Put Your Money Where Your Mouth Is: Dataset and Analysis of Real World Habit Building Attempts

Hitkul Jangra[1], **Rajiv Shah**[1], **Ponnurangam Kumaraguru**[2]

[1]Indraprastha Institute of Information Technology, Delhi, India
[2]International Institute of Information Technology, Hyderabad, India
{hitkuli,rajivratn}@iiitd.ac.in, pk.guru@iiit.ac.in

Abstract

The pursuit of habit building is challenging, and most people struggle with it. Research on successful habit formation is mainly based on small human trials focusing on the same habit for all the participants as conducting long-term heterogonous habit studies can be logistically expensive. With the advent of self-help, there has been an increase in online communities and applications that are centered around habit building and logging. Habit building applications can provide large-scale data on real-world habit building attempts and unveil the commonalities among successful ones. We collect public data on *stickk.com*,[1] which allows users to track progress on habit building attempts called commitments. A commitment can have an external referee, regular check-ins about the progress, and a monetary stake in case of failure. Our data consists of 742,923 users and 397,456 commitments. In addition to the dataset, rooted in theories like *Fresh Start Effect*, *Accountablity*, and *Loss Aversion*, we ask questions about how commitment properties like start date, external accountability, monitory stake, and pursuing multiple habits together affects the odds of success. We found that people tend to start habits on temporal landmarks, but that does not affect the probability of their success. Practices like accountability and stakes are not often used but are strong determents of success. Commitments of 6 to 8 weeks in length, weekly reporting with an external referee, and a monetary amount at stake tend to be most successful. Finally, around 40% of all commitments are attempted simultaneously with other goals. Simultaneous attempts of pursuing commitments may fail early, but if pursued through the initial phase, they are statistically more successful than building one habit at a time.

Introduction

In their endeavor to quit smoking, 92.5% of individuals experience failure (Creamer et al. 2019). Only 20% are able to lose weight (Anderson et al. 2001), and 9% can stick to their new year's resolutions (Norcross and Vangarelli 1988). Building new habits is hard; most people struggle with it.

Recently we have also witnessed a rise in communities and applications centered on helping people in pursuit of habit formations. Reddit communities like *r/loseit*[2] and *r/stopsmoking*.[3] Applications like *Streaks,*[4] *Strides,*[5] *and Way of Life*[6] help users create goals and track progress. Some advanced applications like *StickK, Habitica,*[7] *Habitshare*[8] leverage previous research to build features about accountability, gamification, and incentives which can improve users' propensity to succeed in their goals.

Past research related to habit-forming can be widely divided into two parts: i) Sociology theories relating to behavior change like Operant conditioning (Skinner 1938), loss aversion (Kahneman 1977), and fresh start effect (Dai, Milkman, and Riis 2014); ii) application/effect of these theories in a specific scenario. (Skarupski and Foucher 2018) used peer accountability to improve writing habits. (Giné, Karlan, and Zinman 2010) showed the effectiveness of loss aversion in quitting smoking, and (Dai, Milkman, and Riis 2014) demonstrated that the commitment towards goals, increases when they are started on a new week or month. Habit formation has also been of interest in the computational social science community, with research evaluating the effect of online communities on various habits like physical activity (Althoff, Jindal, and Leskovec 2017), smoking and drinking relapses (Tamersoy, Chau, and De Choudhury 2017), drug consumption (Jangra, Shah, and Kumaraguru 2023), quality of user-generated content (Cheng, Danescu-Niculescu-Mizil, and Leskovec 2014) and involvement in open-source projects (Valiev, Vasilescu, and Herbsleb 2018).

Though current literature explores various habits, none evaluates large-scale heterogeneous attempts of habit building, unveiling the prevalence and effectiveness of guidelines suggested by social science literature. Historically, such studies had a high logistic and monetary cost. However, the advent of habit-tracking applications provides us with data on real-world habit building attempts.

In this study, we collect publicly available data about users' past attempts at habit-forming on *stickk.com*. Grounded in theories of loss aversion (Kahneman 1977), fresh start effect (Dai, Milkman, and Riis 2014), and ac-

[1]https://www.stickk.com/
[2]https://www.reddit.com/r/loseit/
[3]https://www.reddit.com/r/stopsmoking/
[4]https://streaksapp.com/
[5]https://www.stridesapp.com/
[6]https://wayoflifeapp.com/
[7]https://habitica.com/static/home
[8]https://habitshareapp.com/

countability, we perform characterization on temporal patterns (when commitments starts), incentive structures (type and amount of stake), and reporting habits (frequency and length of reporting). Further, we use an unsupervised retrieval method based on Word2Vec (Mikolov et al. 2013) and relevance feedback (Salton and Buckley 1990) to classify commitments into different categories. Finally, we use survival regression to analyze the effect of these confounders, like start date, stake, length of commitment, and category, on the commitment's success rate.

We discover 1) the fresh start effect is very prevalent. Users are 40% more likely to start a commitment on 1^{st} of a month, or Monday, and compared to an average day in the year, four times more commitments are started on New Year's. Though commitments started on New Year's are more likely to fail, commitments started on other temporal landmarks do not affect the likelihood of success; 2) Only 29% and 19% of total commitments have monitor stakes and external accountability attached respectively, but doing so significantly increases the success rate; 3) Success rate increases if a stake is given to an anti-charity instead of a charity on failure; 4) A critical factor in increasing success rate is to keep short-term commitments (7 weeks is the base hazard) and do frequent check-ins on the application; 5) Users pursuing multiple commitments simultaneously are more likely to succeed than users with singular goals.
In summary, our main contributions are:

1. A large-scale heterogeneous dataset of habit building attempts with detailed information about associated parameters and success rates.
2. To quantify the prevalence, patterns, and effect of parameters like start date, accountability, monetary stakes in habit building attempts.

Our work impacts researchers, users, and platform owners by providing a fertile base for developing future research or tools. Our dataset can be used to evaluate the effect of more complex factors like the types of habit overlaps or streaks on commitment success.

