Does Premium Version Adoption in mHealth Improve User Engagement and Health-Related Outcomes?

Yikun Jiang* Kosuke Uetake† Nathan Yang‡

June 7, 2026

Abstract

Freemium upgrade is a primary monetization lever for mHealth apps, yet whether paying for premium features actually sustains user engagement remains an open question. We study this question using large-scale data from a fitness-tracking mobile app and a staggered difference-in-differences design. Premium adoption generates an immediate increase in food and exercise tracking, caloric budget adherence, and exercise calories, but these effects attenuate within several weeks and do not translate into sustained weight loss. Sensitivity analysis and matched-sample designs confirm that the engagement responses are not solely driven by time-varying selection. Heterogeneous engagement lifts by pre-upgrade exposure are more consistent with hedonic decline than with sunk-cost effects or motivational mean reversion: users with limited prior exposure to the free version exhibit substantially larger post-upgrade engagement lifts than those who have already interacted extensively with it. We further demonstrate that failing to account for endogeneity and selection in premium adoption leads to a substantial overstatement of its effects on both engagement and health outcomes.

Keywords: Behavioral Analytics; Freemium; Mobile Health; Subscriptions; Hedonic Decline; Sunk Costs

*Purdue University, Mitch Daniels School of Business. Email: jiangyk@purdue.edu.

†Yale University, School of Management. Email: kosuke.uetake@yale.edu.

‡University of Illinois Urbana-Champaign, Gies College of Business. Email: ncyang@illinois.edu.

We thank Andrew Ching, Laurette Dubé, Ashvin Gandhi, Tongil TI Kim, Jingjing Li, Jūra Liaukonytė, Unnati Narang, Daiva Nielsen, and Vithala Rao, as well as conference/seminar participants at the 15th Annual Conference on Health IT and Analytics, INFORMS Annual Meeting, 3rd Health Care Markets Conference, Temple University, University of Alberta, 11th Triennial Choice Symposium, Canadian Centre for Health Economics, Inaugural George Washington University Intelligence of Things Conference, ET Symposium, Nudges in Health Care Symposium, McGill University, Concordia University, Cornell University, Emory University, University of Guelph, Marketing Science Conference, Southern Ontario Behavioral Decision Research (SOBDR) Conference, and University of Toronto, for helpful comments and suggestions.

1 Introduction

A feature of mobile health (mHealth) applications that distinguishes them from traditional health and wellness programs is their reliance on the freemium business model (Faulkner 2019), which firms have invested heavily on. In a freemium business model, users can access many of the basic core features of the technology for free, but need to pay additional monthly or yearly fees to use enhanced features offered in a premium version. In general, the freemium subscription-based model has emerged as a popular monetization strategy for mobile apps (Narang and Shankar 2019). While past research has helped us better understand the drivers of mHealth adoption (Bojd et al. 2022, Ghose et al. 2022, Kato-Lin et al. 2016, Tang et al. 2015), we know little about the impact of mHealth premium adoption on user engagement dynamics. Establishing this link matters because of a persistent problem faced by mHealth apps: sustained user engagement remains difficult to achieve for most app companies (Murnane et al. 2015), despite the apparent benefits of mHealth usage (Aswani et al. 2019, Kapoor et al. 2024, Zhou et al. 2018, Labonté et al. 2022, Uetake and Yang 2018).

Using large-scale data from a popular fitness-tracking mHealth application, we investigate the impact of premium adoption on user engagement and health-related outcomes. Our empirical analysis focuses on tracking differences in engagement dynamics between premium and basic version users while addressing potential selection concerns (e.g., more motivated users might self-select into upgrading). To ensure comparability, we employ a staggered Difference-in-Differences (DiD) framework that accounts for treatment timing and heterogeneity in adoption patterns (Callaway and Sant'Anna 2021). The findings reveal that premium adopters experience an initial surge in engagement with the app, particularly in food and exercise tracking, but this effect attenuates over time. Despite an initial increase in engagement behavior, there is no sustained improvement in long-term engagement or weight loss outcomes. Moreover, sensitivity analysis (Rambachan and Roth 2023) that explores potential violations of the underlying assumptions behind Callaway and Sant'Anna (2021) confirms that certain outcomes (e.g., food tracking and exercise tracking) are more robust to deviations from the assumptions of conditional parallel trends. This analysis shows the challenges that selection biases pose when analyzing adoption-based treatments. For further robustness, we expand the conditioning variable set to include measures of public attention (i.e., the Google Search Index for keywords related to the app and weight loss) near the time of registration, and show that the estimates from staggered DiD remain quantitatively similar. To account for the most recent engagement and information exposure around the premium upgrade, we implement a matched-sample design that additionally balances

users by recent engagement trajectories and information exposure. Across these robustness exercises, the estimated effects on engagement remain quantitatively similar.[1]

To better understand the behavioral drivers of post-adoption engagement dynamics, we discuss how the evidence is inconsistent with moral licensing and goal liberation, then evaluate two competing mechanisms: sunk costs and hedonic decline. To explore these potential mechanisms, we segment users based on their tracking activity prior to the upgrade and estimate the effects of premium adoption separately for groups with high and low levels of pre-upgrade activity. We use the fact that the free and premium versions of the app share many core user-interactive features, with the premium version largely building on the foundation of the free version. If sunk costs dominate, we would expect comparable effects across both groups, since all users incur the same financial outlay when upgrading. In contrast, if hedonic decline is more salient, engagement lifts should be attenuated for users who have already interacted extensively with the free version before upgrading, as the novelty of the premium features would be attenuated for them. The results are more consistent with the latter: while both groups exhibit engagement lifts immediately following the upgrade, the magnitudes are notably larger among users with less prior experience interacting with the app. As an additional robustness check to account for the most recent engagement and information exposure around the premium upgrade, we focus on users who upgraded in weeks 5 and 6 and classify adopters into low- and high-exposure groups based on pre-upgrade tracking intensity. For each exposure group, we implement a separate matched-sample design that matches adopters with similar non-adopters in demographics, initial goals, engagement, weight-loss outcomes, Google Search Indexes in the first week after registration, and the most recent engagement and Google Search Indexes in the weeks before the premium upgrade. Across both groups, low-exposure users exhibit larger immediate engagement increases that persist longer, whereas high-exposure users show substantially smaller, largely statistically insignificant engagement lifts compared to similar non-adopters. Because both groups are respectively matched with non-adopters who have similar recent pre-adoption dynamics, these heterogeneous responses further support a hedonic-decline mechanism rather than purely motivational mean reversion.

These findings suggest that although the premium upgrade provides an initial behavioral

[1]We also conduct several robustness exercises in the Online Appendix. In Appendix Section A, we replicate the staggered DiD analysis using an alternative sample that excludes 21 week-7 adopters and draws a random 5% sample of never-adopters, yielding nearly identical post-upgrade dynamics and tighter pre-trend alignment. In Appendix Section B, we assess sensitivity to measurement by applying physiologically credible thresholds to food and weight logs, following conventions in nutrition science; the estimated effects on engagement and health outcomes remain largely unchanged. Finally, in Appendix Section C, we implement alternative DiD estimators, including De Chaisemartin and d'Haultfoeuille (2020) and synthetic DiD (Arkhangelsky et al. 2021), which produce treatment effects that are comparable to our preferred estimates and materially smaller than TWFE benchmarks.

nudge, the novelty of its features generates less engagement lift among users with high exposure to the free version before the upgrade. These findings reinforce the idea that sustained engagement in digital health requires more than a one-time intervention; instead, interventions must be designed to mitigate hedonic decline and maintain the novelty or relevance of platform features over time.

1.1 Related Literature

Our research contributes to two main streams of literature: health and marketing, and freemium design in subscription-based services.

Health and marketing. Our empirical analysis contributes towards a growing interest in establishing causal effects of wearables and health apps (Ghose et al. 2022, Kapoor et al. 2024, Liu et al. 2024), while discussing plausible mechanisms (e.g., moral licensing, goal liberation, sunk costs, hedonic decline) behind the observed mHealth engagement dynamics. More generally, we contribute to the broader discussion about the role of marketing in supporting health and wellness. Some notable examples include previous studies on social comparisons (Aral and Nicolaides 2017, Uetake and Yang 2020), self-regulation (Huang et al. 2015), nutrition labeling (Bollinger et al. 2022, Puranam et al. 2017), medical diagnoses (Ma et al. 2013), AI vs human coaching (Kapoor et al. 2024), taxes (Gordon and Sun 2015, Khan et al. 2016, Seiler et al. 2021), financial vs non-financial incentives (Narang 2022), omnichannel grocery shopping (Chintala et al. 2023, Huyghe et al. 2017), variety (Haws et al. 2017), and healthiness claims (Rao and Wang 2017). We add to this discussion by offering insights into the role of mHealth and its design in the marketing and health/wellness ecosystem. We note that the work by Kapoor et al. (2024) is most related to ours. Their study highlights the impact of AI versus human coaching on weight loss among mHealth users. Our work complements their study in the following ways. First, our mHealth app's premium version (at the time of our data period) did not include coaching of any meaningful form, so the actual functional benefits associated with premium versus basic are plausibly more psychologically driven (i.e., the main difference between premium and basic versions is the fact that users pay for a few additional features on premium), in our case. Second, many of the behavioral mechanisms we posit (e.g., sunk costs) rely on our results about the premium version's impact on engagement over time.

Freemium design. Second, our findings add to the general understanding of freemium design. Thus far, much of the attention has been centered around advertising and pricing strategies for the freemium design (Appel et al. 2019, Lambrecht and Misra 2017, Li et al.

2019) and the extent to which they have an impact on customer engagement. Our empirical context is unique in the sense that users need to actively exert effort to remain engaged and to achieve improved health outcomes through the app. Our findings suggest that the subscription-based model of mHealth apps might in itself have an impact on user engagement dynamics in the short term, though its impact appears limited over the long term.

2 Empirical Context

We obtained user data from a leading mHealth fitness app company headquartered in the United States. Upon registering for the app, users are required to enter their basic demographic information, their current weight, height, goal weight, age, and gender. Based on this information, the app generates a daily calorie budget. Users can then track the foods they consume for each meal (the app then supplements caloric values), any exercise they engage in, and their current weight as long as they wish. At the end of each day, users will know whether they stay within their daily calorie budget or not. The notification from the app is purely a simple message, and there is no major change in the app interface if a user consumes beyond the budget.

In particular, the sample provided by the company consists of users who were active for at least 30 days and performed at least five weigh-ins, since these are users who are considered serious in their weight loss effort and are viewed as "focal users" from the company's perspective. In total, 11,873 users aiming to lose weight are included in our sample. We observe their daily activity since registration in late 2015 and over the year 2016. On average, the users were active for 166 days.[2] We focus on the first 15 weeks following each user's registration to analyze behavior during a period when users are generally active in weight loss efforts.[3] Among the 11,873 users, 806 users (about 6.8%) upgraded to the premium version of the app within the first seven weeks since registration, and the remaining 11,067 used the free version during the period.[4] Figure 1 shows the total number of users who upgrade in each week since registration.

Table 1 provides summary statistics for users included in our sample. The users in our sample consist of 29% males and 71% females. An average user in our sample is around 42 years old, with a starting weight of around 204 pounds, and a target weight of 162 pounds,

[2]The number of active days is defined as the number of days between the user's registration date and the last day the user tracked anything on the app for each user.

[3]Users who registered on different dates could have different engagement levels, as mentioned by Oblander and McCarthy (2023). For this reason, we include the registration date as one of the conditioning variables in the Callaway and Sant'Anna (2021) specification later.

[4]The proportion of paid users is comparable to FitBit (a competing and popular fitness app), which has a premium adoption rate of about 4% based on its user/subscriber numbers in 2020.

aiming to lose around 41 pounds, which is around 19% of the starting weight.[5] Over the course of the first 15 weeks, the suggested budget of calories is, on average, around 1626 calories, and users logged around 809 kcal food calories along with around 110 kcal exercise calories. On average, users logged at least one food item on 58% of days, recorded at least one exercise activity on 27% of days, and either tracked food or exercise and remained within their daily caloric budget on 49% of days. Table 2 column 1 presents estimates from a linear probability model of whether users upgraded to the premium version of the mHealth app within the first seven weeks, using the full sample of 11,873 users. The model relates upgrade decisions to observable user characteristics, including demographics, start date, initial goal settings, and first-week engagement. These estimates suggest that a typical premium user can be profiled. In particular, users who upgrade tend to be those who have are older in age, aim to lose a higher proportion of weight, and logged more exercise calories in the first week. However, users who have consistently tracked their food calories in the first week appear less likely to be premium adopters.

Users pay $39.99 for access to the premium version for a year.[6] It is important to note that while the free and premium versions share the core tracking features, including food calorie tracking, exercise calorie tracking, weight tracking, goal setting, and calorie budgeting, the premium version offers additional enhancements such as intuitive food logging, a smart camera for food tracking, detailed nutritional analytics, personalized goal tracking, sync capabilities with other health apps and wearable devices, community support, and more advanced tools for meal and exercise planning, along with gamified rewards and celebrations. These enhancements are designed to provide users with richer insights into their health behaviors and boost engagement through more tailored feedback and recommendations.

3 Staggered DiD Design

Assessing the impact of premium adoption on user engagement and health-related outcomes is challenging due to the potential endogeneity of the premium adoption decision. Moreover, the effect of premium adoption may vary across users. For example, users who are inherently more motivated might be those who are more likely to adopt the premium version as well as be more engaged on the app. Therefore, this inherent and unobserved heterogeneous motivation would confound the actual impact that the premium version has on user engage-

[5] We define the average normalized distance to the goal weight as the distance to the goal weight divided by the starting weight for each user.

[6] During our observation period, the service policy allowed users to cancel the yearly subscription at any time, but they would not receive a refund for unused periods after cancellation. Essentially, all fees and charges were non-refundable at that time.

ment dynamics. Moreover, the users' adoption decision happens at different timing. Thus, understanding the causal effect of premium adoption on user engagement and health-related outcomes requires an empirical framework that accounts for staggered treatment timing and potential treatment heterogeneity. To this end, we use a staggered DiD design that accommodates variation in the timing of premium adoption across users. The methodological approach follows recent advances in staggered DiD estimation, ensuring that treatment effects are properly identified while addressing potential biases arising from heterogeneous adoption dynamics.

We begin by formalizing the empirical problem using the potential outcomes framework, introducing key notation, and defining treatment assignment in a way that reflects the staggered nature of adoption. The empirical analysis primarily focuses on the method developed by Callaway and Sant'Anna (2021), which explicitly models group-time average treatment effects (GATT), thereby allowing us to trace the dynamic evolution of treatment effects.

3.1 Set-Up and Notation

We now formulate our empirical problem using the canonical potential outcomes framework for staggered DiD. The treatment of interest is *premium adoption*, and once a user adopts the premium version, they remain treated in all subsequent periods.

Let G_i denote the *group* that user i belongs to, which corresponds to the time period in which they first adopt the premium version. Specifically, $G_i = g$ indicates that the user first adopts the premium version in period g. For instance, $G_i = 2$ represents users who adopted the premium version in the second week since registering for the app. We define a binary variable $D_{i,g}$ that equals 1 if user i is first treated in period g:

$$D_{i,g} = \mathbb{1}\{G_i = g\}.$$

We index time periods by t, where t represents the number of weeks since the user registered for the app. In particular, a user at $t = 3$ corresponds to being in their third week since registration. Some users never adopt the premium version during the observation period; for such users, we define their group as $G_i = \infty$, which means that they remain untreated throughout.

