

# MyBehavior: Automatic Personalized Health Feedback from User Behaviors and Preferences using Smartphones

Mashfiqui Rabbi<sup>1</sup>, Min Hane Aung<sup>1</sup>, Mi Zhang<sup>2</sup>, Tanzeem Choudhury<sup>1</sup>

<sup>1</sup>Cornell University, <sup>2</sup>Michigan State University

ms2749@cornell.edu, msa242@cornell.edu, mizhang@egr.msu.edu, tanzeem.choudhury@cornell.edu

## ABSTRACT

Mobile sensing systems have made significant advances in tracking human behavior. However, the development of personalized mobile health feedback systems is still in its infancy. This paper introduces MyBehavior, a smartphone application that takes a novel approach to generate deeply personalized health feedback. It combines state-of-the-art behavior tracking with algorithms that are used in recommendation systems. MyBehavior automatically learns a user's physical activity and dietary behavior and strategically suggests changes to those behaviors for a healthier lifestyle. The system uses a sequential decision making algorithm, *Multi-armed Bandit*, to generate suggestions that maximize calorie loss and are easy for the user to adopt. In addition, the system takes into account user's preferences to encourage adoption using the *pareto-frontier* algorithm. In a 14-week study, results show statistically significant increases in physical activity and decreases in food calorie when using MyBehavior compared to a control condition.

## Author Keywords

Mobile Phone Sensing, Machine learning, Mobile Health, Health Feedback

## ACM Classification Keywords

H.1.2 User/Machine Systems; I.5 Pattern Recognition

## General Terms

Systems, Design, Experimentation, Scalability, Performance

## INTRODUCTION

In 2007, the World Health Organization (WHO) declared obesity as a global epidemic [9]. Over one-third of the US adult population is classified as obese [39]. Obesity is now a public health issue and addressing obesity-related problems is beyond the capacity of the healthcare industry [42]. Therefore, scalable solutions that can promote healthier lifestyles outside of clinical settings are desirable.

One way to tackle obesity is to create a calorie deficit by way of decreased food intake and increased physical activity. There is a growing trend in the development of tracking devices and applications to monitor and regulate physical activity and food intake [38][28]. However, existing work makes little use of the tracked data to provide personalized feedback

that fit well into a users's routine. Feedback is often limited to either overall statistics [11][25], visualization of entire self-tracked data [33][15] or generic suggestions [24][46] that are not personalized to a user's behaviors and lifestyle. However, we can go beyond these paradigms and take advantage of more fine-grained information contained in the data. With a deeper analysis of the self-tracked data, patterns that characterize both healthy and unhealthy behavior can be revealed. These patterns then can be leveraged to generate personalized and actionable suggestions that relate to a user's behaviors.

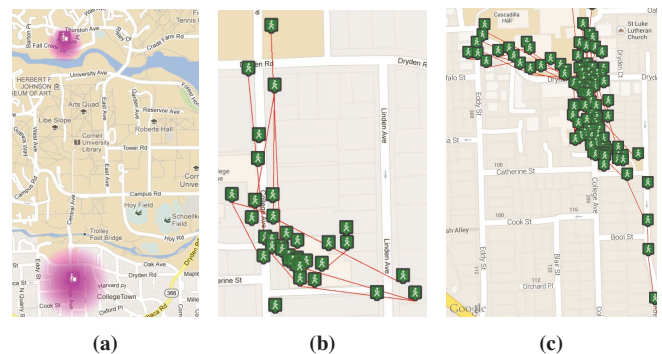


Figure 1: Visualization of user behaviors over a week (a) Heatmap of places a user stayed stationary (b) Location traces of frequent walks for the same user (c) Location traces of frequent walks for another user.

To this end, we created a mobile application called MyBehavior. The novel functionality is an intelligent engine that provides personalized suggestions by learning a user's physical activity and dietary behaviors. For example, Figure 1(a-b) show learnt behaviors of one user's stationary locations and the routes of frequently taken short walks over a week. Then suggestions are issued to take small walks near the stationary locations (Figure 1a) or continue with the existing walk (Figure 1b). Similar contextualization of suggestions can be done on a per person basis. For instance, Figure 1c shows one walking behavior over a week for a different user which can be used for personalized suggestions.

