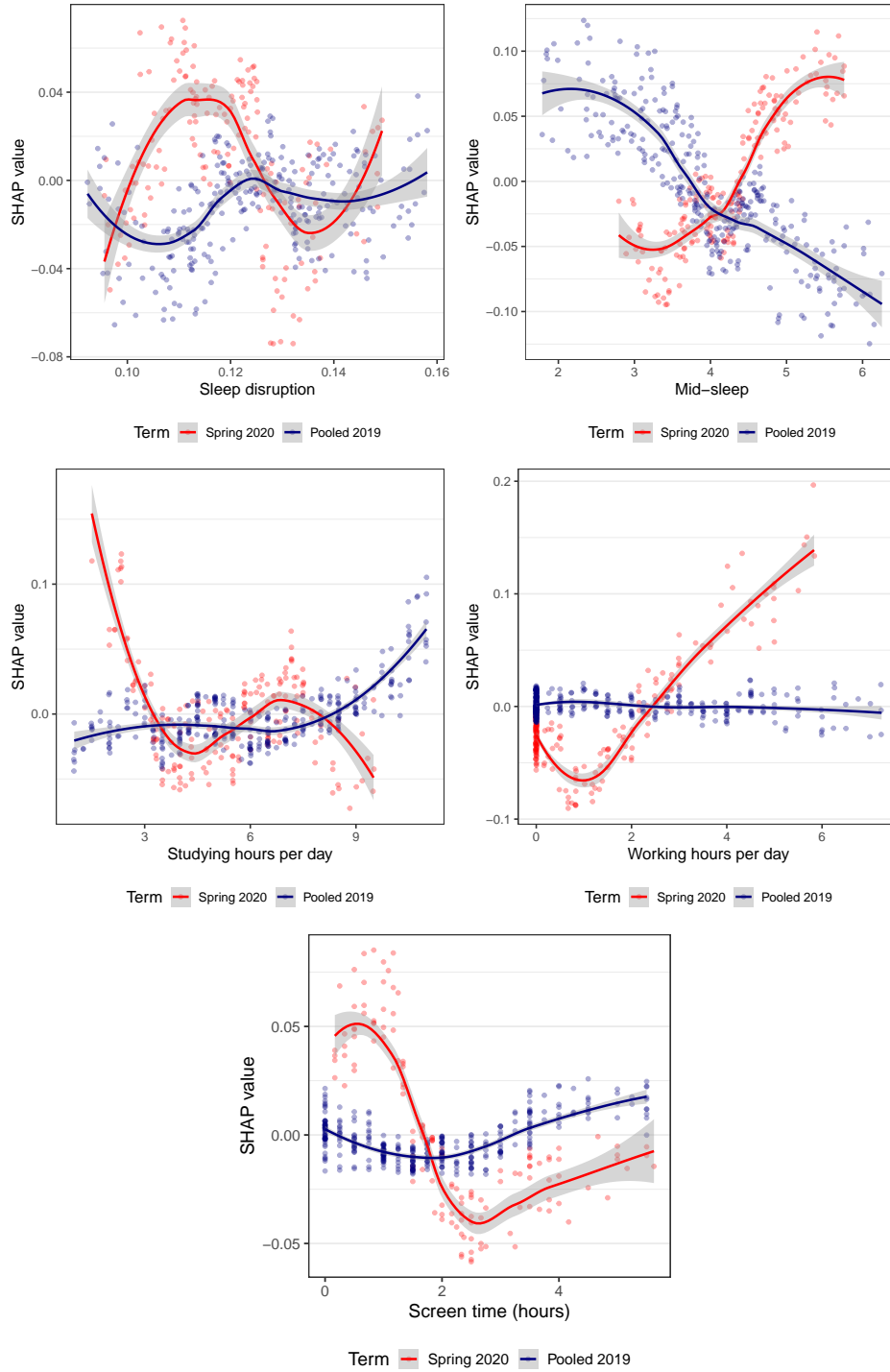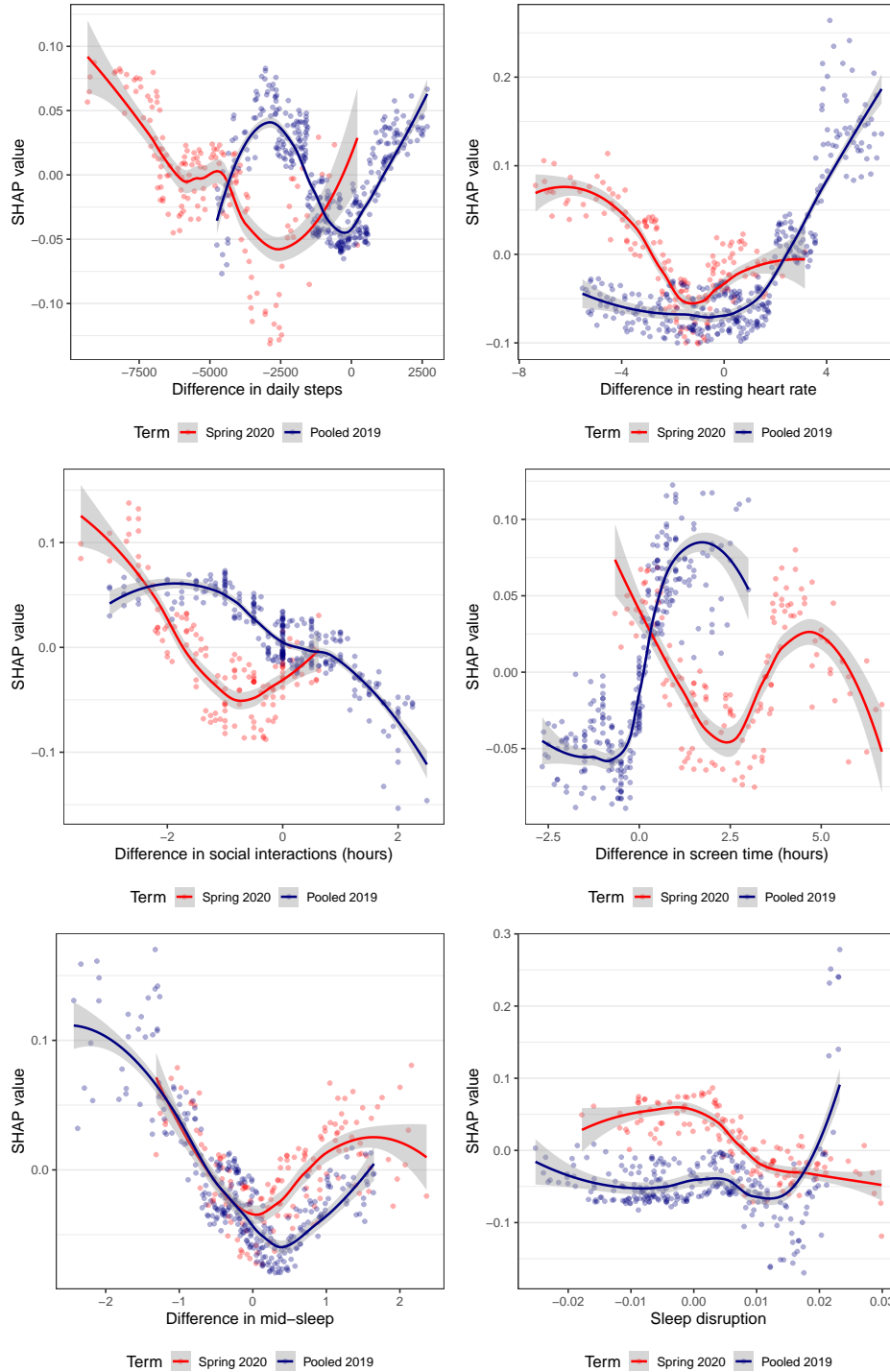

Notes - The figure displays the estimated additive feature contributions (SHAP values) for baseline sleep disruption (percentage of minutes in bed not asleep), mid-sleep (time at which exactly half of the total sleep duration had been completed), studying and working hours per day, and screen time at baseline. A higher SHAP value indicates a higher risk of endline CESD>15.