The primary outcomes of interest, as denoted by $Y_{i,t}$, include user engagement and health. In our empirical setting, user engagement is measured through indicators such as food tracking, exercise tracking, and adherence to a daily calorie budget. Specifically, the food tracking indicator reflects whether a user logs their food intake on a given day, while the exercise tracking indicator captures whether exercise activity is recorded. Additionally, we consider staying

within the daily calorie budget, which measures whether a user's total calorie intake remains within their recommended limit, as well as exercise calories tracked, a continuous variable representing the total daily calories burned through physical activity. Beyond engagement metrics, we also examine health-related outcomes to assess whether increased engagement translates into tangible health benefits. The weight loss indicator is a binary variable that tracks whether a user's recorded weight is lower than their initial weight, whereas the amount of weight lost (in lbs) provides a continuous measure of cumulative weight reduction. Together, these outcome measures enable us to evaluate the extent to which premium adoption influences both user engagement behaviors and long-term health improvements. By including these variables, we determine whether upgrading to the premium version leads to increased interactions with the app and measurable progress toward health goals.

For each user i at time t, we define two potential outcomes $Y_{i,t}(0)$ being the potential outcome under no premium adoption, and $Y_{i,t}(1)$ being the potential outcome under premium adoption. Each user contributes only one observed outcome:

$$Y_{i,t} = D_{i,g}Y_{i,t}(1) + (1 - D_{i,g})Y_{i,t}(0).$$

For users who never adopt the premium version ($G_i = \infty$), we observe $Y_{i,t} = Y_{i,t}(0)$ in all periods. For users who adopt the premium version in period g, the observed outcomes correspond to untreated potential outcomes prior to g and treated potential outcomes from g onward.

3.2 Estimation Strategy

To estimate the causal effect of premium adoption on user engagement and health-related outcomes, we employ a widely used DiD estimator that explicitly accounts for treatment heterogeneity and dynamic effects over time. In particular, we use the estimator developed by Callaway and Sant'Anna (2021), which introduces a framework for defining group-time average treatment effects (GATT). This approach facilitates the estimation of treatment effects at various points in time and enables their aggregation to form an overall effect. Our estimation strategy follows best practices outlined by Baker et al. (2025), who emphasize the importance of accounting for treatment effect heterogeneity in staggered adoption settings. In line with these recommendations, we avoid traditional two-way fixed effects (TWFE) estimators, which can introduce bias in the presence of treatment effect dynamics, and instead employ an estimator that directly addresses variation in treatment timing and effect heterogeneity. This methodological choice ensures that our findings provide a more accurate depiction of how engagement and health outcomes evolve following premium adoption.

Importantly, our application of the Callaway and Sant'Anna (2021) framework directly addresses the limitations of naive regression approaches. Standard TWFE or naive DiD regressions typically assume homogeneous treatment effects and aggregate dynamic effects into a single parameter, thereby failing to capture the evolution of the treatment impact, such as the initial surge in engagement following premium adoption and its subsequent decay as users adapt. Consequently, these naive regressions may yield biased or misleading estimates, with potential issues like negative weighting when treatment effects vary across cohorts.

The approach by Callaway and Sant'Anna (2021) decomposes the overall effect into group-time average treatment effects, explicitly allowing treatment effects to vary with the time elapsed since adoption. This dynamic specification not only reflects the short-run increase in engagement attributable to premium features but also captures the gradual attenuation over time. By adopting this framework, we obtain a more nuanced and accurate understanding of how the impact of premium adoption evolves.

The core feature of their estimator is the GATT, which measures the effect of treatment for users who adopt in a given period and are observed at a specific time afterward. Formally, for users who adopt premium in period g and are observed at time t, the treatment effect is defined as:

$$ATT(g,t) = \mathbb{E}\left[Y_{i,t}(1) - Y_{i,t}(0) \mid G_i = g\right]$$

This approach allows treatment effects to be estimated separately for each adoption cohort and for each time period after adoption, accommodating heterogeneity in how treatment effects evolve over time. In particular, the estimation is implemented via the following steps:

Partitioning users into cohorts. Each user is assigned to a treatment cohort $G_i = g$ based on the period in which they first upgrade to premium. Users who never adopt the premium version during the observation period are classified as never-treated ($G_i = \infty$) and form part of the baseline control group.

Estimating group-specific treatment effects. For each treatment cohort g, GATT is estimated by comparing the observed outcomes of premium users with not-yet-adopters in the same time period. This is formally expressed as:

$$\hat{ATT}(g,t) = \frac{1}{N_g} \sum_{i:G_i=g} Y_{i,t} - \sum_{i:G_i=\infty} \hat{w}_i(g) \cdot Y_{i,t} \tag{1}$$

where N_g is the number of treated individuals in cohort g, and $\hat{w}_i(g)$ are the estimated inverse probability weights that adjust for differences in observed covariates X_i between

treated and untreated individuals. The weights help account for differences in observed characteristics between adopters and not-yet-adopters. This technique adjusts for selection biases by reweighting the not-yet-adopters so as to better resemble the premium users in terms of observed covariates.

Aggregating group-specific estimates. Once the group-specific treatment effects are estimated, the overall Average Treatment Effect on the Treated (ATT) is obtained by aggregating across all treatment cohorts. The weighted ATT is computed as:

$$\hat{ATT} = \sum_{g,t} w_g \hat{ATT}(g,t) \tag{2}$$

where the weights w_g reflect the relative size of each treatment group, ensuring that the aggregate effect appropriately accounts for varying cohort sizes.

Estimating standard errors and confidence intervals. To assess statistical significance and variability, standard errors are computed via a nonparametric bootstrap procedure. Additionally, we report pointwise confidence intervals and fixed-length confidence intervals (FLCI), following the Rambachan and Roth (2023) sensitivity analysis approach, to evaluate robustness against deviations from the conditional parallel trends assumption.

Event-study representation. To visualize how treatment effects evolve over time, we estimate event-time treatment effects by normalizing pre-treatment outcomes and tracking changes in post-treatment periods. This yields dynamic treatment effect curves, illustrating how premium adoption influences engagement patterns over time. These event-study plots help assess whether treatment effects persist, decay, or exhibit delayed impacts, providing deeper insights into user behavior following premium adoption.

3.3 Summary of TWFE Results

Before turning to our preferred estimation approach, we begin with a descriptive benchmark based on a two-way fixed effects (TWFE) specification. While the TWFE framework offers a convenient way to summarize average post-adoption shifts in engagement, it does not account for treatment heterogeneity across adoption cohorts. In particular, if the timing of premium adoption correlates with unobserved characteristics, such as latent motivation or prior engagement trajectories, then pooled estimates may obscure meaningful variation in treatment effects. Moreover, the standard TWFE estimator assumes constant treatment effects across time and cohorts, which may not hold in settings with staggered adoption.

To formalize the empirical approach, we estimate the following event-study specification:

$$Y_{it} = \sum_{\ell} \delta_{\ell} D_{i,t}^{\ell} + \alpha_i + \gamma_t + \varepsilon_{it}, \tag{3}$$

where Y_{it} denotes the outcome of interest for user i in calendar week t; $D_{i,t}^{\ell}$ is an event-time indicator equal to one if the observation falls ℓ weeks before/during/after the user's premium upgrade; α_i and γ_t are user and calendar week fixed effects, respectively. Users who never adopt premium are included as untreated controls.

This specification enables us to track behavioral changes in the weeks surrounding premium adoption while controlling for individual-level and temporal confounds. However, as emphasized by Baker et al. (2025), TWFE estimates may conflate treatment effects across cohorts and time horizons in the presence of treatment effect heterogeneity. For this reason, we view the TWFE results as a preliminary diagnostic and rely on the group-time average treatment effects (GATT) framework in subsequent sections to more flexibly capture dynamic treatment responses.

Table D1 in the Online Appendix presents the results from TWFE specification, which provides a baseline assessment of how premium adoption influences user engagement. For each user, we assess engagement using four metrics: (1) the proportion of days per week the user tracked at least one exercise activity; (2) the proportion of days per week the user tracked calories for at least one food item; (3) the average number of exercise calories tracked per week; and (4) the proportion of days per week the user either tracked food or exercise and remained within their daily calorie budget.

Note that all of our results use estimation samples that remove immediate adopters (i.e., those adopted in week 1), and thus are estimated on those adopted in weeks 2-7 and non-adopters. This sample construction ensures that there is at least one pre-treatment period, so as to satisfy the requirements of any DiD estimators (including Callaway and Sant'Anna (2021), which we explain below).

While our preferred approach will ultimately rely on Callaway and Sant'Anna (2021) , examining the naive TWFE estimates serves as a useful benchmark. These estimates capture the immediate and sustained changes in engagement post-adoption while controlling for individual and time-invariant confounds. However, as with any DiD specification that pools across treatment cohorts, TWFE estimates may obscure treatment heterogeneity across adoption cohorts, motivating the need for the GATT approach.

The TWFE estimates reveal a sharp increase in engagement during the week of premium adoption across all four outcomes: exercise tracking, food tracking, adherence to caloric budgets, and exercise calories. This immediate spike is expected, as premium users gain

access to enhanced tracking tools and are likely more motivated following their upgrade. However, engagement levels remain elevated in the short and medium term, though with slight attenuation over time. Notably, engagement does not fully return to pre-adoption levels even after 14 weeks, suggesting that at least some behavioral change persists.

The health-related outcomes in Table D2 offer a complementary perspective on the impact of premium adoption. For each user, we define weight loss outcomes using two metrics: (1) the proportion of days per week the user exhibited weight loss relative to their initial weight at registration and (2) the average amount of weight lost (in pounds) compared to their initial weight per week.

While premium users are more likely to experience positive weight loss post-adoption, the magnitude of weight loss exhibits a similar pattern of initial improvement followed by gradual stabilization. This suggests that while increased tracking may correlate with healthier behaviors, it does not necessarily translate into sustained weight loss improvements. Moreover, the pre-trend estimates indicate that premium adopters already exhibit some weight loss improvements in the weeks leading up to adoption, which underscores the importance of accounting for selection effects.

These results highlight both the potential and limitations of TWFE estimates in assessing the impact of premium adoption. The observed engagement increases suggest that premium features drive meaningful short-term behavioral changes. However, the attenuation of these effects over time raises questions about the long-term efficacy of premium adoption as a standalone engagement strategy. In our context, where later adopters show rising pre-adoption engagement, the TWFE estimates are potentially upward-biased. These findings underscore the importance of accommodating for treatment heterogeneity, which we explore in greater depth using the subsequent discussion about the Callaway and Sant'Anna (2021) results. While no observational design can fully eliminate selection concerns, the staggered DiD framework provides a more conservative benchmark that explicitly accounts for treatment timing heterogeneity.

3.4 Summary of Callaway and Sant'Anna (2021) Results

Our discussion is organized based on the outcomes that relate to engagement (Figure 2), and the outcomes that relate to health (Figure 3). We use the same set of engagement and health-related outcomes as in the TWFE specification.

To adjust for potential selection bias in users' premium upgrade decisions, we implement a doubly robust estimator within the Callaway and Sant'Anna (2021) framework. Specifically, we estimate the likelihood of upgrading to the premium version using a logistic classifier. The

model incorporates a range of conditioning variables, including user demographics (starting weight, initial BMI, height, age, and gender), initial goal settings (goal weight and goal BMI), early engagement patterns (engagement and weight loss outcomes in the first week since registration).

Further, premium adopters may differ from non-adopters along unobserved dimensions (e.g., intrinsic motivation, self-discipline) that could be jointly related to both their decision to upgrade and their subsequent engagement or health outcomes.

To address these concerns more directly, we enrich the set of conditioning variables used in the logistic classifier with measures capturing plausibly exogenous short-run fluctuations in the informational environment faced by users at the time of registration. Specifically, we construct weekly measures of national Google Trends search intensity for mHealth- and premium-related keywords, which proxy for changes in the salience of premium availability driven by nationwide factors such as media coverage, app-store promotion, wellness awareness, or advertising cycles. By capturing aggregate attention shocks that influence adoption decisions, these variables help account for systematic differences across cohorts in the propensity to upgrade that may differentially shape counterfactual engagement trends for users registering at different points in time. In other words, national search intensity is plausibly unrelated to individual users' latent motivation, so conditioning on it helps isolate adoption timing variation that is potentially driven by external salience than by intrinsic motives.

Incorporating these attention measures into the conditioning set strengthens the plausibility of the conditional parallel trends assumption by improving the comparability of treated users and not-yet-treated controls. In the inverse-probability weighting step of the Callaway and Sant'Anna (2021) framework, the inclusion of these variables shifts identifying variation toward comparisons in which adoption timing is more likely driven by external salience rather than by evolving private predispositions.

Column 2 of Table 2 reports a linear regression of the upgrade decision on user demographics, initial goal settings, early engagement behavior, and measures constructed from the weekly Google Trends Search Index, which ranges from 0 to 1. We focus on three keywords: "app name + premium," "app name + weight loss", and "weight loss premium." The first term directly reflects awareness of the premium version of the app, whereas the second captures broader attention to the focal app, and the third proxies for general interest in premium weight-loss tools. For each user, we compute the average national search intensity for these keywords during the first week after registration, thereby capturing the informational environment users face when they initially learn about the app and its upgrade options.

The estimates in Table 2 column 2 indicate that higher search intensity for "app name

+ premium" in the first post-registration week is positively and significantly associated with the probability of upgrading during weeks 2–7, even after conditioning on demographics, initial goal settings, calendar time, and early engagement behavior. This finding is informative for two reasons. First, it confirms that variation in public attention maps into meaningful variation in adoption behavior. Second, incorporating this information improves the estimation of the first-stage adoption propensity. By doing so, the inverse-probability weighting procedure places greater emphasis on periods in which upgrading is more likely to be triggered by external visibility rather than by latent user traits, thereby reducing the extent to which selection on unobservables contaminates the identifying variation.

Table 3 presents the overall ATT estimates. The results show that the TWFE estimates can overestimate the overall ATT compared to estimates from Callaway and Sant'Anna (2021). Also, the upgrading generally leads to more engagement (food tracking, exercise tracking, staying within the budget, exercise calories), while the effects on weight outcomes are statistically significant but small in magnitude.

For the cohort-level results, the effects of premium upgrades are consistent across all four engagement outcome variables. As illustrated in Figure 2, the proportion of days per week a user tracks food, tracks exercise, and remains within their daily calorie budget, along with the average exercise calories tracked per week, all show an immediate post-adoption spike, followed by a gradual reversion toward baseline levels in the weeks that follow.

Compared to users who have not upgraded, those who upgrade are 11.4 percentage points more likely to track exercise during the first week, with the effect decreasing to 7.9 percentage points by the seventh week and losing statistical significance thereafter. This indicates that the initial motivation to monitor physical activity, likely boosted by premium functionalities, diminishes over time.

A similar temporal pattern emerges for food tracking. Premium adopters are 10.3 percentage points more likely to track food in the first week following adoption. This effect steadily declines to 7.8 percentage points by the seventh week and becomes statistically insignificant after the eighth week. These results suggest that while premium features initially promote stronger dietary tracking behavior, this effect also wanes relatively quickly.

The pattern persists in users' adherence to their daily calorie budget. In the first week, premium adopters are 7.1 percentage points more likely to both track and remain within their calorie budget. However, this effect decreases to 6.4 percentage points by the eighth week and becomes statistically insignificant thereafter. These findings point to a short-term improvement in caloric discipline following premium adoption, but also underscore the difficulty in sustaining consistent adherence over the long term.

Moreover, premium adoption also leads to an initial increase in exercise calorie tracking,

with users recording an additional 49.2 exercise calories in the first week. This increase declines to 42.8 calories by the seventh week and becomes statistically insignificant thereafter. The observed short-term spike suggests that premium features may initially encourage more diligent tracking of physical activity, though this effect, like the others, gradually fades.

Ultimately, these patterns reveal a consistent short-term boost across multiple engagement metrics following premium adoption, followed by a gradual return toward baseline levels. This pattern points to the difficulty of translating short-term behavioral changes induced by premium features into sustained long-term habits.

