Online Appendix

A. Data

Table §A.1 describes the survey variables that were constructed from the survey responses.

Table §A.1 Survey Description Table

Name	Description	Number of Survey Questions, Scale	Mean (SD)
Self-Efficacy for Exercise	Scale measuring whether an individual is effective at engaging in regular exercise (Bandura, 2006)	18 questions. Average of 0-100%	59.19 (19.72)
Self-Regulation for Exercise	Scale measuring whether an individual is effective at regulating their exercise behavior (Deci and Ryan, 2008)	16 questions. Weighted average of 7-point scale	8.78 (4.06)
Self-Esteem	Scale measuring whether an individual has confidence and satisfaction in oneself (Rosenberg, 1965)	10 questions. Average of 5-point Likert scale	3.33 (.69)
Trust	Scale measuring whether an individual is willing to rely on the character, ability, strength, or truth of someone else (Hetherington, 1998)	6 questions. Average of 5-point Likert Scale	3.18 (.53)
Depression	Scale measuring whether an individual exhibits signs of depression (Hann et al., 1999; Radloff, 1977)	20 questions. Sum of 4-point scale (Rarely or None of the time – Most or all of the time)	25 (9.10)
Anxiety	Scale measuring whether an individual exhibits anxiety (apprehensive uneasiness or nervousness) (Spielberger et al., 1983)	20 questions. Sum of 4-point scale (Almost never... Almost always)	39.92 (10.36)
Mobil Messenger App Usage	The number of smartphone apps used to message with other people (Top responses: iMessage, Facebook Messenger, WhatsApp, GroupMe)	N/A	2.64 (1.17)
Social Media Usage	The number of social media applications used to post and view text, pictures, and video (Top responses: Facebook, Instagram, Twitter, Snapchat)	N/A	3.05 (1.37)
Discuss Health	Frequency at which individuals discussed health with others	Single Survey Questions, 5-point scale, Not at all (0), Less than 1-2 times a month, 1-2 times a month, 1-2 times a week, Three times a week or more (4)	2.19(1.15)
Discuss Politics	Frequency at which individuals discussed politics with others		2.58 (1.13)
Discuss Religion	Frequency at which individuals discussed religion with others		2.24 (1.16)
Exercise Alone	Frequency at which individuals exercise alone		2.22 (1.26)
Exercise with Others	Frequency at which individuals exercise with others		2.32 (1.38)
Visit Friends	Amount of time individuals spend daily sitting/hanging out/talking with friends	Single survey questions, 8-point scale, Do not do (1) – More than 4 hours (8)	4.88 (1.37)
Gender	The individual's gender	One survey question, Two option scale, Male (1), Female (0)	0.48 (0.5)
Parent Income	Measure of the focal user's parents' income	One survey question, 8-point scale, (1) <$25,000, .., (8) $25,000-$250,000 or more	5.55 (2.18)
Catholic	Whether the individual is of the Catholic faith	One survey question, 2 options, Catholic (1), Non-Catholic (0)	0.74 (0.44)
Body Mass Index	Calculated based on answers to two survey questions about the individual's height and weight	Two survey questions (weight and height), free text answer	23.11 (3.46)
U.S. Citizen	Whether the individual is a U.S. Citizen	One survey question, 3 option scale (U.S. Citizen, Permanent Resident, Other)	0.92 (0.27)
Native English Speaker	Whether English is their native language	One survey question, 2 option scale (Yes, No)	0.87 (0.33)

B. Main Effects with Survey Controls

In this subsection, we recalculate the main effects presented in Table 2 by adding time-varying control variables extracted from survey data.

$$Steps_{it} = \beta_0 + \beta_1(LeaderBoard_{it}) + \alpha(\mathbf{X_{it}}) + \theta_\mathbf{i} + \lambda_\mathbf{t} + \gamma_\mathbf{i} \times \mathbf{t} + \phi_\mathbf{i} \times \mathbf{t^2} + \epsilon_\mathbf{it} \quad \text{(§B.1)}$$

While our base model does not control for time-varying individual characteristics, we also estimated the enhanced specification §B.1 that includes a vector of controls, X_{it}, which are time-varying characteristics extracted from the individuals' surveys. These controls include measures of an individual's ability to self-regulate and their efficacy in maintaining a regular exercise routine, their mental health states (e.g., depression, self-esteem, anxiety), their sociability (frequency of discussing personal topics with friends), and their technology use (e.g., number of social media applications used). The various controls may help us account for time-varying heterogeneity in an individuals' motivation to exercise, to adopt a leaderboard, or both.

Table §B.2 Survey Data and Negative Controls

	(1) steps b/se	(2) steps b/se	(3) steps b/se	(4) steps b/se	(5) steps b/se
Leaderboard	515.964**	691.827*			
	(213.233)	(351.630)			
Active LB					333.841**
					(146.505)
Dormant LB			-183.837	-345.776	
			(1123.769)	(1137.145)	
Psych and Tech Use Controls (Survey)	Yes	Yes	No	No	No
Social and Exercise Controls (Survey)	No	Yes	No	No	No
IPTW	No	No	No	Yes	No
Individual Fixed Effects	Yes	Yes	Yes	Yes	Yes
Week Fixed Effects	Yes	Yes	Yes	Yes	Yes
Individual Linear Trends	Yes	Yes	Yes	Yes	Yes
Individual Quadratic Trends	Yes	Yes	Yes	Yes	Yes
Observations	19,689	15,821	14,921	14,700	27,758
Individuals	325	298	516	501	516
Adjusted R-Squared	0.33	0.30	0.34	0.38	0.33
VCE	Robust	Robust	Robust	Robust	Robust

* $p < 0.10$, ** $p < 0.05$.

