Ayo Animashaun, Bo Brandt, Rohan Kapuria, Joyce Lee
INFO 290 Final Project: Experiments and Causal Inference

Grayscale & Screen Time

I. Background

The rise of smartphones and continuous connectivity has given way to what some describe as "compulsive" use.[1] This has been purported to yield negative outcomes in terms of productivity, such diminished attention spans, due to frequent interruptions.[2] Some research suggests that the reduction in ability to concentrate can be long lasting or permanent. In fact, researchers have found that even the mere presence of one's smartphone – when not in use – may induce a "brain drain," occupying limited-capacity cognitive resources for purposes of attentional control.[3] Researchers have also found smartphone use to have detrimental effects on enjoyment of face-to-face interactions[4] and the associated mental health benefits of social connections: popular media coverage has often focused on the damaging effects of smartphones among youth users, with sensationalist headlines such as "Have Smartphones Destroyed a Generation?"[5]

This technology backlash has led to the emergence of the "digital wellness" movement, pioneered by the Center for Humane Technology, which aims to "realign technology with humanity's best interests."[6] The notion of "responsible" technology development has placed mounting public pressure on large technology companies to take action and curtail profiteering off user engagement. At the beginning of this year, for instance, a contingent of Apple shareholders wrote a public letter asking for better controls for parents to help protect their kids from the risks of digital addiction and the side effects of social media: "The average American teenager who uses a smart phone receives her first phone at age 10 and spends over 4.5 hours a day on it (excluding texting and talking)," the investors wrote, adding that "78 percent of teens check their phones at

[1] Park, B.W., Lee, K.C. (2011) The Effect of Users' Characteristics and Experiential Factors on the Compulsive Usage of the Smartphone. In: Kim T., Adeli H., Robles R.J., Balitanas M. (eds) Ubiquitous Computing and Multimedia Applications. UCMA 2011. Communications in Computer and Information Science, vol 151. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-20998-7_52.

[2] Darmoul, S., Ahmad, A., Ghaleb, M., Alkahtani, M. "Interruption Management in Human Multitasking Environments," IFAC-PapersOnLine, Volume 48, Issue 3, 2015, 1179-1185. https://doi.org/10.1016/j.ifacol.2015.06.244.

[3] Ward, A.F., Duke, K., Gneezy, A., & Bos, M.W. (2017). "Brain Drain: The Mere Presence of One's Own Smartphone Reduces Available Cognitive Capacity," Journal of the Association for Consumer Research 2, no. 2 (April 2017): 140-154. https://doi.org/10.1086/691462.

[4] Dwyer, R.J., Kushlev, K., Dunn, E.W. Smartphone use undermines enjoyment of face-to-face social interactions, Journal of Experimental Social Psychology, Volume 78, 2018, 233-239. https://doi.org/10.1016/j.jesp.2017.10.007.

[5] Twenge, Jean. "Have Smartphones Destroyed a Generation?" The Atlantic, September 2017. https://www.theatlantic.com/magazine/archive/2017/09/has-the-smartphone-destroyed-a-generation/534198.

[6] Center for Humane Technology. 2018. http://humanetech.com.

least hourly and 50 percent report feeling 'addicted' to their phones."[7] In response to such dialogue, both major smartphone companies announced features to discourage use at their annual developer conferences earlier this year: Google's "Digital Wellbeing" and Apple's "Screen Time" offer engagement data dashboards to users, in addition new capabilities to limit device usage either overall or for specific applications.[8]

II. Motivation

While the introduction of digital wellness tools offers the opportunity for greater self-awareness and self regulation, we questioned the efficacy of such offerings: does becoming aware of one's mobile phone usage affect future behavior? From our own personal experiences, we did not find that usage information was sufficient to diminish phone usage. Peer-reviewed research on estimated versus real-world smartphone use is somewhat lacking, although a study published in 2015 found that test subjects (n=23) were good at guessing the time spent on devices, but tended to underestimate the number of times they check their phones on a daily basis by more than half.[9] Another 2015 study (n=12) found that disabling notifications left some test subjects feeling less stressed and more productive during working hours but anxious in their off-time, for fear of missing important information and violating expectations of others.[10]

With the new availability of usage data, we decided to examine the effect of a grayscale screen, which – before the advent of Digital Wellbeing and Screen Time features – had gained traction as a way to limit smartphone "addiction" after being endorsed by Tristan Harris, founder of the Center for Humane Technology.[11] Our research question thus evolved to: *does changing a grayscale screen cause less time spent on mobile devices?* We hypothesized that using devices in grayscale would indeed result in decreased usage, as Harris suggested.

This research question is well suited to an experimental approach, given that usage habits vary per individual, and that they are affected by various factors (e.g. age, gender, day of the week, lifestyle, etc.). The use of random assignment – and thus, a good covariate balance between treatment and control groups – would be instrumental in isolating the treatment effect and ultimately discovering whether or not a causal

[7] Chappell, B. "Large Shareholders Ask Apple To Help Wean Digital-Addicted Youths." National Public Radio, January 8, 2018. https://www.npr.org/sections/thetwo-way/2018/01/08/576541828/large-shareholders-ask-apple-to-help-wean-digital-addicted-youths.