MyBehavior's intelligent suggestion engine is built upon two well-known decision theory models. The first is the multi-armed bandit (MAB) [45] which dynamically learns and influences user behaviors by suggesting actions that maximize the chances of achieving calorie loss goals. Maximization is achieved by strategically suggestion a combination of frequent and infrequent healthy behaviors. The frequent vs. infrequent behavior suggestions map to the "*exploit* vs. *explore*" principle that often underpins MABs. For example, if the user makes a 20 minute walk to work 4-5 days a week and goes to the gym once or twice a month then MyBehavior would more often suggest that the user walk to work and would occasionally suggest to increase gym visits. The assumption is that walking to work is more regular and will be

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 ACM or the author must be honored. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.  
UbiComp '15, September 7-11, 2015, Osaka, Japan.  
Copyright 2015 ACM 978-1-4503-1770-2/13/09...\$15.00.  
<http://dx.doi.org/10.1145/2750858.2805840>

*lower-effort* to adopt while also yielding more aggregate calorie loss compared to going to the gym. Prioritizing frequent behaviors also means that these behaviors are practiced and therefore the user is likely to be good at those actions (i.e., users have *mastery* or *self-efficacy*). Such low-effort change and self-efficacy are well-grounded in persuasion [16] and behavior change theories [4]. A further function of the system is to keep users in the loop by giving users control to prioritize suggestions that they prefer to follow. User preferences are then balanced with the machine generated suggestions using the second decision theory model, the pareto-frontier [48].

The blending of algorithms with behavioral theories into a usable and deployable mobile application required several iterations of research and development. We previously published early ideas of health feedback automation along with a feasibility pilot study [47]. The previous version used MAB to generate suggestions without considering user preference. This paper presents MyBehavior 2.0 which builds and improves on our earlier work and conducts more extensive testing and evaluation. Specifically the contributions include:

1. The design of an improved system to create actionable suggestions that takes into account both users behaviors and preferences. MyBehavior interfaces allowed users to easily input their preferences. User preferences and behaviors are utilized to generate a set of suggestions using Multi-armed bandit and pareto-frontier. Both of these models operationalizes the principles of behavior change theories.
2. A energy efficient, deployable android application that provides automated feedback based on real-time activity tracking, food logging and user preferences,
3. A 14-week study with 16 participants to demonstrate MyBehavior's efficacy quantitatively. Participants using MyBehavior followed more suggestions with more calorie loss (increased activity or decreased calorie intake) compared to a control condition with prescribed recommendations from health experts. These improvements lasted beyond the initial novelty period and continued over 5-9 weeks.

### MOTIVATING DESIGN OF MYBEHAVIOR

In this section, we discuss the motivation and vision that led to the development of MyBehavior.

**Low effort and self-efficacy:** Social cognitive theory [4] suggests that in order to voluntarily initiate an action, a person needs a sense of self-efficacy or confidence to perform the action. The more frequently a person performs an action in a certain context the more self-efficacy increases and the less effortful the action is perceived to be. The Fogg behavior model applies this theoretical principle to technology design by creating tools to prompt low effort actions that can be triggered even when motivation is low [16]. MyBehavior leverages the principles of low effort and self-efficacy to create suggestions that focus on repeated actions in distinct contexts.

**Personalization of suggestions:** Proponents of small data [14] and N-of-1 interventions [53][17] argue that each individual is unique and heterogeneous. This uniqueness means personalized intervention should perform better than one-size-fits-all interventions that may fail to satisfy a person's

specific requirements. To our knowledge, so far such personalization is provided only through human health coaches. MyBehavior aims to build an automated suggestion generation system that personalizes without a human health coach.

**Mobile recommender system for health feedback:** Over the last decade, search engines (e.g., Google, Bing) have transformed the way we acquire information. Similarly, movie [37] or news [27] recommendation systems influence our media consumption. However, no automated health recommendation system has yet utilized the vast amount of personal data collected using mobile or wearable devices. To our knowledge, MyBehavior is a first step in filling this gap and is the first adaptive health suggestion generator. MyBehavior also tackles the practical challenges of usability, privacy and battery life.

### MYBEHAVIOR APPLICATION DEVELOPMENT

The development of automated health suggestion generation from logged data has little precedence. We had to overcome both technical and user-centered challenges in order to create a system that can promote change for real world users. To this end, MyBehavior was honed using an iterative development process that spanned nearly 2.5 years. During this period, several MyBehavior prototypes were created and deployed. These deployments revealed a core set of requirements that an automated health feedback application needs to satisfy. Here we describe the two significant versions of MyBehavior from this iterative process.