Data and Code: Dataset, and code is available at *https://precog.iiit.ac.in/research/put-your-money/.*

Theories and Research Questions

The *Fresh Start Effect* (Dai, Milkman, and Riis 2014) is defined as the human tendency to take action towards your goals starting a specific key date. Dates that stand out as more meaningful, such as the new week or month, birthday, or a holiday, signal the start of a distinct period. These "temporal landmarks" make it easier for people to mentally separate from their past imperfections and failures. (Dai, Milkman, and Riis 2015) showed that the searches related to dieting, visits to the gym, and self-reported motivation towards the goals increases after temporal landmarks. Another study by (Alter and Hershfield 2014) showed that the rate of exercise, extramarital affairs, and suicide increased when adults approach a new decade in their age, i.e., ages 29, 39, 49. Authors claim that certain numerical ages show a greater sense of self-reflection than others. Thus we expect users would be more likely to start new commitments on key dates and ask our first question:

RQ1. [Key Dates] *What is the prevalence of the fresh start effect in commitment start dates? Do commitments started on these dates have a higher rate of success?*

Best-selling books Tiny Habits (Fogg 2019) and Better than Before (Rubin 2015) claim accountability to be a key component in successful habit building. Past research has also shown a consensus with these claims. A 15-week study of 704 participants showed increased weight loss when paired with a support buddy (Dailey et al. 2018). Similarly, people working in an accountability group showed improvement in writing habits (Skarupski and Foucher 2018). (Renfree et al. 2016) showed that check-in reminders in habit formation applications improve adherence, though they also create a dependency. Accountability can be of many kinds, like self (check-in to an application), external (buddy to validate your progress), and social (support community or social media announcements). Grounded in these, we ask:

RQ2. [Accountability] *What is the extent of different accountability methods (external and social) and their effect on commitment success?*

People are motivated or deterred from doing things based on the power of incentive or fear of loss, colloquially known as the method of "Carrots and Sticks" (Ayres 2010). This effect is rooted in *Operant Conditioning*, stating the probability of acting in the future is a function of the stimuli received in the past (Skinner 1938). Stimuli can be *appetitive* or *aversive*. Appetitive stimuli are those one voluntarily approaches, while aversive stimuli are those one try to avoid or escape. Analyzing aversive stimuli, (Kahneman 1977) found that the pain of losing is psychologically twice as powerful as the pleasure of gaining. This was termed as *Loss aversion*, and have shown wide applications ranging from designing insurance products (Putler 1992), to help people abstain from smoking (Giné, Karlan, and Zinman 2010). Methods of reward and punishment have been at the center of habit building recommendations. In the context of habit building, the most common method of incentive/loss is putting monetary amounts at stake to an entity in case of failure. Considering this, we ask:

RQ3. [Stake] *Does monetary stake affect the probability of success in a commitment? Does the nature of the entity with which money is staked affect the success rate?*

Guidelines on simultaneous habit forming are conflicting. On the one hand, it is suggested that an individual should focus on one goal at a time (Dalton and Spiller 2012) and strive to reach the state of "automaticity" (Lally et al. 2010). This means, initially, a new habit needs conscious effort, but after a while, it becomes an automatic routine, after which the person can move on to other commitments. On the other hand, concepts of habit stacking/anchoring (Fogg 2019) talk about linking a series of habits to one after the other. For example, working out makes you more likely to have a healthy meal. This is rooted in Behavioral Momentum Theory (Nevin and Shahan 2011). Once you are in a flow of doing things, momentum will carry you through the series of habits. To evaluate the effect simultaneity has on success, we ask:

RQ4. [Simultaneity] *What is the extent of the user trying*

to pursue simultaneous goals? How does it affect the success rate of a user?

Data Source and Description

We use data from the habit building platform *stickk.com*. We chose StickK because of the availability of large-scale public data, heterogeneous types of habits, and a diversity of features like allowing users to have referee and put money on stake. A user can have multiple habits called commitments on the platform. Commitments can be active, i.e., being pursued right now or completed. Optionally, users can put money on stake for each commitment, which needs to be paid in case of failure. We use the entire historic dump of the platform to create a final dataset of 742,923 users who created 397,456 commitments with a collective $35.5 Million on stake. Table 1 shows our dataset statistics.

Each user profile has a unique numeric id, username, and joining date. Optionally, users can also add display pictures, location, interests, and a bio message. Users can decide to keep their profiles public or private. Out of the total platform users, 6.4% have designated their profiles as private. The unique numeric id allows us to identify the chronological order of private or deleted profiles but not any information associated with them. For such users, we approximate their joining date using a public profile before or after them in the sequence. [9]

Figure 1 shows the cumulative distribution function (CDF) of the joining dates for users. We observe a linear growth in the number of users over the year, with a slight bump in the rate of private account creation between 2011 to 2013. We also observed an abnormal number of account deletions in 2011, potentially caused by a platform-level glitch. Since our analysis are only performed over public profiles, this does not affect our findings.

Our dataset has 397,456 commitments created by 244,313 unique users (32.8% of total users). Among users with commitments, 187,333 (76.6% of users with commitment) have only one completed commitment; 95% of users have less than four completed commitments in the past. Each commitment has a title, optional description, length, and reporting interval, which defines how often users would report their status. At every reporting period, the user declares whether or not they succeeded in achieving the goal. The user decides the total length and reporting interval.

Users can optionally assign referees and supporters to a commitment. A referee's job is to audit the user's performance and mark the status for each reporting period. [10] Whereas supporters can provide encouragement and social accountability to the user. Unlike referee, supporters can not audit a users performance. Each commitment can have only one referee but any number of supporters. Only 19% and 8.6% of total commitments had referee and supporters assigned, respectively.

Observation 1 (RQ2: Accountability Extent) *Most users do not utilize accountability methods in their pursuit of building habits. Only 19% of total commitments had external accountability (referee), and 8.6% had social accountability (supporters) attached to them.*

Finally, the user can also attach a monetary stake to the commitment. The user chooses a stake amount per reporting period, leading total money at stake to be *stake per period* $\times$ *# of reporting periods*. When a user fails to achieve the goal during a reporting period, the stake for that period is awarded to a selected entity (charity, anti-charity, friend, or StickK). It is worth noting that the complete freedom in allowing users to set parameters of the commitments does lead to outlier cases in the dataset, e.g., commitments with unusually long lengths or very high stakes. All analysis in this paper is performed after removing outliers from the dataset using the Interquartile range (IQR) method (Dekking et al. 2005), as most attributes in our data follow a skewed distribution.

[9] For most cases joining dates of the public profile before and after is the same, ensuring the deleted/private profile was also created on the same date. In cases deviating from this pattern, we assign the joining date of the previous public profile.

Total # of users	742,923
# of public users	655,750
# of private users	47,716
# of deleted users	39,457
# of total commitments	397,456
# of users with commitments	244,313
Total $ at stake	$35,598,253.74
Minimum $ at stake	$0.5
Maximum $ at stake	$20,000
Date range	19 Oct 2007 - 17 Aug 2023

Table 1: Dataset details. 32.8% of total users have created commitments, with total $35 Million at stake.

Temporal Analysis

Temporal patterns have proven to help analyze trends. This section first discusses the temporal patterns observed in the commitments' total length and reporting period. Then in the context of **RQ1**, we look at the relations between the commitment start date and temporal landmarks.

Commitment Length and Reporting Interval

Any habit building exercise aims to reach the state of automaticity (Lally et al. 2010). Initially, the user must invest effort towards the commitment until it becomes routine. This makes the initial length of commitment an essential parameter for habit building. Figure 2 shows a histogram of lengths of commitments in our dataset. We observe prominent peaks at weeks 4, 8, and 12 and relatively high frequencies at weeks 16 and 20. This shows that users think about habit building planning more often in a quantum of months, with three months being the most popular choice.