The weight loss outcomes present an interesting contrast to the engagement metrics. Specifically, there is no significant immediate effect following premium adoption. However, by the third week post-adoption, users who upgraded to the premium version report, on average, 0.68 pounds more weight loss compared to those who did not upgrade. This difference grows to 1.34 pounds by the tenth week but becomes statistically insignificant thereafter. The loss of significance beyond this point may be attributed to premium users discontinuing weight updates in the app, which aligns with the observed return of engagement levels to baseline around the ninth week after adoption. Furthermore, there is no significant difference between premium and non-premium users in the probability of maintaining a weight lower than their initial weight within any week throughout the observation period. This suggests that the observed larger weight loss amount among premium users may be driven by a small subset of individuals experiencing substantial weight reductions. Overall, these results indicate that while premium adoption may yield modest, delayed improvements in weight loss outcomes, the overall impact remains limited.

The findings highlight the role of premium adoption as a potentially relevant intervention for weight loss. However, the lack of a strong weight loss effect despite increased engagement and efforts reinforces the importance of sustaining long-term behavioral changes in intervention design.

To get a sense of the potential bias from using TWFE, we now compare the estimates across these approaches. Note that the TWFE estimates reported in Tables D1 and D2 provide a baseline assessment of how engagement and weight outcomes evolve following premium adoption. These estimates point to substantial increases across all outcomes, with effects appearing both immediate and persistent. However, when compared to the Callaway and Sant'Anna (2021) estimates shown in Figures 2 and 3, we see that the TWFE estimates tend to overstate both the magnitude and duration of the post-adoption response.

For engagement outcomes, TWFE suggests relatively large and stable increases in food and exercise tracking that persist for at least three months. By contrast, the Callaway and Sant'Anna (2021) estimates reveal that while engagement does rise sharply following

premium adoption, these effects attenuate more quickly than the pooled TWFE estimates imply. This pattern is particularly evident for exercise calories and budget adherence, where the Callaway and Sant'Anna (2021) effects begin to dissipate within several weeks.

The discrepancy is even more pronounced for weight-related outcomes. While TWFE estimates in Table D2 suggest steady gains in both weight loss probability and amount lost, the Callaway and Sant'Anna (2021) estimates in Figure 3 indicate that these effects are considerably smaller, with confidence intervals that widen over time. By the tenth week post-adoption, the Callaway and Sant'Anna (2021) estimates for weight loss amount are no longer statistically distinguishable from zero. Moreover, throughout the entire observation period, there is no significant difference between premium and non-premium users in the probability of maintaining a weight lower than their initial weight.

4 Sensitivity Analysis and Robustness

In this section, we examine the robustness of our staggered DiD estimates by assessing potential violations of key identification assumptions. For additional analysis that explores alternative subsamples of adoption cohorts (Section A), sensitivity of our results to food/weight log accuracy (Section B), as well as estimator choice (Section C), we refer readers to the Online Appendix.

Before describing the sensitivity analysis, we first describe the relevant assumptions underlying the Callaway and Sant'Anna (2021) estimator. Specifically, we explore the plausibility of the *limited anticipation* and *conditional parallel trends* assumptions in Callaway and Sant'Anna (2021). Given the possibility that adoption timing may be influenced by pre-adoption engagement or health-related outcomes, we present descriptive analyses to characterize potential selection dynamics and motivate the use of sensitivity tests (Rambachan and Roth 2023) to further assess validity.

4.1 Assumptions and Potential Biases

The Callaway and Sant'Anna (2021) estimator relies on several key assumptions related to the nature of treatment assignment, potential outcomes, and identification strategy. These assumptions collectively ensure that the estimated treatment effects meaningfully capture the impact of premium adoption on user engagement and health-related outcomes.

Some assumptions, such as the *irreversibility of treatment* and *random sampling*, are relatively straightforward in our empirical setting and are likely to hold. Others, such as *limited treatment anticipation* and *conditional parallel trends*, require closer scrutiny to assess

whether unobserved confounders or dynamic treatment effects could bias our estimates. Additionally, the *overlap assumption* ensures that treatment assignment remains well-supported across observed covariates, which we validate by examining propensity score distributions. Below, we formally state each assumption and discuss its relevance in the context of our empirical setting.

Irreversibility of treatment. This assumption states that once a user adopts the premium version, they do not switch back to the basic version in subsequent periods. In our data, premium subscriptions are annual, and we do not observe any users reverting to the free version within our entire observation period. Thus, we believe this assumption holds.

Random sampling. This assumption states that conditional on observables, the treatment assignment is as good as random.

Limited treatment anticipation. This assumption posits that users do not fully anticipate their eventual adoption of the premium version before actually adopting. If anticipation occurs, it could influence pre-adoption engagement levels. To investigate this, we examine pre-trends in user engagement. Specifically, we discuss the extent to which engagement outcomes differ in the weeks leading up to adoption.

Conditional parallel trends based on never-treated group. This assumption requires that, conditional on covariates, the outcome trajectories of never-treated users provide a valid counterfactual for treated users. Given potential deviations from this assumption, we perform robustness checks to assess the sensitivity of our results.

Conditional parallel trends based on not-yet-treated groups. This assumption extends the previous one to users who eventually adopt but have not yet done so in a given period. As with the previous assumption, we discuss possible violations and conduct the relevant sensitivity checks.

Overlap. This assumption requires that each adoption cohort contains sufficient observations and that the propensity scores are non-degenerate. To validate this, we plot the distribution of predicted propensity scores for each cohort.

In Figure 4, each plot corresponds to a different adoption cohort (weeks 2 to 7), showing the density of the predicted propensity score. The densities confirm that the propensity scores are non-degenerate across adoption cohorts.

4.2 Pre-Adoption Trajectories

An important concern in our empirical setting is whether the timing of premium adoption is endogenous to user behavior. To explore these selection dynamics, we examine pre-adoption trends in our outcomes of interest across different adoption cohorts. To better understand how engagement and health-related outcomes evolve across different user types, we categorize individuals into four groups based on their adoption timing.

The first group, early adopters, consists of users who upgrade to the premium version within the first week of app registration. These users, represented by the red line in our analysis, likely demonstrate strong initial interest in the premium features and may differ in their motivations from later adopters. The second group, middle adopters, includes users who upgrade in weeks 2 and 3. These individuals may have spent some time exploring the free version before deciding to invest in premium features, suggesting a more deliberative adoption pattern. The third group, late adopters, represents users who upgrade in weeks 5, 6, and 7. Their delayed adoption suggests that they may have taken longer to assess the app or may have been gradually increasing their engagement with the app before making the decision to upgrade. Finally, never adopters (purple line) are users who continue using the free version throughout the observation period. This group is an important comparison set: their engagement and health trajectories let us assess how premium adoption affects user behavior relative to those who never upgrade.

Figure 5 depicts pre-adoption outcome trajectories for these groups, with vertical dashed lines indicating the time of upgrade for each cohort. Each plot presents the mean outcome for each group, first averaged within users per week and then across users over each week since registration. First, the engagement-related outcomes reveal systematic pre-trend dynamics, indicating the presence of potential selection biases in the timing of premium adoption. Notably, middle and late adopters display a clear upward trajectory in engagement leading up to adoption. This pattern suggests that these users may have been increasingly motivated to engage with the app prior to upgrading, which could bias our estimates if not properly accounted for. In contrast, never adopters follow a downward trajectory, indicating a gradual decline in engagement over time. This trend suggests that users who do not upgrade may systematically differ from those who do, further complicating causal inference by introducing potential confounding factors related to intrinsic motivation and app usage behavior.

Interestingly, early adopters and never adopters exhibit relatively similar decreasing trends, though early adopters consistently maintain higher levels of food tracking compared to never adopters. This similarity implies that early adopters may have different adoption motives as compared with middle and late adopters, reinforcing the need to account for

heterogeneous treatment effects when interpreting the impact of premium adoption. These patterns indicate that middle and late adopters might have greater intrinsic motivation to upgrade, leading to a *positive bias* in estimated treatment effects if unaccounted for.

The pre-trend dynamics for health-related outcomes differ from those observed for engagement metrics. Notably, the probability of weight loss leading up to adoption appears similar across all adoption cohorts, with each group displaying an upward trend. This suggests that, regardless of when users upgrade to the premium version, many were already experiencing gradual weight loss in the periods prior to adoption. However, when considering the amount of weight lost, a distinct pattern emerges. Late adopters exhibit a noticeably steeper upward trend in weight loss compared to both middle adopters and never adopters. This suggests potential inertia in weight loss efforts, where users who delay premium adoption may have already been making sustained progress toward their weight loss goals before upgrading.

The pronounced pre-adoption progress among late adopters implies that the estimated premium adoption effects for this cohort may appear more persistent than for early or middle adopters. If late adopters were already on a strong weight loss trajectory, their post-adoption engagement and outcomes might reflect an extension of pre-existing behavioral patterns rather than a direct causal effect of premium adoption. This highlights the importance of carefully accounting for pre-trends when interpreting treatment effects on health-related outcomes.

The observed pre-trends raise important concerns about the conditional parallel trends assumption in Callaway and Sant'Anna (2021). A key concern in our empirical setting is the potential endogeneity of premium adoption, particularly if users decide to upgrade based on their prior engagement levels. If more engaged users are systematically more likely to adopt the premium version, our treatment effect estimates may be overstated, as the observed post-adoption engagement increases could partially reflect pre-existing behavioral patterns rather than a causal impact of premium adoption. Furthermore, the presence of selection dynamics implies that the estimated effects of the treatment may differ depending on the time of adoption. Users who adopt early may have different motivations and engagement trajectories compared to those who adopt later, potentially leading to heterogeneous treatment effects in adoption cohorts. This variation raises challenges in interpreting the average effect of treatment, as it can mask important differences in how different groups respond to premium adoption.

To assess these concerns more systematically, we employ the Rambachan and Roth (2023) sensitivity test, which allows for controlled deviations from the parallel trends assumption in staggered DiD settings.

4.3 Formal Sensitivity Test

The robustness of our staggered DiD estimates depends on the extent to which deviations from the parallel trends assumption influence treatment effect estimates. To assess the stringency of this assumption for our empirical context, we implement the sensitivity analysis framework developed by Rambachan and Roth (2023), which provides a structured approach for evaluating potential violations of parallel trends in event-study-based staggered DiD settings (Roth et al. 2023). For our sensitivity analysis, we focus our analysis on the engagement outcomes, as the health-related outcomes are inconclusive to begin with even in the absence of parallel pre-trend violations.

Following Rambachan and Roth (2023), we define $\overline{M}$ as a deviation parameter that reflects linear departures from parallel trends and estimate fixed-length confidence intervals (FLCIs) for a range of $\overline{M}$ values. When $\overline{M} = 0$, the standard parallel trends assumption holds, whereas larger $\overline{M}$ values correspond to decreasing reliability about the validity of this assumption. Figure 7 presents FLCIs across varying $\overline{M}$, illustrating how sensitive our estimates are to deviations from parallel trends, the potential magnitude of selection bias, and whether engagement or health-related outcomes exhibit greater sensitivity.

The sensitivity analysis establishes the extent to which deviations from the parallel pre-trend assumption can be accommodated while maintaining statistically significant treatment effects. Results indicate that for exercise tracking, the estimated effects remain robust to assumption violations up to $\overline{M} \leq 1.6$, while for food tracking, the threshold is slightly lower at $\overline{M} \leq 1.1$. This suggests that the significant effects are robust to violations of the parallel trends assumption up to 1.6 and 1.1 times the maximum deviation observed during the pre-treatment period, respectively. Adherence to budget constraints and the exercise calories exhibit slightly larger sensitivity, with significance holding for deviations up to $\overline{M} \leq 1.0$. This result suggests that the significant effects are robust to violations of the parallel trends assumption up to 1.0 times the maximum deviation observed during the pre-treatment period. These patterns indicate that exercise and food tracking are relatively resilient to moderate violations, whereas budget adherence and exercise calories exhibit slightly greater dependence on strict pre-trend alignment. These findings underscore that while our results are robust to some degrees of violation of the parallel trend assumption, extreme violations will render the estimates to be insignificant.

4.4 Robustness to Time-Varying Selection

To further assess the role of time-varying selection in shaping treatment timing, we complement the baseline staggered DiD design with a propensity score matching approach that

conditions on users' recent engagement trajectories and Google Search intensities, in addition to demographics, initial goal settings, and early engagement and weight-loss outcomes. The pre-adoption patterns documented earlier indicate that many adopters upgrade during periods of rising activity, suggesting that short-lived motivational surges may jointly influence both the decision to upgrade and subsequent engagement. Because these dynamics reflect within-user, time-varying heterogeneity, a matching design that balances users on recent engagement histories and information exposure can help isolate variation that is more plausibly attributable to upgrading itself.

In this robustness exercise, we focus on adopters in weeks 5 and 6 (102 users), where we observe both a sufficient sample size and clear evidence of pre-adoption increases in activity. We estimate a propensity score of premium adoption based on demographic characteristics, initial goal setting, week 1 engagement, Google Search intensities, and a set of variables capturing recent engagement levels and weight-loss outcomes in weeks 3 and 4. Each adopter is matched to ten non-adopters with comparable recent engagement dynamics. This procedure yields a control group that experiences similar short-run momentum at the time treatment occurs, and the key matching variables are well balanced in the resulting sample. We then estimate a two-way fixed-effects model on the matched sample to recover event-study coefficients that adjust for these time-varying differences.

The matched-sample results in Figure 8 remain broadly consistent with the baseline findings. For exercise tracking, food tracking, caloric budgeting, and exercise calories, we continue to observe an immediate increase following the upgrade, followed by a gradual attenuation over subsequent weeks. Pre-treatment differences are also noticeably flatter in the matched specification, indicating that conditioning on recent engagement trajectories successfully balances short-run trends prior to treatment. Weight-related outcomes display a similar degree of robustness. The trajectory of weight loss amounts continues to show a gradual buildup over time, whereas the probability of any weight loss remains largely unchanged relative to baseline. Across different empirical designs, the qualitative dynamic patterns are stable.

The matching analysis provides additional reassurance that the main results are not solely an artifact of short-lived motivational spikes that simultaneously affect adoption timing and behavior. Conditioning on recent engagement histories reduces sensitivity to such time-varying selection, yet the key dynamic features of the baseline analysis persist. We view these patterns as supportive of the interpretation that upgrading generates meaningful short-run adjustments in engagement, while still acknowledging the limits of observational data for establishing definitive point-identified causal magnitudes.

5 Potential Behavioral Mechanisms

5.1 Mechanism Analysis

We begin by positing behavioral mechanisms that may account for the empirical patterns observed in mHealth engagement dynamics. Specifically, we consider the roles of moral licensing, goal liberation, sunk costs, and hedonic decline as candidate drivers. We then develop an empirical test to distinguish between two competing explanations.

Moral licensing. The literature on moral licensing (Chiou et al. 2011, Khan and Dhar 2006, Sachdeva et al. 2009, Wilcox et al. 2009) suggests that after engaging in virtuous behavior, individuals may feel justified in subsequently reducing effort or engagement. If subscribing to the premium version of the mHealth app is perceived as a virtuous health commitment, users may grant themselves implicit permission to decrease tracking efforts post-upgrade. This mechanism predicts a *decline* in engagement following premium adoption. However, because our empirical results reveal a *short-term increase* in engagement immediately after upgrading, moral licensing does not appear to be the primary driver of the observed effects.

Goal liberation. The goal liberation hypothesis (Fishbach and Dhar 2005) posits that individuals who perceive themselves as having made progress toward a goal may subsequently reduce effort toward achieving it. If premium adoption signals progress toward health and wellness goals, users may experience a reduced need to actively engage with tracking features, leading to a decline in engagement. Yet, given the observed immediate engagement increase post-adoption, goal liberation is unlikely to be the dominant force.