#### MyBehavior 1.0: An Automated Health Suggestion Engine with Multi-armed Bandit Algorithm

MyBehavior 1.0 solves the important hurdle of transforming raw log data into personalized health suggestions. This system is comprised of two modules: (1) a logging and behavior mining module to track and mine user behaviors, (2) an automated suggestion generating module that utilizes the behavior data to suggest small changes that maximize chances of calorie loss. Details of MyBehavior 1.0 along with results from a 3-week pilot deployment to determine feasibility of automated feedback and usability concerns can be found in [47]. This paper significantly extends the previous work by developing the MyBehavior 2.0 system and by quantitatively evaluating its effectiveness through a 14-week user study. To provide context, we briefly describe key aspects and findings of MyBehavior 1.0 with more technical details.

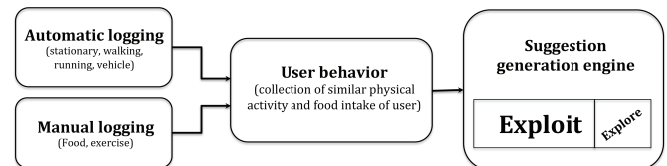
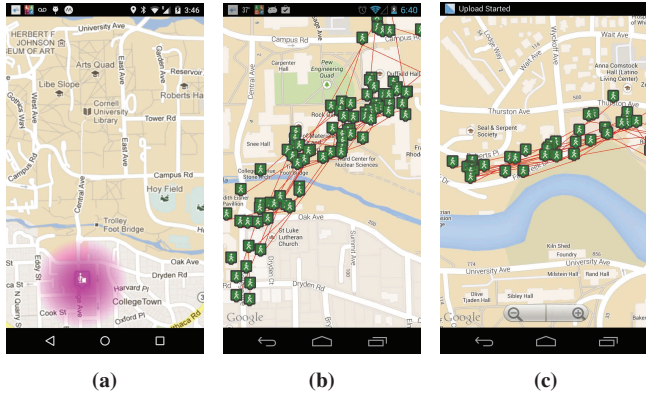


Figure 2: MyBehavior 1.0 processing pipeline

#### Logging and Behavior mining module

MyBehavior 1.0 uses a combination of automatic sensing and manual logging to record user's food intake and physical activity. Stationary, walking, running, and driving activities are automatically inferred using the technique described in [28].

The inferred activities are also tagged with corresponding locations. Physical activities that are not automatically tracked can be manually logged from a database of 800 activities [1]. Foods can be manually logged using the USDA database which contains more than 8000 food items [19].



**Figure 3: A few clusters representing different user behaviors (a) a stationary cluster (b) a walking cluster (c) another walking cluster**

Unsupervised clustering algorithms are then used to find different user behaviors from log data. Food categories are clustered based on common ingredients. For example, different types of burgers would be clustered together if they share a common bread or meat type. Similarly, manually logged exercises are also grouped by type. For instance, yoga or other types of gym workout would be clustered together. The automatically classified physical activities, namely stationary, running and walking, are also sub-categorized into similar behavior classes using clustering. For stationary events, GPS distance is used to separate different stationary episodes, for example being stationary at home would be a different cluster from being stationary at work. For walking and running activities, discretized Fréchet distance [52] is used to match similar walking or running trajectories in a computationally efficient manner. A dataset of activities collected from 20 users is utilized to derive appropriate thresholds for clustering. The BIRCH online clustering algorithm [55] is used to efficiently cluster activities. Figure 3 shows a few generated clusters representing different user behaviors.

#### *Suggestion generation module*

Once MyBehavior learns user behaviors, it determines a set of high calorie loss suggestions that involve small changes to the user's existing behaviors. In the following, we describe how we model this goal as an algorithmic optimization problem.

Any optimization algorithm requires an appropriate objective function. In MyBehavior, this function is grounded in principles of persuasion and behavior change theories. Users often take actions that are easy to do [16]; from a psychological perspective, an action is easy if it relates to a user's lifestyle [1] and has been frequently done before [4]. Given this insight, MyBehavior sets up the objective function as the multiplication of frequency of a user behavior and average calorie benefit when that behavior is performed. Frequency of a behavior is simply the size of the behavior's corresponding cluster. For dietary behaviors, the average calorie count is the mean of all foods in its corresponding cluster. For activity clusters, we

use the metabolic equivalent of task (MET) scale [26] to get calorie count for each activity instance in the clusters. Finally, the mean calorie loss is calculated over the whole cluster.