[10] The platform does not provide a method for referees to audit. It is managed offline between the user and the referee. The latter's

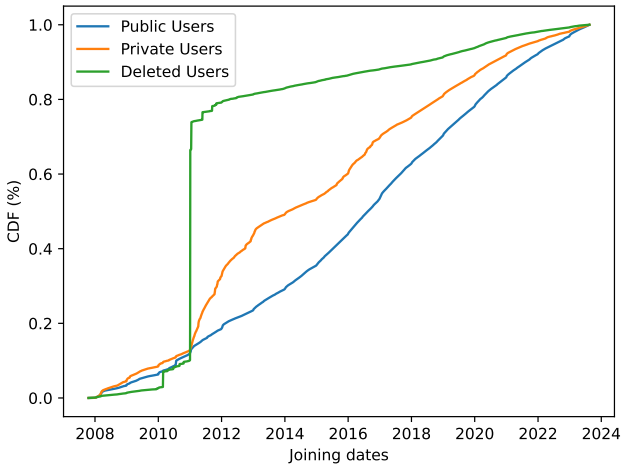


Figure 1: CDF of joining date's to the platform. We see a linear increase in number of users over the year. 2011 shows a spike in the number of deleted users potentially caused by platform-level data loss/malfunction.

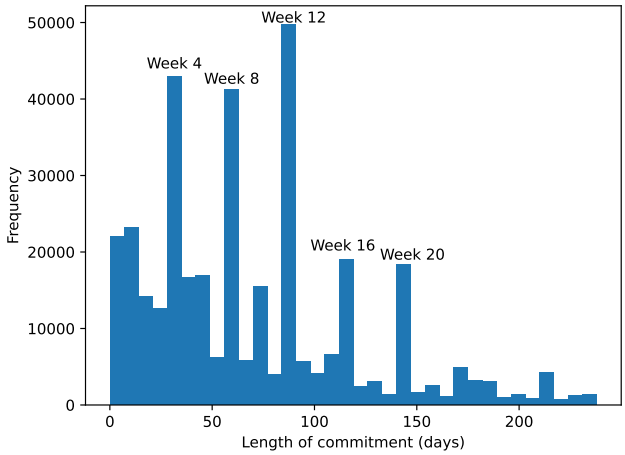


Figure 2: Distribution of length of commitments. Lengths in quantum of months like week 4, 8, 12 are most frequent.

Further, we look at the length of reporting intervals chosen by users. Out of the total commitments, 72% (286,206) and 10.6% (42,263) are set up to report progress weekly and daily, respectively. We compare the median commitment length for commitments with daily and weekly reporting intervals. On average, commitments with daily reporting had a length of 35 days, compared to 84 days (12 weeks) for the ones with weekly reporting, showing users shift to more coarser reporting intervals as the total commitment length increases.

Observation 2 (Commitment Length) *Users plan habit building in the quantum of months, i.e., 4, 8, and 12 weeks, with the daily or weekly reporting intervals being the most popular.*

Commitment Start Date

The Fresh Start Effect (Dai, Milkman, and Riis 2014) is a cognitive bias which is the user's tendency to start a new goal on specific dates known as temporal landmarks. These landmarks can be general, like New Year, new week/month, or specific, like birthdays, start of a new job/semester. Considering the data available, we keep our analysis limited to the effects of general temporal landmarks, specifically, the start of a new week i.e. Monday, 1^{st} of a new month, and New Year.

Figure 3 shows the distribution of commitment start dates across (a) weekdays, (b) days of the months, and (c) days of the year. A skew towards Monday, 1^{st} date, and New Year is apparent. Activities stay relatively high for the first 15 days of the year, with the highest on 1^{st} January. Users are four times more likely to start a commitment on New Year than on an average day of the year. Similar observations are made in the patterns of new weeks and months too. Compared to an average day, users are 40% and 38% are more likely to start a commitment on Monday or 1^{st} of a month, respectively.

Observation 3 (RQ1: Key Dates Extent) *The fresh start effect is prevalent among users starting new habits. Users are 40% more likely to start on Monday or 1^{st} day of the month. Four times more commitments are started on New Year's than on an average day.*

Stake Analysis

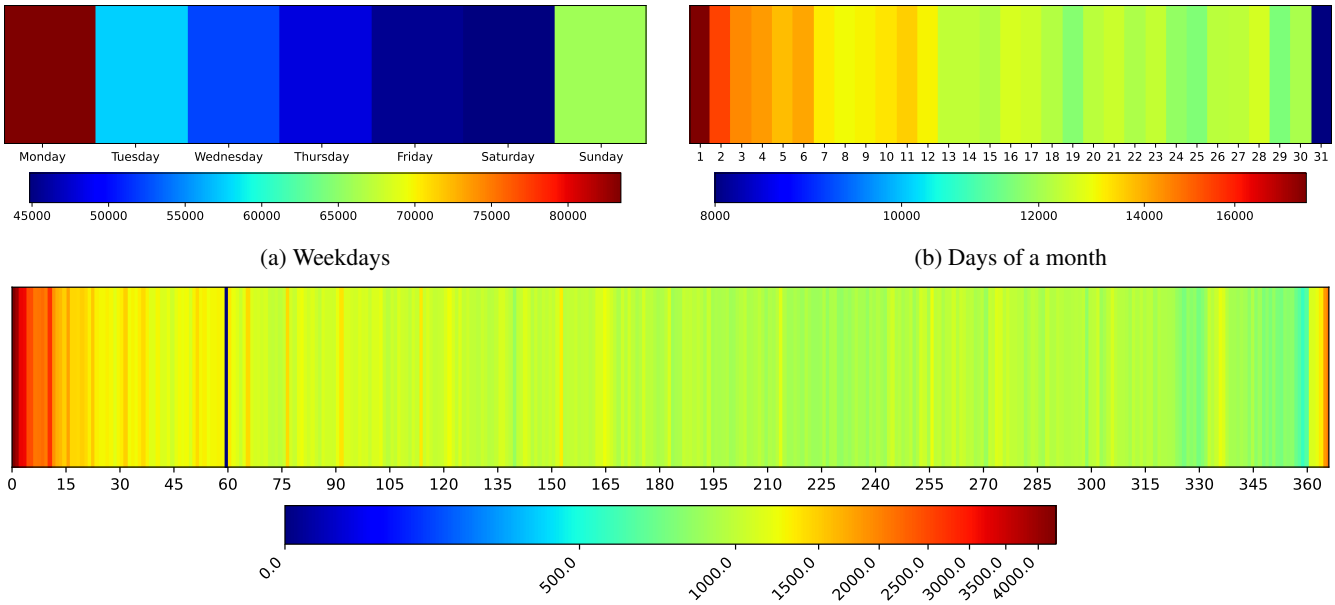
This section discusses the extent part of **RQ3**. The theory of loss aversion tells us that the physiological pain of losing is very powerful and can induce behavioral change (Kahneman 1977). Stickk allows its users to leverage this in their habit building journey by allowing users to attach a monetary stake to the commitments if they wish to do so. In case of failure to achieve the goal for a specific reporting interval, the stake is transferred to an entity chosen prior by the user. StickK allows users to choose from 4 different kinds of entities:

- **StickK**: Stake is passed on to the platform itself.
- A **friend** chosen by the user.
- **Charity** of user's choice.
- **Anti-charity**: An Anti-charity[11] is an organization whose views user strongly oppose. The assumption being a user would want to avoid extending monetary value to such an organization, increasing the motivation to succeed in the commitment.

