

Not Now, Ask Later: Users Weaken Their Behavior Change Regimen Over Time, But Expect To Re-Strengthen It Imminently (Supplementary Materials)

GEZA KOVACS, ZHENGXUAN WU, MICHAEL S. BERNSTEIN

ACM Reference Format:

Geza Kovacs, Zhengxuan Wu, Michael S. Bernstein. 2021. Not Now, Ask Later: Users Weaken Their Behavior Change Regimen Over Time, But Expect To Re-Strengthen It Imminently (Supplementary Materials). In *CHI '21: ACM CHI Conference on Human Factors in Computing Systems, May 08–13, 2021, Yokohama, Japan*. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3411764.3445695>

1 SUPPLEMENT A: EFFECTIVENESS OF INTERVENTIONS

1.1 Methodology

In this study we seek to determine which interventions are most effective, and how well our raters' ratings of intervention difficulty correspond to their effectiveness.

We tested this in a study where on each visit to Facebook, a randomly chosen intervention (or no intervention) is shown. The reason for choosing Facebook for this analysis is that some HabitLab interventions are site-specific, and the largest number of interventions are available for Facebook. Additionally, Facebook is also the most widely-chosen goal site among our users. We then measure time spent on Facebook in the presence of that intervention. A total of 14,139,727 exposure samples were used in this study, from 14,834 users.

1.2 Results

Median times of Facebook session lengths in the presence of each intervention are shown in Figure 1. Interventions where session lengths are the shortest are the most effective at reducing time on Facebook. We see that all interventions are more effective than no intervention, and that the most effective intervention is closing the tab automatically after 60 seconds. Note that the "close tab after 60 seconds" intervention includes an "add time" button, hence why time spent on Facebook in its presence can exceed 60 seconds.

We asked three independent raters (HabitLab users who had been using it for over a month) to rate their perceived difficulty level of each intervention as either easy, medium, or hard. We opted for a 3-level difficulty categorization, as our studies ask users to choose difficulty levels and we did not want to overwhelm them with too many choices. We took the intervention's difficulty to be the median of its ratings. Our raters' intervention difficulty ratings are shown in Table 1.

To investigate whether interventions perceived as more difficult by our raters are also more effective, we group the samples according to the raters' intervention difficulty. Median times of Facebook session lengths in the presence of each intervention difficulty are shown in Figure 2. The most time is spent when there was no intervention (median of 199 seconds per session), followed by easy (185 seconds), medium (161 seconds), and hard (135 seconds) interventions, as shown in Figure 2. There is a significant effect of difficulty on effectiveness according to a Kruskal-Wallis H test

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.

