Online Appendix

Table of Contents

A Alternative Sample of Adoption Cohorts

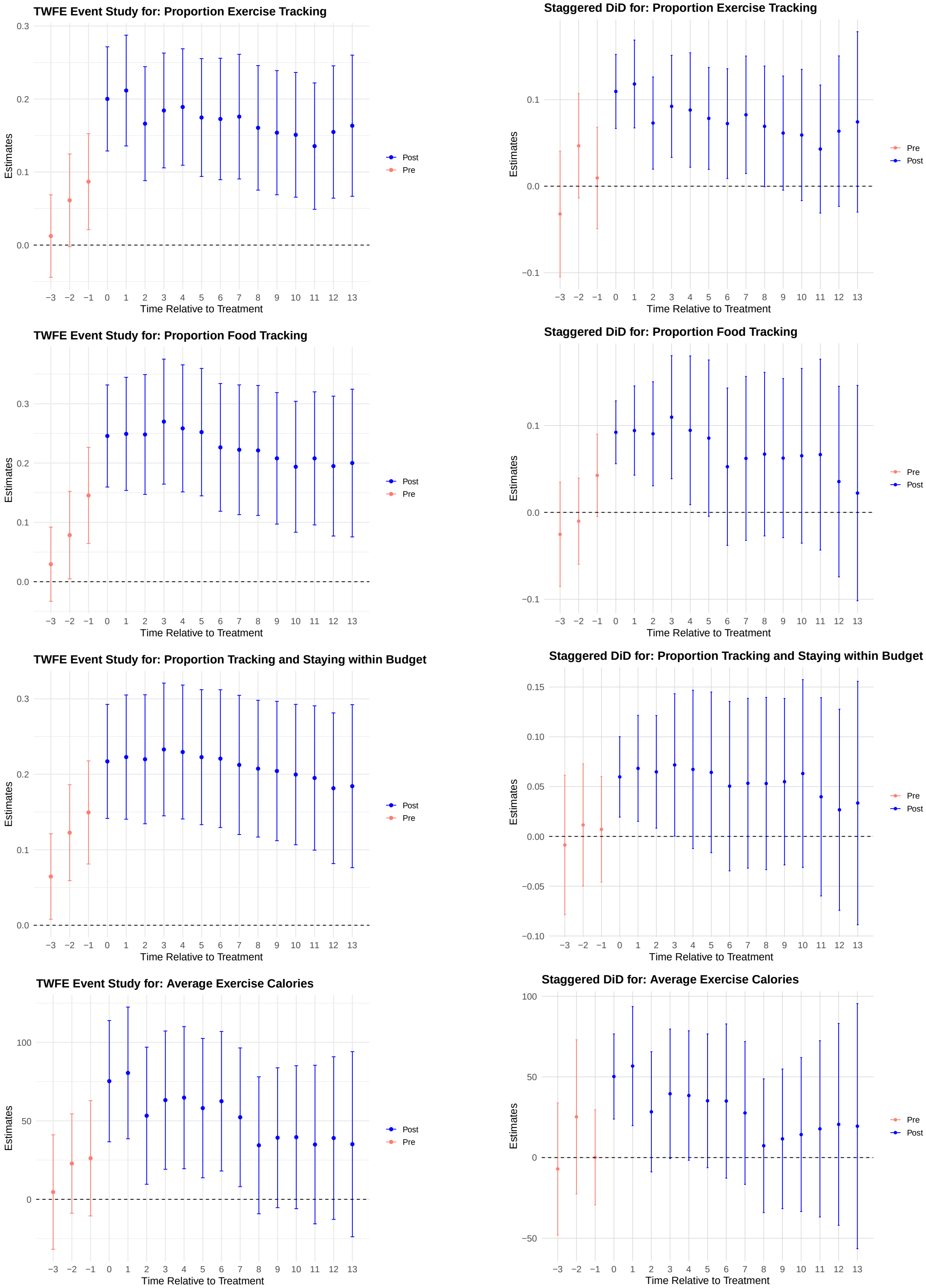
Since only 21 individuals adopted in week 7, this section replicates our estimation in 3.4 for user cohorts who adopted between weeks 2 and 6 and a random sample of 3% users who do not upgrade over our entire observation period of 15 weeks. The qualitative patterns that emerge from the alternative sample estimates in Figures A1 and A2 closely resemble those in our main specification (Figures 2 and 3). A first, encouraging difference is that the pre-treatment trajectories of treated and control users align almost perfectly in the alternative sample. That is, every pre-period GATT in Figure A1 fluctuates within a narrow confidence band around zero, whereas in Figure 2 a mild anticipatory uptick is visible for food tracking.