Table 2 shows a distribution of different types of stakes in our dataset. For 71% of total commitments, there are no stakes attached. Anti-charity is the most popular for the commitments with stakes, followed by charity, friend, and StickK. $5 is the most common stake amount, followed by $10, $50, and $20, irrespective of the stake type.

report is considered in case of a conflict.

[11] https://stickk.zendesk.com/hc/en-us/articles/206833337



(a) Weekdays

(b) Days of a month

(c) Days of a year

Figure 3: Distribution of commitment start dates. Users are four times more likely to start a commitment on New Year (c) and 40% more likely to start on Monday (a) or 1^{st} of the month (b).

Type of stake	Frequency	Percentage
No stake	282,366	71.0%
Anti-charity	60,368	15.2%
Charity	27,053	6.8%
Friend	21,711	5.5%
StickK	5,958	1.5%

Table 2: Distribution of different types of stakes. 71% of total commitments do not have any stakes attached to them.

Observation 4 (RQ3: Stake Extent) *29% of the total commitments have a monetary stake attached to them. Anti-charity is the most common stake type and $5, $10 the most common stake amounts per period.*

Commitment Classification and Simultaneity

In this section, we start with classifying commitments into different classes of habits. This helps us understand what habits and goals are prevalent among users and how they have changed over time. Further, as part of **RQ4**, how common it is for users to pursue multiple commitments simultaneously.

Title of a commitment talks about what habit users are trying to build. These are generally short combinations of tokens like *"Lose weight", "Study two hours"*, or *"exercise 3x a week"* instead of complete sentences. Considering this semantic property of titles, we used a Word2Vec-based relevance feedback (Balsamo et al. 2021) method for retrieval/-classification rather than a supervised neural classifier. [12] We start with a set of classes and respective query terms based on common occurrences observed in the manual inspection of the dataset. While retrieval, we match a title to the class if it has query terms or terms semantically similar to query terms. Pretrained Word2Vec (Mikolov et al. 2013) embeddings were used for semantic matching. Similar terms were added to the query of the respective class as part of the feedback for the subsequent retrieval cycle. These steps are repeated till convergence. Table 5 in Appendix contains the initial list of classes and related query terms.

Table 3 shows the frequency of the top 15 classes identified, and Figure 4 shows the proportion of these classes over the year. In total, our Word2Vec-enabled relevance feedback algorithm was able to classify 73% of total commitments successfully. Habits related to health make 53.94% of total commitments. Weight-related commitments were most common at 25.81%, followed by exercise (22.31%) and study (5.54%).

From Figure 4, we can observe that though the most famous, proportion of commitments related to weight has gone down over the years, an increase in commitments related to food has gone up. This probably indicates a shift in user mindset about how they perceive food and weight management. We also observe a rise in the proportion of commitments related to reading, sleep, meditation, and digital (limiting phone and internet usage), indicating a shift in emphasis towards mental wellness and self-development.

[12] A neural classifier would probably work well for this task. However, relevance feedback works well enough without requiring large annotations and computational power.

Habit class	Frequency	Percentage
Weight	102,584	25.81%
Exercise	88,659	22.31%
Study	22,040	5.55%
Food	14,202	3.57%
Smoking	10,251	2.57%
Sleep	8,956	2.25%
Read	8,735	2.19%
Meditate	7,886	1.98%
Money	7,639	1.92%
Write	5,981	1.50%
Business	4,535	1.14%
Alcohol	4,194	1.05%
Digital	2,826	0.71%
Pornography	2,486	0.62%
Self-care/ Personal hygiene	1,761	0.44%

Table 3: Fifteen most common habit classes in our dataset. Habits related to health (Weight, Exercise, Food, Sleep) make 53.94% of total commitments.

Simultaneity

In our effort to answer **RQ4**, we want to see how common it is for users to pursue multiple habits in parallel. During our manual inspection, we observed two ways users were structuring multiple habits on the platform. 1) Multiple commitments running during the same time period, 2) User's listed multiple goals in a singular commitment, e.g., *"Lose weight and exercise"* or *"quit smoking and study for finals"*. Our analysis considers both types of structures since our relevance feedback method can perform multi-label classification. Despite being advised against (Dalton and Spiller 2012), 42% of total commitments (167,511) are pursued with other commitments, with similar habits like weight and exercises often paired together. §Survival Analysis discusses the effect of the simultaneity on commitment success in detail.

Observation 5 (RQ4: Simultaneity Extent) *Users tend to pursue multiple habits together, with 41% habit building attempts paired with other goals.*

Survival Analysis

A key component common across all our research questions is measuring the effect of commitment properties (e.g., start date and length) on users' success rate in the habit building pursuit. We perform survival analysis (Miller 2011) on our data to answer the "effect" part in **RQ1-4**. Specifically, we used the Kaplan–Meier estimator (Kaplan and Meier 1958) for measuring the effect of categorical variables and Cox regression (Cox 1972) for continuous variables. In this section, we first define how we measure the success rate of a commitment, followed by the details about our survival analysis experiment and the variables used.

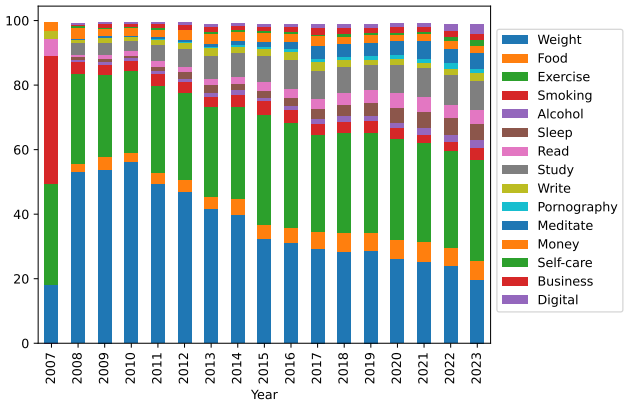


Figure 4: The proportion of types of habit over the years. We observe a decline in weight-related habits and an increase in habits related to sleep (brown), meditation (blue), reading (pink), and digital technology (purple).