Manuscript submitted to ACM

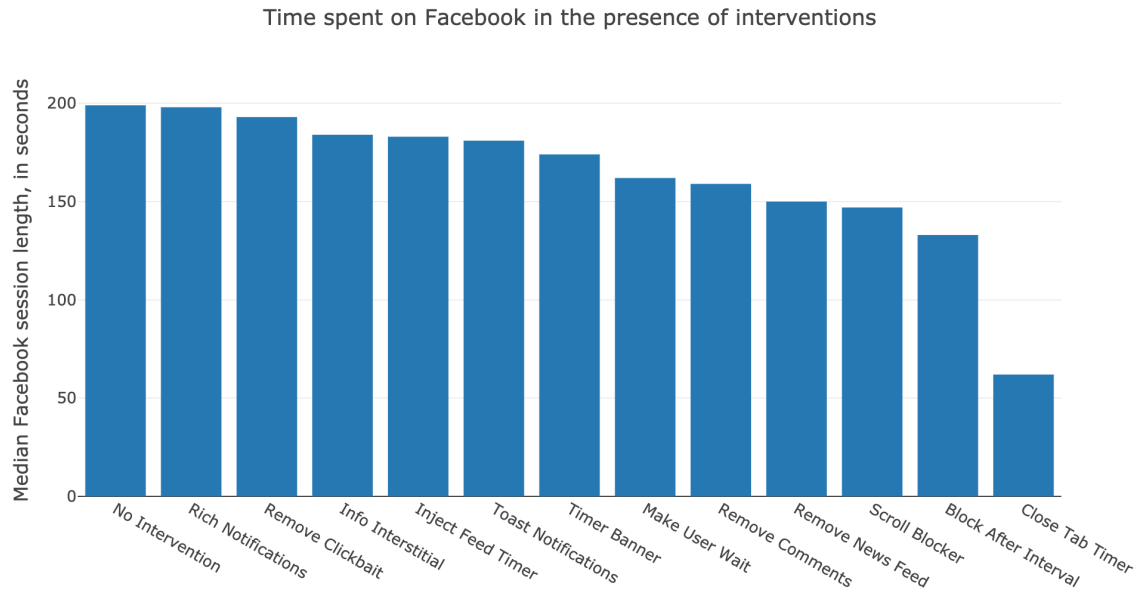


Fig. 1. Median Facebook session durations in the presence of each intervention. Intervention descriptions follow:

Rich Notifications: Notifies you of time spent in the corner of your desktop

Removes clickbait: Removes feed items from sources such as BuzzFeed

Info Interstitial: Show time spent and visit count each visit

Inject Timer: Injects timer into the Facebook feed

Toast Notifications: Notifies you of time spent every minute

Timer Banner: Shows time spent on site at the top of screen

Make User Wait: Makes you wait a few seconds before visiting

Remove Comments: Hides comments in posts

Remove News Feed: Hides the news feed

Scroll Blocker: Freezes scrolling after a certain amount of scrolls

Block After Interval: Asks how long you want to spend on site this visit

Close Tab Timer: Closes tab after 60 seconds

($H=37654$, $p < 0.001$). Differences between pairs of groups are all statistically significant ($p < 0.001$) according to pairwise Mann-Whitney U tests shown in Table 2. From this result, we conclude that the difficulty labels capture not only raters' opinions, but also are associated with monotonically increasing time savings when deployed, suggesting that they are in practice more aggressive.

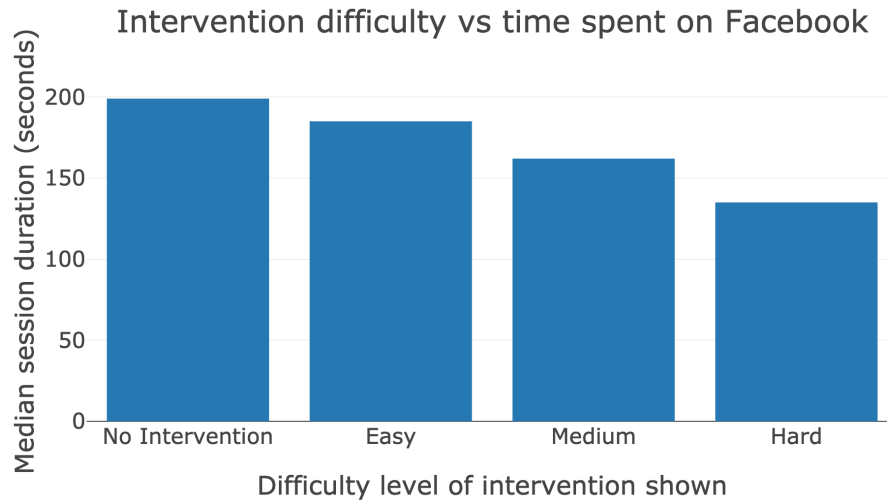


Fig. 2. Facebook session durations are significantly shorter in the presence of more difficult interventions. Any intervention difficulty level is more effective than no intervention.