Sunk costs. An alternative explanation is that premium adoption enhances subsequent engagement through the sunk-cost effect. Upon paying for the premium version, users may feel psychological pressure to justify their financial investment by increasing engagement. This mechanism can generate an immediate engagement boost following premium adoption, although the lift is inherently transient. The transient nature of sunk-cost-driven engagement has been widely documented across diverse empirical contexts (Arkes and Blumer 1985, Augenblick 2016, Camerer and Weber 1999, Gourville and Soman 1998, Ho et al. 2018, Goli et al. 2022).

Hedonic decline. A related but distinct mechanism is hedonic decline, whereby repeated exposure to the same experience yields diminishing marginal responses (Brickman 1971,

Galak and Redden 2018, Sevilla et al. 2019). In the context of premium adoption, hedonic adaptation implies that the novelty and perceived value of premium features attenuate over time, leading to a gradual reduction in engagement. This framework is consistent with our empirical findings: premium adoption initially elevates engagement by offering access to novel features, but user engagement eventually decays with continued exposure.

Among the mechanisms discussed above, both hedonic decline and the sunk-cost effect predict an immediate boost in engagement from premium adoption, followed by a gradual decline. To determine which mechanism plays a more dominant role, we exploit the fact that the free and premium versions of the app share numerous core user-interactive features, with premium features largely building upon the free version. In particular, the user interface design remains consistent across versions with respect to food and exercise tracking. That is, both the aesthetic design and the presentation of consumption and exercise information are virtually identical across versions.[7] Consequently, users who extensively interacted with the free version before upgrading would likely have already experienced many features similar to those of the premium version.

Hedonic decline predicts that the engagement lift should be lower among users with high pre-upgrade exposure, as the novelty of the premium features would be attenuated. In contrast, the sunk-cost effect implies that the engagement lift should be similar across users, because all adopters incur the same financial outlay and may feel compelled to justify the expense. Distinguishing between these mechanisms has important implications for designing interventions to sustain user engagement over time.

To assess the relative importance of hedonic decline versus the sunk-cost effect, we segment users based on their pre-adoption exposure and estimate separate treatment effects for each subgroup. Specifically, premium adopters are classified as *low*- or *high*-exposure users based on whether their pre-upgrade tracking frequency falls below or above the sample median. Each subgroup is then separately grouped with non-adopters, and we implement staggered DiD analyses comparing low-exposure adopters versus non-adopters, and high-exposure adopters versus non-adopters. If the effects are statistically indistinguishable and of similar magnitude across the two groups, it would suggest that hedonic decline is unlikely to be the primary driver.

To sharpen inference about underlying mechanisms, we focus on immediate and short-term treatment effects: the week of adoption and the first week thereafter. This design choice is motivated by two considerations. First, engagement effects measured during this

[7]The primary difference between the premium and free versions lies in the display of nutritional information: the free version presents only total food and exercise calories, whereas the premium version provides a more detailed breakdown, including macronutrients and micronutrients (e.g., carbohydrates, fats, proteins, vitamins).

early window are less likely to be confounded by downstream outcomes, such as weight loss, which typically materialize with delay. Second, isolating early effects provides a cleaner lens into users' immediate behavioral responses to premium adoption, before habit formation or reinforcement mechanisms influence usage trajectories. In each case, treatment effects are estimated separately for low- and high-exposure adopters to test the competing predictions from hedonic decline and sunk-cost theories.

Table 4 presents summary statistics for key user characteristics across low- and high-exposure adopters, along with corresponding t-statistics and p-values from two-sample mean tests. By construction, low-exposure adopters tracked food or exercise for an average of 9.7 days prior to upgrading, significantly fewer than high-exposure adopters, who tracked for an average of 27.3 days before upgrading. However, the two groups are otherwise comparable across a range of characteristics, including demographics, starting weight, and goal progress on the day of upgrade, suggesting that differences in the effect between the groups are unlikely to be driven by these characteristics.

Table 5 presents the estimated causal effects on engagement outcomes separately for low- and high-exposure adopters. Across most outcomes, low-exposure adopters exhibit a significantly larger post-upgrade lift compared to their high-exposure counterparts, particularly during the first week after adoption. For instance, food tracking increases by 0.19 among low-exposure adopters compared to 0.04 among high-exposure adopters. Similar patterns emerge for exercise tracking and caloric budgeting outcomes. Although week-of-upgrade effects are directionally consistent, the magnitude separation becomes even more pronounced by the first week post-upgrade, suggesting that engagement among high-exposure adopters may decline more rapidly.

These findings are broadly consistent with hedonic decline as the behavioral mechanism. Users with limited prior exposure to the free version respond to the premium upgrade as if encountering genuinely new features, whereas users who have already interacted extensively with those same core features arrive at the upgrade having largely exhausted whatever novelty the premium tier can offer. The attenuation of the engagement lift with prior exposure is precisely what hedonic decline predicts: the marginal stimulus from a new experience declines with accumulated exposure to similar ones. The sunk-cost interpretation is unlikely the primary driver of this pattern. With sunk-cost effects, the act of paying the premium fee creates psychological pressure to justify the expenditure through increased engagement, and this pressure should be uniform across users because all adopters incur the same financial outlay regardless of their prior activity. The fact that high-exposure users exhibit substantially smaller engagement lifts (in many cases statistically indistinguishable from zero) is therefore difficult to reconcile with sunk costs as the primary driver.

5.2 Robustness of Mechanism Analysis

To further examine the robustness of the mechanism analysis, we restrict attention to users who upgraded in week 5 and study whether engagement responses vary systematically with prior exposure to the free version of the app. This focus allows us to assess whether the post-adoption decline in engagement is better explained by hedonic decline than by motivational mean reversion. If hedonic decline plays a meaningful role, one would expect the largest engagement responses among users with relatively low prior exposure, whereas users with high prior exposure to the free version before upgrading should show smaller, more transitory responses.

We begin by computing each user's pre-adoption exposure levels. Using a median split, we classify adopters into low- and high-exposure groups based on the number of days they tracked food or exercise during the four weeks preceding the upgrade. For each group, we estimate a propensity score of premium adoption using demographic characteristics, initial goal settings, recent Google Search intensities (week 4), recent engagement trajectories (weeks 3 and 4), and recent weight-loss outcomes (week 4). We then match each adopter to non-adopters whose pre-adoption patterns closely resemble those within the corresponding exposure group, ensuring that pre-treatment dynamics do not differ meaningfully across matched pairs. The resulting matched samples balance the key matching variables for both adopters and non-adopters.

The event-study results that emerge from this design in Figures 9 and 10 reveal a clear divergence across exposure groups. For all engagement measures, users with low prior exposure exhibit pronounced increases immediately following the upgrade, with effects that are both economically sizable and statistically precise. By contrast, high-exposure users display much smaller and often statistically indistinguishable changes, with post-adoption responses that remain close to zero throughout the analysis window. These differences are consistent with the interpretation that the perceived novelty of premium features is greater among users with limited prior exposure, triggering a more salient engagement response, whereas those who upgrade after already having high exposure to the free version experience a dampened response.

These findings provide additional support that hedonic decline, rather than motivational mean reversion, plays an important role in shaping the observed post-adoption dynamics. Because the analysis additionally matches users on recent engagement histories, both exposure groups begin with comparable motivational and behavioral trajectories to matched non-adopters. The fact that only low-exposure adopters exhibit large initial engagement responses that subsequently attenuate is therefore consistent with a decline in novelty effects.

In Figures D1 and D2 in the Online Appendix, we obtain very similar results when conducting the matching-based analysis for users who upgraded in week 6: low-exposure users exhibit sharp increases in exercise tracking, food tracking, caloric budget adherence, and exercise calories upon upgrading, followed by gradual attenuation, whereas high-exposure users display much smaller and generally statistically indistinguishable changes throughout the observable period.

In summary, the week-6 results further support prior exposure as an important channel underlying the effect of premium adoption on engagement: users with *limited exposure* to the free version respond *more strongly* upon upgrading, whereas highly exposed users exhibit muted responses, consistent with hedonic decline shaping post-adoption engagement dynamics.

6 Conclusion

This paper examines how premium adoption influences user engagement and health outcomes in an mHealth setting. Using a staggered difference-in-differences framework, we find that premium features yield an immediate but short-lived increase in mHealth app engagement, particularly in food and exercise logging. These engagement gains attenuate within weeks, consistent with hedonic decline. We observe limited downstream effects on weight loss, suggesting that elevated engagement does not necessarily translate into sustained health improvements.

To understand the behavioral mechanisms of this transient engagement response, we analyze heterogeneity in adoption effects by users' pre-upgrade exposure levels to the free version. The engagement lift is disproportionately larger among users with low prior exposure, whereas users who have already interacted extensively with the free version before upgrading exhibit more muted responses. This asymmetry aligns more closely with hedonic decline, where the novelty of premium features temporarily boosts engagement, than with sunk cost effects, as sunk costs would predict a similar engagement lift across users when they pay the same fee to upgrade. In particular, marginal engagement gains taper off quickly regardless of prior activity, implying that repeated exposure reduces the effectiveness of premium features in boosting engagement over time.

These findings have implications for digital health design. First, pricing-based one-time upgrades can generate short-run engagement increases but are unlikely to sustain long-run adherence in the absence of complementary reinforcement mechanisms. The estimated effects suggest that engagement gains following upgrading are transient and attenuate quickly within weeks. Second, the limited association between higher user engagement and weight-

loss outcomes cautions against equating increased engagement with long-term health improvement. The evidence indicates that elevated short-term tracking activity alone does not reliably translate into sustained weight changes in this setting. Finally, the exposure-based heterogeneity results suggest that novelty could be a relevant lever for the app. With the hedonic decline mechanism, prolonged free trials may reduce the incremental impact of upgrading by exhausting the novelty of premium features before users subscribe. Introducing new premium features over time may therefore be a more effective lever for sustaining behavioral change.

We conclude by discussing the limitations of our empirical framework and the scope of the conclusions that can be drawn from the data. As in many empirical applications of staggered DiD, our framework relies on a conditional parallel trend assumption. The pre-adoption engagement trajectories (particularly among middle and late adopters) indicate that users often upgrade during periods of rising engagement, consistent with time-varying selection. Although we condition on user characteristics, early and recent engagement measures and information exposure proxies, and implement formal sensitivity analyses with extensive robustness checks, these approaches mitigate but cannot fully eliminate concerns about endogenous treatment timing. Importantly, however, any remaining selection is likely to bias the estimates upward, as users exhibiting increasing engagement trajectories are more likely to upgrade. As a result, our findings should be viewed as conservative evidence for our central conclusion: analyses that ignore individual self-selection into premium adoption systematically overstate its effects. Accordingly, our contribution lies in documenting a transparent set of empirical results and articulating the limitations of TWFE estimators under endogenous adoption, rather than in claiming definitive point-identified causal magnitudes.

Funding and Competing Interests. All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript. The authors have no funding to report.

References

Appel G, Libai B, Muller E, Shachar R (2019) Retention and the monetization of apps. *International Journal of Research in Marketing* 20:79.

Aral S, Nicolaides C (2017) Exercise contagion in a global social network. *Nature communications* 8(1):1–8.

Arkes HR, Blumer C (1985) The psychology of sunk cost. *Organizational behavior and human decision processes* 35(1):124–140.

Arkhangelsky D, Athey S, Hirshberg DA, Imbens GW, Wager S (2021) Synthetic difference-in-differences. *American Economic Review* 111(12):4088–4118.

Aswani A, Kaminsky P, Mintz Y, Flowers E, Fukuoka Y (2019) Behavioral modeling in weight loss interventions. *European journal of operational research* 272(3):1058–1072.

Augenblick N (2016) The sunk-cost fallacy in penny auctions. *The Review of Economic Studies* 83(1):58–86.

Baker A, Callaway B, Cunningham S, Goodman-Bacon A, Sant'Anna PHC (2025) Difference-in-differences designs: A practitioner's guide.

Bojd B, Song X, Tan Y, Yan X (2022) Gamified challenges in online weight-loss communities. *Information Systems Research* .

Bollinger B, Liebman E, Hammond D, Hobin E, Sacco J (2022) Educational campaigns for product labels: Evidence from on-shelf nutritional labeling. *Journal of Marketing Research* 59(1):153–172.

Brickman P (1971) Hedonic relativism and planning the good society. *Adaptation level theory* 287–301.

Callaway B, Sant'Anna PH (2021) Difference-in-differences with multiple time periods. *Journal of econometrics* 225(2):200–230.

Camerer CF, Weber RA (1999) The econometrics and behavioral economics of escalation of commitment: A re-examination of staw and hoang's nba data. *Journal of Economic Behavior & Organization* 39(1):59–82.

Chintala SC, Liaukonyte J, Yang N (2023) Browsing the aisles or browsing the app? how online grocery shopping is changing what we buy. *Marketing Science* forthcoming.

Chiou WB, Yang CC, Wan CS (2011) Ironic effects of dietary supplementation: illusory invulnerability created by taking dietary supplements licenses health-risk behaviors. *Psychological science* 22(8):1081–1086.

De Chaisemartin C, d'Haultfoeuille X (2020) Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review* 110(9):2964–2996.

Faulkner C (2019) Fitbit's new premium subscription service hopes to sway you with personalized data, challenges, and more. *The Verge* .

Finkler E, Heymsfield SB, St-Onge MP (2012) Rate of weight loss can be predicted by patient characteristics and intervention strategies. *Journal of the Academy of Nutrition and Dietetics* 112(1):75–80.

Fishbach A, Dhar R (2005) Goals as excuses or guides: The liberating effect of perceived goal progress on choice. *Journal of Consumer Research* 32(3):370–377.

Galak J, Redden JP (2018) The properties and antecedents of hedonic decline. *Annual review of psychology* 69(1):1–25.

Ghose A, Guo X, Li B, Dang Y (2022) Empowering patients using smart mobile health platforms: Evidence from a randomized field experiment. *Management Information Systems Quarterly* .

Goli A, Chintagunta PK, Sriram S (2022) Effects of payment on user engagement in online courses. *Journal of Marketing Research* 59(1):11–34.

Gordon BR, Sun B (2015) A dynamic model of rational addiction: Evaluating cigarette taxes. *Marketing Science* 34(3):452–470.

Gourville JT, Soman D (1998) Payment depreciation: The behavioral effects of temporally separating payments from consumption. *Journal of consumer research* 25(2):160–174.

Haws KL, Liu PJ, Redden JP, Silver HJ (2017) Exploring the relationship between varieties of variety and weight loss: When more variety can help people lose weight. *Journal of Marketing Research* 54(4):619–635.

Ho TH, Png IP, Reza S (2018) Sunk cost fallacy in driving the world's costliest cars. *Management Science* 64(4):1761–1778.

Huang G, Khwaja A, Sudhir K (2015) Short-run needs and long-term goals: A dynamic model of thirst management. *Marketing Science* 34(5):702–721.

Huyghe E, Verstraeten J, Geuens M, Van Kerckhove A (2017) Clicks as a healthy alternative to bricks: how online grocery shopping reduces vice purchases. *Journal of Marketing Research* 54(1):61–74.

Kapoor A, Manchanda P, Narayanan S (2024) Can a human coach help you lose more weight than an ai coach: Empirical evidence from a mobile fitness tracking app. *Working paper* .

Kato-Lin YC, Abhishek V, Downs JS, Padman R (2016) Food for thought: The impact of m-health enabled interventions on eating behavior. *Available at SSRN 2736792* .

Khan R, Misra K, Singh V (2016) Will a fat tax work? *Marketing science* 35(1):10–26.

Khan U, Dhar R (2006) Licensing effect in consumer choice. *Journal of Marketing Research* 43(2):259–266.