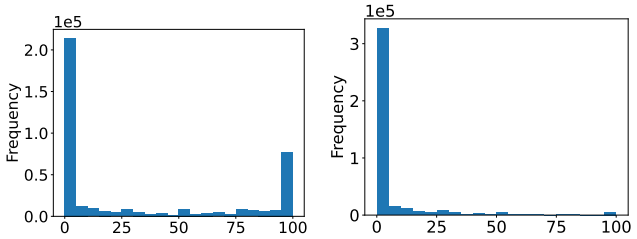
Commitment Success Rate

An advantage of using data from StickK is that success for a component is not binary. During a commitment, users check-in at pre-chosen intervals (weekly and daily, most common) to update if they could stick with the habit. A time-stamped record of this checks-in is maintained and is available in our data. A reporting interval can have three statuses, *successful*, *not successful*, or *not reported*. Figure 5 shows the proportion distribution of all three statutes across commitments. We observe that the peak frequencies for successful status are in $< 5\%$ bucket or $> 95\%$ bucket, showing that users tend to fail early or do well. Interestingly, the frequency of high rates of not successful is low, but not reported is high, indicating the users who had trouble pursuing the commitments tend to discard the pursuit. Historically, of the total $35.5 Million on stake, $4.2 Million and $1.5 Million have been lost due to intervals being marked as *not reported* and *not successful*, respectively.

Observation 6 (Success Rate) *Users either fail early or stick through the entire commitment, indicating that the initial phase of habit building is the most critical. Users with low success levels tend to refrain from returning to the platform for reporting.*

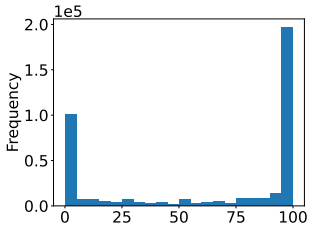
Experiment Details

Typically data for survival analysis experiments are set up in terms of the time it took for an event of interest to occur, e.g., in a study of patients with critical cancer, how many days did a patient survive after the initial diagnosis? In our case, the lengths of commitments are widely different. Hence, to standardize the comparison, instead of measuring success in an absolute number of days, we measure it in terms of the proportion of reports marked as successful. In the terminology of survival analysis, the timeline of our experiment becomes 0 to 100, and the event of interest for a commitment occurs at a timestamp represented by the proportion of reports marked as a success by the user. Commitments with a 100%



(a) Proportions of status marked *successful*.

(b) Proportions of status marked *not successful*.



(c) Proportions of status marked *not reported*.

Figure 5: Distribution of reporting interval statuses. Users tend to either fail early or do really well (a). The frequency of high rates of not successful is low (b), but not reported is high (c), indicating user's abandonment of the commitment.

success rate are marked as right censored entries. In order to prevent our experiments from being biased by commitments that were made frivolously, we excluded all commitments which did not have even one report marked as successful.

We used Kaplan–Meier estimator (Kaplan and Meier 1958) to study the effect on survival in cases where the treatment variable is categorical. It is used to estimate the survival function, which measures the fraction of users surviving for a particular time after treatment. To measure the effect of a categorical variable, a comparison is made across the Kaplan–Meier estimates for shards of data divided based on the values the variable can take. Logrank test (Bland and Altman 2004) is used to validate if estimates generated from different variable values are statistically significant. In cases where the treatment variable is continuous, we use the Cox proportional hazard model (Cox 1972), which fits a regression model to evaluate how changes in a continuous variable's value affect a user's survival.

In the context of **RQ1**, we conduct three experiments where the start date of the commitment is the treatment variable. Specifically, we compare the success rate of commitments that started on New Year's, 1^{st} of any month, and Monday with their respective counterparts. Extending on the theme of temporal properties, we also evaluate the effect commitment length and reporting interval have on success rate. Though reporting interval is a continuous variable, 82.6% of all the commitments are set up to report weekly or daily. Hence, we treat reporting interval as a categorical variable with possible values of weekly, daily, and others. For **RQ2**, we compare the success rates of commitments with an external referee vs. self-referring and the effect of the number of supporters (social support) on commitment success. To answer **RQ3**, we measure the effects changes in monetary value of stake per reporting period and who it is bet against (Charity, Anti-charity, Friend, or StickK) has on the commitment success rate. Finally, for **RQ4**, we compare the survival functions of commitments pursued in simultaneity with others vs. individually.

Covariate	**HR**	**95% CI**
# of reports	1.0447***	[1.0437, 1.0457]
Length of commitment	1.0045***	[1.0043, 1.0046]
# of supporters	0.9797***	[0.9725, 0.9870]
$ on stake per period	0.9474***	[0.9467, 0.9482]

$^{***}p <= 0.005$

Table 4: Cox Regression results. An increase in monetary stake and the number of supporters increase the success rate. In contrast, increasing the length and number of reports is more hazardous for the user.

Results

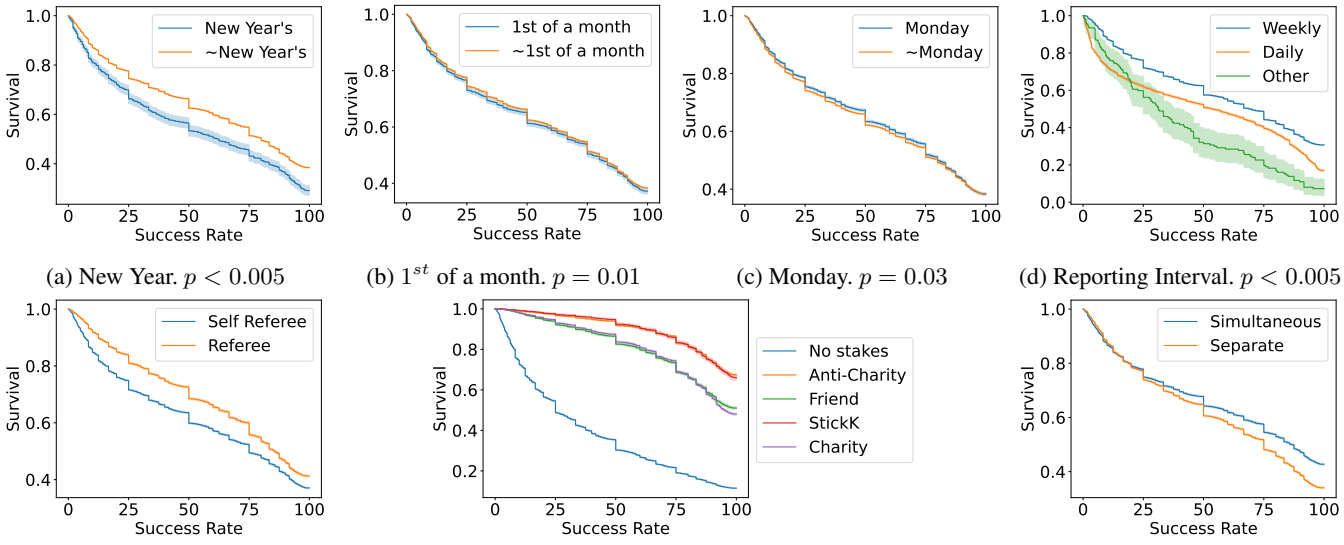
Figure 6 and Table 4 show all the results for our survival analysis. As seen in Figure 6(b), 6(c), commitments started on temporal landmark days like 1^{st} of a month or Monday did not perform any better compared to the ones started on any other day. Conversely, commitments that are started on New Year's have a much worse survival rate than those started later in the year (Figure 6(a)).