Intervention	R1	R2	R3	Difficulty (Median)
Injects timer into the Facebook feed	1	1	1	Easy (1)
Removes clickbait	1	1	1	Easy (1)
Notifies you of time spent in the corner of your desktop	1	2	1	Easy (1)
Shows time spent on site at the top of screen	1	1	3	Easy (1)
Notifies you of time spent every minute	1	2	3	Medium (2)
Removes Facebook comments	2	1	3	Medium (2)
Removes the Facebook news feed	2	2	2	Medium (2)
Makes you wait a few seconds before visiting	3	2	2	Medium (2)
Asks how long you want to spend on site this visit	3	2	2	Medium (2)
Freezes scrolling after a certain amount of scrolls	3	3	3	Hard (3)
Show time spent and visit count each visit	3	3	3	Hard (3)
Closes tab after 60 seconds	3	3	3	Hard (3)

Table 1. Perceived intervention difficulty ratings from our 3 raters, where 1 = Easy, 2 = Medium, 3 = Hard.

Pair of conditions being compared	Mann-Whitney U statistic	p value
No Intervention vs Easy	U=5955826705939	p < 0.001
No Intervention vs Medium	U=5109405780672	p < 0.001
No Intervention vs Hard	U=1102958826925	p < 0.001
Easy vs Medium	U=10987237324854	p < 0.001
Easy vs Hard	U=2373577140176	p < 0.001
Medium vs Hard	U=2207988223327	p < 0.001

Table 2. Pairwise comparisons of the time spent in each condition under the Mann-Whitney U test indicates that being served interventions of higher difficulty levels significantly reduces time spent on Facebook.

2 SUPPLEMENT B: EFFECTS OF USER CONTROL

Our previous experiments have focused on being able to accurately model users' intervention difficulty preferences, while attempting to minimize adverse effects of prompting users. However, the implicit assumption we have made is that giving users choices and respecting their preferences are fundamentally good. That is, we have assumed that once we are able to accurately model users' intervention difficulty preferences, we should give them those interventions.

That said, there are reasons to suspect that blindly following user choices may often be suboptimal from a behavior change perspective. Users are myopic in their choices, and will often make short-term choices that are not in line with their long-term goals [1, 3, 5, 6]. Our results from our first study, where we observed that users' intervention selections decline in difficulty over time, are perhaps a reflection of this short-term myopic decision making. The solution may be analogous to a personal fitness coach – a coach may push us to go beyond what we think we can do or want in the moment; with the agreement on our part to cede some of our autonomy to the coach, we actually pay people to disrespect our in-the-moment wishes. Hence, reducing short-term user choices may potentially be beneficial for helping users achieve their long-term goals.

2.1 Methodology

In this experiment we study the effects of removing user choices, compared to providing them with choices. We do so by randomly assigning new users to one of the following six conditions when they install, two of which provide users with some level of choice as to the level of difficulty of their interventions, and four of which randomly assign users to be shown interventions of some difficulty level:

Non-choice conditions where users are randomly assigned a difficulty level for interventions:

- No interventions
- Easy interventions
- Medium interventions
- Hard interventions

Choice conditions:

- Initial choice: User initially chooses difficulty but does not subsequently update it
- Continuous choice: User chooses a difficulty initially and can subsequently update it via the difficulty prompting interface, with the choice of when to ask again

We then compare these conditions based on retention as well as efficacy. We compare retention rates based on a Cox Hazard Regression model [2], and use a Linear Mixed Model (LMM) [4] modeling time spent per day on Facebook, where the user is a random effect, and condition is a fixed effect, as a means of comparing efficacy between the conditions. We chose to analyse results for Facebook because it is the site that is most commonly chosen as a goal, and we have the most interventions for it. For our intervention efficacy analysis, we only considered days where the user saw at least one intervention.

This study was conducted with 743 users, of whom 413 had Facebook as a goal site, over the course of 59 days. The mean install lifetime of a user was 17.2 days, median is 12 days, max install lifetime is 59 days, and standard deviation is 16.8.