[8] Gartenberg, C. "How do Apple's Screen Time and Google Digital Wellbeing stack up?" The Verge, June 5, 2018. https://www.theverge.com/2018/6/5/17426922/apple-digital-health-vs-google-wellbeing-time-well-spent-wwdc-2018

[9] Andrews S, Ellis DA, Shaw H, Piwek L (2015) Beyond Self-Report: Tools to Compare Estimated and Real-World Smartphone Use. PLoS ONE 10(10): e0139004. https://doi.org/10.1371/journal.pone.0139004.

[10] Pielot, M. and Rello, L. (2015) The Do Not Disturb Challenge: A Day Without Notifications. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (CHI EA '15). ACM, New York, NY, USA, 1761-1766. https://doi.org/10.1145/2702613.2732704.

[11] Harris, T. "Distracted in 2016? Reboot Your Phone with Mindfulness." TristanHarris.com, January 27, 2016. https://www.tristanharris.com/2016/01/distracted-in-2016-welcome-to-mindfulness-bootcamp-for-your-iphone.

relationship exists. Furthermore, a randomized experiment design would help examine our research question in a more longitudinal fashion and in the context of real-life device use, rather than within the unnatural experiment of a laboratory or the short-term measurement of a survey, which could measure attitudes but not be reflective of actual behavior.

III. Experiment Design

The experiment took place over the course of two weeks, from November 5th to 19th, 2018, with 21 participants. It was preceded by a three-day pilot with four different participants, from October 31st to November 2nd. Participants were split into two groups, R1 and R2, to designate assignment to treatment or control per day (Table 1). R1 participants were managed by Ayo and Joyce, while R2 participants were managed by Bo and Rohan.

Each subject was assigned both treatment and control on each day of the week, with half having treatment both at the beginning and end of the study (R1) while the other half had consecutive days in the middle (R2). The experiment design was intended to allow for both analysis of within subjects and across subjects comparisons. Screen time measurements occurred on Mondays, given that our data source, the Apple Screen Time app, only retains data for seven days.

Table 1. Experiment design.

	Week 1							**Week 2**							
	Mo	**Tu**	**We**	**Th**	**Fr**	**Sa**	**Su**	**Mo**	**Tu**	**We**	**Th**	**Fr**	**Sa**	**Su**	**Mo**
R1	🟦	X	X	X	O	O	O	▒	O	O	O	X	X	X	▒
R2	🟦	O	O	O	X	X	X	▒	X	X	X	O	O	O	▒

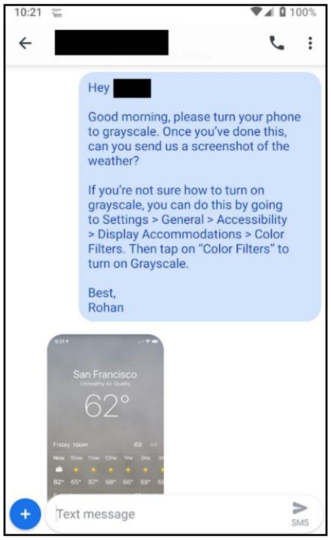
X	Treatment (Grayscale)	🟦	Pre-test measurement
O	Control (Default color mode)	▒	Measurement days

Sampling. Our sample was limited to iPhone users only, given that the Screen Time application is native to the iOS operating system. Though this imposes limits of generalizability, we omitted Android users from the study to avoid priming participants by asking them to download a screen tracking app. In our pilot study, however, we quickly encountered ethical concerns of not disclosing the nature of the study to participants beforehand, which is fairly apparent from the procedure. We thus opted to be transparent with our research subjects, adopting a position of informed consent.

Many participants were also fellow students in our graduate program: while their behavioral patterns may not generalize to non-student populations, their physical proximity during the course of the experiment enabled better means of accountability to complying with treatment.

Pre-test measurement. As a part of onboarding into the study, participants were asked to update their devices to iOS 12, the operating system that includes the Screen Time application, if they had not already. The majority of participants (81%) had at least one day of pre-experiment data, having already updated to iOS 12. When available, we included pre-experiment phone usage into our model; however, we recognize that individuals who may automatically update their device's operating systems may be more active users than those who do not.

Treatment. Subjects in treatment were asked to switch their phones into grayscale mode via text message in the morning. Upon first treatment assignment, participants were provided instructions on how to access this mode in their settings. Participants were then asked for a screenshot of the weather to ensure compliance. In our pilot study, we came to learn that even if the participant is in grayscale, the screenshot sent does not render in grayscale; however, we still retained responsiveness to the researcher as a proxy for compliance.