Increased commitment length and number of reporting intervals lead to a decrease in success rate (Table 4). The baseline hazard for commitment length was found to be at seven weeks. Regarding reporting intervals, commitments with weekly check-ins have the highest success rate, followed by daily and then others. We observe a sharp decline in the survival function for commitments with daily reporting, indicating that many commitments of this type fail with less than a 20% success rate. However, commitments that pass this threshold tend to achieve greater success rates, as depicted by a reduction in the slope of the survival curve later on (Figure 6(d)). Both external and social accountability have a significant positive impact on success rates. Commitments with external referees achieve higher success rates compared to one self-referred (Figure 6(e)). Similarly, increasing the number of supporters (social accountability) on the commitment reduces the hazard.

Observation 7 (RQ1: Key Dates Effect) *Starting a commitment on a temporal landmark day like Monday or* 1^{st} *of a month does not affect the success rate of habit building.*

Observation 8 (RQ2: Accountability Effect) *Having external and social accountability attached to a habit building pursuit in terms of referee and supporters significantly increase the odds of success.*

Commitments with no monetary stakes perform much worse than those with money on the line. An increase in the stake amount per period positively affects the success rate.



(a) New Year. $p < 0.005$ (b) 1^{st} of a month. $p = 0.01$ (c) Monday. $p = 0.03$ (d) Reporting Interval. $p < 0.005$

(e) Referee. $p < 0.005$ (f) Type of stake. $p < 0.005$ (g) Simultaneity. $p < 0.005$

Figure 6: Survival analysis results (Kaplan–Meier curves). (a) Commitments started on New Year's have a lower survival probability. Starting on 1^{st} of a month (b) or Monday (c) does not affect the survival function. (d) Reporting every week increases survival. Commitments with an external referee (e) and monitory stake have better survival (f). (g) Finally, Pursuing multiple goals may fail early, but if pursued, it is better than pursing one commitment at a time.

Interestingly, along with the amount, the entity which the stake is a bet against, also strongly influences users' success rate. Commitments where lost money went to anti-charity or the platform (StickK) had better survival than those with money going to charity or a friend (Figure 6(f)). Finally, Figure 6(g) shows the effect of pursuing simultaneous commitments on success. Though initially, users pursuing individual goals have a mildly better survival function, but pass a success threshold (approximately 20% success rate), users pursuing multiple commitments tend to have statistically better survival than those pursuing an individual goal.

Observation 9 (RQ3: Stake Effect) *Amount of monetary stake and who it is bet against strongly affect the success of habit building pursuit. Adherence increases with the amount at stake and when it is a bet against an entity that may induce a greater sense of loss, like anti-charities.*

Observation 10 (RQ4: Simultaneity Effect) *Though sustaining multiple habits initially may be challenging, passing a threshold, users pursuing multiple habits together tend to perform better than those building one habit at a time.*

Discussion and Conclusion

Research Questions

In **RQ1**, we analyze the *Fresh Start Effect* (Dai, Milkman, and Riis 2014), a cognitive bias that defines the human tendency to start taking action towards a goal on specific dates, also known as temporal landmarks. These dates can be general, e.g., New Year's, the start of a new month/week, or specific to the user, like birthday or the start of a new job/semester. In this paper, we study the effects of only the general landmarks. We compare the likelihood of a commitment starting on temporary landmarks and how the success rate of such commitments is different. We found that the fresh start effect is highly prevalent, with 40% more commitments starting on 1^{st} date or Mondays compared to the average day of the month or week. Our analysis does not show any benefits of taking action on these landmark days, with the success rate of commitments started on these dates statistically the same as others. However, such behavior does add an opportunity cost for the user by introducing a delay between the decision to pursue a habit and taking action toward it. These delays can lead to overindulgence (justified as "one last time"), distractions, or loss of motivation. In the absence of any statistical edge, users should take immediate action and not wait for specific days. Further, users are four times more likely to start a commitment on New Year's than on an average day, but commitments started on New Year's are much more likely to fail.

Accountability is an essential factor for successful habit building. In **RQ2**, we study the two accountability options in our data. Referee adds external accountability to a commitment by verifying the user's progress, and supporters can provide social accountability to the user. We found that though users leveraging accountability is rare, with only 19% commitments having a referee and 8.6% having supporters, it is a strong determiner of success. The presence of an external referee and an increased number of supporters lead to a higher success rate. Often, habit building is perceived as an individual pursuit. Adherence increases when others are involved, maybe because users attach success to social standing, and individuals try to present an idealized version of themselves (Goffman 1959).

Theory of *Loss Aversion* tells us that the psychological pain of losing is twice as powerful as the pleasure of gain. In **RQ3**, we explored how the deterrence of loss can help users in habit building. On StickK, this is manifested by allowing users to assign an optional monetary stake to the commitment. In case of failure to achieve the goal during a reporting period, the amount on stake is passed on to a pre-chosen entity, which can be a friend, the platform itself, or a choice of charity/anti-charity. Like accountability, users' leveraging stakes is uncommon, with only 29% of commitments having it, but it is a strong determinant of success rate. An increase in stake amount causes an increase in success rates. Commitments with no stakes have a steep decaying survival function. Effects of loss aversion are also observed in the types of stakes, with entities that induce a higher sense of loss like anti-charity or platform, leading to a statistically prolonged survival (higher success rate) function for the commitment.

We used a Word2Vec-enabled relevance feedback system to classify commitment into various classes. Fifty four percent of all habits are related to the health of the user. Over the years, we have observed a reduction in the proportion of commitments related to weight and an uptick in the habits related to food, sleep, meditation, and reading. This shows a shift in users' perspective towards a more holistic approach to health.

Finally, in answering **RQ4**, we found that users lean towards developing multiple habits at a time, with 41% of total habit building attempts being made in simultaneity with others. The effect of simultaneous habit building has been unclear. On the one hand, some research suggests users should focus on one thing at a time (Dalton and Spiller 2012), but theories like habit stacking/anchoring (Fogg 2019) suggest using a combination of related habits at a time. Survival analysis of our data showed that habits practiced in simulating lead to a higer success rate. Through simulations, habits initially have a mildly lower success rate; if a user survives this phase, later on, the success rate is statistically higher, showing the benefit of behavioral momentum (Nevin and Shahan 2011) in habit building. Most people want to create multiple habit changes, and our analysis shows that this is not only possible, but it can be a better approach. The worst survival of simulation commitments in regions with low success rates indicates that starting with multiple habits together can be challenging and requires proper planning. Further analysis is required to answer planning-related questions such as what kinds of habits go together well or the optimal number of habits to build simultaneously.