Table S.12: Relative Importance of Predictors: Differences Model

Differences Model	2019 Pooled	2020
Mental Health - Baseline	**0.559**	**0.321**
CESD	0.358	0.086
GAD7	0.083	0.141
Life Satisfaction	0.070	0.005
Resilience	0.046	0.089
Time Use - Changes	**0.137**	**0.227**
Studying	0.032	0.049
Eating/Cooking	0.024	0.042
Social Interactions	0.028	0.042
Screen	0.034	0.040
Working	0.020	0.056
Physical Activity - Changes	**0.101**	**0.209**
Steps	0.033	0.045
Active Hours	0.027	0.120
Resting Heart Rate	0.042	0.044
Sleep - Changes	**0.121**	**0.149**
Mid-Sleep	0.038	0.040
Sleep Duration	0.032	0.060
Sleep Disruption	0.051	0.048
Demographics	**0.063**	**0.065**
Female	0.002	0.006
Age	0.007	0.006
Financial Aid	0.010	0.005
Parents' Highest Education	0.011	0.021
Ethnicity: Black	0.002	0.001
Ethnicity: Hispanic or Latino	0.000	0.000
Ethnicity: Asian or Pacific Islander	0.007	0.002
Ethnicity: Caucasian/White	0.007	0.004
Year in College: 1	0.002	0.010
Year in College: 2	0.002	0.004
Year in College: 3	0.002	0.002
Year in College: 4	0.002	0.001
Field of Study: Social Science	0.002	0.002
Field of Study: Humanities	0.005	0.000
Field of Study: Business	0.002	0.000
Field of Study: Medical	0.001	0.000
Other	**0.018**	**0.030**
Self-reported Health	0.014	0.021
Incentivized to Sleep	0.005	0.009

Figure S.9: Additional SHAP Figures from Differences Model

*Notes -* The figure displays the estimated additive feature contributions (SHAP values) for differences in daily steps, resting heart rate, social interactions, screen time, mid-sleep, and sleep disruption. This difference compares individuals' average values for each feature after March 23 to their corresponding average values at baseline. A higher SHAP value indicates a higher risk of endline CESD>15.

2.5 Figures and Regression with alternative data definition

Alternative Data Definition

In this section, we repeat the analysis from the main text but use a different data definition. In the main text, we compare our measures of lifestyle and time use from the baseline period (before participants were randomized to treatments on February 17th) to data collected starting March 23rd, the day on which students started their classes online after a prolonged spring break (March 9–March22. As a robustness check, here we compare baseline data (through February 16th) to the combined post-baseline data (starting February 17th).

Figure S.10: Screen Time and Social Interactions Before and After Feb 17
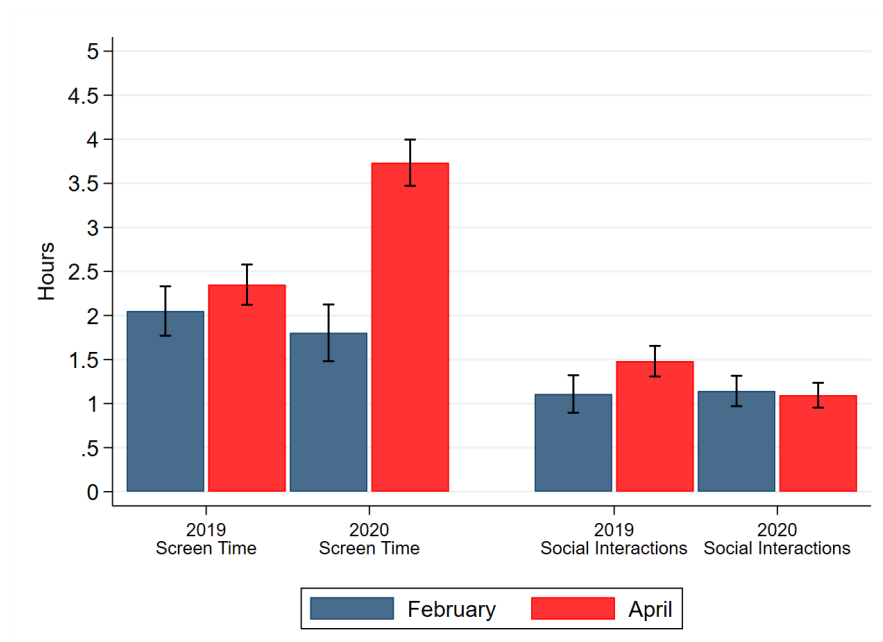
Notes - The figures above display the average time spent with friends (social time) and the average screen time before and after the beginning of the lockdown during the Spring term of 2019 and 2020. Screen time does not include time spent working or studying on a device, but it includes playing games, watching TV, or surfing the Internet.