Table §B.2 columns 1 and 2 present this recalculation. For column 1, the model includes measures drawn from various psychological scales (self-efficacy, self-regulation, etc.) and measures of technology use. For column 2, the estimate also includes controls for social interaction such as discussions of politics, religion, and health as well as the frequency at which the individual exercises alone vs. with others. We added these two sets of controls sequentially as missing survey data exclude some participants from our estimation when the controls are included. Also, participants had to have completed at least two waves of the survey to have variance over time in these measures. In both columns 1 and 2, we continue to find significant and positive effects of leaderboards on daily steps walked. Estimates from models presented in Tables 2 and §B.2 suggest that students' leaderboard adoption led to an average daily increase of 338–691 steps.

C. Inverse Probability of Treatment Weighting (IPTW)

IPTW methods require the estimation of the propensity score $\hat{p}(X_i)$, where X_i is the vector of pre-treatment covariates for the individual i (Hernan and Robins, 2018; Abadie and Cattaneo, 2018). The outcome of the treated individuals is weighted by $1/\hat{p}(X_i)$, whereas the outcome of the untreated individuals is weighted by $1/(1-\hat{p}(X_i))$. The goal of weighting observations is to achieve covariate balance across the treated and untreated groups. Thus, it is crucial that the propensity score estimation method is optimized for covariate balance across the treated and untreated groups. To estimate the propensity of leaderboard adoption, we use the Toolkit for Weighting and Analysis of Nonequivalent Groups (TWANG), which implements a generalized boosted regression model (GBM). GBM is an ensemble machine learning method, which uses boosting to assign higher weights to misclassified observations as it iterates through multiple functional specifications (please see McCaffrey et al. (2013) for details). The propensity score estimated by TWANG optimizes covariate balance across those users who adopt a leaderboard and those who do not. Table §C.3 shows substantive improvement in the covariate balance post-weighting such that the absolute standardized mean

difference (SMD) for observed covariates is less than or equal to 0.2 (which is better than the accepted threshold of 0.25).

Table §C.4 and Figure §C.I provide measures on the importance of covariates in predicting leaderboard adoption. We now briefly describe the metrics presented in Table §C.4 and Figure §C.I. Gain is the improvement in accuracy that can be attributed to a particular covariate, cover is the related number of observations related to that covariate, and frequency is the percentage of times that covariate occurs in all of the trees (gradient boosted method is a tree-based ensemble algorithm (Lesmeister, 2019)). Shapley values are a theoretical construct from cooperative game theory, which allows the fair distribution of payoffs to players. SHAP (SHapley Additive exPlanation) values, an implementation of Shapley values for complex machine learning models, provide measure of covariate importance (Lundberg and Lee, 2017). The model explanations or covariate importance ranking may differ across the measures (Hall and Gill, 2019) but they are roughly similar in our analysis . For instance, Body Mass Index (BMI) is ranked second by all four metrics, viz., gain, cover, frequency, and mean absolute SHAP value. The individual-level breakdown of BMI's impact on leaderboard adoption shows significant variation. For instance, the breakdown in Figure §C.I suggests that the BMI's impact on leaderboard adoption is negative when the value of this covariate is high. Another example is the covariate *trust*, in which the SHAP values suggest that lower values of trust tend to have a negative impact on leaderboard adoption. Using SHAP values, the top five covariates are anxiety, BMI, exercise with others, social (app) use, and discuss politics.

Using propensity score to balance pre-treatment covariates is one method to make the DID's common trends assumption more plausible (Xu, 2017). Abadie (2005) proposed an estimator which combines DID with direct weighting on the propensity score, and provided inspiration for our analysis in this paper.

D. Falsification Tests Using Negative Controls

We also construct a negative control treatment (NCT)[23] to further probe the possibility of bias in our main results due to unmeasured time-varying confounders. Our hypothesized mechanisms for leaderboards require the presence of other active users, the "active ingredient" of the leaderboard treatment. Without other active users, a leaderboard is essentially neutralized under our hypothesized mechanisms.

Leaderboards become inactive when non-focal users of a leaderboard either stop using their Fitbit devices altogether or stop uploading their data to the Fitbit platform. Figure §D.II shows such a leaderboard in which Fitbit suppresses the inactive users as these inactive users have no recent activity data and cannot provide competition or reference points. However, an inactive or dormant leaderboard would be subject to the same sources of bias as an active leaderboard. We define a negative control treatment, *Dormant LB*, as the focal user's leaderboard with no other active users. Thus, we obtain two sets of control observations: first, in which no leaderboard is present, and second, in which a dormant leaderboard is present. We can then contrast these two sets of control observations to test the counter-hypothesis of a time-varying confounder driving the observed results. In particular, if a change in an individual's motivation for physical activity

[23] An NCT is a treatment in which the "active ingredient" has been inactivated. Thus, the NCT and the actual treatment will share the same unmeasured confounders, if any. Consequently, a *null* NCT effect makes the actual treatment effect more plausible.