Implications and Use Cases

Any habit-tracking tool aims to enable the users in achieving their goals. We believe the findings in this paper provide direct, actionable insights for both users and platform owners. Research related to habit building exercises can be logistically challenging. Most of the past literature is based on small-scale single-habit trials. Our data will enable researchers to observe patterns and validate theories on a large-scale dataset of real-life heterogeneous habit building attempts. In this paper, we measure the extent and effects of various commitment properties on success rate. However, this data allows us to study multiple aspects of habit building not touched on in our analysis. Firstly, we define success rate as the proportion of the reporting period marked as successful. We do not account for the chronology of the reports. It would be interesting to explore the relationship between user success rate and their commitment stage. Further, how do streaks relate to the overall success rate of the attempt? Second, since we find simultaneous habit building attempts are very prevalent, their scope of exploring the specifics of such attempts. What commitments are most often linked together? Are certain combinations more favorable than others? Finally, since weight management is one of the most common commitments, the platform allows users to add specific information to such commitments like start weight, end weight, rate of change in weight, etc. Though not utilized in our analysis, this information is in our data. It can enable researchers to explore weight management specific questions such as what is the optimal rate of weight loss from the perspective of adherence? Further, insights from our analysis can help users to plan their goals and to platforms for designing features and interventions.

Threats to Validity

Ensuring generalizability and data accuracy is always challenging while using online data to analyze offline human behavior. Our analysis is also susceptible to these challenges. Firstly, we define success rate as the proportion of reporting periods marked as successful by the user (by the referee in case one is present). We do not have a way to validate the authenticity of these report, ensuring that the user was able to achieve the goal. The converse is also true; lack of a report does not necessarily mean a relapse in habit but can be a function of other factors such as not liking the platform. We use large-scale longitudinal data consisting of hundreds of thousands of commitments, reducing the possibility of large-scale tampered data. Success at habit building is also affected by the user's environmental factors, like social/family support and motivation behind the pursuit, which are not captured in our data.

Our data is collected online, not leading to a representative sample. A total of 38,828 users, which accounts for 5.22% of all users, have opted to make their location public on the platform. Most of our users are in the USA. Followed by Europe and Southeast Asia. Regions like Russia, Africa, China, Mongolia, and South America have limited presence in our data. This is probably caused by various cultural, political, or economic reasons. Further work is required to ensure our findings are applicable globally. Lastly, our unsupervised relevance feedback-based classifier could not assign labels to 27% of the data. In this study, we curated classes and associated keywords (Table 5) based on frequent occurrences during manual inspection. A structured recursive annotation process will be required to ensure better coverage and validity of labels.

References

Alter, A. L.; and Hershfield, H. E. 2014. People search for meaning when they approach a new decade in chronologi-

cal age. *Proceedings of the National Academy of Sciences*, 111(48): 17066–17070.

Althoff, T.; Jindal, P.; and Leskovec, J. 2017. Online actions with offline impact: How online social networks influence online and offline user behavior. In *WSDM*.

Anderson, J. W.; Konz, E. C.; Frederich, R. C.; and Wood, C. L. 2001. Long-term weight-loss maintenance: a meta-analysis of US studies. *The American journal of clinical nutrition*, 74(5): 579–584.

Ayres, I. 2010. *Carrots and sticks: Unlock the power of incentives to get things done*. Bantam.

Balsamo, D.; Bajardi, P.; Salomone, A.; and Schifanella, R. 2021. Patterns of Routes of Administration and Drug Tampering for Nonmedical Opioid Consumption: Data Mining and Content Analysis of Reddit Discussions. *Journal of Medical Internet Research*.

Bland, J. M.; and Altman, D. G. 2004. The logrank test. *Bmj*, 328(7447): 1073.

Cheng, J.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. 2014. How community feedback shapes user behavior. In *ICWSM*.

Cox, D. R. 1972. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*.

Creamer, M. R.; Wang, T. W.; Babb, S.; Cullen, K. A.; Day, H.; Willis, G.; Jamal, A.; and Neff, L. 2019. Tobacco product use and cessation indicators among adults—United States, 2018. *Morbidity and mortality weekly report*, 68(45): 1013.

Dai, H.; Milkman, K. L.; and Riis, J. 2014. The fresh start effect: Temporal landmarks motivate aspirational behavior. *Management Science*, 60(10): 2563–2582.

Dai, H.; Milkman, K. L.; and Riis, J. 2015. Put your imperfections behind you: Temporal landmarks spur goal initiation when they signal new beginnings. *Psychological science*, 26(12): 1927–1936.

Dailey, R.; Romo, L.; Myer, S.; Thomas, C.; Aggarwal, S.; Nordby, K.; Johnson, M.; and Dunn, C. 2018. The buddy benefit: Increasing the effectiveness of an employee-targeted weight-loss program. *Journal of health communication*, 23(3): 272–280.

Dalton, A. N.; and Spiller, S. A. 2012. Too much of a good thing: The benefits of implementation intentions depend on the number of goals. *Journal of Consumer Research*, 39(3): 600–614.

Dekking, F. M.; Kraaikamp, C.; Lopuhaä, H. P.; and Meester, L. E. 2005. *A Modern Introduction to Probability and Statistics: Understanding why and how*, volume 488. Springer.

Fogg, B. J. 2019. *Tiny habits: The small changes that change everything*. Eamon Dolan Books.

Giné, X.; Karlan, D.; and Zinman, J. 2010. Put your money where your butt is: a commitment contract for smoking cessation. *American Economic Journal: Applied Economics*, 2(4): 213–235.

Goffman, E. 1959. The Presentation of Self in. Butler, Bodies that Matter. *The Presentation of Self in. Butler, Bodies that Matter*.

Jangra, H.; Shah, R.; and Kumaraguru, P. 2023. Effect of Feedback on Drug Consumption Disclosures on Social Media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, 435–446.

Kahneman, D. 1977. Prospect theory: An analysis of decision under risk. *Journal of Political Economy*, 85: 97–122.

Kaplan, E. L.; and Meier, P. 1958. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282): 457–481.

Lally, P.; Van Jaarsveld, C. H.; Potts, H. W.; and Wardle, J. 2010. How are habits formed: Modelling habit formation in the real world. *European journal of social psychology*, 40(6): 998–1009.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Miller, R. 2011. *Survival Analysis*. Wiley.