Turning to the short-run effects on engagement, Figure A1 replicates the same burst-and-fade profile observed in Figure 2. The initial lift in daily food-logging probability peaks at roughly 0.09 percentage points in both samples; exercise logging exhibits an identical hump that returns to baseline within six to seven weeks; and adherence to the calorie budget again displays only a brief, statistically fragile improvement. If anything, the decline toward zero is marginally faster once week-7 adopters are excluded. Hence, the evidence for a rapidly decaying engagement response is robust to this alternative cohort definition.

Figure A2 echos the modest and fading health effects reported in Figure 3. Estimated impacts on the probability of any weight loss and on cumulative pounds lost remain minimal and largely drift insignificantly around zero for the full twelve-week horizon, and the tighter pre-period match rules out the possibility that differential trends mask a delayed benefit. Therefore, in both the baseline and the restricted sample, the transient surge of participation does not translate into sustained, compounding improvements in body weight.

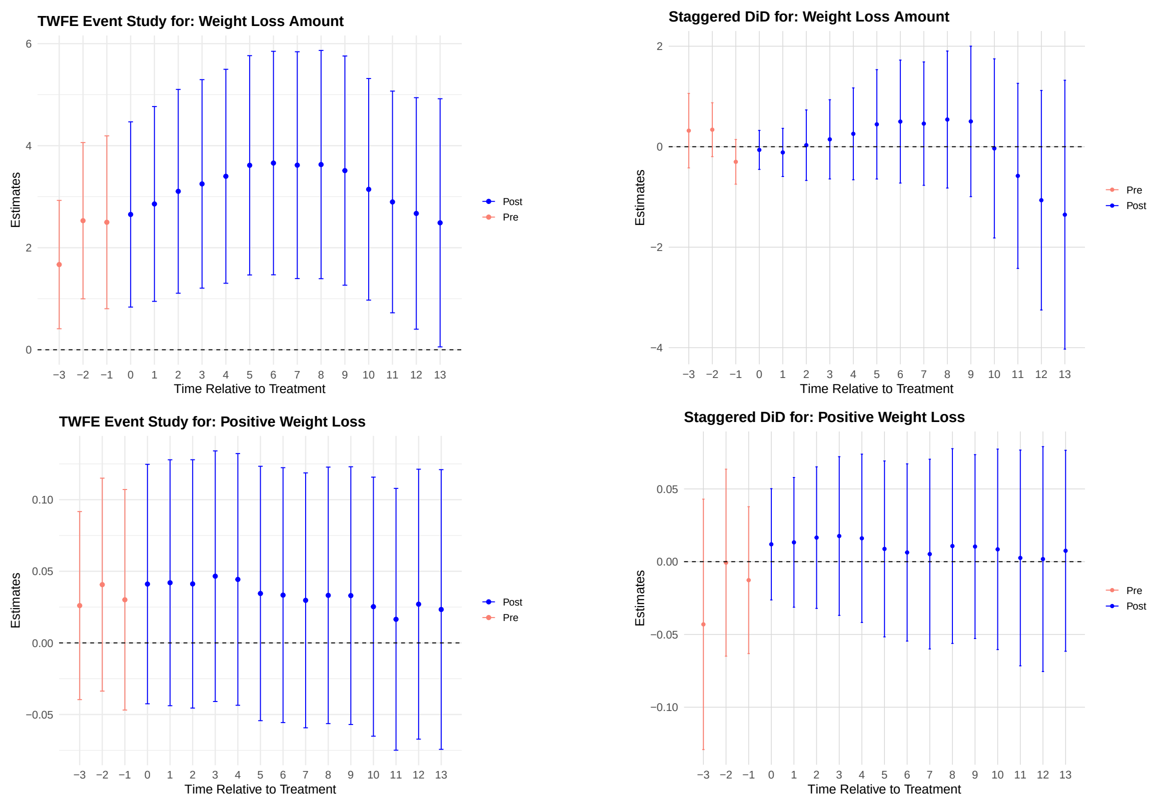
Together, the comparison reinforces the main conclusion of the paper. Premium adoption generates only an ephemeral spike in self-monitoring that dissipates within two months, and this fleeting behavioral change yields no detectable health gains. The fact that removing week 7 adopters produces virtually identical post-upgrade trajectories, while simultaneously strengthening the evidence in favor of parallel pre-trends, further underscores the internal validity of our main findings.