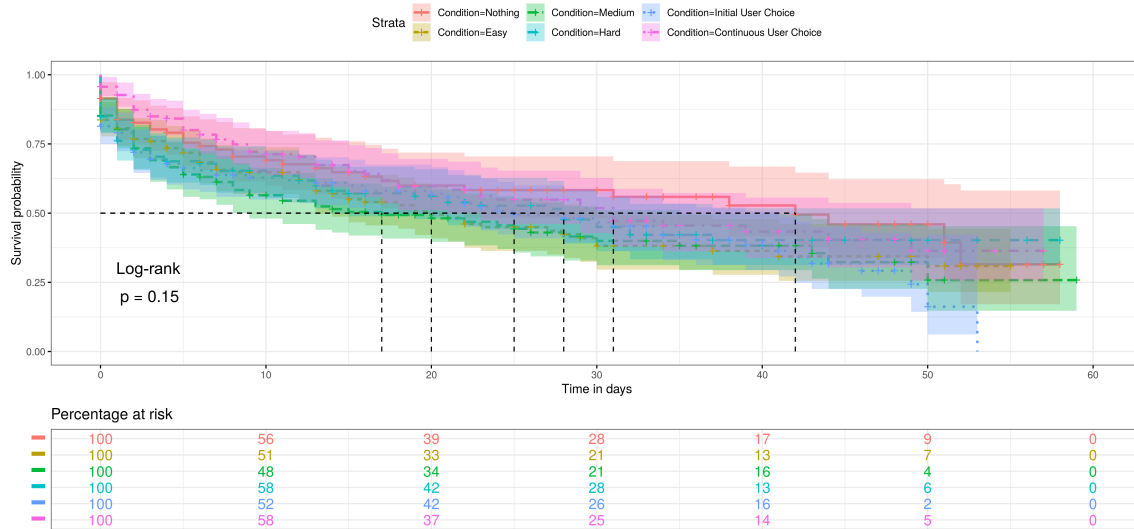


Fig. 3. No statistically significant differences were observed between conditions in terms of retention rates, between the difficulty levels to which users were randomized.

2.2 Results

The retention results are shown in Figure 3. We observe that there are no significant differences between the conditions in terms of retention rates. Hence, randomly assigning users to harder or easier interventions by default, does not increase attrition.¹

The efficacy results are shown in Table 3. We observe that the most effective condition – that is, the condition where time online is reduced the most – is asking users for their initial intervention difficulty preference and giving them that difficulty. We suspect that the reason this is the most effective is that users initially choose more difficult interventions during onboarding than what they would choose long-term, as we saw in Study 1 (Figure 5 of main paper). As for why it would be more effective than assigning hard interventions by default, given that HabitLab allows users to disable interventions, we suspect that users assigned to interventions harder than they would prefer would end up disabling interventions at higher rates.

¹Figure 3 shows results from all users in the experiment, but if we restrict to only users who had Facebook as a goal, the results are identical.

<i>Dependent variable:</i>	
Log daily time on Facebook	
Assigned to Easy	0.182 (0.302)
Assigned to Medium	0.223 (0.310)
Assigned to Hard	0.538 (0.307)
Initial User Choice	0.615 (0.298)
Continuous User Choice	0.495 (0.298)
Constant	5.323 (0.228)
Observations	3,524
Log Likelihood	6,688.215
Akaike Inf. Crit.	13,392.430
Bayesian Inf. Crit.	13,441.770

Note: p<0.05; p<0.01; p<0.001

Table 3. Daily time spent on Facebook in each of the conditions – initial user choice, continuous user choice, vs random assignments to various different difficulty levels. The baseline condition is where no intervention is shown. Units are log-normalized seconds.

3 SUPPLEMENT C: EFFECTS OF PROMPTING ON RETENTION

We observed in Study 2 of the main paper that attrition is significantly higher if users are shown the difficulty choice prompt on every visit, as opposed to only 25% of visits. This led us to believe that low-frequency prompting can be beneficial for retention, due to striking the right balance between giving users choices while not overwhelming them with choices. In this study we investigate the roles of choice and prompting frequency further, by comparing not just high and low frequency prompting – in this case, once per visit vs once per day – but also a condition where users can choose the prompting frequency, and a condition where the user can choose to ignore the difficulty prompt.