Labonté K, Knäuper B, Dubé L, Yang N, Nielsen DE (2022) Adherence to a caloric budget and body weight change vary by season, gender, and bmi: an observational study of daily users of a mobile health app. *Obesity Science & Practice* .

Lambrecht A, Misra K (2017) Fee or free: When should firms charge for online content? *Management Science* 63(4):1150–1165.

Li H, Jain S, Kannan P (2019) Optimal design of free samples for digital products and services. *Journal of Marketing Research* 56(3):419–438.

Liu CW, Wang W, Gao G, Agarwal R (2024) The value of virtual engagement: Evidence from a running platform. *Management Science* 70(9):6179–6201.

Ma Y, Ailawadi KL, Grewal D (2013) Soda versus cereal and sugar versus fat: drivers of healthful food intake and the impact of diabetes diagnosis. *Journal of Marketing* 77(3):101–120.

Murnane EL, Huffaker D, Kossinets G (2015) Mobile health apps: adoption, adherence, and abandonment. *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*, 261–264.

Narang U (2022) Incentives to work out: Evidence from field experiments. *Working paper* .

Narang U, Shankar V (2019) Mobile marketing 2.0: State of the art and research agenda. *Marketing in a Digital World* .

Oblander S, McCarthy DM (2023) Frontiers: Estimating the long-term impact of major events on consumption patterns: Evidence from covid-19. *Marketing Science* .

Puranam D, Narayan V, Kadiyali V (2017) The effect of calorie posting regulation on consumer opinion: A flexible latent dirichlet allocation model with informative priors. *Marketing Science* 36(5):726–746.

Rambachan A, Roth J (2023) A more credible approach to parallel trends. *Review of Economic Studies* 90(5):2555–2591.

Rao A, Wang E (2017) Demand for "healthy" products: False claims and ftc regulation. *Journal of Marketing Research* 54(6):968–989.

Roth J, Sant'Anna PH, Bilinski A, Poe J (2023) What's trending in difference-in-differences? a synthesis of the recent econometrics literature. *Journal of Econometrics* .

Sachdeva S, Iliev R, Medin DL (2009) Sinning saints and saintly sinners: The paradox of moral self-regulation. *Psychological science* 20(4):523–528.

Seiler S, Tuchman A, Yao S (2021) The impact of soda taxes: Pass-through, tax avoidance, and nutritional effects. *Journal of Marketing Research* 58(1):22–49.

Sevilla J, Lu J, Kahn BE (2019) Variety seeking, satiation, and maximizing enjoyment over time. *Journal of Consumer Psychology* 29(1):89–103.

Tang J, Abraham C, Stamp E, Greaves C (2015) How can weight-loss app designers' best engage and support users? a qualitative investigation. *British journal of health psychology* 20(1):151–171.

Uetake K, Yang N (2018) Harnessing the small victories: Goal design strategies for a mobile calorie and weight loss tracking application. *Available at SSRN 2928441* .

Uetake K, Yang N (2020) Inspiration from the "biggest loser": Social interactions in a weight loss program. *Marketing Science* 39(3):487–499.

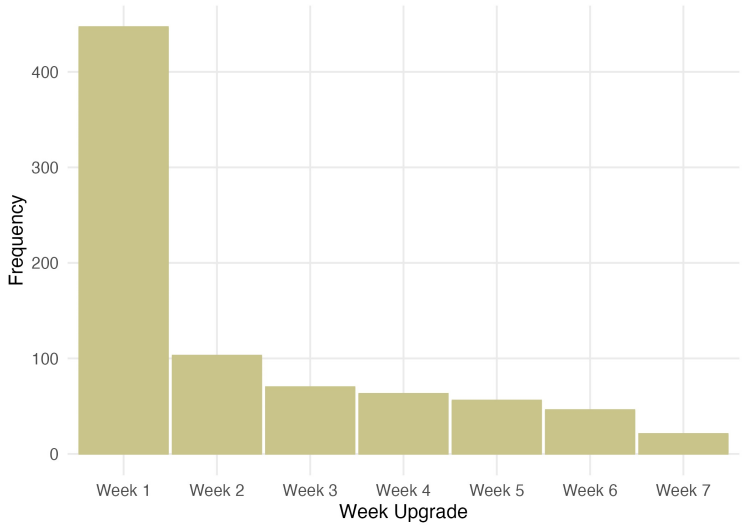
Wilcox K, Vallen B, Block L, Fitzsimons GJ (2009) Vicarious goal fulfillment: When the mere presence of a healthy option leads to an ironically indulgent decision. *Journal of Consumer Research* 36(3):380–393.

Willett W (2012) *Nutritional epidemiology*, volume 40 (Oxford university press).

Zhou M, Fukuoka Y, Mintz Y, Goldberg K, Kaminsky P, Flowers E, Aswani A, et al. (2018) Evaluating machine learning–based automated personalized daily step goals delivered through a mobile phone app: Randomized controlled trial. *JMIR mHealth and uHealth* 6(1):e9117.

FIGURES AND TABLES

Figure 1: Number of Upgrades Each Week Since Registration



Notes: This histogram shows the number of users who adopted the paid premium version of the mHealth app each week after registration.

Table 1: Summary Statistics for Full Sample

Variable	Mean	Std. Dev.
Male	0.294	0.456
Age	41.553	14.330
Starting Weight (lb)	204.132	52.628
Target Weight (lb)	162.300	36.207
Initial Distance to Goal Weight (lb)	41.488	32.345
Proportional Distance to Goal Weight	0.189	0.112
Suggested Budget of Calories (kcal)	1626.050	421.538
Food Calories Logged (kcal)	809.441	490.913
Exercise Calories Logged (kcal)	109.809	158.230
Average Proportion of Days Tracked Food	0.584	0.268
Average Proportion of Days Tracked Exercise	0.273	0.254
Average Proportion of Days Tracked and Stayed within Budget	0.485	0.246
Number of Active Days	165.782	82.472
N	11,873	

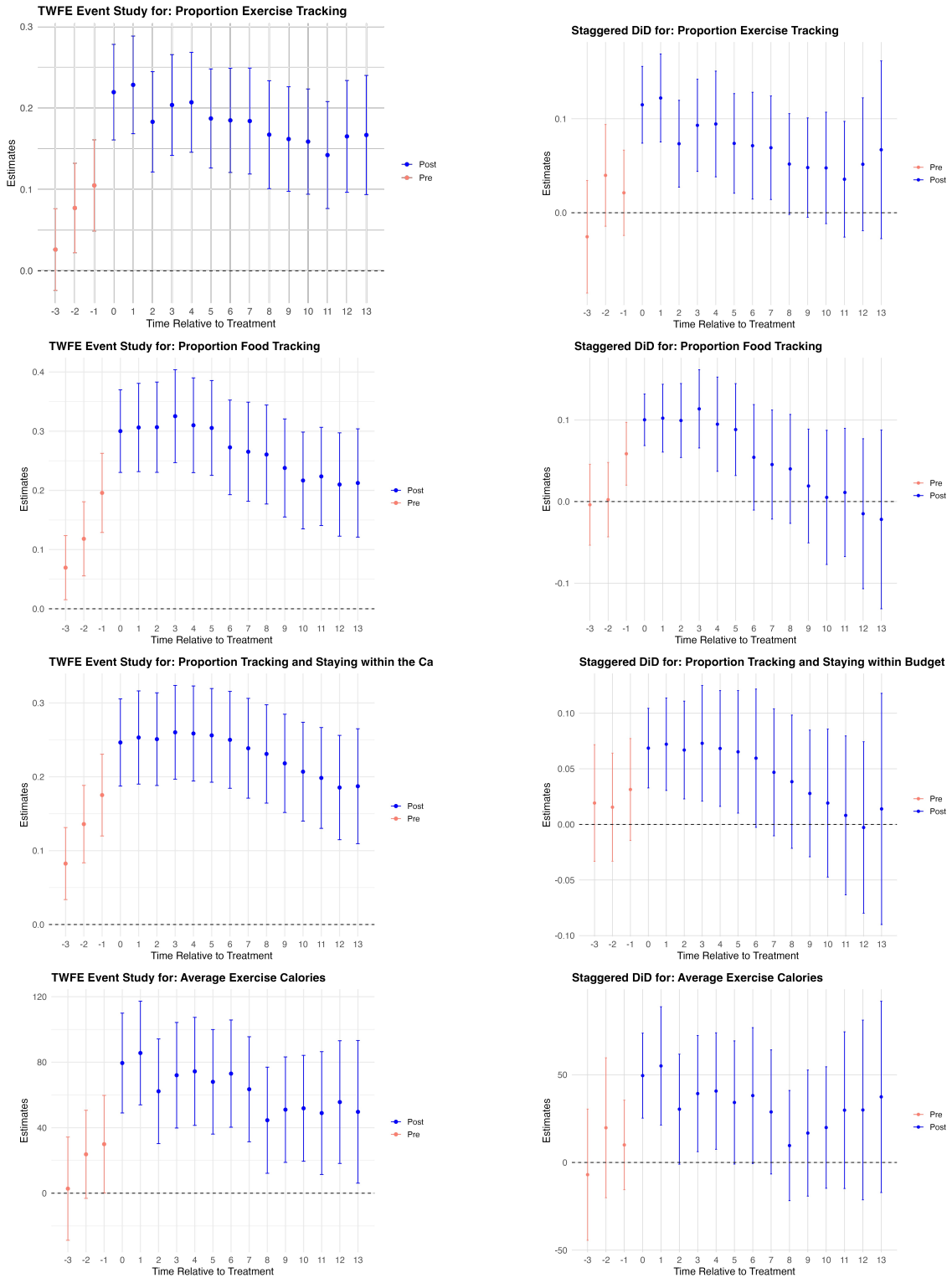
Notes: This table presents summary statistics for key variables in the full sample.

Table 2: Linear Regression of Upgrade Decision on User Demographics, Initial Goal Settings, Engagement, and Google Search Intensities

	Dependent variable:	
	Ever Upgraded	
	Full Sample	Restricted Sample
Starting Weight	−0.002	−0.003***
	(0.001)	(0.001)
Height	−0.001	−0.002
	(0.004)	(0.003)
Age	0.002***	0.001***
	(0.0002)	(0.0001)
Start Date	−0.00003	−0.00002
	(0.00004)	(0.00004)
Initial Goal Weight	0.002	0.002**
	(0.001)	(0.001)
Initial Goal BMI	−0.009	−0.014**
	(0.007)	(0.005)
Initial BMI	0.133	0.200***
	(0.089)	(0.071)
Proportional Distance to Goal Weight	0.186***	0.069
	(0.056)	(0.044)
Male	−0.007	−0.010
	(0.008)	(0.006)
Prop. Days Tracked Food (Week 1)	−0.046*	−0.055***
	(0.024)	(0.019)
Food Calories (Week 1)	0.00001	0.00000
	(0.00001)	(0.00001)
Prop. Days Tracked Exercise (Week 1)	−0.004	−0.009
	(0.011)	(0.008)
Prop. Days Within Budget (Week 1)	−0.016	−0.009
	(0.015)	(0.012)
Exercise Calories (Week 1)	0.00003**	0.00002*
	(0.00002)	(0.00001)
Prop. Days Lower than Initial Weight (Week 1)	0.010*	0.006
	(0.006)	(0.005)
Average Weight Change (Week 1)	0.00002	−0.001
	(0.001)	(0.001)
App Name Premium Searches (Week 1)		0.023**
		(0.010)
App Name Weight Loss Searches (Week 1)		−0.018
		(0.027)
Weight Loss Premium Searches (Week 1)		−0.011
		(0.008)
Constant	0.183	0.233
	(0.334)	(0.304)
Observations	11,873	11,426
R^2	0.017	0.017

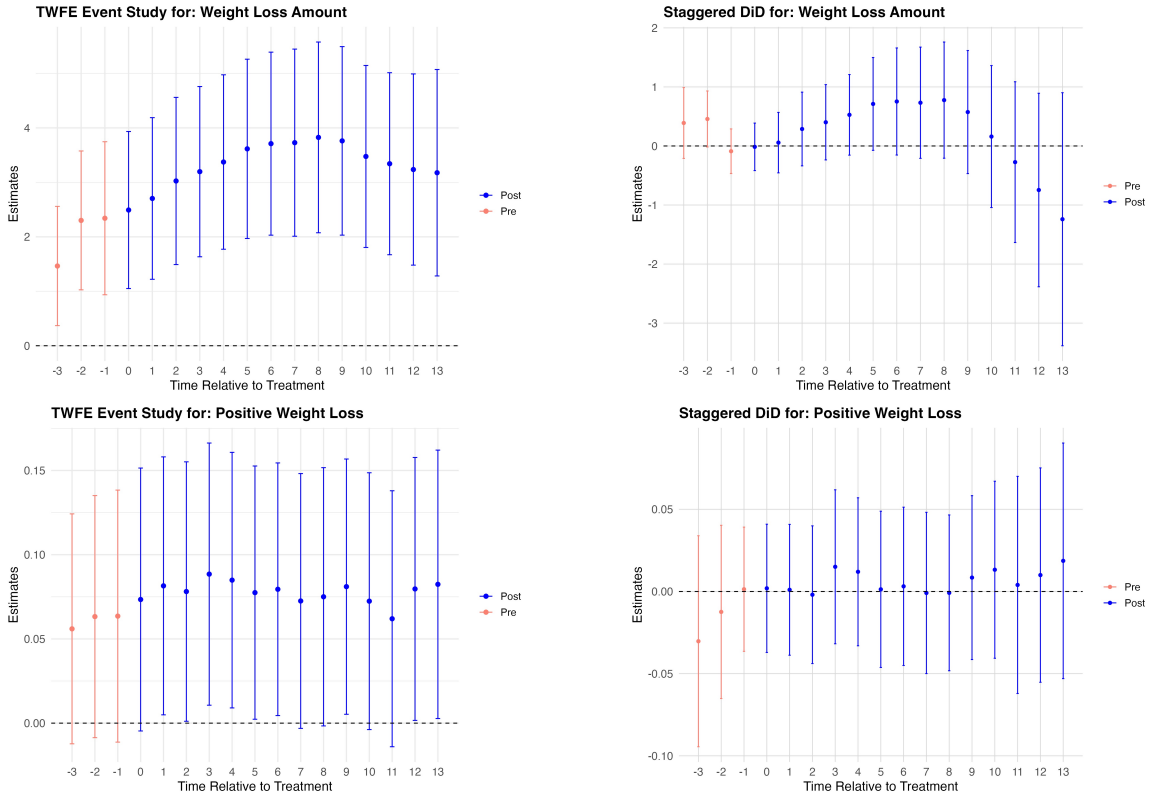
Notes: This table reports linear probability model estimates for whether users upgraded to the premium version of the mHealth app within the first seven weeks. Controls include demographics, start date, initial goal settings, and first-week engagement. Column 2 additionally includes national Google search intensities for app- and weight-loss-related keywords in week 1. Column 1 uses the full sample of 11,873 users; Column 2 uses the TWFE and staggered DiD sample, consisting of adopters in weeks 2–7 and non-adopters.* denotes $p < 0.10$, ** denotes $p < 0.05$, and *** denotes $p < 0.01$.