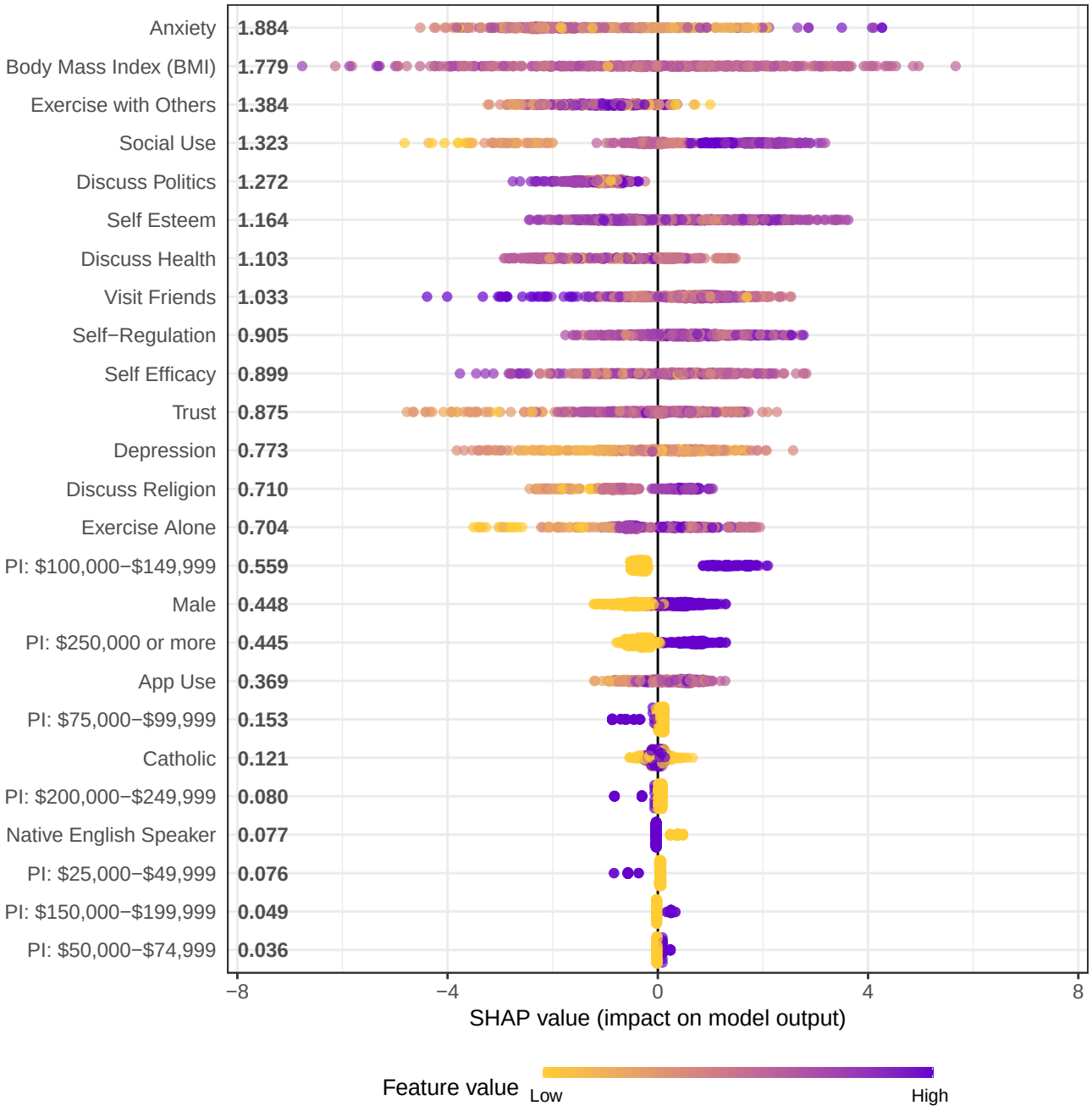
Table §C.3 Covariate Balance Before and After Weighting

	Treated mean	Control mean (weighted)	SMD (weighted)	Control mean (unweighted)	SMD (unweighted)
Self Efficacy	59.74	57.85	0.11	58.76	0.06
Self-Regulation for Exercise, Overall	8.77	8.20	0.16	8.56	0.06
Self Esteem	3.42	3.43	-0.02	3.32	0.16
Trust	3.19	3.20	-0.02	3.11	0.16
Depression	1.23	1.20	0.08	1.25	-0.08
Anxiety	1.93	1.92	0.04	2.01	-0.17
App Use	2.79	2.68	0.10	2.47	0.29
Social Use	3.44	3.21	0.20	2.87	0.52
Discuss Health	2.81	2.73	0.08	2.58	0.22
Discuss Politics	2.49	2.54	-0.05	2.46	0.02
Discuss Religion	2.31	2.14	0.16	2.08	0.21
Exercise Alone	2.57	2.49	0.08	2.29	0.28
Exercise with Others	2.42	2.44	-0.02	2.29	0.12
Visit Friends	5.02	5.12	-0.07	5.03	-0.00
Female	0.51	0.51	0.00	0.52	-0.01
Male	0.49	0.49	0.00	0.48	0.02
Body Mass Index (BMI)	22.85	23.01	-0.05	22.93	-0.02
Parent Income < $25,000	0.05	0.02	0.14	0.05	0.00
Parent Income $25,000–$49,999	0.05	0.07	-0.06	0.07	-0.07
Parent Income $50,000–$74,999	0.06	0.05	0.04	0.08	-0.07
Parent Income $75,000–$99,999	0.10	0.12	-0.07	0.08	0.06
Parent Income $100,000–$149,999	0.21	0.17	0.10	0.18	0.08
Parent Income $150,000–$199,999	0.11	0.09	0.08	0.11	0.02
Parent Income $200,000–$249,999	0.08	0.12	-0.17	0.10	-0.10
Parent Income $250,000 or more	0.30	0.33	-0.06	0.27	0.07
Not Catholic	0.24	0.31	-0.16	0.30	-0.12
Catholic	0.75	0.69	0.15	0.70	0.11
Not US Citizen	0.08	0.08	-0.01	0.10	-0.08
US Citizen	0.92	0.92	0.02	0.89	0.11
Non-native English	0.10	0.08	0.07	0.15	-0.14
Native English Speaker	0.90	0.92	-0.06	0.84	0.17

SMD abbreviates standardized mean difference. Number of observations is 501.