Figure A1: Impact of Premium Adoption on Engagement Outcomes



Notes: This table compares event-study estimates using the TWFE with those obtained with the Callaway and Sant'Anna (2021) DiD estimator. In the TWFE event-study specification, the omitted reference category consists of user-weeks more than three weeks before premium adoption, and all user-weeks for users who never adopt premium. The comparison group for Callaway and Sant'Anna (2021) DiD estimator is not-yet-treated users (including never adopters). The estimation is based on the sample that contains users who upgrade to the premium in weeks 2 - 6 and a random sample of 3% users who do not upgrade over our entire observation period of 15 weeks.

Figure A2: Impact of Premium Adoption on Health-Related Outcomes



Notes: This table compares event-study estimates using the TWFE with those obtained with the Callaway and Sant'Anna (2021) DiD estimator. In the TWFE event-study specification, the omitted reference category consists of user-weeks more than three weeks before premium adoption, and all user-weeks for users who never adopt premium. The comparison group for Callaway and Sant'Anna (2021) DiD estimator is not-yet-treated users (including never adopters). The estimation is based on the sample that contains users who upgrade to the premium in weeks 2 - 6 and a random sample of 3% users who do not upgrade over our entire observation period of 15 weeks.

B Sensitivity to Measurement

This section explores the sensitivity of our main findings to potential accuracy issues with respect to food calorie and weight loss metrics. Table B1 summarizes the results from this sensitivity analysis.

Food log accuracy. The food calorie information is provided by users. Users typically take photos of their food and update them to their app so that the app infers calorie. As such, there might be questions about the validity of this volunteered information. Unfortunately, there is no way of knowing what the "ground truth" is for food calorie information in our specific empirical context. However, we note the following points that might assuage such concerns. First, our focus of analysis is on the extensive margin of food tracking (i.e., whether or not users tracked food consumption). For this reason, any user-generated errors in the numeric value of food log calories will likely only materialize at the intensive margin, rather than the extensive margin. Second, and most importantly, we are able to demonstrate the robustness of the key data patterns using a food log filtering approach originally used in nutrition sciences. To isolate credible tracking events, we consider an alternative specification that counts a user as tracking food calories on a given day only if the total food calories entered that day are credible. Essentially, observations for which daily caloric intake is lower than 500 kcal/day or greater than 4000 kcal/day are counted as not tracking food calories on the given day, as caloric intake of this amount per day is considered to be extreme and incredible (Willett 2012). This filtering rule has been applied in other nutrition science research involving the same empirical context (Labonté et al. 2022). In this sensitivity analysis, the dependent variable can be understood as indicating whether a user tracked food calories within a plausible range on a given day. The results from this analysis confirm that for the sub-sample of physiologically credible food log entries, the main qualitative patterns still exist.

Weight log accuracy. Similar to our sensitivity analysis for food calorie reliability, we also assess the sensitivity of our results related to weight loss. To get a sense of what is credible, we turn to findings from Finkler et al. (2012), which demonstrate that realistic weight loss ranges from 0.5 to 1 kg per week. Accordingly, in this analysis, if a user reports a weight loss exceeding 1 kg compared to the previous week, the reported value is capped at 1 kg (2.2 lbs). From this analysis, we observe that the alternative variable construction for plausible weight loss yields very similar results as in our baseline results.

Table B1: Overall ATT Estimates from Alternative Measurements

Callaway and Sant'Anna (2021) Estimates			
Variable	ATT	SE	95% CI
Exercise Tracking	0.074	0.013	[0.048, 0.100]
Food Tracking	0.060	0.015	[0.030, 0.089]
Tracking and Staying within Budget	0.048	0.014	[0.022, 0.075]
Average Exercise Calories	32.183	8.930	[14.680, 49.685]
Positive Weight Loss	0.026	0.014	[-0.002, 0.054]
Weight Loss Amount	0.487	0.269	[-0.040, 1.014]
Estimates with Alternative Measurements			
Exercise Tracking	0.074	0.013	[0.048, 0.099]
Food Tracking	0.054	0.016	[0.024, 0.085]
Tracking and Staying within Budget	0.048	0.014	[0.022, 0.075]
Average Exercise Calories	32.034	9.052	[14.293, 49.774]
Positive Weight Loss	0.027	0.014	[-0.001, 0.055]
Weight Loss Amount	0.406	0.311	[-0.204, 1.015]

Notes: This table reports overall ATT estimates from alternative measurements. **Panel A** uses the approach by Callaway and Sant'Anna (2021), and **Panel B** uses the estimator from Callaway and Sant'Anna (2021) with alternative measurements. Both analyses are based on users who upgrade between weeks 2–7 and all those who never upgrade, observed for 15 weeks. To isolate credible tracking events, the specification in **Panel B** only counts a user as tracking food calories on a given day if the user's total food calories entered on the day are credible. Essentially, observations for which daily caloric intake is lower than 500 kcal/day or greater than 4000 kcal/day are counted as not tracking food calories on the given day, as caloric intake of this amount per day is considered to be extreme and incredible (Willett 2012). For weight loss outcomes in **Panel B**, if a user reports a weight loss exceeding 1 kg compared to the previous week, the reported value is capped at 1 kg (2.2 lbs), as findings from Finkler et al. (2012) demonstrate that realistic weight loss ranges from 0.5 to 1 kg per week. Standard errors are clustered at the user level.