Figure S.10 shows data for time use (screen time and social interactions) using this alternative data definition. As shown in the text, we still see a spike in screen time. However, with this conservative data definition we no longer see a substantial decline in social interactions. Figure

S.11 is analogous to Figure 7 from the main text, and displays depression rates (CES-D>15) for participants with smaller (below median) or larger (above median) disruptions in the number of total steps, active hours, sleep duration, wake-up time, screen time, and social interactions. As in the main text, we see that directionally people with larger disruptions are more likely to be depressed.

Figure S.11: Habit Disruptions and Depression Before and After Feb 17
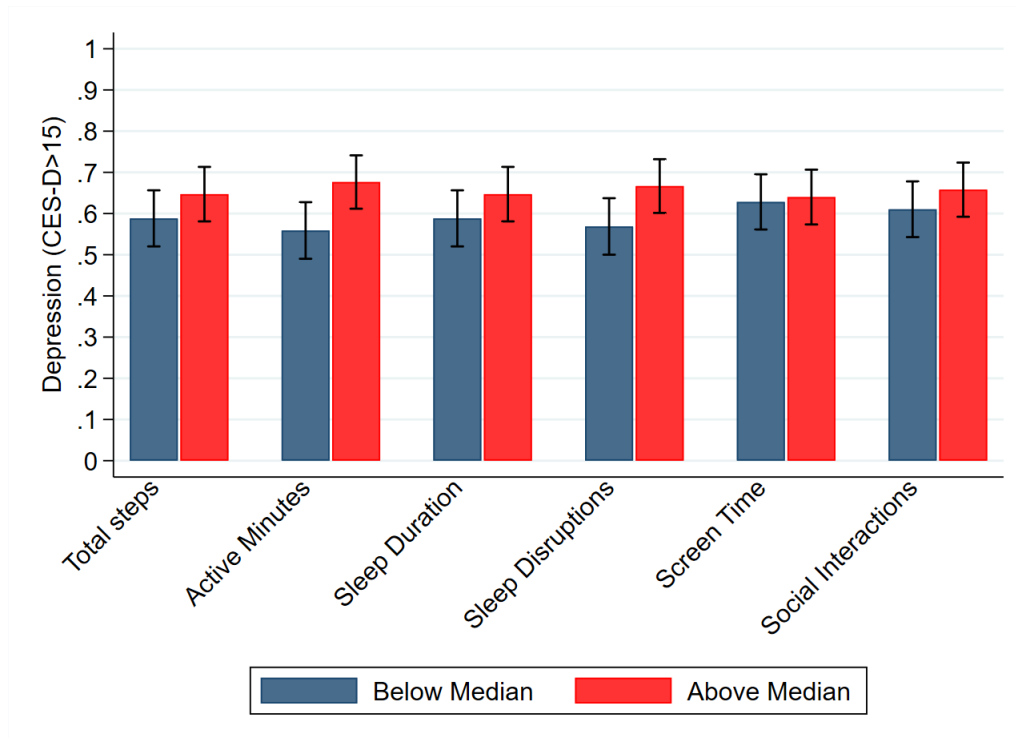
Notes - The figure reports the proportion of individuals reporting clinical depression (CES-D>15) below (Smaller Disruptions) and above (Larger Disruptions) median change in steps, sleep, social, and screen time.

Tables S.13 and S.14 display the difference-in-differences and fixed effects regressions using this alternative data definition. Overall, the results are not substantially different from those reported in the main analysis (see also Figure S.13).

Table S.13: Difference in difference analysis: Spring 2019 & Spring 2020 (Baseline vs. Post-Baseline)

	(1)	(2)	(3)	(4)
Panel A: Mental Health	CES-D	CES-D>15	GAD-7	GAD-7>7
end of term 2020	4.940***	0.234***	0.639	0.055
	(1.041)	(0.058)	(0.478)	(0.058)
end of term	1.932***	0.064	1.072***	0.113***
	(0.698)	(0.043)	(0.343)	(0.042)
Observations	779	779	779	779
R-squared	0.152	0.109	0.074	0.060
Number of individuals	460	460	460	460
Mean of Dep. Var.	14.59	0.408	6.241	0.331
std. dev.	9.640	0.492	4.449	0.471
Panel B: Activity and Sleep	Total Steps	Active Hours	Sleep Hrs	Wake Up Time
end of term 2020	-4,595.117***	-1.067***	0.230***	0.379***
	(237.886)	(0.086)	(0.071)	(0.098)
end of term	-149.610	-0.046	0.067	-0.068
	(138.882)	(0.047)	(0.054)	(0.053)
Observations	864	853	875	865
R-squared	0.439	0.257	0.111	0.117
Number of individuals	491	486	474	474
Mean of Dep. Var.	8214	4.164	7.083	7.755
std. dev.	3503	1.104	0.769	1.141
Panel C: Time Use	Social Interactions	Screen Hours	Work Hours	Study Hours
end of term 2020	-0.959***	3.333***	-0.265	-1.545***
	(0.153)	(0.219)	(0.240)	(0.313)
end of term	0.096	-0.236	0.185	-0.683***
	(0.127)	(0.157)	(0.198)	(0.252)
Observations	739	739	739	739
R-squared	0.091	0.304	0.191	0.189
Number of individuals	391	391	391	391
Mean of Dep. Var.	1.188	2.854	1.888	4.774
std. dev.	1.202	2.171	2.153	2.730