Table §C.4 Influence of Covariates on Leaderboard Adoption

	Covariate	Gain	Cover	Frequency	Mean Abs SHAP
1	Discuss Politics	0.2202	0.0130	0.0052	1.2715
2	Body Mass Index (BMI)	0.1381	0.1551	0.1248	1.7793
3	Trust	0.0927	0.0403	0.0185	0.8753
4	Self Esteem	0.0660	0.3974	0.5973	1.1641
5	Anxiety	0.0628	0.0553	0.0366	1.8843
6	Social Use	0.0564	0.0275	0.0099	1.3234
7	Self-Regulation for Exercise, Overall	0.0510	0.0480	0.0400	0.9049
8	Self Efficacy	0.0478	0.0536	0.0381	0.8991
9	Depression	0.0463	0.0471	0.0293	0.7735
10	Visit Friends	0.0451	0.0326	0.0226	1.0328
11	Discuss Health	0.0341	0.0234	0.0121	1.1033
12	Exercise with Others	0.0312	0.0265	0.0118	1.3840
13	Exercise Alone	0.0311	0.0240	0.0108	0.7039
14	App Use	0.0198	0.0092	0.0050	0.3693
15	Parent Income $100,000–$149,999	0.0148	0.0045	0.0020	0.5591
16	Catholic	0.0107	0.0036	0.0016	0.1210
17	Discuss Religion	0.0104	0.0090	0.0060	0.7105
18	Male	0.0068	0.0129	0.0196	0.4476
19	Parent Income $250,000 or more	0.0049	0.0070	0.0067	0.4453
20	Parent Income $75,000–$99,999	0.0044	0.0035	0.0006	0.1528
21	Parent Income $200,000–$249,999	0.0036	0.0019	0.0002	0.0798
22	Parent Income $25,000–$49,999	0.0009	0.0021	0.0005	0.0755
23	Native English Speaker	0.0005	0.0013	0.0003	0.0773
24	Parent Income $50,000–$74,999	0.0003	0.0004	0.0002	0.0355
25	Parent Income $150,000–$199,999	0.0002	0.0008	0.0003	0.0488



To conserve space, we abbreviate Parent Income as "PI:," and "Self-Regulation for Exercise, Overall" by "Self-Regulation."

Figure §C.I SHAP Values for the Effect of Covariates on Leaderboard Adoption

is driving their leaderboard adoption as well as increased physical activity, we should observe an effect for dormant leaderboards similar to the effect in our main analysis.

Table §B.2, column 3 shows null result for the negative control treatment *Dormant LB*, which we estimate by dropping the observations that have active leaderboards. Table §B.2, column 4 repeats the analysis with an IPT weighted sample, and we again estimate a null effect. Thus, the null effect with a negative control treatment supports our main analysis. As a further check, Table §B.2, column 5 provides the results in which

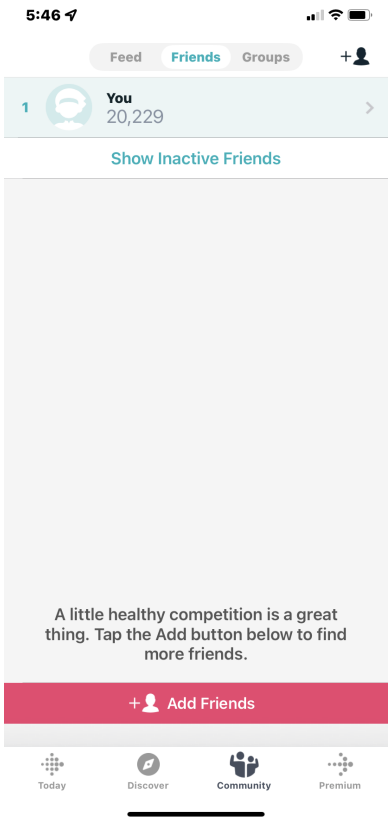


Figure §D.II A Leaderboard with No Active Users

we consider only participants on active leaderboards to be treated, and those with no leaderboard or dormant leaderboards to be untreated. Reassuringly, we find that active leaderboards lead to an increase of ≈ 334 steps for participants, an effect similar to the one observed in our main analysis.[24]

E. Extended Analysis of Fitbit Compliance

Section 5.3 presents summary results to address two potential concerns related to Fitbit compliance. Here, we provide details for the aforementioned summary results.

E.1. Null Effect of Leaderboards on Compliance

The first concern is the possibility that rather than increasing steps, leaderboard adoption increases compliance, which may lead us to observe higher step count purely because of better measurement. To address this concern, we explore if the leaderboard adoption has an increasing effect on daily compliance, i.e., whether daily compliance percentage increases substantially for leaderboard adopters in the post-adoption period. We estimate this potential effect using a model similar to equation (1) but with daily percentage compliance as the dependent variable. We estimate this model for a number of samples—the entire sample as well as

[24] In Table §B.2, column 5, the number of observations is 27,758 as we do not drop any observations. The estimated effect has slightly higher magnitude and better precision if we drop the dormant leaderboard observations.

Table §E.5 Leaderboard Null Effect on Compliance with Fitbit Use

	(1) % Comp b/se	(2) % Comp b/se	(3) % Comp b/se	(4) % Comp b/se	(5) % Comp b/se	(6) % Comp b/se	(7) % Comp b/se	(8) % Comp b/se	(9) % Comp b/se
Leaderboard	2.01	0.35	0.49	-0.05	-0.78	-0.77	-1.97	0.16	-0.25
	(1.57)	(1.54)	(1.59)	(1.31)	(1.29)	(1.37)	(1.63)	(1.37)	(0.79)
Mean Compliance	All	>60pct	>65pct	>70pct	>75pct	>80pct	>85pct	>90pct	>95pct
Individual Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Week Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual Linear Trends	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	236,336	133,595	120,676	105,756	91,329	68,398	44,071	22,011	5,344
Individuals	516	278	246	210	177	133	84	43	8
Adjusted R-Squared	0.17	0.13	0.11	0.11	0.11	0.08	0.07	0.04	0.01
VCE	Robust	Robust	Robust	Robust	Robust	Robust	Robust	Robust	Robust

* $p < 0.10$, ** $p < 0.05$.

for sub-samples at various mean compliance levels (ranging from 60 percent to 95 percent).[25] Table §E.5 presents the results of this analysis using daily percentage compliance as the dependent variable. The various columns in the table use samples that differ in the mean compliance levels of the participants. All of these estimates have low magnitude and are not statistically significant, suggesting that leaderboards do not have an effect on participants' daily percent compliance.