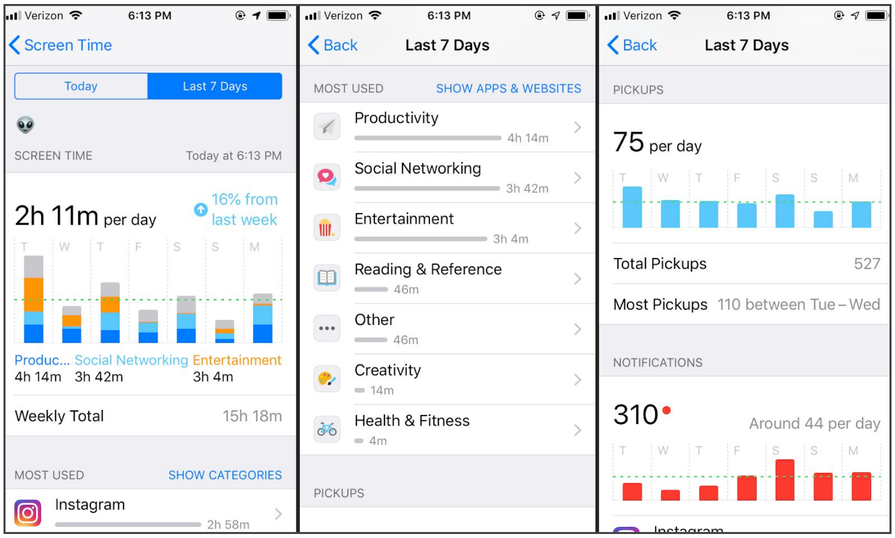

[redacted]

Hey [redacted]

Good morning, please turn your phone to grayscale. Once you've done this, can you send us a screenshot of the weather?

If you're not sure how to turn on grayscale, you can do this by going to Settings > General > Accessibility > Display Accommodations > Color Filters. Then tap on "Color Filters" to turn on Grayscale.

Best,
Rohan



SCREEN TIME Today at 6:13 PM

2h 11m per day — 16% from last week

Produc... 4h 14m Social Networking 3h 42m Entertainment 3h 4m

Weekly Total 15h 18m

MOST USED SHOW CATEGORIES

Instagram 2h 58m

MOST USED SHOW APPS & WEBSITES

Productivity	4h 14m
Social Networking	3h 42m
Entertainment	3h 4m
Reading & Reference	46m
Other	46m
Creativity	14m
Health & Fitness	4m

PICKUPS

PICKUPS

75 per day

Total Pickups 527

Most Pickups 110 between Tue – Wed

NOTIFICATIONS

310 Around 44 per day

Instagram

Figure 1. Sample text message during the first day of treatment (left)
Figure 2. Sample screenshots of data available within the iOS Screen Time app (right)

Outcome and covariate measurements. Data was manually pulled from the iOS Screen Time app, with measurements collected in person or over the phone, if meeting in person was not possible. While our main focus was screen time usage (in minutes per day), other available outcome units included pick-ups per day and notifications per day. Also available from the Screen Time app was most used categories, of which we collected the top three.

Based on data collected, we generated the covariate, “Most_Used_Cat_Social,” a dummy variable for subjects who had social networking as the most frequently used category both prior to the experiment and on the last day of the experiment. We hypothesize that the grayscale treatment would make content shared on social media applications less attention piquing, and thus decrease screen time for subjects who use social media heavily. As previously mentioned, we also hypothesized that those who had automatically updated may be heavier device users compared to participants who we needed to ask to update for the purpose of the experiment; as we result, we created a “Pre_Installed_Screen_Time” dummy variable as a proxy for tech savviness. Where available, pre-experiment measures for screen time, pickups, and notifications were additionally included in the analysis (noted by the suffix "_pm").

Furthermore, at the end of the two-week study, participants were asked to report their age, gender, commute method, daily commute time, and whether or not they traveled over Veteran’s Day weekend, which took place during the experimental timeframe. A correlation matrix of variables considered can be seen below, in Figure 3.