C Alternative Estimators

To explore the sensitivity of our results to different estimation approaches, we consider alternative TWFE estimators that differ in reweighting strategies. First, the switcher-difference-in-differences estimator of De Chaisemartin and d'Haultfoeuille (2020) isolates the average jump that occurs precisely when a user's treatment status flips, stripping out periods in which treated units serve as controls for themselves. Second, the synthetic DiD of Arkhangelsky et al. (2021) relaxes parallel trends by matching each cohort to a synthetic control whose pre-paths mimic its own. If the premium effect remains under both estimators, we can discount artifacts of TWFE weighting and attribute any persistence to genuine behavioral change.

C.1 De Chaisemartin and d'Haultfoeuille (2020) Estimator

One alternative approach, proposed by De Chaisemartin and d'Haultfoeuille (2020), addresses heterogeneous treatment effects by computing weighted average treatment effects across treatment cohorts. This method ensures that the estimated treatment effect properly reflects variations in adoption timing, making it particularly useful in staggered adoption settings where different users upgrade at different points in time. Note that the assumptions this estimator requires are subtly different than the ones in Callaway and Sant'Anna (2021), namely assumptions about sharp design and independent groups.

Sharp design. This assumption implies that the adoption of premium does not vary within cohort and time. Note that based on our definition of time (i.e., weekly level rather than daily), the premium adoption treatment will by definition vary within cohort and time (i.e., there are multiple days within the same week for which adoption can take place).

Independent groups. This assumption allows for the possibility that the premium adoption treatment and potential user engagement and health-related outcomes of a group may be correlated over time, but with the main requirement that the potential outcomes and treatments across different adoption cohorts are independent. What this essentially requires, under our empirical setting, is that the adoption of premium is not strongly driven by social/peer considerations. As the "social functions and referrals" in the mHealth app are severely lacking in the user interface design, we believe that user-to-user contamination is highly unlikely thereby making this assumption trivially satisfied.

Strict exogeneity. While both De Chaisemartin and d'Haultfoeuille (2020) and Callaway and Sant'Anna (2021) rely on parallel trends (conditional or unconditional), the De Chaisemartin and d'Haultfoeuille (2020) estimator requires a stronger assumption about strict exogeneity as compared with Callaway and Sant'Anna (2021). This assumption means that the unobserved shocks that impact the pre-adoption period outcomes are mean independent of the adoption sequence. In contrast, Callaway and Sant'Anna (2021) allow for violations of strong exogeneity if valid control groups (e.g., never-treated or not-yet-treated) can be used.

C.1.1 Summary of Results

Across engagement outcomes, the overall ATT estimates from the estimator by De Chaisemartin and d'Haultfoeuille (2020), shown in Table C1, replicate the qualitative patterns observed in both TWFE and estimates from Callaway and Sant'Anna (2021): a significant increase in food and exercise tracking following premium adoption. However, the magnitude of these effects is notably more muted relative to TWFE and closely aligns with the estimates from Callaway and Sant'Anna (2021). This divergence highlights how TWFE can overstate treatment effects when dynamic effects are pooled across adoption cohorts with differing underlying trends. By assigning more balanced weights across cohorts, the De Chaisemartin and d'Haultfoeuille (2020) estimator yields a more conservative assessment of the post-adoption engagement lift than TWFE. Finally, we note that the point estimates for the ATTs from Callaway and Sant'Anna (2021) and De Chaisemartin and d'Haultfoeuille (2020) are highly comparable and the confidence intervals mostly exhibit overlap.