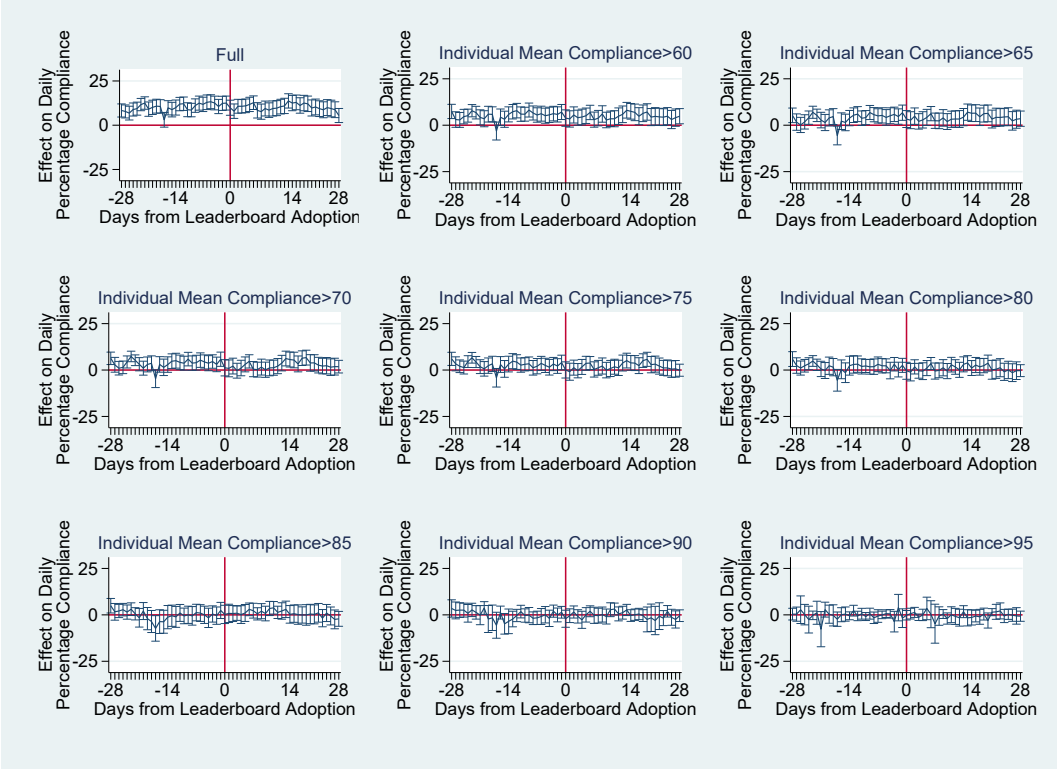


Figure §E.III Effect on Daily Percentage Compliance by Days from Leaderboard Adoption

[25] Given that these Fitbit devices need to be taken off for a variety of reasons, e.g., battery charging, risk of water exposure, etc., it is not likely that any individual will wear their Fitbit for the entire study period. In fact, even the number of participants with 85% mean compliance is relatively small—please see Table §E.5, column 7.

Finally, we probe the compliance issue by estimating a leads-lags model of leaderboard adoption, similar to equation (2), with daily comply percentage as the dependent variable, and the time measured in days (rather than weeks). If we observed more data for an individual, we collapsed it into extreme periods. We set this aggregated extreme pre-treatment period as the baseline period and withhold it from the specification to avoid the "dummy variable trap." Figure §E.III shows these charts for corresponding columns in Table §E.5. Given the null results (i.e., no effect on compliance) reported in Table §E.5, a sharp increase in the lags and leads plot at or after the time of adoption may still be a cause for concern. A visual examination of these charts allays this concern as no sharp shifts are discernible at or after adoption. One aspect of the chart for the entire sample needs further explanation—while the overall effect of leaderboard on compliance is null, the chart for the full sample suggests that adopter compliance may be higher than non-adopter compliance. However, this difference in levels of compliance in itself is not a cause for concern for two reasons. First, the compliance effect is null and there is no sharp increase in compliance at or after adoption which could contribute to increased measurement of steps. Second, as long as the common trend assumption (CTA) for the main outcome, i.e., *steps*, is plausible, compliance level difference is not likely affecting the estimation of our main effects. If compliance level differences were indeed making a difference, i.e., non-adopters were not trending similar to adopters, the CTA diagnostics for steps would have indicated it. We have provided diagnostics for CTA for *steps* in other sections, which all pointed to the plausiblity of the CTA assumption.

E.2. Are Leaderboard Effects Discernible at Higher Compliance Levels?

Table §E.6 Leaderboard Effect on Steps by Compliance

	(1) steps b/se	(2) steps b/se	(3) steps b/se	(4) steps b/se	(5) steps b/se	(6) steps b/se
Leaderboard	590.48**	552.63**	461.12*	478.84**	407.74	597.93*
	(223.05)	(232.06)	(238.75)	(235.92)	(250.47)	(315.68)
Mean Compliance	>60pct	>65pct	>70pct	>75pct	>80pct	NA
Weekly Pct Compliance	NA	NA	NA	NA	NA	>95pct
Individual Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Week Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Individual Linear Trends	Yes	Yes	Yes	Yes	Yes	Yes
Individual Quadratic Trends	Yes	Yes	Yes	Yes	Yes	Yes
Observations	18,183	16,570	14,641	12,715	9,646	6,479
Individuals	278	246	210	177	133	395
Adjusted R-Squared	0.37	0.36	0.37	0.38	0.38	0.40
VCE	Robust	Robust	Robust	Robust	Robust	Robust

* $p < 0.10$, ** $p < 0.05$.