Figure 2: Impact of Premium Adoption on Engagement Outcomes



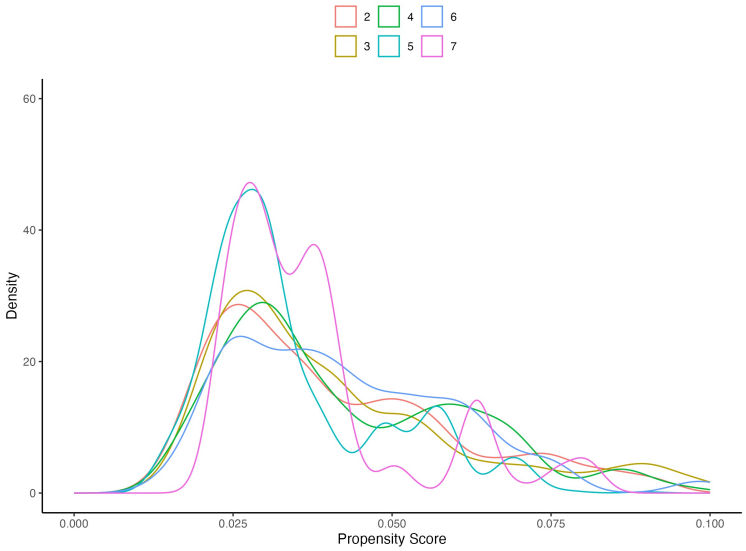
Notes: This figure compares event-study estimates using the TWFE with those obtained with the Callaway and Sant'Anna (2021) DiD estimator. The reference period for the TWFE specification is more than three weeks prior to premium adoption. The comparison group for Callaway and Sant'Anna (2021) DiD estimator is not-yet-treated users. The estimation is based on the sample that contains users who upgrade to the premium in weeks 2 - 7 and users who do not upgrade over our entire observation period of 15 weeks.

Figure 3: Impact of Premium Adoption on Health-Related Outcomes



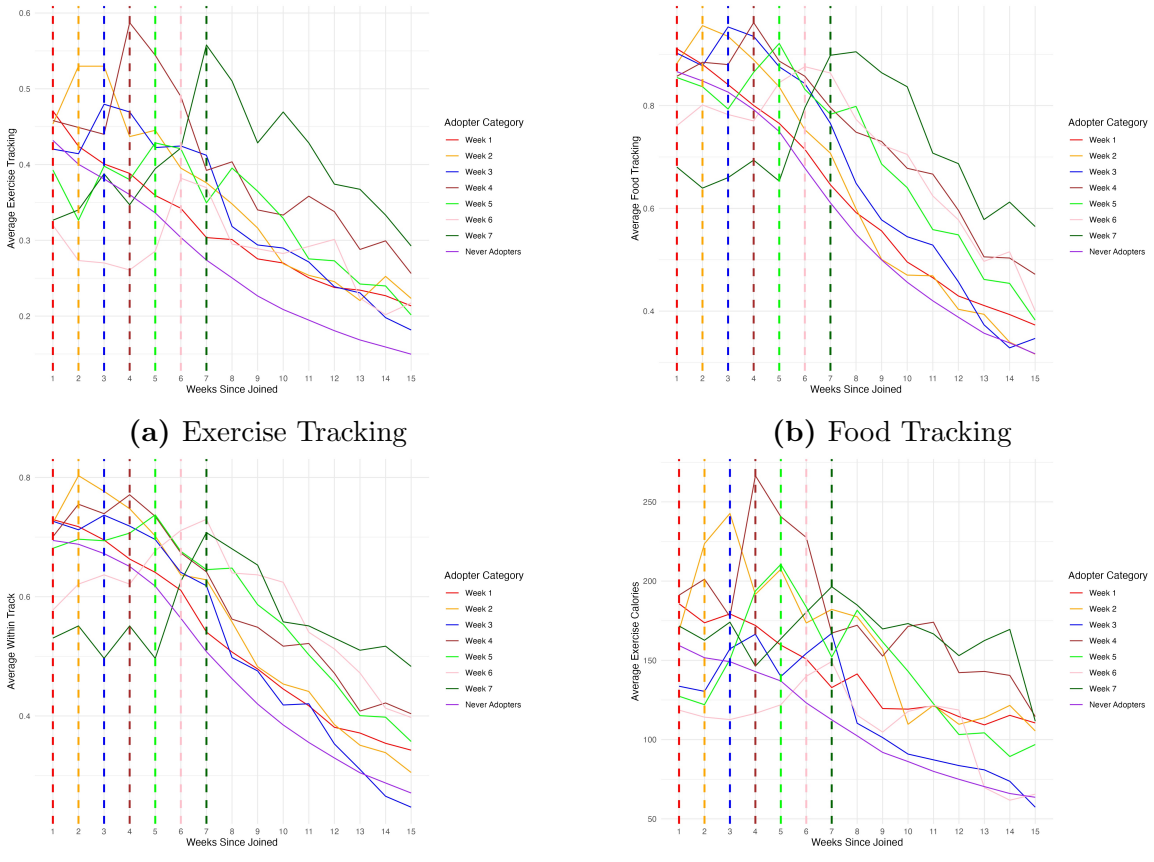
Notes: This figure compares event-study estimates using the TWFE with those obtained with the Callaway and Sant'Anna (2021) DiD estimator. The reference period for the TWFE specification is more than three weeks prior to premium adoption. The comparison group for Callaway and Sant'Anna (2021) DiD estimator is not-yet-treated users. The estimation is based on the sample that contains users who upgrade to the premium in weeks 2 - 7 and users who do not upgrade over our entire observation period of 15 weeks.

Figure 4: Propensity Score Distributions Across Adoption Cohorts



Notes: This figure presents the density of propensity score predicted by the random forest model across adoption cohorts.

Figure 5: Pre-Adoption Trends for Engagement Outcomes



(a) Exercise Tracking

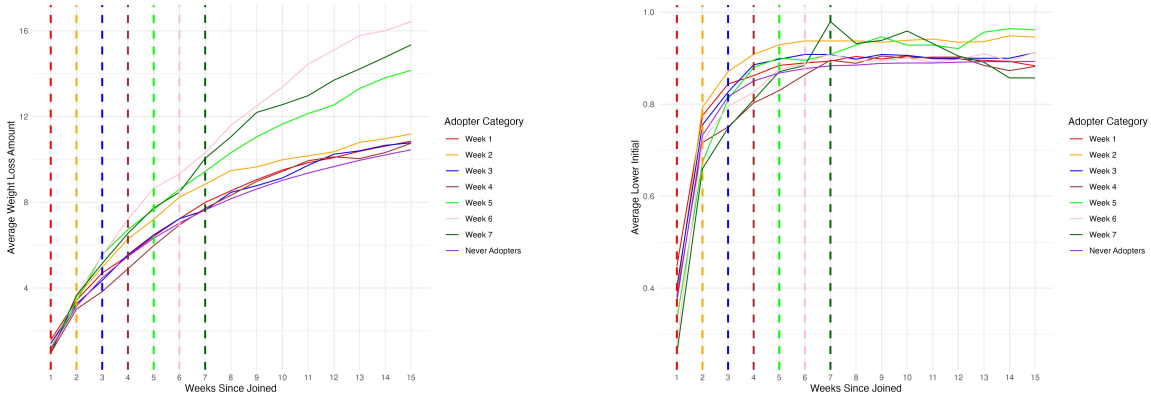
(b) Food Tracking

(c) Staying within Budget

(d) Exercise Calories

Notes: This figure displays pre-trend plots for the engagement outcomes across adoption cohorts.

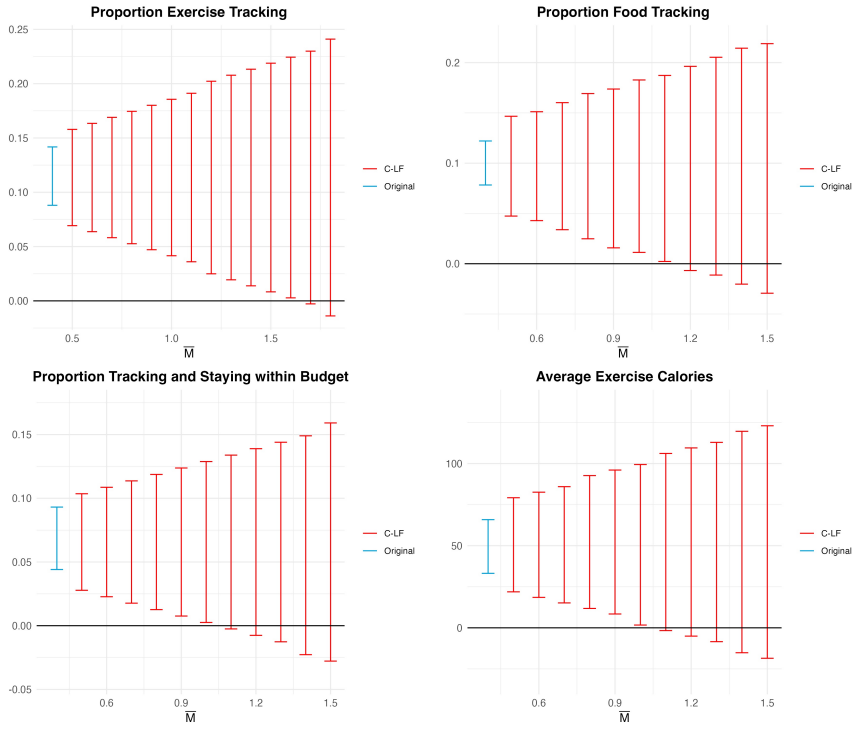
Figure 6: Pre-Adoption Trends for Health Outcomes



(a) Weight Loss Amount **(b)** Positive Weight Loss

Notes: This figure displays pre-trend plots for weight loss outcomes across adoption cohorts.

Figure 7: Sensitivity Analysis Using Fixed-Length Confidence Intervals



Notes: This figure displays fixed-length confidence intervals (FLCIs) estimated using Rambachan and Roth (2023). The horizontal axis corresponds to $\overline{M}$, and the vertical axis represents the FLCIs given $\overline{M}$. The blue-colored interval represents the case when the standard parallel trends assumption holds (i.e., $\overline{M} = 0$), and the red confidence intervals are for deviations from this assumption.

Table 3: Overall ATT Estimates: TWFE vs. Callaway and Sant'Anna (2021)

TWFE Estimates			
Variable	ATT	SE	95% CI
Exercise Tracking	0.124	0.017	[0.089, 0.158]
Food Tracking	0.163	0.021	[0.121, 0.204]
Tracking and Staying within Budget	0.124	0.018	[0.089, 0.159]
Average Exercise Calories	47.730	9.958	[28.21, 67.25]
Positive Weight Loss	0.027	0.020	[-0.011, 0.066]
Weight Loss Amount	1.573	0.383	[0.822, 2.325]
Callaway and Sant'Anna (2021) Estimates			
Variable	ATT	SE	95% CI
Exercise Tracking	0.075	0.015	[0.046, 0.104]
Food Tracking	0.062	0.016	[0.030, 0.093]
Tracking and Staying within Budget	0.050	0.013	[0.024, 0.076]
Average Exercise Calories	33.034	9.414	[14.582, 51.485]
Positive Weight Loss	0.005	0.015	[-0.024, 0.035]
Weight Loss Amount	0.329	0.278	[-0.215, 0.873]

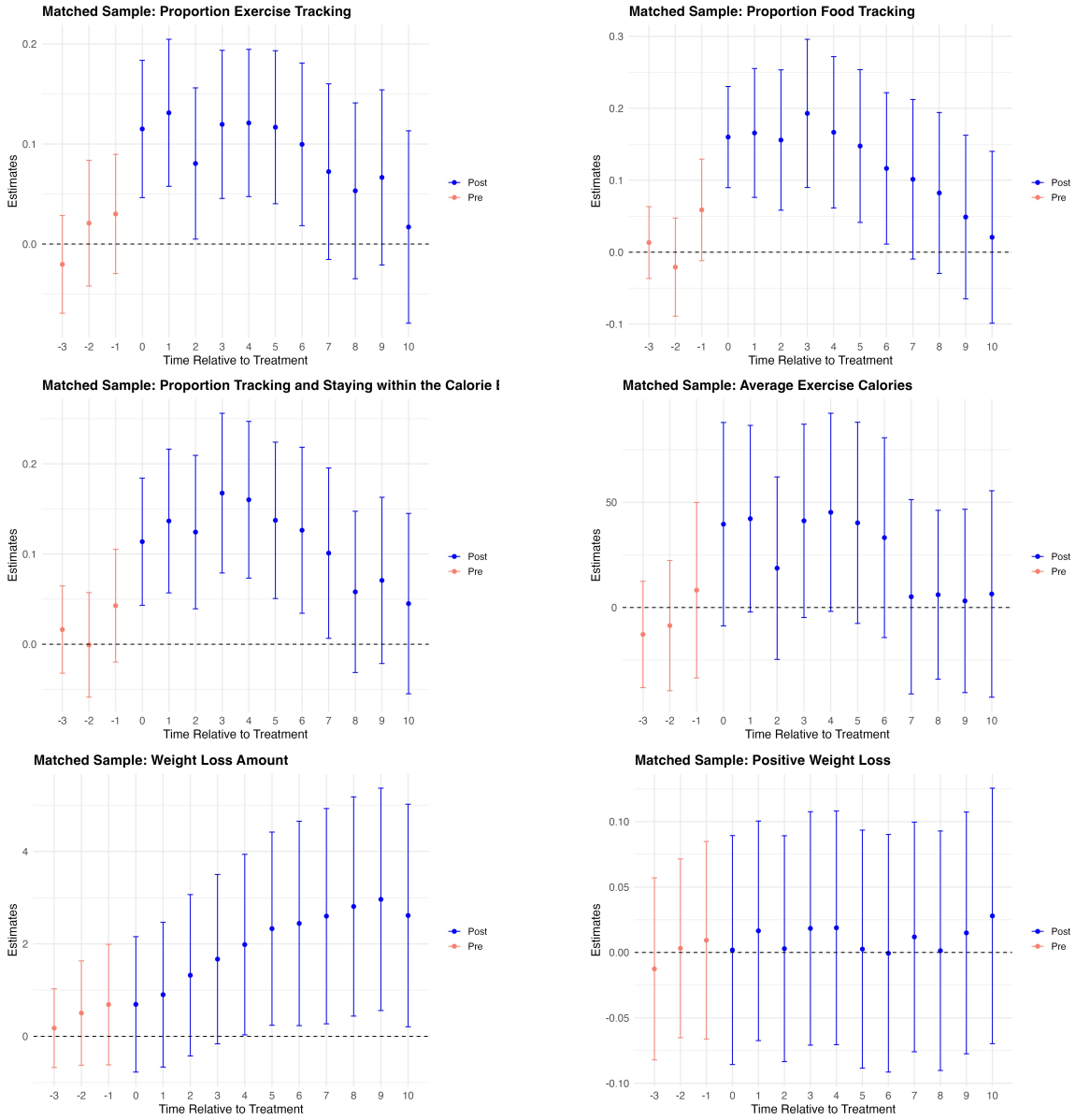
Notes: This table reports overall ATT estimates from two models. **Panel A** uses the two-way fixed effects (TWFE) approach. **Panel B** uses the estimator from Callaway and Sant'Anna (2021). Both are based on users who upgrade to premium between weeks 2–7 and all users who never upgrade during a 15-week observation period. Standard errors are clustered at the user level.

Table 4: Comparison of Means Between Low- and High-Exposure Adopters

Variable	Low Exposure	High Exposure	T-Statistic	P-Value
Number of Prior Days Tracked Food/Exercise	9.721	27.332	-28.555	0.000
Male	0.285	0.299	-0.303	0.762
Age	45.477	48.235	-1.841	0.067
Starting Weight (lb)	214.243	218.262	-0.726	0.468
Height (inch)	66.512	66.342	0.435	0.664
Initial Distance to the Goal Weight (lb)	167.256	164.783	0.643	0.521
Initial Goal BMI	26.386	26.119	0.581	0.562
Current Weight (lb)	211.196	209.764	0.272	0.785
Current Distance to the Goal Weight (lb)	44.813	44.949	-0.037	0.971
Number of Observations	172	187		

Notes: This table summarizes the key variables for the low- and high-exposure adopters and reports the corresponding t-statistics and p-values from two-sample mean tests.