3.1 Methodology

In this experiment we study the effects of prompting frequency and user choice on user attrition rates. We do so by randomly assigning new users to one of the following four conditions when they install:

Choose Frequency: Every time the users are shown the difficulty selection prompt, they are also then shown a prompt asking them when they would like to be asked again about intervention difficulty. The options are next visit, next hour, next day, or next week.

Every visit, forced choice: Users are shown a prompt asking them what difficulty they would like to have this visit. They must make a choice in order to dismiss the dialog. This condition is equivalent to the high frequency, “100% of visits” condition from Study 2 of the main paper.

Every visit, optional choice: Users are shown a prompt asking them what difficulty they would like to have this visit. A counter is also shown – if they do not answer within 10 seconds, the dialog will dismiss itself and an intervention difficulty will be randomly chosen for that visit.

Once per day, forced choice: Users are shown a prompt asking them what difficulty they would like to have this visit. They must make a choice to dismiss the dialog. It is shown at most once every 24 hours – for the remainder of the day, the chosen result will be preserved.

We run this study over the course of 116 days, with 1562 users. We compare retention rates based on a Cox Hazard Regression model [2].

3.2 Results

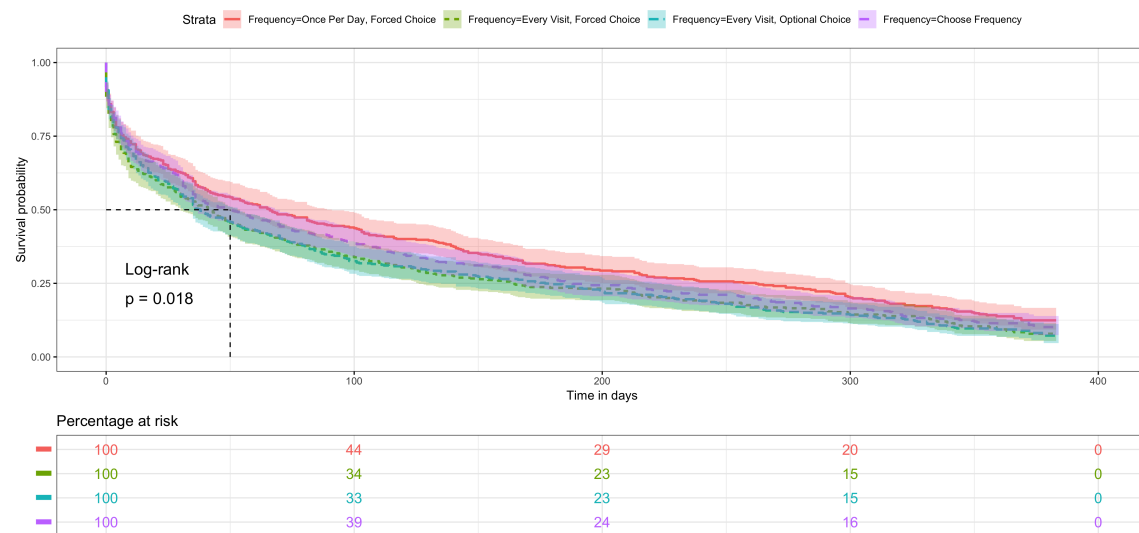


Fig. 4. No statistically significant differences in retention rates between the conditions of asking every visit vs every day, or making responses optional, or allowing users to choose when to be asked next.