In the main article, we posit that our empirical analysis would be more convincing if the leaderboard's effect on participants' activity was clearly discernible for participants with high compliance levels. We stated that the impact of leaderboards for participants with high levels of compliance ranged from 408 to 598 steps and referred to this appendix for details. Table §E.6 presents the results of this analysis.[26] Columns 1–5

[26] To facilitate comparison with the main results, we present analysis with weekly data. The results are similar with daily data.

present results for sub-samples selected on the basis of mean compliance ranging from 60% to 80%. The number of individuals in the sub-samples at even higher levels of mean compliance is very small (please see Table §E.5, columns 7–9). However, we also estimated the effect using only weeks when individual compliance was greater than 95%. The leaderboard effect at higher levels of compliance is estimated to have even larger effect sizes, thus supporting the claims from our main results.

F. Sample Attrition

To examine concerns related to sample attrition, we first analyzed whether leaderboard adoption was related to attrition from the sample. Specifically, we created a longitudinal measure of attrition that was coded as 1 when a user stopped reporting Fitbit data (i.e., this variable was set to 1 after the last day that individuals reported Fitbit data and was always 0 if they kept reporting data until the end of the observation period). Utilizing this dependent variable, we estimate a model similar to our main model but focusing on attrition for low performers in the sample. We find that there is a near 0 and insignificant ($p = 0.74$) relationship between leaderboard adoption and attrition (Table §F.7, column 1). Thus, we don't find support for the problematic trend of leaderboard adoption driving attrition for low performers.

Table §F.7 Attrition Analysis

	(1) attrit b/se	(2) steps b/se	(3) steps b/se	(4) steps b/se
Leaderboard	-0.012		334.061*	474.959†
	(0.037)		(200.758)	(323.809)
Attrited		-27.964		
		(359.050)		
Individual Fixed Effects	Yes	No	Yes	Yes
Week Fixed Effects	Yes	Yes	Yes	Yes
Individual Linear Trends	Yes	Yes	Yes	Yes
Individual Quadratic Trend	Yes	Yes	Yes	Yes
Observations	18,698	10,788	20,348	7,410
VCE	Robust	Robust	Robust	Robust

† $p < 0.15$, * $p < 0.10$, ** $p < 0.05$.

Moreover, we evaluated whether leaderboard adopters who eventually leave the sample exhibit problematic pre-treatment trends in physical activity prior to leaving the sample. For example, leaderboard adopters who leave the sample may also be individuals who had lower levels of physical activity to begin with. To evaluate this, we identify individuals who eventually leave the sample (*Attrited*==1), and evaluate that attrition is related to differences in steps prior to leaderboard adoption. To ensure we are comparing weeks that individuals wore their devices, we constrain our analysis to weeks that individuals wore their Fitbit devices at least 60 percent of the time (Table §F.7, Column 2). We observe no significant differences in physical activity for those who eventually leave the sample in this analyses.

We also considered another issue related to attrition. Perhaps individuals who adopt leaderboards and then observe little benefit (i.e., have smaller treatment effects) from them are those who drop out, which then influences our results. Thus, we also considered whether the impact of leaderboards differs for those who stay

throughout the entire sample vs. those who eventually abandon their Fitbit device. In Table §F.7, column 3, we estimate the impact of leaderboard only for those who don't exit the sample by the end of the observation period and find the same significant and consistent positive effect of leaderboards. We also evaluate the impact of leaderboards only for those who eventually drop out of the sample (Table §F.7, column 4). We identify similar treatment effects of leaderboard for both groups and these treatment effects are consistent with the main analysis. Although the treatment effect is actually larger using those who eventually leave the sample, it is marginally insignificant. We attribute this insignificance to the reduced sample size for these estimations (e.g., more than 2/3 of the original sample is excluded in the analysis in column 4).

G. Sensitivity to Outliers—"Leave Out One" Procedure

We also examine the robustness of our results using the "leave out one" procedure, which is a way to guard against potential outliers driving our results. The "leave out one" procedure works as follows for individuals (and analogously for weeks): first, we omit a distinct individual in every iteration of the procedure and re-estimate our main model 516 times to produce vectors of estimates on *Leaderboard*. We then summarize the vectors of effect sizes, and the p-values in Figure §G.IV (using box plots) and Table §G.8 (table also includes t-statistics). We find that effect sizes range from 331–404, the t-statistics stay over 1.97, and the p-values range from 0.017–0.049. Thus, the "leave out one" estimates remain tight around the main effect of 370 and are all statistically significant at the 5% level. The "leave out one" procedure for weeks is analogous and similar results are obtained when we repeat the procedure for weeks.

Table §G.8 Summary Statistics for Outlier Analyses

	(1) Individuals			(2) Weeks		
	min	mean	max	min	mean	max
Effect size	331.1622	370.4632	404.1664	335.7431	370.1841	406.0687
T-statistic	1.9753	2.1687	2.4054	1.9766	2.1639	2.3197
P-value	0.0165	0.0307	0.0488	0.0207	0.0312	0.0486
Observations	516			103		

H. Extension of Heterogeneous Effects Analysis

H.1. Summary Statistics by Prior Activity Levels and Correlations between Being Sedentary and Individual Characteristics

Table §H.9 presents summary statistics by prior activity levels, whereas Table §H.10 presents correlations between being sedentary and individual characteristics.