Figure 8: TWFE Estimates on Matched Sample Balancing on Recent Engagement and Google Search Intensities



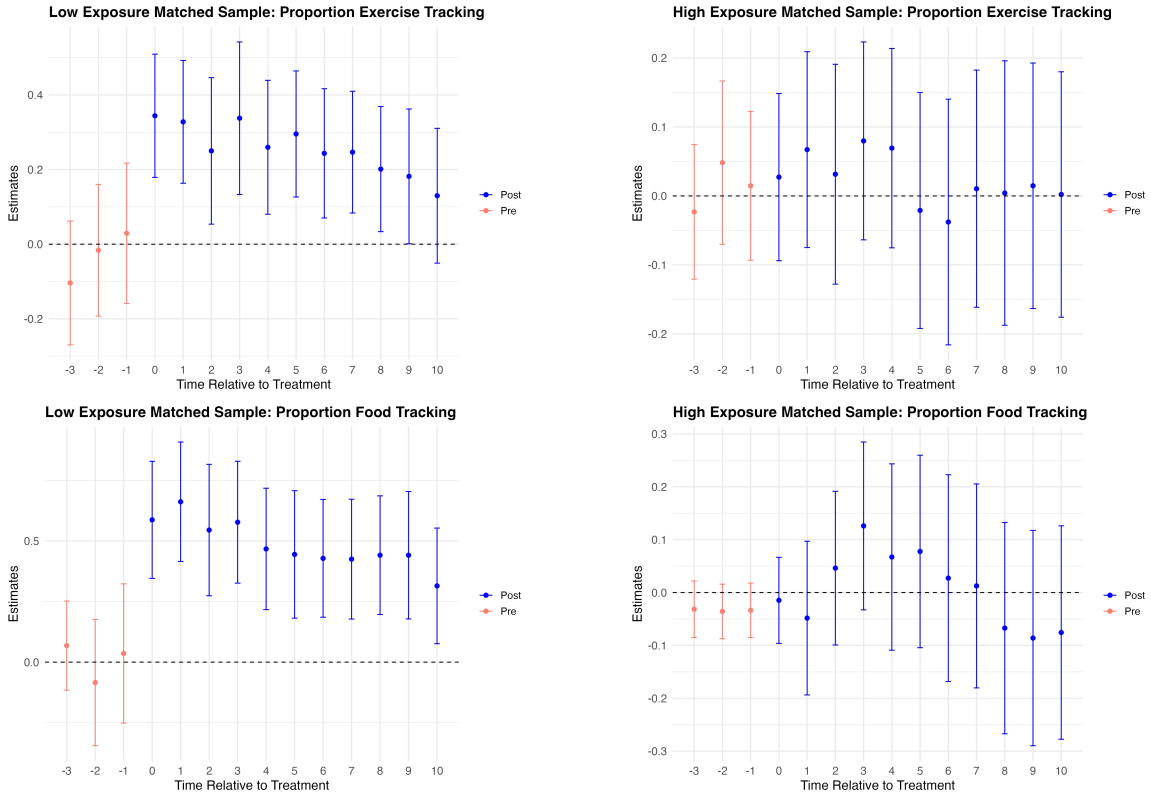
Notes: Each subfigure displays TWFE event-study coefficients for the indicated outcome estimated on the matched sample (adopters in weeks 5–6 matched to non-adopters on demographics, initial goal settings, week 1 engagement, week 3 and week 4 engagement and weight-loss outcomes, and week 3 and week 4 Google Trends intensities). Red points show pre-treatment leads; blue points show post-treatment lags. Error bars are 95% confidence intervals with standard errors clustered at the user level.

Table 5: Heterogeneous Effects by Pre-Adoption Exposure

	Immediate (Week 0)		Short-Term (Week 1)	
Outcome	Estimate	Std. Error	Estimate	Std. Error
Exercise Tracking (Low Exposure)	0.142	0.022	0.164	0.027
Exercise Tracking (High Exposure)	0.091	0.017	0.086	0.020
Food Tracking (Low Exposure)	0.140	0.018	0.165	0.024
Food Tracking (High Exposure)	0.064	0.013	0.043	0.018
Within Budget (Low Exposure)	0.106	0.020	0.117	0.021
Within Budget (High Exposure)	0.034	0.017	0.029	0.019
Exercise Calories (Low Exposure)	65.212	12.296	89.494	18.882
Exercise Calories (High Exposure)	35.673	11.777	24.313	15.847

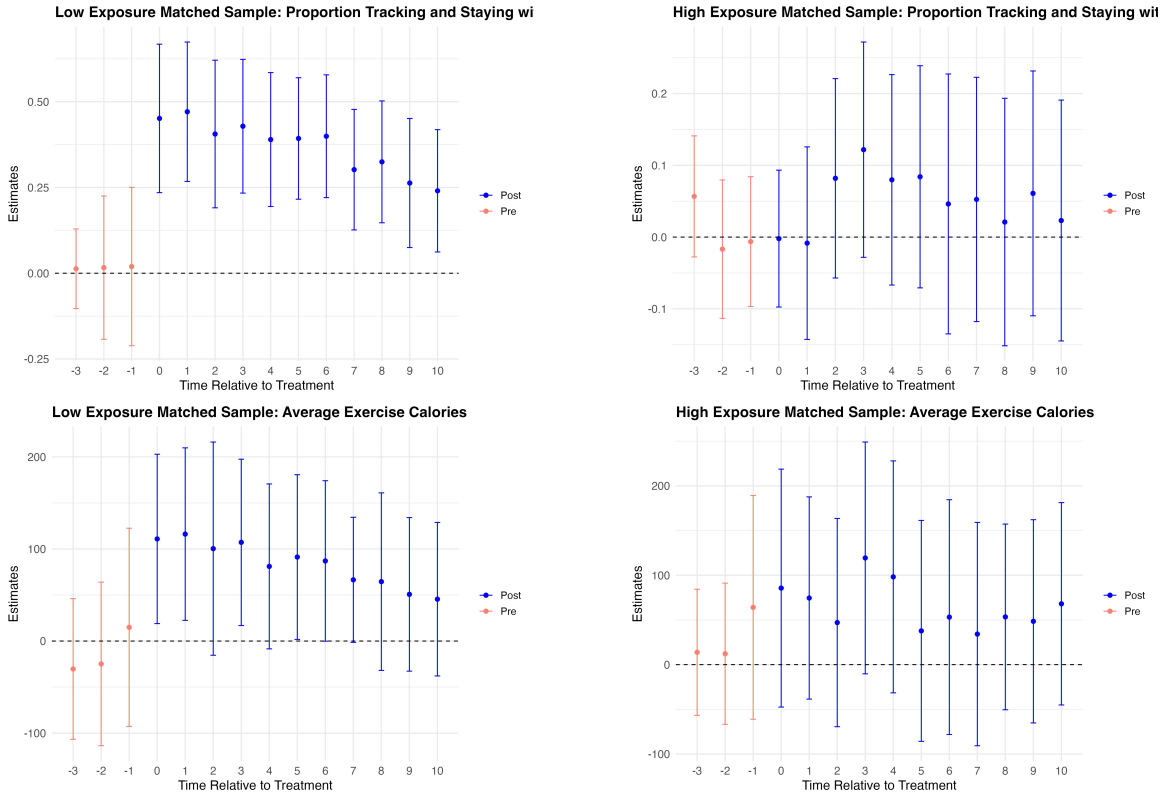
Notes: This table reports subgroup-specific treatment effects based on pre-adoption engagement levels. Users are classified as high or low exposure based on whether their food/exercise-tracking activity prior to premium adoption falls above or below the sample median. Estimates correspond to the immediate (week-of-upgrade) and short-term (one week post-upgrade) effects of premium adoption, with separate estimations run for each outcome and subgroup. All models control for user and calendar week fixed effects. Standard errors are clustered at the user level.

Figure 9: TWFE Estimates on Matched Sample Balancing on Additional Recent Variables and Segmented on Pre-adoption Exposure (Week 5 Adopters Matched with Non-Adopters)



Notes: Panels compare TWFE event-study estimates for week-5 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 4), early and recent engagement trajectories (weeks 3 and 4), and recent weight-loss outcomes (week 4). 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 10: TWFE Estimates on Matched Sample Balancing on Additional Recent Variables and Segmented on Pre-adoption Exposure (Week 5 Adopters Matched with Non-Adopters) [continued...]



Notes: Panels compare TWFE event-study estimates for week-5 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 4), early and recent engagement trajectories (weeks 3 and 4), and recent weight-loss outcomes (week 4). 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.

Online Appendix

Table of Contents

A Alternative Sample of Adoption Cohorts

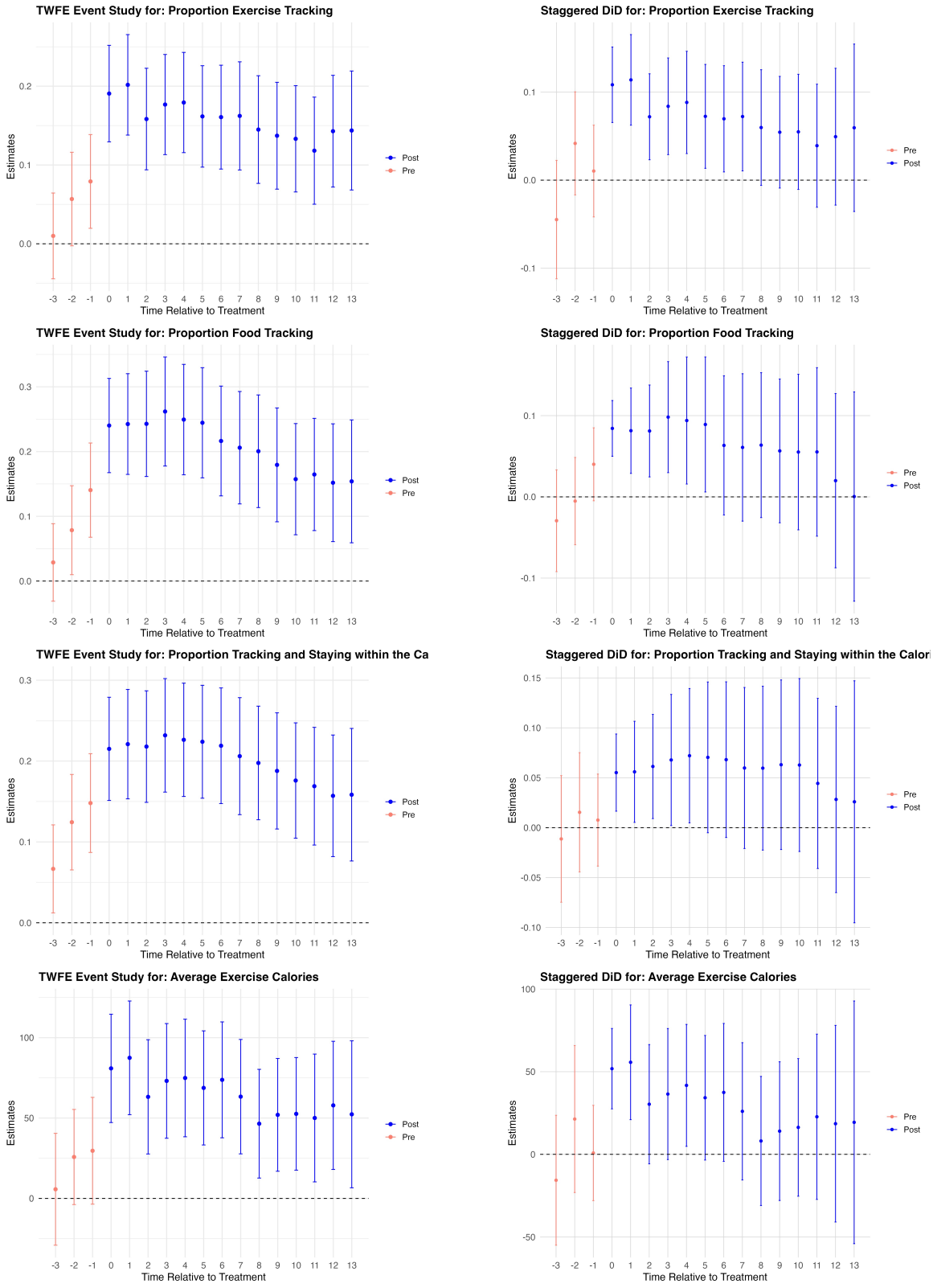
Since only 21 individuals adopted in week 7, this section replicates our estimation in 3.4 for user cohorts who adopted between weeks 2 and 6 and a random sample of 5% users who do not upgrade over our entire observation period of 15 weeks. The qualitative patterns that emerge from the alternative sample 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.

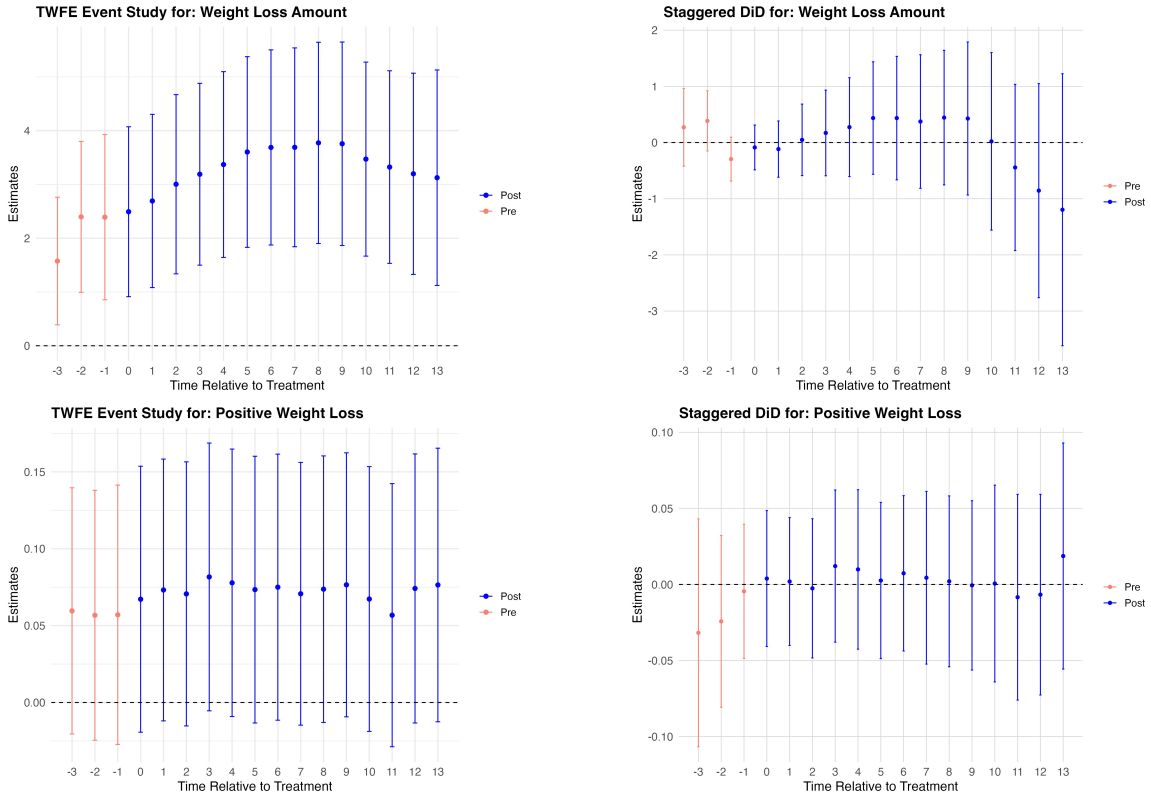
Together, the comparison reinforces the main conclusion of the paper. Premium adoption generates only an ephemeral spike in self-monitoring that dissipates within two months, and this fleeting behavioral change yields no detectable health gains. The fact that removing week 7 adopters produces virtually 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. The reference period for the TWFE specification is more than three weeks prior to premium adoption. The comparison group for Callaway and Sant'Anna (2021) DiD estimator is not-yet-treated users. The estimation is based on the sample that contains users who upgrade to the premium in weeks 2 - 6 and a random sample of 5% 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. The reference period for the TWFE specification is more than three weeks prior to premium adoption. The comparison group for Callaway and Sant'Anna (2021) DiD estimator is not-yet-treated users. The estimation is based on the sample that contains users who upgrade to the premium in weeks 2 - 6 and a random sample of 5% 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	0.0794	0.0176	[0.0449, 0.114]
Food Tracking	0.0784	0.0178	[0.0435, 0.1132]
Tracking and Staying within the Calorie Budget	0.0601	0.0161	[0.0285, 0.0916]
Average Exercise Calories	35.5726	10.5616	[14.8723, 56.273]
Positive Weight Loss	0.0028	0.0194	[-0.0352, 0.0409]
Weight Loss Amount	0.9524	0.2778	[0.4079, 1.497]
Estimates with Alternative Measurements			
Exercise Tracking	0.073	0.017	[0.0426, 0.1033]
Food Tracking	0.074	0.018	[0.0387, 0.1093]
Tracking and Staying within the Calorie Budget	0.06	0.0161	[0.0285, 0.0916]
Average Exercise Calories	35.576	10.6087	[14.7832, 56.3687]
Positive Weight Loss	0.0028	0.0194	[-0.0353, 0.0408]
Weight Loss Amount	0.9509	0.2779	[0.4062, 1.4955]