Table C1: Overall ATT Estimates from Alternative DiD estimators

Callaway and Sant'Anna (2021) Estimates			
Variable	ATT	SE	95% CI
Exercise Tracking	0.074	0.013	[0.048, 0.100]
Food Tracking	0.060	0.015	[0.030, 0.089]
Tracking and Staying within Budget	0.048	0.014	[0.022, 0.075]
Average Exercise Calories	32.183	8.930	[14.680, 49.685]
Positive Weight Loss	0.026	0.014	[-0.002, 0.054]
Weight Loss Amount	0.487	0.269	[-0.040, 1.014]
de Chaisemartin and D'Haultfœuille (2020) Estimates			
Proportion Exercise Tracking	0.081	0.017	[0.048, 0.113]
Proportion Food Tracking	0.080	0.018	[0.045, 0.115]
Proportion Tracking and Staying within Budget	0.075	0.015	[0.047, 0.104]
Average Exercise Calories	35.970	11.021	[14.369, 57.572]
Positive Weight Loss	0.025	0.018	[-0.009, 0.060]
Weight Loss Amount	0.945	0.282	[0.393, 1.497]

Notes: This table reports overall ATT estimates from the two DiD estimators. **Panel A** uses the approach by Callaway and Sant'Anna (2021), and **Panel B** uses the estimator from De Chaisemartin and d'Haultfoeuille (2020). **Panel A** and **Panel B** analyses are based on users who upgrade between weeks 2–7 and all those who never upgrade, observed over 15 weeks.

C.2 Synthetic Difference-in-Differences

As a complementary specification test, we estimate ATT with the *synthetic difference-in-differences* (SDɪD) estimator of Arkhangelsky et al. (2021). SDɪD retains the double–differencing logic of a classical DiD while using synthetic–control weights to relax parallel-trend requirements.

Let Y_{it} denote the outcome for user i on day t, and let $W_{it} = 1$ from the upgrade date onward. For each adoption cohort g we first construct unit and time weights, ω_i^{SDID} and

λ_t^{SDiD}, that align the pre-treatment paths of never-upgraders with those of cohort g up to an additive constant and that balance pre- and post-period means for every control unit. Using these weights, we obtain a cohort-specific treatment effect

$$\widehat{\tau}_g^{\text{SDiD}} = \arg\min_{\tau} \sum_{i,t} \big(Y_{it} - \mu - \alpha_i - \beta_t - \tau W_{it}\big)^2 \, \omega_i^{\text{SDiD}} \lambda_t^{\text{SDiD}},$$

where (μ, α_i, β_t) are the usual two-way fixed effects. To recover the overall average treatment effect on the treated, we can aggregate these cohort estimates with cohort-size weights,

$$\widehat{\tau}^{\text{SDiD}} = \frac{\sum_g n_g \, \widehat{\tau}_g^{\text{SDiD}}}{\sum_g n_g},$$

where n_g is the number of users who upgrade in cohort g. Inference relies on the cohort-level cluster multiplier bootstrap proposed by Arkhangelsky et al. (2021).

Note that we confine the SDiD exercise to the overall ATT for illustration purposes. Our main purpose in running the SDiD is to show that the SDiD method provides consistent estimates compared to other methods we use. Collapsing to a single, cohort-size–weighted ATT facilitates this comparison. Moreover, generating event-study graphs is computationally demanding because of the staggered adoption design and the large sample size.

C.2.1 Summary of Results

The SDiD robustness exercise replicates our main findings: a premium upgrade delivers a short-lived burst in self-monitoring yet fails to translate into sizable health gains. Table C2 juxtaposes the overall ATTs from Callaway and Sant'Anna (2021) with those from Arkhangelsky et al. (2021). Point estimates and standard errors are very similar, but most SDiD coefficients are slightly larger in magnitude than the estimates from Callaway and Sant'Anna (2021), but still notably smaller than estimates from TWFE in Table 3. The slightly larger SDiD estimates, compared to those reported by Callaway and Sant'Anna (2021), may stem from limitations in the number of covariates we are able to include. Incorporating additional covariates substantially increases computational burden, which constrains model specifications we can consider.