Table §H.9 Average Steps by Prior Activity Levels

	Steps (Avg.)	Steps (SD)	Individuals
Sedentary	8,021.81	(3,543.18)	129
Middle-50	10,292.42	(3,616.20)	258
Highly Active	13,110.96	(4,532.922)	129

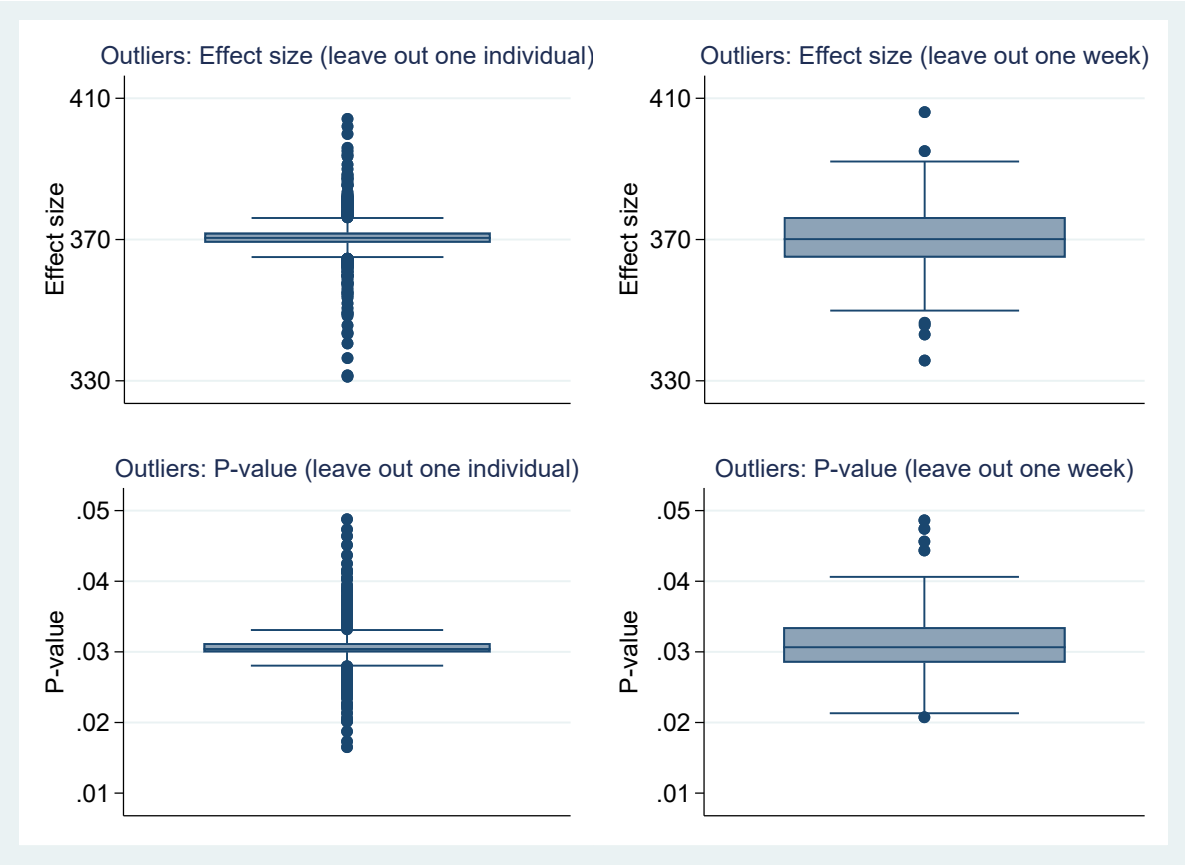


Figure §G.IV Outlier Analysis

Table §H.10 Correlation between Being Sedentary and Individual Characteristics

	(1) Self Efficacy b/se	(2) Self Regulation b/se	(3) Self Esteem b/se	(4) Depression b/se	(5) Anxiety b/se	(6) Exercise Alone b/se
Sedentary	-12.20**	-1.96*	-0.54**	4.68*	6.11**	1.76**
	(5.36)	(1.10)	(0.24)	(2.41)	(3.06)	(0.61)
Week Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Individual Linear Trends	Yes	Yes	Yes	Yes	Yes	Yes
Individual Quadratic Trends	Yes	Yes	Yes	Yes	Yes	Yes
Observations	4,354	4,354	5,497	5,360	5,497	4,089
Individuals	134	134	169	167	169	116
VCE	Robust	Robust	Robust	Robust	Robust	Robust

* $p < 0.10$, ** $p < 0.05$.

H.2. Heterogeneous Effects by Prior Activity Levels

In Table §H.11, we evaluate heterogeneous treatment effects of leaderboards based on prior activity levels using the full dataset (cf. sub-sample analysis). Specifically, we evaluate differential treatment effects for sedentary (column 1), highly active (column 2), and middle quartiles (column 3) individuals. These results are consistent with our main analysis. We also find consistent results when only utilizing the sedentary and highly active individuals (column 4) or when estimating all treatment effects simultaneously (column 5).

Table §H.11 Heterogeneous Effect by Prior Activity (Interaction Models)

	(1) steps b/se	(2) steps b/se	(3) steps b/se	(4) steps b/se	(5) steps b/se
Leaderboard	268.25	938.06**	-295.30	-648.45**	-794.91**
	(180.16)	(194.28)	(260.70)	(290.06)	(286.40)
Leaderboard × Sedentary	945.13*			1,933.97**	2,005.21**
	(536.18)			(586.83)	(580.92)
Leaderboard × Mid Activity			1,179.56**		1,680.52**
			(331.95)		(352.18)
Leaderboard × Highly Active		-1,733.48**			
		(343.48)			
Sample	Full	Full	Full	S & HA	Full
Individual Fixed Effects	Yes	Yes	Yes	Yes	Yes
Week Fixed Effects	Yes	Yes	Yes	Yes	Yes
Individual Linear Trends	Yes	Yes	Yes	Yes	Yes
Individual Quadratic Trends	Yes	Yes	Yes	Yes	Yes
Observations	27,758	27,758	27,758	13,465	27,758
Individuals	516	516	516	258	516
Adjusted R-Squared	0.33	0.33	0.33	0.34	0.33
VCE	Robust	Robust	Robust	Robust	Robust

* $p < 0.10$, ** $p < 0.05$.