Nevin, J. A.; and Shahan, T. A. 2011. Behavioral momentum theory: Equations and applications. *Journal of Applied Behavior Analysis*, 44(4): 877–895.

Norcross, J. C.; and Vangarelli, D. J. 1988. The resolution solution: Longitudinal examination of New Year's change attempts. *Journal of substance abuse*, 1(2): 127–134.

Putler, D. S. 1992. Incorporating reference price effects into a theory of consumer choice. *Marketing science*, 11(3): 287–309.

Renfree, I.; Harrison, D.; Marshall, P.; Stawarz, K.; and Cox, A. 2016. Don't Kick the Habit: The Role of Dependency in Habit Formation Apps. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '16, 2932–2939. New York, NY, USA: Association for Computing Machinery. ISBN 9781450340823.

Rubin, G. 2015. *Better than before: Mastering the habits of our everyday lives*. Hachette UK.

Salton, G.; and Buckley, C. 1990. Improving retrieval performance by relevance feedback. *Journal of the American society for information science*, 41(4): 288–297.

Skarupski, K. A.; and Foucher, K. C. 2018. Writing accountability groups (WAGs): A tool to help junior faculty members build sustainable writing habits. *The Journal of Faculty Development*, 32(3): 47–54.

Skinner, B. F. 1938. The behavior of organisms: an experimental analysis. In *Appleton-Century*.

Tamersoy, A.; Chau, D. H.; and De Choudhury, M. 2017. Analysis of smoking and drinking relapse in an online community. In *Proceedings of the 2017 international conference on digital health*, 33–42.

Valiev, M.; Vasilescu, B.; and Herbsleb, J. 2018. Ecosystem-level determinants of sustained activity in open-source projects: A case study of the PyPI ecosystem. In *ESEC/FSE*.

Ethics Checklist

All our analysis is performed on public data. We ensure that none of our analysis utilizes any personally identifiable information (PII) or critical characteristics like age, gender, race, and religion of the user. All our findings and insights are derived at a population level and do not target any singular user. Finally, we ensure that our dataset adheres to the FAIR principles.

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, all our analysis is performed on public data. We do not use any critical characteristics like age, gender, race, and religion of the user.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, Our central claim is the characterization of heterogeneous habit building attempts at a large scale. To this end, our paper analyzes 397,456 attempts belonging to 15+ categories across properties like start date, effect of monitory stake, accountability, and simultaneous habit building.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, Each section in the paper starts by explaining the question we aim to answer in that part, followed by the methodology used and the reasoning behind the decision.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, part of our work being a characterization study, we look at the distributions and patterns in various properties of a habit building attempt. Our data also has many outliers. Details about the cause and prevention steps are at the end of §Data Source and Description. In §Threats to Validity, we also show a distribution of user location to argue that our analysis may not be globally applicable. However, none of the user features are used while analyzing the habit building attempts.
   (e) Did you describe the limitations of your work? Yes, a detailed discussion is present in §Threats to Validity.
   (f) Did you discuss any potential negative societal impacts of your work? No, as a characterization paper, our contribution is to uncover patterns from past habit building attempts. This paper aims to provide fertile insights for Researchers and Platform owners to build upon. §Implications goes into the details of the same. We believe the paper can not cause any potential negative social impact.
   (g) Did you discuss any potential misuse of your work? No, habit building is a primarily individual pursuit. Most of our actionable insights are about steps a user can take for their benefit. Our dataset also does not have any personally identifiable information (PII) or critical characteristics like age, gender, race, and religion of the user.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, our analysis is performed on public data. We ensure that none of our analyses utilizes PII information. All our work's processed data and code are available at *https://doi.org/10.5281/zenodo.8347427* (Anonymous link, also present at the end of §Introduction).
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, we do not train a machine learning model but use the pretrained Word2Vec embeddings in our relevance feedback retrieval. All our work's processed data and code are available at *https://doi.org/10.5281/zenodo.8347427* (Anonymous link, also present at the end of §Introduction).
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, Our model is a classic relevance feedback-based model; hence, no training is involved. Table 5 of the Appendix shows the initial list of classes and related query terms.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA, since it is not a trained model, but a deterministic algorithm.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA, our model is a unsupervised retrieval model.

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA, though we do talk about limitation of our method in §Threats to Validity.

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? Yes, we do not use any previously available dataset or code, but our data is collected from the platform StickK. This information is mentioned in the Abstract, §Introduction, §Data Source, and Description.

(b) Did you mention the license of the assets? NA

(c) Did you include any new assets in the supplemental material or as a URL? Yes, All our work's processed data and code are available at *https://doi.org/10.5281/zenodo.8347427* (Anonymous link, also present at the end of §Introduction).

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, we limit ourselves to using publicly available data.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, our data do not contain any PII information. Further, only the data made public by the user is used. Our data is about habit building attempts. Some habits may be around topics that can be perceived as sensitive by people, such as pornography or drugs, but the dataset in itself does not contain any explicit content.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? Our proposed dataset adheres to the four FAIR data principles: Findable, Accessible, Interoperable and Reusable, as follows: (i) The complete dataset is publicly available at the following link: *https://doi.org/10.5281/zenodo.8347427* (Anonymous link) making our dataset easily Findable and Accessible, (ii) All the data is stored in JSON format files that can be viewed and parsed easily. In addition to that, JSON files can be easily exported to other data formats like CSV (Comma Separated Values), making our dataset Interoperable and Reusable.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? Yes, various information required by a dataset datasheet are present in various sections of this paper like §Introduction has motivation, §Data Source and Description has collection details, and §Future Work has recommended use cases of our data.

6. Additionally, if you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA

Appendix

Class	**Query**
Weight	weight, diet, fat, pound, kg, calories, kilos, pounds, kcal
Exercise	exercise, run, walk, race, cycling, work-out, workout, bicyclng, gym, km, steps, miles, fitness, yoga, cardio, squats, deadlift, climbing, hike, pushup, pullup, healthy
Study	study, exam, diploma, phd, assignment, math, gmat, homework, gre, sat, school, learn, thesis, degree, certification, preparation, dissertation, class, course, english, french, spanish, java, experiments
Food	eat, chocolate, water, food, sugar, softdrinks, candy, desserts, veggies, gluten, lactose, snacking, coffee, beverage, shakes, caffeine
Smoking	smoking
Sleep	sleep, bed, wake, asleep, nap
Read	read, book
Meditate	meditate, journal
Money	money, finance, saving, expense, spending, earn, save, budget, buy, invest, cash, debt
Write	write, draft, screenplay, scripts, copywriting
Business	client, job, business, network, inbox, emails, career
Alcohol	alcohol, drink, beer, wine, booze
Digital	internet, electronics, tv, phone, mobile, games
Porngraphy	mastrubate, mastrubation, porn, masturbation, nofap, fap, porngraphy
Self-care	nail, hair, brush, floss, shower

Table 5: List of classes and related query terms.