Notes - All estimates include dummies for year and wave, and controls for gender, age, college year, parental highest education, financial aid, race, ethnicity, a dummy for whether the individual stayed in the Pittsburgh area after the beginning of remote learning, and dummies controlling for treatment assignment. Standard errors are clustered at the individual level.

*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Table S.14: Fixed-Effects Analysis (Baseline vs. Post-Baseline): Spring 2019 & Spring 2020

|	(1)	(2)	(3)	(4)
Panel A: Mental Health	CES-D	CES-D>15	GAD-7	GAD-7>7
end of term * 2020	5.682***	0.238***	1.079**	0.074
	(0.969)	(0.059)	(0.440)	(0.057)
end of term	1.895***	0.075*	1.052***	0.113***
	(0.639)	(0.042)	(0.319)	(0.042)
Observations	779	779	779	779
R-squared	0.303	0.177	0.167	0.090
Number of individuals	460	460	460	460
Mean of Dep. Var.	14.59	0.408	6.241	0.331
std. dev.	9.640	0.492	4.449	0.471
Panel B: Activity and Sleep	Total Steps	Active Hours	Sleep Hrs	Wake Up Time
end of term * 2020	-4,580.660***	-1.112***	0.385***	0.545***
	(199.454)	(0.075)	(0.089)	(0.119)
end of term	-139.423	-0.039	-0.037	-0.098
	(136.178)	(0.046)	(0.073)	(0.083)
Observations	928	913	875	875
R-squared	0.750	0.537	0.105	0.080
Number of individuals	491	486	474	474
Mean of Dep. Var.	8079	4.125	7.115	7.803
std. dev.	3544	1.132	0.825	1.252
Panel C: Time Use	Social Interactions	Screen Hours	Work Hours	Study Hours
end of term * 2020	-1.008***	3.324***	-0.341	-1.411***
	(0.154)	(0.205)	(0.232)	(0.291)
end of term	0.136	-0.348**	0.254	-0.760***
	(0.129)	(0.147)	(0.192)	(0.244)
Observations	719	719	719	719
R-squared	0.232	0.614	0.007	0.359
Number of individuals	391	391	391	391
Mean of Dep. Var.	1.100	3.212	1.854	4.408
std. dev.	1.188	2.488	2.113	2.749

Notes - All estimates include a year dummy and individual fixed effects. Standard errors are clustered at the individual level.
*** p<0.01, ** p<0.05, * p<0.1

Predicting Clinical Depression, Alternative Data Definition

Below we replicate the prediction exercise with the alternative data definition discussed in the previous section.