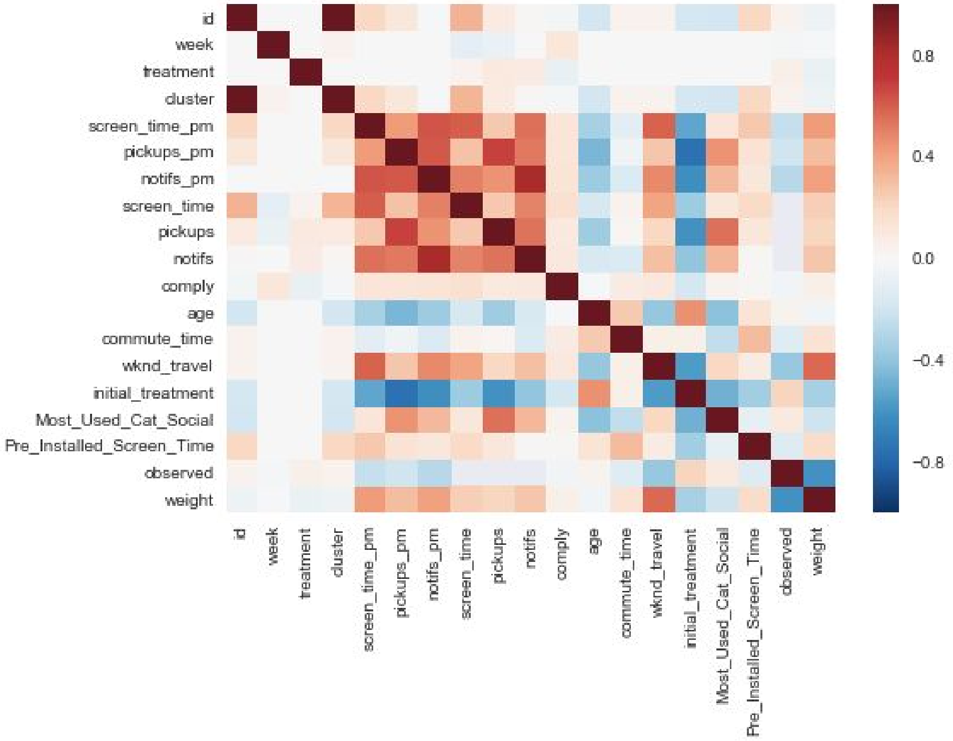
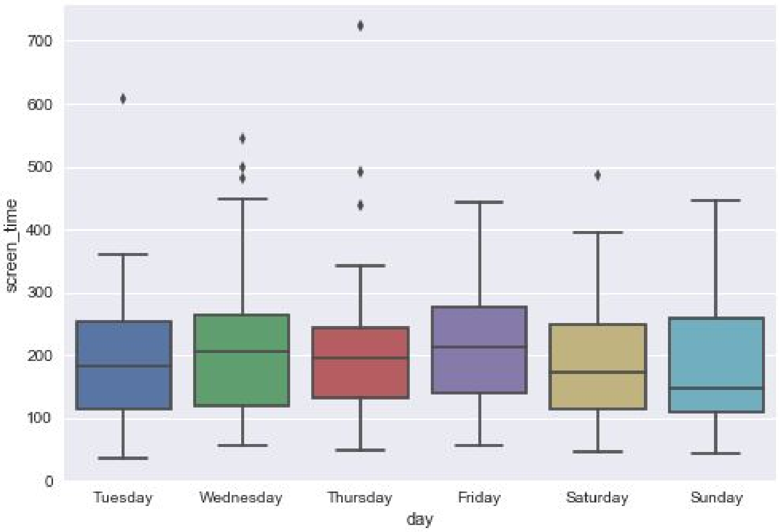
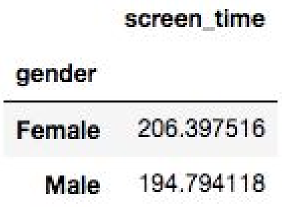


Figure 3. Correlation matrix of model variables.

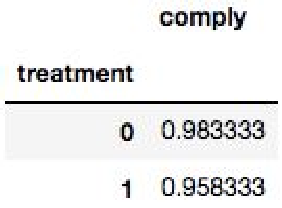
IV. Outcomes

Before the experiment, average daily measures were about 3.5 hours of screen time, 104 pick ups and 126 notifications. Pre-treatment measurements were available for 92 observations from 16 participants. When including data from the experimental timeframe, we found that screen time tended to be lower on weekends and that females tended to have higher screen time measures. Compliance was also generally equal among control and treatment groups.





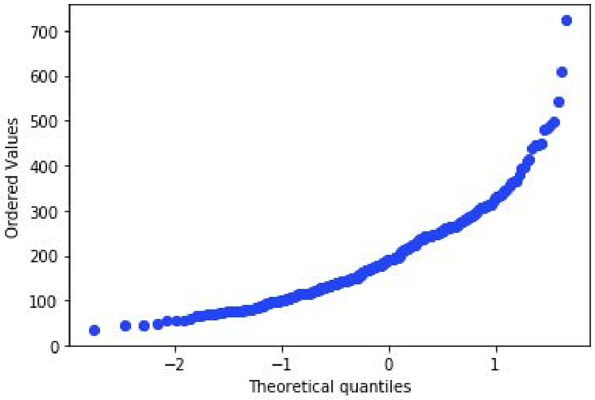
	screen_time
gender	
Female	206.397516
Male	194.794118

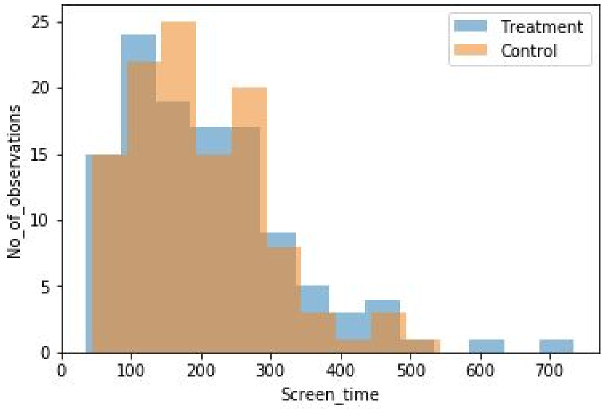


	comply
treatment	
0	0.983333
1	0.958333

Figure 4. Boxplot of screen time, by day of the week (left)
Figure 5. Comparison of screen time values, by gender (right, top)
Figure 6. Comparison of compliance rates, by treatment assignment (right, bottom)