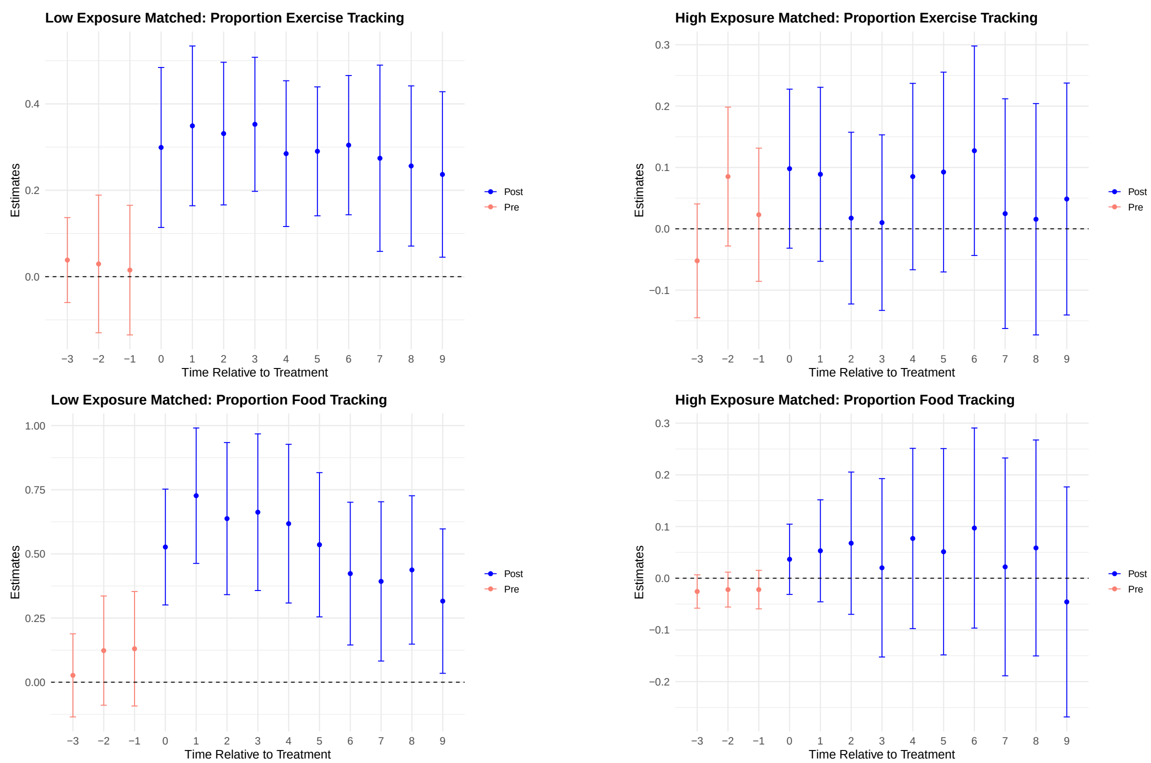
Table C2: Overall ATT Estimates from Alternative DiD estimators

Callaway and Sant'Anna (2021) Estimates			
Variable	ATT	SE	95% CI
Exercise Tracking	0.074	0.013	[0.048, 0.100]
Food Tracking	0.060	0.015	[0.030, 0.089]
Tracking and Staying within Budget	0.048	0.014	[0.022, 0.075]
Average Exercise Calories	32.183	8.930	[14.680, 49.685]
Positive Weight Loss	0.026	0.014	[-0.002, 0.054]
Weight Loss Amount	0.487	0.269	[-0.040, 1.014]
Synthetic Difference-in-Differences Estimates			
Exercise Tracking	0.091	0.014	[0.062, 0.119]
Food Tracking	0.094	0.016	[0.061, 0.126]
Tracking and Staying within Budget	0.075	0.015	[0.047, 0.104]
Average Exercise Calories	45.567	10.326	[25.329, 65.806]
Positive Weight Loss	0.017	0.018	[-0.017, 0.052]
Weight Loss Amount	0.895	0.297	[0.313, 1.476]

Notes: This table reports overall ATT estimates from the two DiD estimators. **Panel A** uses the approach by Callaway and Sant'Anna (2021), and **Panel B** uses the estimator from Arkhangelsky et al. (2021). Both analyses are based on users who upgrade between weeks 2–7 and those who never upgrade, observed over 15 weeks. The confidence intervals for synthetic DiD estimates are based on bootstrap.