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 over 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. Finally, we note that the point estimates are highly comparable and the confidence intervals for the ATT estimates from Callaway and Sant'Anna (2021) and De Chaisemartin and d'Haultfoeuille (2020) 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	0.0794	0.0176	[0.0449, 0.114]
Food Tracking	0.0784	0.0178	[0.0435, 0.1132]
Tracking and Staying within the Calorie Budget	0.0601	0.0161	[0.0285, 0.0916]
Average Exercise Calories	35.5726	10.5616	[14.8723, 56.273]
Positive Weight Loss	0.0028	0.0194	[-0.0352, 0.0409]
Weight Loss Amount	0.9524	0.2778	[0.4079, 1.497]
de Chaisemartin and D'Haultfœuille (2020) Estimates			
Exercise Tracking	0.088	0.019	[0.051, 0.126]
Food Tracking	0.090	0.022	[0.048, 0.133]
Tracking and Staying within the Calorie Budget	0.070	0.020	[0.032, 0.108]
Average Exercise Calories	31.983	12.322	[7.833, 56.133]
Positive Weight Loss	-0.008	0.022	[-0.052, 0.035]
Weight Loss Amount	0.632	0.345	[-0.044, 1.308]

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** analyses are based on users who upgrade between weeks 2–7 and all those who never upgrade, observed over 15 weeks. **Panel B** presents analyses based on users who upgraded between weeks 2 and 3, and all those who never upgrade, observed over 15 weeks. We focus on week 2 and week 3 adopters because the estimator proposed by De Chaisemartin and d'Haultfoeuille (2020) does not accommodate staggered adoption.

C.2 Synthetic Difference-in-Differences

As a complementary specification test, we estimate ATT with the *synthetic difference-in-differences* (SDiD) estimator of Arkhangelsky et al. (2021). SDiD 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.

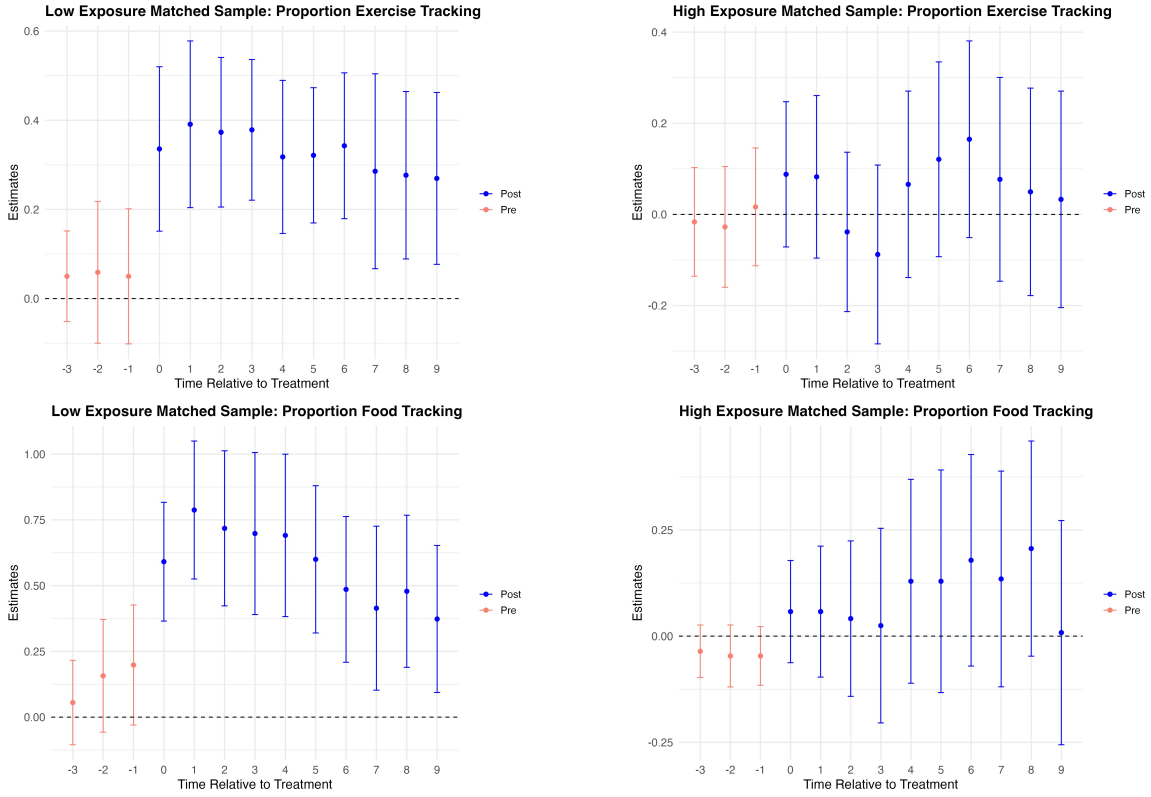
Table C2: Overall ATT Estimates from Alternative DiD estimators

Callaway and Sant'Anna (2021) Estimates			
Variable	ATT	SE	95% CI
Exercise Tracking	0.0794	0.0176	[0.0449, 0.114]
Food Tracking	0.0784	0.0178	[0.0435, 0.1132]
Tracking and Staying within the Calorie Budget	0.0601	0.0161	[0.0285, 0.0916]
Average Exercise Calories	35.5726	10.5616	[14.8723, 56.273]
Positive Weight Loss	0.0028	0.0194	[-0.0352, 0.0409]
Weight Loss Amount	0.9524	0.2778	[0.4079, 1.497]
Synthetic Difference-in-Differences Estimates			
Exercise Tracking	0.093	0.014	[0.0663, 0.1199]
Food Tracking	0.094	0.019	[0.0570, 0.1301]
Tracking and Staying within the Calorie Budget	0.075	0.015	[0.0459, 0.1045]
Average Exercise Calories	45.567	10.304	[25.3715, 65.7628]
Positive Weight Loss	0.012	0.020	[-0.0272, 0.0510]
Weight Loss Amount	0.894	0.316	[0.2753, 1.5134]

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.

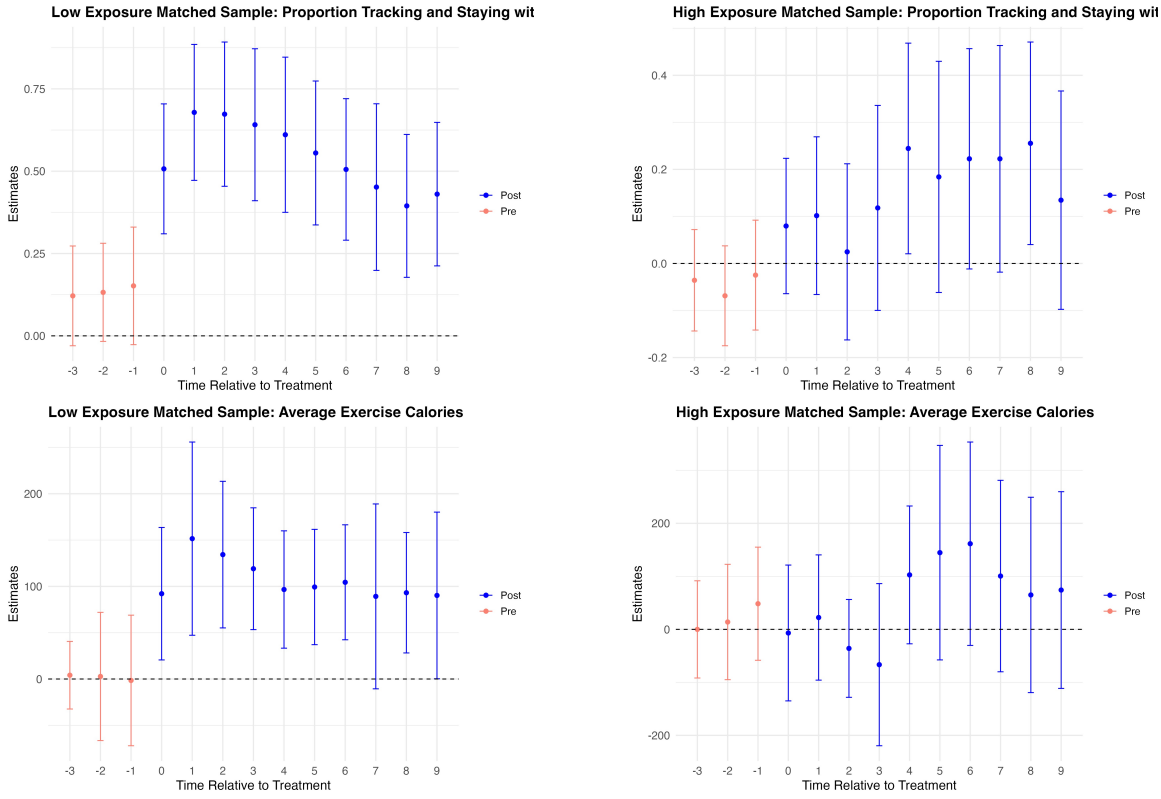
D Additional Tables and Figures

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	0.0242	0.0686**	1.851	0.0830***
	(0.0255)	(0.0276)	(16.10)	(0.0249)
Two weeks before upgrade	0.0753***	0.1179***	22.68*	0.1370***
	(0.0280)	(0.0319)	(13.71)	(0.0267)
One week before upgrade	0.1027***	0.1956***	28.57*	0.1768***
	(0.0286)	(0.0340)	(15.19)	(0.0282)
Immediate Post-Treatment (Week of Upgrade)				
Week of upgrade	0.2172***	0.2999***	78.11***	0.2481***
	(0.0299)	(0.0355)	(15.54)	(0.0300)
Short-Term Post-Treatment (1–6 Weeks After Upgrade)				
One week after upgrade	0.2262***	0.3060***	84.24***	0.2548***
	(0.0305)	(0.0379)	(16.14)	(0.0321)
Two weeks after upgrade	0.1808***	0.3064***	60.87***	0.2525***
	(0.0314)	(0.0388)	(16.29)	(0.0319)
Three weeks after upgrade	0.2014***	0.3250***	70.67***	0.2617***
	(0.0316)	(0.0399)	(16.42)	(0.0323)
Four weeks after upgrade	0.2047***	0.3096***	73.02***	0.2601***
	(0.0313)	(0.0408)	(16.82)	(0.0328)
Five weeks after upgrade	0.1849***	0.3050***	66.62***	0.2575***
	(0.0310)	(0.0408)	(16.29)	(0.0323)
Six weeks after upgrade	0.1826***	0.2723***	71.63***	0.2515***
	(0.0326)	(0.0407)	(16.68)	(0.0335)
Medium-Term Post-Treatment (7–12 Weeks After Upgrade)				
Seven weeks after upgrade	0.1817***	0.2649***	62.07***	0.2401***
	(0.0332)	(0.0427)	(16.33)	(0.0345)
Eight weeks after upgrade	0.1598***	0.2595***	41.51**	0.2352***
	(0.0337)	(0.0425)	(16.64)	(0.0340)
Nine weeks after upgrade	0.1593***	0.2366***	47.53***	0.2225***
	(0.0334)	(0.0422)	(16.56)	(0.0344)
Ten weeks after upgrade	0.1497***	0.2180***	44.50***	0.2105***
	(0.0331)	(0.0418)	(15.85)	(0.0348)
Eleven weeks after upgrade	0.1484***	0.2343***	50.45**	0.2087***
	(0.0344)	(0.0428)	(19.77)	(0.0357)
Twelve weeks after upgrade	0.1543***	0.2010***	49.97***	0.1907***
	(0.0360)	(0.0447)	(19.02)	(0.0372)
Long-Term Post-Treatment (13 Weeks After Upgrade)				
Thirteen weeks after upgrade	0.1535***	0.1928***	26.87	0.1677***
	(0.0413)	(0.0481)	(24.23)	(0.0419)
User FE	Yes	Yes	Yes	Yes
Week FE	Yes	Yes	Yes	Yes
Observations	171,495	171,495	171,495	171,495
R^2	0.60837	0.57058	0.61481	0.55340
Within R^2	0.00192	0.00223	0.00095	0.00158

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. 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	0.0560	1.465***
	(0.0348)	(0.5584)
Two weeks before upgrade	0.0633*	2.297***
	(0.0367)	(0.6499)
One week before upgrade	0.0636*	2.336***
	(0.0382)	(0.7161)
Immediate Post-Treatment (Week of Upgrade)		
Week of upgrade	0.0735*	2.487***
	(0.0398)	(0.7350)
Short-Term Post-Treatment (1-6 Weeks After Upgrade)		
One week after upgrade	0.0815**	2.699***
	(0.0391)	(0.7563)
Two weeks after upgrade	0.0782**	3.021***
	(0.0393)	(0.7823)
Three weeks after upgrade	0.0885**	3.192***
	(0.0397)	(0.7967)
Four weeks after upgrade	0.0849**	3.369***
	(0.0387)	(0.8162)
Five weeks after upgrade	0.0775**	3.611***
	(0.0384)	(0.8391)
Six weeks after upgrade	0.0795**	3.706***
	(0.0383)	(0.8564)
Medium-Term Post-Treatment (7-12 Weeks After Upgrade)		
Seven weeks after upgrade	0.0725*	3.724***
	(0.0386)	(0.8757)
Eight weeks after upgrade	0.0751*	3.811***
	(0.0391)	(0.8918)
Nine weeks after upgrade	0.0813**	3.741***
	(0.0387)	(0.8803)
Ten weeks after upgrade	0.0727*	3.477***
	(0.0389)	(0.8501)
Eleven weeks after upgrade	0.0624	3.269***
	(0.0388)	(0.8550)
Twelve weeks after upgrade	0.0802**	3.271***
	(0.0398)	(0.8934)
Long-Term Post-Treatment (13+ Weeks After Upgrade)		
Thirteen weeks after upgrade	0.0837**	3.222***
	(0.0406)	(0.9612)
User FE	Yes	Yes
Week FE	Yes	Yes
Observations	171,495	171,495
R^2	0.61452	0.80565
Within R^2	0.00017	0.00149

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. Standard errors are clustered at the user level. * denotes $p < 0.10$, ** denotes $p < 0.05$, and *** denotes $p < 0.01$.