Treatment-specific measures. The overall distribution of screen time during the experiment includes a right skew, though the peak does seem close to our average pretreatment screen time measure, 206 minutes. The distribution skew is demonstrated a non-linear Q-Q plot of screen time (Figure 7). Examining a histogram that splits treatment and control (Figure 8), we see that the outliers are in the treatment group. Upon further investigation, we found that both were from one individual, who explained that they had a sleep app that they use most days, which required their screen to be on. As a result, we dropped this subject’s data from our analysis to prevent a biased estimator of the average treatment effect.





(left) Figure 7. Q-Q plot of screen time.
(right) Figure 8. Histogram of screen time values, by treatment assignment.

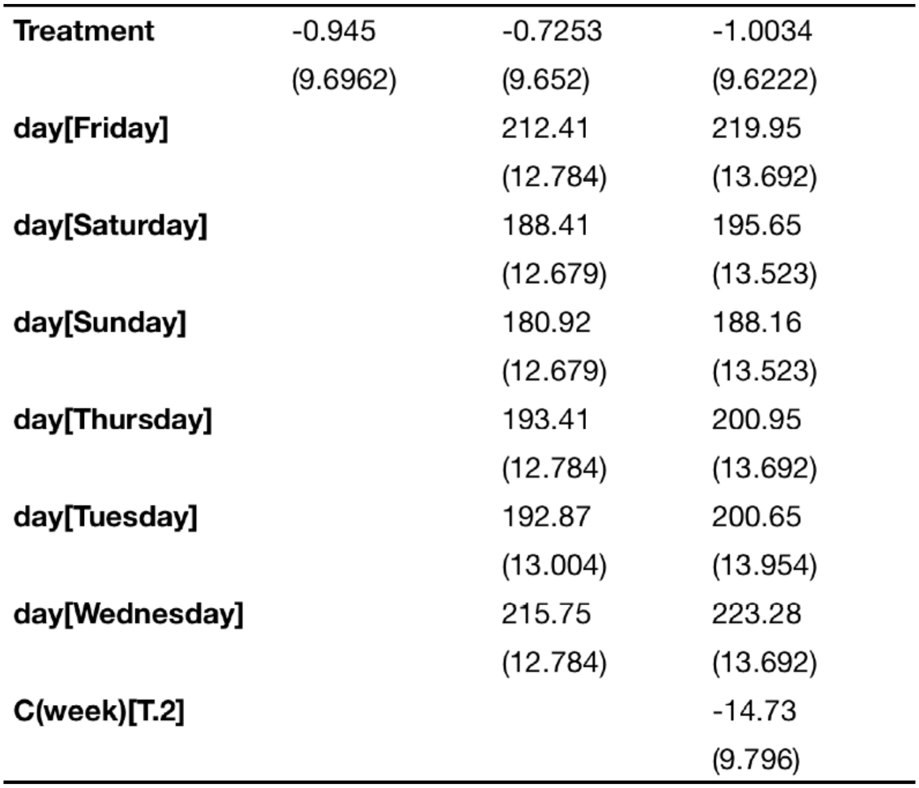
V. Analysis

Model selection. In order to analyze the impact of grayscale on screen time usage we ran a Panel OLS regression as well as clustered and non-clustered OLS regressions to measure the treatment effect. Panel OLS allows us to look at observations across time but within each specific individual. Since we can think about individuals as different in this model we can remove unobserved heterogeneity. Therefore, we can disregard that there are differences in average levels of screen time usage between individuals and assume that the differences are due to individual specific characteristics that don't change over time. Finally, we implemented clustered OLS regression (clusters are based on treatment assignment, meaning each individual contained 4 clusters given they experienced treatment twice and control twice) because we felt the potential outcomes would be similar within each cluster given that treatment or control occurred over a three-day period.

Within subjects treatment effects: fixed effects. In the panel OLS regression (Table 2) we included the variables 'day' and 'week' to control for effects the day of week and week in question may have on screen_time within each subject. After paneling our data, by multi-indexing with participant_id and date of observation, we applied an F-test t for the poolability across cross sections.

Results. The within-subject coefficient slope for the treatment effect is statistically insignificant, with a coefficient of -1.0034 minutes (9.622), p = 0.9171. The test statistic, F-test for Poolability: 14.264 and p-value: 0.0000, allow us to reject the null hypothesis of poolability that assumes homogeneous treatment effects within subjects. The R^2 overall (0.0197) is close to the R^2 within (0.0457) however, so we see that individual heterogeneity is low, indicating that our specification using a pooled regression could work well. Adding day and week variables reduce the SE in the treatment coefficient marginally from 9.69 minutes to 9.62 minutes, giving us a 95% CI of (-19.984, 17.976) minutes.