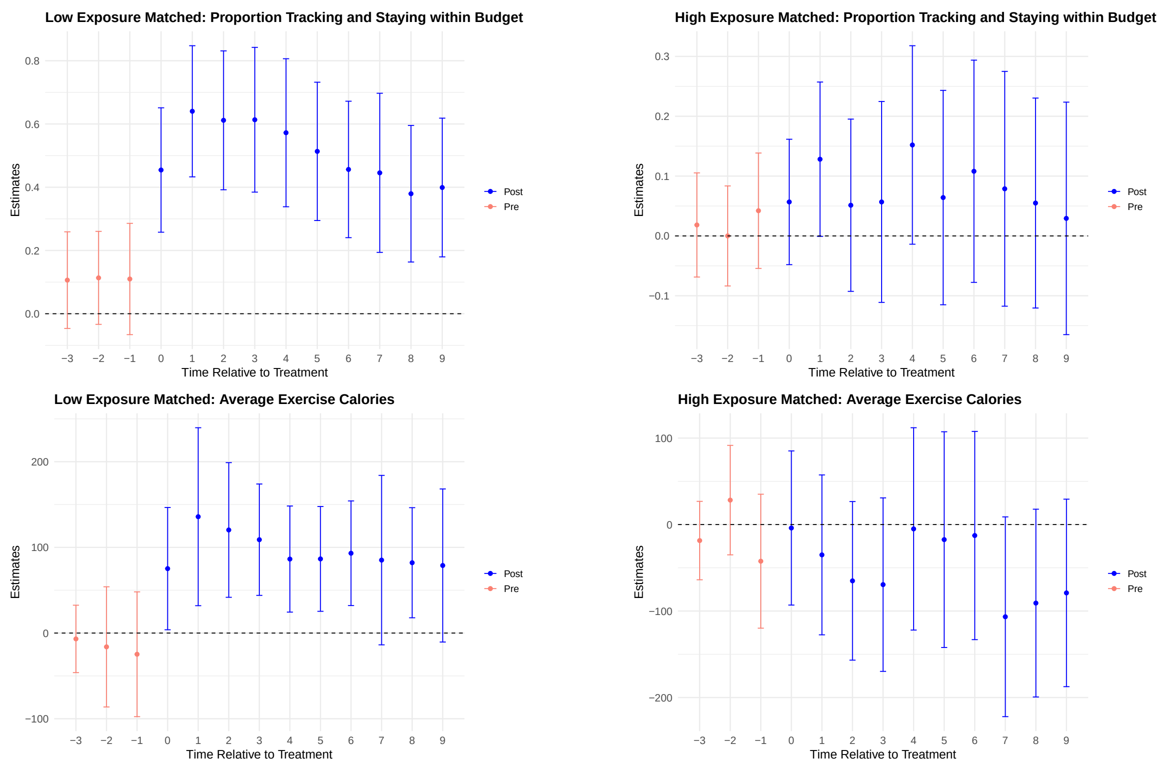
D Additional Tables and Figures

Figure D1: TWFE Estimates on Matched Sample Balancing on Additional Recent Variables and Segmented on Pre-adoption Exposure (Week 6 Adopters Matched with Non-Adopters)



Notes: Panels compare TWFE event-study estimates for week-6 adopters split by pre-adoption exposure (median split), matching with similar non-adopters, respectively. The matching is based on demographic characteristics, initial goal settings, recent Google Search intensities (week 5), early and recent engagement trajectories (weeks 4 and 5), and recent weight-loss outcomes (week 5). Left column shows matched estimates for low pre-adoption exposure; right column shows matched estimates for high pre-adoption exposure. Points are coefficient estimates; error bars are 95% confidence intervals with standard errors clustered at the user level.

Figure D2: TWFE Estimates on Matched Sample Balancing on Additional Recent Variables and Segmented on Pre-adoption Exposure (Week 6 Adopters Matched with Non-Adopters) [continued...]



Notes: Panels compare TWFE event-study estimates for week-6 adopters split by pre-adoption exposure (median split), matching with similar non-adopters, respectively. The matching is based on demographic characteristics, initial goal settings, recent Google Search intensities (week 5), early and recent engagement trajectories (weeks 4 and 5), and recent weight-loss outcomes (week 5). Left column shows matched estimates for low pre-adoption exposure; right column shows matched estimates for high pre-adoption exposure. Points are coefficient estimates; error bars are 95% confidence intervals with standard errors clustered at the user level.