Figure S.12: Baseline Measures and Depression - Alternative Definition
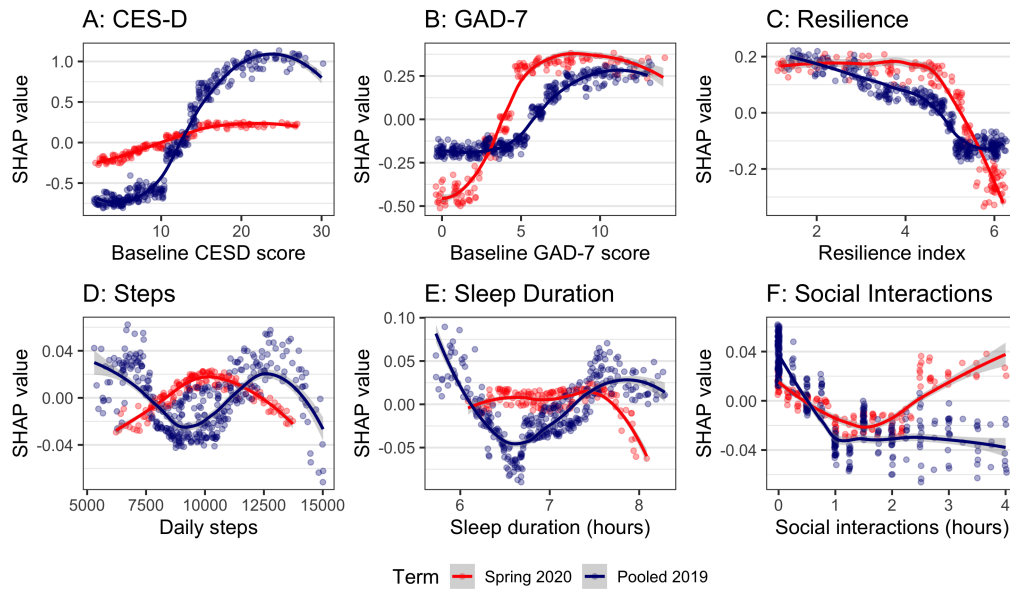
Notes - The figure displays the estimated additive feature contributions (SHAP values) for baseline values of CES-D score, GAD-7 score, resilience index, daily steps, daily sleep duration, and daily social interactions, respectively. This figure uses the alternative definition of "pre" and "post" based on the date of the University of Pittsburgh's announcement of remote instruction, March 11. A higher SHAP value indicates a higher risk of endline CESD>15.

71

Figure S.13: Changes in Activities and Depression - Alternative Definition
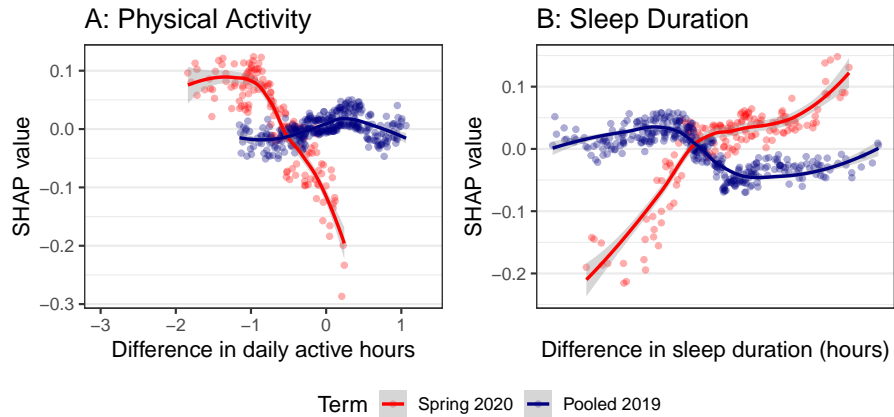
A: Physical Activity   B: Sleep Duration

Term ▬ Spring 2020 ▬ Pooled 2019

Notes - The figure displays the estimated additive feature contributions (SHAP values) for differences in daily active hours and daily sleep duration. This difference compares individuals' average values for each feature after March 11 (the alternative definition of "pre" and "post") to their corresponding average values at baseline. A higher SHAP value indicates a higher risk of endline CESD>15.

Bounce back and Incentives for Exercising RCT

2.6   Bounce back

The wellness study was set to conclude at the end of the semester, the last week of April. We obtained approval from the University of Pittsburgh Institutional Review Board to invite participants to continue their participation in a second phase of the study. A total of 204 participants (78.8%) elected to participate in exchange for a minimum payment of $50. We continued following these participants through May and June. Throughout this period we sent reminders to sync and administered time use and experience sampling surveys every week, as we did during the semester. The weekly time use surveys were filled out by 75.3% of participants (ranging from 89.2% in week 18 to 60.1% in week 17).

On May 29th, 2020, we elicited measures of mental health (CES-D and GAD-7). A total of 176 participants filled out the survey (86.3%). We observe an improvement in mental health. The average CES-D score among these participants was 17.54, significantly lower than in April ($p < 0.001$), and the fraction of participants at risk of depression declined to 50%. We compare

average biometrics and time use measures for those participants at the onset of the pandemic (March 23rd to April 20th) to averages during Phase 2, from April 20th to May 30th. Our analyses reveal a significant increase in steps and active minutes, and a directional decline in sleep. As for time use, social interactions increase by approximately 10 minutes ($p < 0.010$). Screen time does not bounce back. Rather, it increases by approximately 1 hour ($p < 0.010$) after the end of the semester, see Table S.15.