Table 2. Panel OLS Regression Results

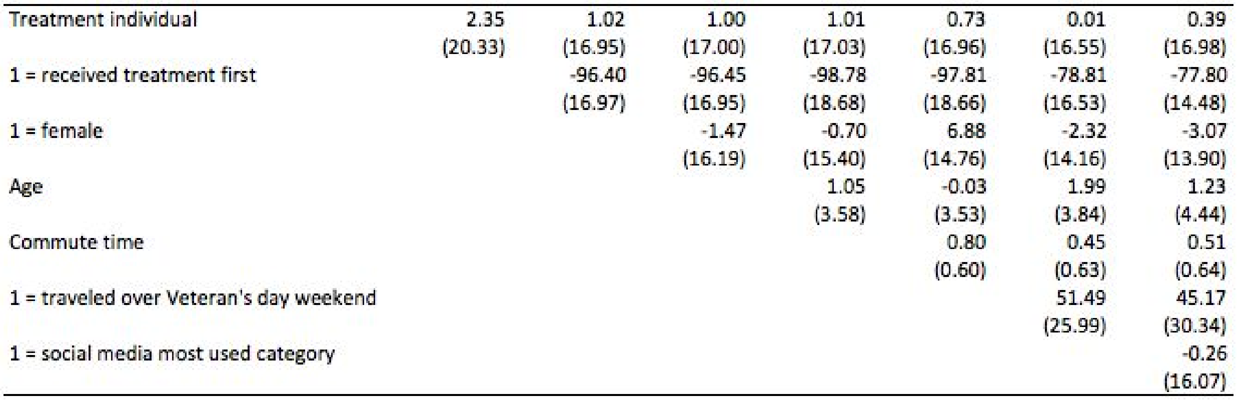


Treatment	-0.945	-0.7253	-1.0034
	(9.6962)	(9.652)	(9.6222)
day[Friday]		212.41	219.95
		(12.784)	(13.692)
day[Saturday]		188.41	195.65
		(12.679)	(13.523)
day[Sunday]		180.92	188.16
		(12.679)	(13.523)
day[Thursday]		193.41	200.95
		(12.784)	(13.692)
day[Tuesday]		192.87	200.65
		(13.004)	(13.954)
day[Wednesday]		215.75	223.28
		(12.784)	(13.692)
C(week)[T.2]			-14.73
			(9.796)

Between-subjects treatment effect: zero. While our hypothesis predicted that grayscale would negatively impact screen time, our findings actually suggest no treatment effect after adding in covariates and clustering at the individual level. Adding pre-experiment measurements reduces our number of observation but improves the model's AIC by about 629. Other covariates such as wknd_travel and Most_Used_Cat_Social reduce the models AIC by 186 and 51 respectively.

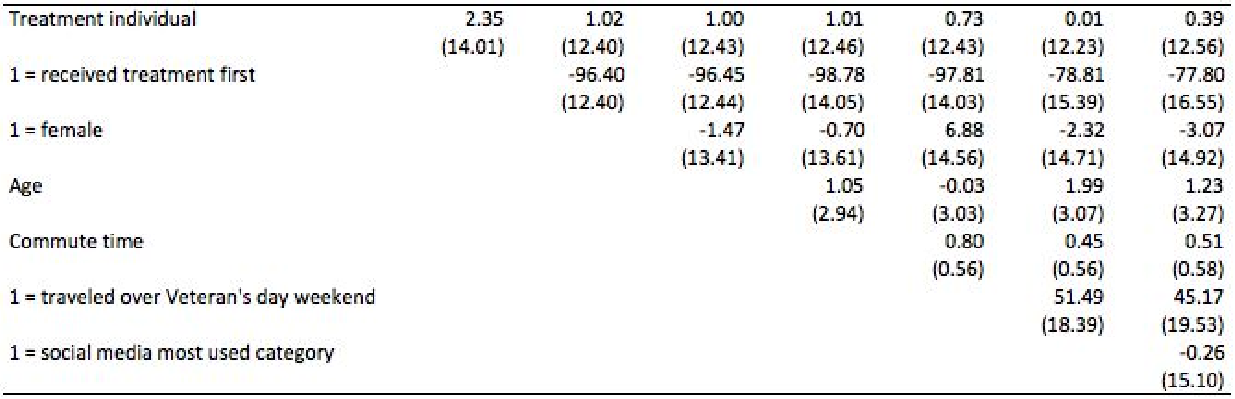
*Results.*The clustered OLS regression (Table 3) shows a positive treatment effect of grayscale on screen time usage of 2.35 minutes, but the effect is not significant, given the standards error of 20.33 minutes. As we continue to add covariates, the treatment effect drops even closer to zero, although the insignificance of the results remains the same. Without individual-level clustering (Table 4) we can see that the standard errors decrease while the coefficients remain the same. This is an expected result because it does not incorporate the higher variance from the similarity within clusters that exceeds the similarity between clusters.