Further, Table §H.12 replicates the analyses in Table 4, Columns 4–5 and Table 5. As expected, the results in Table §H.12 are very similar to the ones presented in the main body of the paper.[27]

Table §H.12 Heterogeneous Effect by Active User, Rank, and Prior Activity (Interaction Models)

	(1) steps b/se	(2) steps b/se	(3) steps b/se
Leaderboard × Sedentary	1,178.45**	920.79*	624.58
	(502.20)	(514.18)	(554.63)
Leaderboard × Sedentary × FirstonLB	796.69*	6.55	
	(438.38)	(318.42)	
Leaderboard × Sedentary × LB Active Users		230.27**	694.48**
		(50.98)	(255.52)
Leaderboard × Sedentary × LB Active Users × FirstonLB		465.01**	
		(82.19)	
Leaderboard × Sedentary × LB Active Users^2			-68.89**
			(34.77)
Leaderboard × Highly Active	-826.82**	-1,183.45**	-1,266.34**
	(300.75)	(329.50)	(317.25)
Leaderboard × Highly Active × FirstonLB	496.26**	213.93	
	(152.43)	(189.09)	
Leaderboard × Highly Active × LB Active Users		187.11**	388.50**
		(59.20)	(85.83)
Leaderboard × Highly Active × LB Active Users × FirstonLB		154.22**	
		(54.26)	
Leaderboard × Highly Active × LB Active Users^2			-20.88**
			(7.60)
Sample	S & HA	S & HA	S & HA
Individual Fixed Effects	Yes	Yes	Yes
Week Fixed Effects	Yes	Yes	Yes
Individual Linear Trends	Yes	Yes	Yes
Individual Quadratic Trends	Yes	Yes	Yes
Observations	13,465	13,465	13,465
Individuals	258	258	258
Adjusted R-Squared	0.34	0.35	0.35
VCE	Robust	Robust	Robust

* $p < 0.10$, ** $p < 0.05$.

[27] We thank an anonymous reviewer for suggesting that we include this result.

H.3. Leaderboard Competition Measures

We identified two instances of leaderboards that capture different elements of leaderboard competition.

H.3.1. No Competition Leaderboards We created *NoCompetitionLB* as a binary indicator of leaderboards without credible competition because the focal user was sandwiched between other Fitbit users (and thus was neither first nor last), and the users above and below were both too distant from the focal user to credibly compete with them. To generate the thresholds for this measure, we considered the within difference in steps from week to week. We found that the standard deviation for this difference was 596 which informed our threshold. Specifically, a 1,000-step threshold approximates a greater than 2 standard deviation change in steps from week to week. This threshold is useful because it is highly unlikely that any of the users on the leaderboard will have a 1,000 step swing (upward or downward) in one week, and it substantiates the notion that competition between the focal user and others near them is not credible.

H.3.2. High Competition Leaderboards To identify leaderboards where the focal user was engaged in relatively intense competition for the top spot on the leaderboard, we start by capturing competition intensity as the frequency of the focal user alternating as being ranked first on the leaderboard with another user week to week. We then a create a measure labeled "High Competition Leaderboard" defined as a leaderboard in the top quartile of competition intensity. In essence, this measure identifies leaderboards where the focal user is regularly being displaced and then reclaiming the top spot on the leaderboard.

References for Online Appendix

[Appendix1] Abadie, A. (2005). Semiparametric difference-in-differences estimators. *The Review of Economic Studies 72*(1), 1–19.

[Appendix1] Abadie, A. and M. D. Cattaneo (2018). Econometric methods for program evaluation. *Annual Review of Economics 10*, 465–503.

[Appendix3] Bandura, A. (2006). Guide for constructing self-efficacy scales. *Self-efficacy Beliefs of Adolescents 5*(1), 307–337.

[Appendix4] Deci, E. L. and R. M. Ryan (2008). Self-determination theory: A macrotheory of human motivation, development, and health. *Canadian Psychology 49*(3), 182.

[Appendix5] Hall, P. and N. Gill (2019). *An introduction to machine learning interpretability*. O'Reilly Media, Incorporated.

[Appendix6] Hann, D., K. Winter, and P. Jacobsen (1999). Measurement of depressive symptoms in cancer patients: Evaluation of the Center for Epidemiological Studies Depression Scale (CES-D). *Journal of Psychosomatic Research 46*(5), 437–443.

[Appendix7] Hernan, M. A. and J. M. Robins (2018). *Causal Inference* (1 edition ed.). CRC Press.

[Appendix8] Hetherington, M. J. (1998). The political relevance of political trust. *American Political Science Review 92*(4), 791–808.

[Appendix9] Lesmeister, C. (2019). *Mastering Machine Learning with R: Advanced machine learning techniques for building smart applications with R 3.5*. Packt Publishing Ltd.

[Appendix10] Lundberg, S. M. and S.-I. Lee (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pp. 4768–4777.

[Appendix11] McCaffrey, D. F., B. A. Griffin, D. Almirall, M. E. Slaughter, R. Ramchand, and L. F. Burgette (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine 32*(19), 3388–3414.

[Appendix12] Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement 1*(3), 385–401.

[Appendix13] Rosenberg, M. (1965). *Society and the Adolescent Self-image*. Princeton, New Jersey: Princeton University Press.

[Appendix14] Spielberger, C. D., R. L. Gorsuch, R. Lushene, P. R. Vagg, and G. A. Jacobs (1983). Manual for the state-trait anxiety inventory (Palo Alto, CA, Consulting Psychologists Press).

[Appendix44] Xu, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis 25*(1), 57–76.