2.6.1 Randomization and Procedures

We randomized participants to a control and a treatment group blocking on CES-D levels at the end of May. We pre-registered the intervention in the AEA RCT Registry (RCT ID #AEARCTR-0005949). On June 1st we informed participants in the treatment group that we would pay them for walking a minimum of 10,000 steps per day for the next 14 days. Participants received a $5 transfer for each day in which they reached this goal (up to a maximum payment of $70). Every morning, we notified participants of whether they had achieved, or failed to achieve, this goal. To address potential income effects, we gave participants in the control group the same distribution of payments in the treatment group. We anonymously paired each treatment group subject with a control group subject who had a similar CES-D score in May in order to balance the conditional distribution of payments on pre-interventiuon depression. This was done using an optimal pair matching algorithm [Hansen and Klopfer, 2006]. The last day of intervention was June 14th. The final payments were processed on June 15th. On June 16th, we surveyed our participants and elicited the mental health measures (CES-D and GAD-7). A total of 181 participants filled out the survey (89.2% of all participants who elected to participate in phase 2).

Results

Table S.16 shows the results. As shown in the table, the intervention significantly increased the number of steps and active hours ($p < .001$). However, as shown in Table S.17, the increase in physical activity did not result in a decrease in depression. The intervention led to a small directional but not significant decrease in average CES-D score. When looking at the fraction of participants at or above the threshold of 16 in the CES-D scale, we see that the fraction of people

Table S.15: Time Use and Activity - Bounce Back

Dep.var.	(1) Steps	(2) Active (hrs)	(3) Sleep (hrs)	(4) Wake up time	(5) Screen time	(6) Social interactions
Δ May 30–April 20	1,202.763***	0.524***	-0.130**	-0.195***	1.009***	0.162*
	(140.284)	(0.074)	(0.062)	(0.067)	(0.205)	(0.088)
Observations	552	514	539	535	388	388
R-squared	0.033	0.043	0.005	0.006	0.032	0.006
Mean of Dep. Var.	4984	3.443	7.078	8.523	5.573	0.589
std. dev.	3319	1.269	0.971	1.313	2.822	1.084

Notes - These estimates report the difference in the number of steps, active hours, sleep, wake up time, screen time, social interactions between the end of the term (April 20) and the end of May 2020.

*** p<0.01, ** p<0.05, * p<0.1

74

above the threshold is directionally higher for participants in the treatment group. However, this is driven by very few observations, as 86% of the participants do not change status.

Table S.16: Incentive to Increase Physical Activity

Dep. Var. as of June 15	(1) Total Steps	(2) Total Steps	(3) Total Steps	(4) Total Steps	(5) Activity (hrs)	(6) Activity (hrs)	(7) Activity (hrs)	(8) Activity (hrs)
treatment	2,016.335*** (330.994)	2,315.276*** (335.302)	1,927.313*** (647.242)	1,915.127*** (648.792)	0.591*** (0.139)	0.610*** (0.135)	0.539** (0.261)	0.535** (0.262)
Above median recovery in activity in May		−263.366 (342.769)	−440.608 (466.634)	−463.577 (468.952)		−0.041 (0.138)	−0.098 (0.188)	−0.105 (0.189)
CESD>15 as of May 2020		213.611 (332.011)	62.600 (466.634)	210.133 (522.446)		0.032 (0.134)	0.037 (0.188)	0.080 (0.211)
treatment * above median recovery in activity in May			385.461 (691.725)	413.636 (694.504)			0.126 (0.279)	0.134 (0.280)
treatment * CESD>15 as of May 30 2020			305.859 (667.445)	245.577 (675.499)			−0.012 (0.269)	−0.030 (0.273)
CESD>15 at baseline				−262.969 (415.545)				−0.077 (0.168)
Observations	190	160	160	160	190	160	160	160
Mean of Dep. Var.	979.051	979.051	979.051	979.051	0.332	0.332	0.332	0.332
Std.Dev.	2135.807	2135.807	2135.807	2135.807	0.884	0.884	0.884	0.884

Notes - The table above reports differences between subjects in the treatment and control group in the change in steps and active minutes during the intervention relative to the period going from April 20 to May 31st. The sample size decreases from 190 to 160 when adding controls because 30 individuals had Fitbit data from the intervention period but did not respond to the May CES-D survey.