Table 3. OLS regression of screen time (minutes), clustering at the individual level



Treatment individual	2.35	1.02	1.00	1.01	0.73	0.01	0.39
	(20.33)	(16.95)	(17.00)	(17.03)	(16.96)	(16.55)	(16.98)
1 = received treatment first		-96.40	-96.45	-98.78	-97.81	-78.81	-77.80
		(16.97)	(16.95)	(18.68)	(18.66)	(16.53)	(14.48)
1 = female			-1.47	-0.70	6.88	-2.32	-3.07
			(16.19)	(15.40)	(14.76)	(14.16)	(13.90)
Age				1.05	-0.03	1.99	1.23
				(3.58)	(3.53)	(3.84)	(4.44)
Commute time					0.80	0.45	0.51
					(0.60)	(0.63)	(0.64)
1 = traveled over Veteran's day weekend						51.49	45.17
						(25.99)	(30.34)
1 = social media most used category							-0.26
							(16.07)

Table 4. OLS regression of screen time (minutes), without individual-level clustering



Treatment individual	2.35	1.02	1.00	1.01	0.73	0.01	0.39
	(14.01)	(12.40)	(12.43)	(12.46)	(12.43)	(12.23)	(12.56)
1 = received treatment first		-96.40	-96.45	-98.78	-97.81	-78.81	-77.80
		(12.40)	(12.44)	(14.05)	(14.03)	(15.39)	(16.55)
1 = female			-1.47	-0.70	6.88	-2.32	-3.07
			(13.41)	(13.61)	(14.56)	(14.71)	(14.92)
Age				1.05	-0.03	1.99	1.23
				(2.94)	(3.03)	(3.07)	(3.27)
Commute time					0.80	0.45	0.51
					(0.56)	(0.56)	(0.58)
1 = traveled over Veteran's day weekend						51.49	45.17
						(18.39)	(19.53)
1 = social media most used category							-0.26
							(15.10)

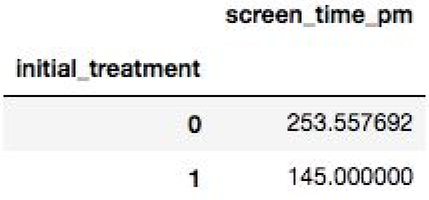
Attrition. There were 11 instances of attrition, mostly between two subjects. In the control condition, 94.2% of outcomes were measured, and 96.67% of outcomes were measured in the treatment condition. Conducting a chi-squared t-test to test the null hypothesis of no difference between these proportions, we find a p-value of 0.3603, indicating that the difference between proportions of observed outcome values both conditions is statistically significant.

We realize that the unbiased estimate of treatment effect for “always reporters:” subjects who report their outcomes regardless of group assignment, may not be the same as the overall subject pool ATE. As a result, we implemented inverse probability weighting to tackle this problem and recover the true ATE of the subject pool. The results after implementing weighted estimates show that our estimates are similar to those in our original specification.

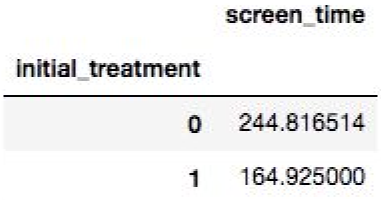
VI. Limitations and Future Work

Even among our relatively homogenous population (similar lifestyle, age, etc.), we found quite a bit of noise in phone usage – we suspect this may be due to the timing of our experiment period, which may have been affected by Veteran's Day and Thanksgiving holidays. Forest fires during this period also led to recommendations to stay inside, which may have affected mobile phone use. An experiment of longer duration and with more subjects would be key to truly understand effects reliably.

In terms of limitations that were within our control, one key improvement that we would make in the future is to have more robust randomization engineering. A more robust approach would have involved analyzing pre-experimental measurements of the outcome variable, to achieve a more even balance of this measure between subjects groups who start the experiment in the treatment condition (R1, or initial_treatment = 1), and subjects groups who started the experiment in the control condition (R2, or initial_treatment = 0). Examining the screen time values both before and during the experiment (Figures 9 and 10, respectively), we see that the groups have substantially different values and therefore device usage behaviors. If we had collected additional covariates (like age, gender, etc.) before and not after the experiment, we could have also improved the balance of baseline covariates on average through better randomization engineering.