Results of the Cox Hazard Regression model are shown in Table 4, and user retention is also depicted visually in Figure 4. There is a significant effect of condition on retention according to the log-rank test ($p < 0.05$). We see in Table 4 that the hazard ratio between the “Every Visit, Forced Choice” and “Once Per Day, Forced Choice” conditions is above 1 (1.24), and is statistically significant ($p < 0.01$). This means that user attrition is higher with high-frequency prompting (every visit), compared to low-frequency prompting (only once per day). We also see the same for the hazard ratio between “Every Visit, Optional Choice” and “Once Per Day, Forced Choice”, indicating user retention is higher with low-frequency prompting, than high-frequency prompting with optional choices .

Frequency	Beta (SE)	Hazard Ratio (95% CI)	p
Once Per Day, Forced Choice (ref)	-	-	-
Every Visit, Forced Choice	0.21 (0.08)	1.24 (1.06, 1.44)	0.006
Every Visit, Optional Choice	0.20 (0.08)	1.23 (1.06, 1.42)	0.008
Choose Frequency	0.11 (0.08)	1.11 (0.96, 1.29)	0.16

Number of events	1,414
Observations	1,581
Concordance	0.525 (SE = 0.009)
Likelihood ratio test	10.2 (df=3, p=0.02)
Wald test	10.06 (df=3, p=0.02)
Log-rank test	10.09 (df=3, p=0.02)

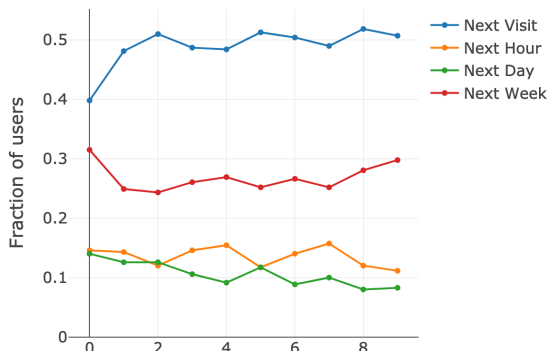
Table 4. Cox Hazard Regression table indicating effects of different prompting strategies on retention. Prompting every visit significantly increases attrition compared to once per day.

4 SUPPLEMENT D: CHANGES IN PROMPTING FREQUENCY CHOICES OVER TIME

In this study, we investigate whether users' prompting frequency preferences change over time. A possible reason why we might expect it to change over time is that users may become tired of answering the prompts after repeatedly answering them, and therefore gravitate towards a lower prompting frequency as time passes. The data used to generate these figures is from Study 3 in the paper, representing 31,240 exposure samples from 643 users over the course of 385 days.

In Figure 5, we show the first 10 choices of users who answered the "Ask me again about difficulty" prompt at least 10 times. There were 349 such users. The proportion of users who choose each time option does not change much over time, except for a brief increase in the fraction choosing to be asked again the next visit during the first 3 steps. If we look at the individual choices made by each users, shown in Figure 6, where each user is represented as a row, and each of their

Changes in user choices of when to be asked again



i-th time the user is choosing when to be asked next

Fig. 5. The first 10 choices of when to be prompted again, among the 349 users who made at least 10 choices. User preferences for prompting frequency do not change much over time.

Changes in user choices of when to be prompted again

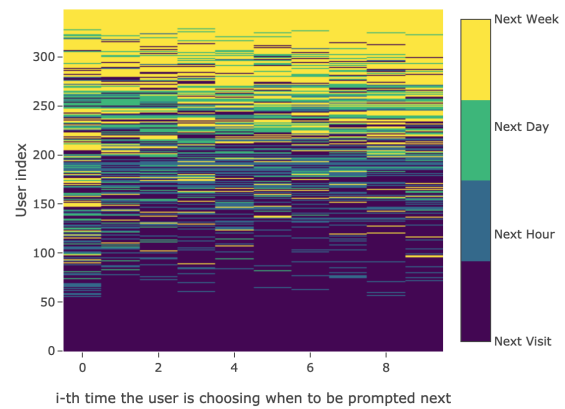


Fig. 6. The first 10 choices of when to be prompted again, among the 349 users who made at least 10 choices. Each of the rows represents a user, and each colored box on the grid represents a choice by a user.