Table D1: Naive TWFE Estimates for Engagement Outcomes

	Track Exercise	Track Food	Exercise Calories	Within Budget
Pre-Treatment Period (Weeks Before Upgrade)				
Three weeks before upgrade	0.026	0.069**	2.790	0.083***
	(0.026)	(0.028)	(16.075)	(0.025)
Two weeks before upgrade	0.077***	0.118***	23.764*	0.136***
	(0.028)	(0.032)	(13.719)	(0.027)
One week before upgrade	0.105***	0.196***	29.965**	0.175***
	(0.029)	(0.034)	(15.189)	(0.028)
Immediate Post-Treatment (Week of Upgrade)				
Week of upgrade	0.219***	0.300***	79.495***	0.246***
	(0.030)	(0.036)	(15.551)	(0.030)
Short-Term Post-Treatment (1–6 Weeks After Upgrade)				
One week after upgrade	0.228***	0.306***	85.620***	0.253***
	(0.031)	(0.038)	(16.147)	(0.032)
Two weeks after upgrade	0.183***	0.307***	62.252***	0.251***
	(0.031)	(0.039)	(16.314)	(0.032)
Three weeks after upgrade	0.204***	0.325***	72.054***	0.260***
	(0.032)	(0.040)	(16.446)	(0.032)
Four weeks after upgrade	0.207***	0.310***	74.416***	0.259***
	(0.031)	(0.041)	(16.825)	(0.033)
Five weeks after upgrade	0.187***	0.305***	68.032***	0.256***
	(0.031)	(0.041)	(16.286)	(0.032)
Six weeks after upgrade	0.185***	0.273***	73.046***	0.250***
	(0.033)	(0.041)	(16.704)	(0.033)
Medium-Term Post-Treatment (7–12 Weeks After Upgrade)				
Seven weeks after upgrade	0.184***	0.265***	63.489***	0.239***
	(0.033)	(0.043)	(16.357)	(0.034)
Eight weeks after upgrade	0.167***	0.261***	44.576***	0.231***
	(0.034)	(0.043)	(16.533)	(0.034)
Nine weeks after upgrade	0.162***	0.238***	51.034***	0.218***
	(0.033)	(0.042)	(16.425)	(0.034)
Ten weeks after upgrade	0.159***	0.217***	51.854***	0.207***
	(0.033)	(0.042)	(16.496)	(0.034)
Eleven weeks after upgrade	0.142***	0.224***	48.949**	0.198***
	(0.034)	(0.042)	(19.138)	(0.035)
Twelve weeks after upgrade	0.165***	0.210***	55.651***	0.185***
	(0.035)	(0.045)	(19.115)	(0.036)
Long-Term Post-Treatment (13 Weeks After Upgrade)				
Thirteen weeks after upgrade	0.167***	0.213***	49.703**	0.187***
	(0.037)	(0.047)	(22.221)	(0.040)
User FE	Yes	Yes	Yes	Yes
Week FE	Yes	Yes	Yes	Yes
Observations	171,390	171,390	171,390	171,390
R^2	0.615	0.576	0.623	0.562

Notes: This table reports event-study estimates from TWFE on engagement outcomes. The data includes users who upgrade to premium between weeks 2–7 and all users who never upgrade during a 15-week observation period. The omitted reference category consists of user-weeks more than three weeks before premium adoption, and all user-weeks for users who never adopt premium. Standard errors are clustered at the user level. * denotes $p < 0.10$, ** denotes $p < 0.05$, and *** denotes $p < 0.01$.

Table D2: Naive TWFE Estimates for Weight-Related Outcomes

	Positive Weight Loss	Weight Loss Amount
Pre-Treatment Period (Weeks Before Upgrade)		
Three weeks before upgrade	0.040	1.464***
	(0.029)	(0.559)
Two weeks before upgrade	0.062*	2.301***
	(0.032)	(0.650)
One week before upgrade	0.055*	2.341***
	(0.033)	(0.717)
Immediate Post-Treatment (Week of Upgrade)		
Week of upgrade	0.075**	2.492***
	(0.034)	(0.736)
Short-Term Post-Treatment (1-6 Weeks After Upgrade)		
One week after upgrade	0.077**	2.704***
	(0.035)	(0.757)
Two weeks after upgrade	0.080**	3.026***
	(0.034)	(0.783)
Three weeks after upgrade	0.089***	3.197***
	(0.034)	(0.797)
Four weeks after upgrade	0.088***	3.374***
	(0.033)	(0.817)
Five weeks after upgrade	0.080**	3.615***
	(0.034)	(0.840)
Six weeks after upgrade	0.080**	3.711***
	(0.034)	(0.857)
Medium-Term Post-Treatment (7-12 Weeks After Upgrade)		
Seven weeks after upgrade	0.076**	3.729***
	(0.034)	(0.877)
Eight weeks after upgrade	0.078**	3.826***
	(0.034)	(0.893)
Nine weeks after upgrade	0.082**	3.762***
	(0.034)	(0.883)
Ten weeks after upgrade	0.075**	3.476***
	(0.034)	(0.852)
Eleven weeks after upgrade	0.069**	3.342***
	(0.034)	(0.853)
Twelve weeks after upgrade	0.081**	3.236***
	(0.035)	(0.896)
Long-Term Post-Treatment (13+ Weeks After Upgrade)		
Thirteen weeks after upgrade	0.080**	3.177***
	(0.036)	(0.967)
User FE	Yes	Yes
Week FE	Yes	Yes
Observations	171,390	171,390
R^2	0.678	0.805

Notes: This table reports event-study estimates from TWFE on weight loss outcomes. The data includes users who upgrade to premium between weeks 2–7 and all users who never upgrade during a 15-week observation period. The omitted reference category consists of user-weeks more than three weeks before premium adoption, and all user-weeks for users who never adopt premium. Standard errors are clustered at the user level. * denotes $p < 0.10$, ** denotes $p < 0.05$, and *** denotes $p < 0.01$.