*** p<0.01, ** p<0.05, * p<0.1

Table S.17: Physical Activity Incentives and Depression

Dep. Var. as of June 15	(1) CESD	(2) CESD	(3) CESD	(4) CESD	(5) CESD>15	(6) CESD>15	(7) CESD>15	(8) CESD>15
treatment	-0.292 (1.555)	-0.342 (1.207)	-0.128 (1.523)	0.664 (1.493)	0.058 (0.075)	0.053 (0.055)	0.078 (0.067)	0.097 (0.067)
recovery in activity in May		-0.011 (0.022)	0.558 (1.848)	1.019 (1.739)		-0.001 (0.001)	0.068 (0.081)	0.079 (0.078)
CESD>15 as of May 30 2020		14.482*** (1.230)	14.481*** (1.235)	11.662*** (1.212)		0.727*** (0.056)	0.727*** (0.056)	0.657*** (0.070)
treatment * above median recovery in activity in May			-0.566 (1.841)	-1.036 (1.732)			-0.068 (0.081)	-0.080 (0.078)
treatment * CESD>15 as of May 30 2020			0.004 (0.045)	-0.005 (0.042)			0.001 (0.002)	0.000 (0.002)
CESD>15 at baseline				6.431*** (1.435)				0.160** (0.072)
Observations	179	179	179	179	179	179	179	179
Mean of Dep. Var	15.468	15.468	15.468	15.468	0.44	0.44	0.44	0.44
Std.Dev.	10.360	10.360	10.360	10.360	0.50	0.50	0.50	0.50

Notes - The table above reports estimates of the differences between the treatment and the control group in CES-D score and depression (CES-D >15). All estimates include an indicator for whether information on May 16 CESD was missing.

*** p<0.01, ** p<0.05, * p<0.1

77

Table S.18: Other time use variables, Spring 2019 & Spring 2020

Dep. Var.	(1) Eating and preparing food	(2) House errands	(3) Personal care	(4) Commuting (car or bus)
end of term * 2020	0.510***	0.055	-0.040	-0.218***
	(0.085)	(0.044)	(0.058)	(0.068)
end of term	-0.133**	-0.033	-0.045	-0.005
	(0.067)	(0.028)	(0.046)	(0.051)
Observations	719	719	719	719
R-squared	0.115	0.110	0.127	0.112
Mean of Dep. Var.	1.580	0.270	0.876	0.313
std. dev.	0.707	0.375	0.524	0.506

Notes - All estimates include dummies for year and wave, and controls for gender, age, college year, major, parental highest education, financial aid, race, ethnicity, a dummy for whether the individual stayed in the Pittsburgh area after the beginning of remote learning, and dummies controlling for treatment assignment. Standard errors are clustered at the individual level. *** $p<0.01$, ** $p<0.05$, * $p<0.1$

Table S.19: Selection into Phase 2 of the Study

Dep. Var.:	(1) Total steps	(2) Active minutes	(3) Sleep (hrs)	(4) Wake up time	(5) Social time (hrs)	(6) Screen time (hrs)	(7) Female	(8) Age	(9) College year	(10) CES-D	(11) GAD-7	(12) Resilience index
Did not participate to phase 2	-288.227 (342.531)	-0.143 (0.124)	-0.131 (0.090)	-0.152 (0.157)	-0.013 (0.242)	-0.690** (0.277)	-0.059 (0.055)	0.177 (0.259)	0.097 (0.132)	1.483 (1.089)	0.651 (0.525)	1.146 (0.871)
Observations	302	301	295	295	238	238	305	305	305	305	305	304
R-squared	0.003	0.005	0.009	0.004	0.000	0.028	0.004	0.002	0.002	0.007	0.005	0.005
Mean of Dep. Var.	9731	4.516	7.003	7.451	1.562	2.543	0.725	19.35	2.889	12.67	5.659	22.07
std. dev.	2641	0.960	0.663	1.172	1.251	1.592	0.447	1.996	1.089	8.747	4.339	7.699

Notes - The estimates above report univariate regression of the outcomes on an indicator for whether the subject agreed to participate to the second phase of the study. Robust standard errors are in parentheses.

*** $p<0.01$, ** $p<0.05$, * $p<0.1$