	screen_time_pm
initial_treatment	
0	253.557692
1	145.000000



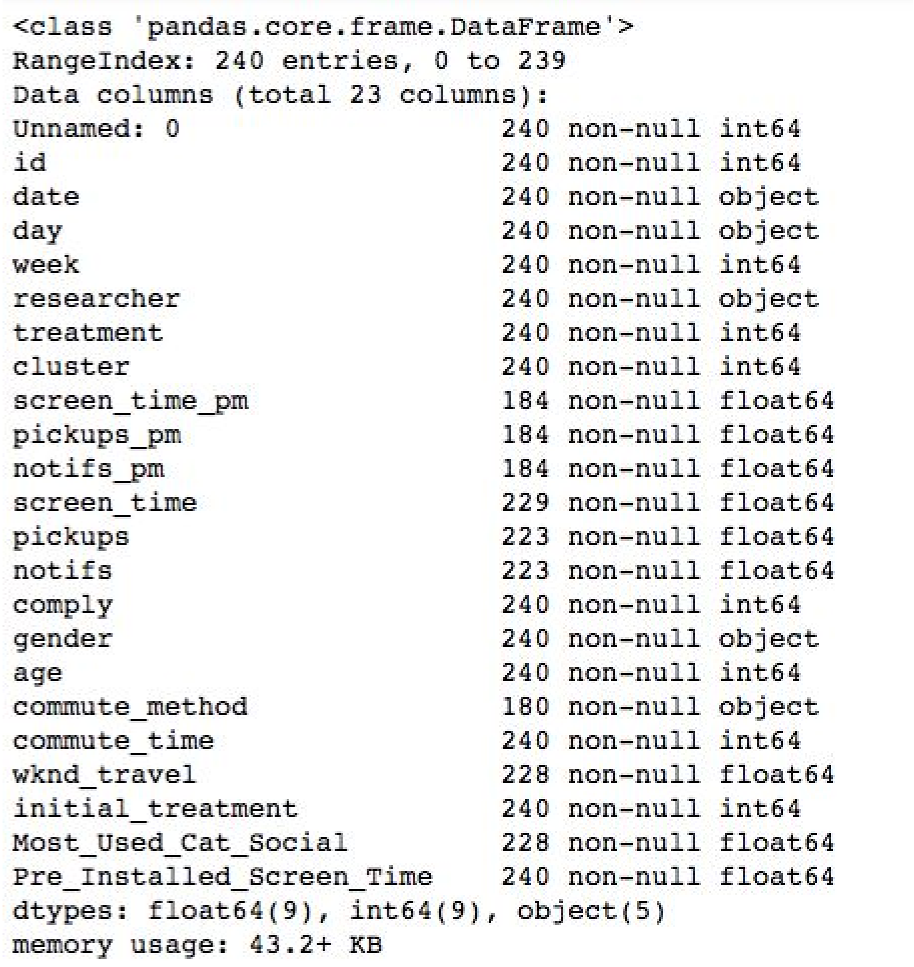
	screen_time
initial_treatment	
0	244.816514
1	164.925000

Figure 9. Comparison of pre-experiment screen time, by treatment group (left)
Figure 10. Comparison of in-experiment screen time, by treatment group. (right)

Another area of improvement could have been in more consistent research protocol: some measurements were collected over the phone or in person, and treatment-assignment text message were not sent at the same time for each day of the experiment. Each researcher also managed the same group of participants throughout the study, so any pre-existing relationship between the researcher and the participant may have also affected behavior in the experiment; a more refined approach may have anonymized the researcher to prevent this effect. Given the confounding nature of the results, future research may benefit supplemental collection of attitudinal survey data, to understand how participants feel about the treatment condition, even it may not affect their behavior.

Appendix

I. Type of features and number of observations



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 240 entries, 0 to 239
Data columns (total 23 columns):
Unnamed: 0                     240 non-null int64
id                             240 non-null int64
date                           240 non-null object
day                            240 non-null object
week                           240 non-null int64
researcher                     240 non-null object
treatment                      240 non-null int64
cluster                        240 non-null int64
screen_time_pm                 184 non-null float64
pickups_pm                     184 non-null float64
notifs_pm                      184 non-null float64
screen_time                    229 non-null float64
pickups                        223 non-null float64
notifs                         223 non-null float64
comply                         240 non-null int64
gender                         240 non-null object
age                            240 non-null int64
commute_method                 180 non-null object
commute_time                   240 non-null int64
wknd_travel                    228 non-null float64
initial_treatment              240 non-null int64
Most_Used_Cat_Social           228 non-null float64
Pre_Installed_Screen_Time      240 non-null float64
dtypes: float64(9), int64(9), object(5)
memory usage: 43.2+ KB
```

II. Feature definitions

id: individual id
date: date of measurement
day: day of week (e.g. Monday, Tuesday, etc.)
week: 1 = first week; 2 = second week
researcher: researcher in charge of communicating and measuring subject (e.g. Joyce)
treatment: 1 = treatment; 0 = control
cluster: individual cluster assignment
screen_time: screen time in minutes
screen_time_pm: premeasurement screentime in minutes
pickups_pm: premeasurement pickups
notifs_pm: premeasurement notifications

commute_time: time commuting in minutes
wknd_travel: 1 if traveled over Veterans day weekend
inital_treatment: 1 if person received treatment on first day of the experiment (e.g. Joyce's and Ayo's cohort)
Most_Used_Cat_Social: 1 = if most used category was social media
Pre_Installed_Screen_Time: 1 = screen time app pre-installed