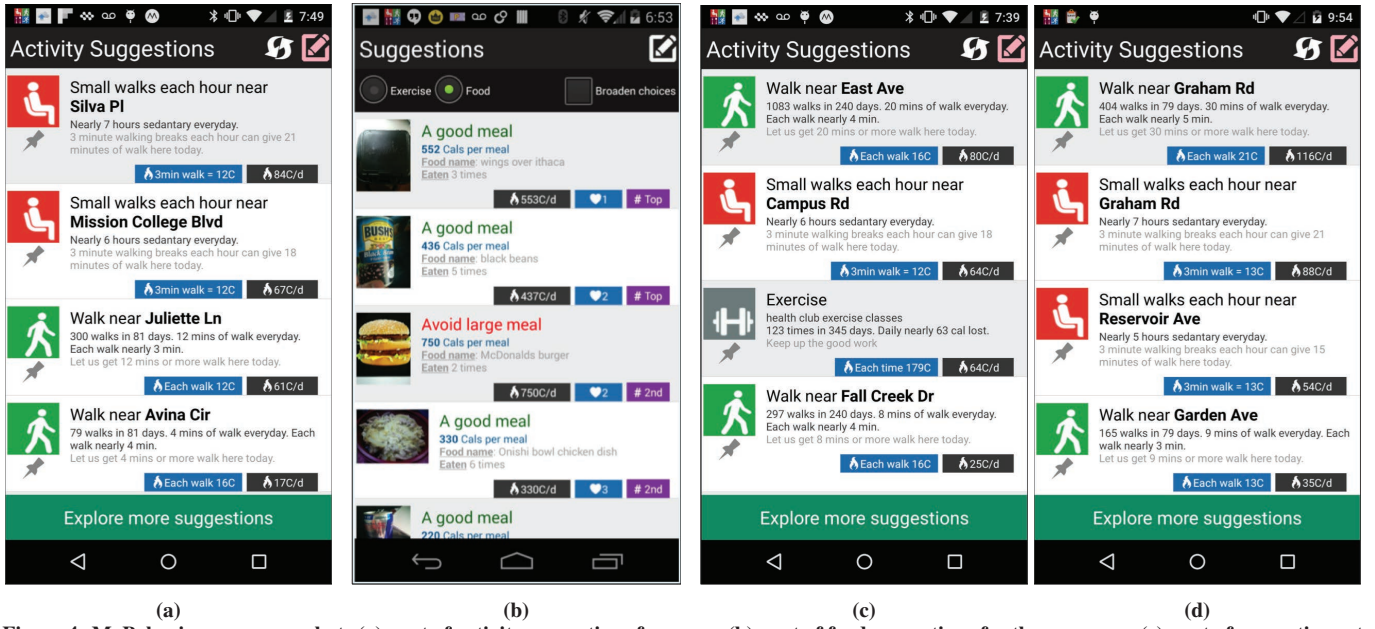
A simple suggestion making strategy could be a list of suggestions ranked according to their frequency and average caloric benefit. However such a simple strategy would not take into account a person's lifestyle changes over time (e.g., seasonal changes or major life events). Moreover, MyBehavior's influence in itself may also cause changes. Formally, this would mean that past frequent behavior would not be entirely comprehensive and future proof. One approach to protect against such future scenarios is to *exploit* most frequent past calorie loss behaviors, while including a few suggestions that *explore* past infrequent behaviors to see whether the user starts doing them frequently. Finally, personal data is limited and scarce. Thus the algorithm should not be highly parameterized nor should it require a lot of data in order to generate useful suggestions.

The multi-armed bandit (MAB) algorithm can address the aforementioned issues in its core optimization process. To illustrate this process, we briefly introduce the classic MAB problem. Consider a scenario where a gambler needs to sequentially choose from a set of slot machines with different reward distributions initially unknown to the gambler. Each time an arm is selected (pulled), a reward is drawn from that arm's reward distribution. The goal is to maximize the long term cumulative rewards obtained from the slot machines. Stated this way, the tradeoffs between explore and exploit is straightforward. Clearly, the long term reward would be maximized by pulling the arm whose mean payoff is the highest (exploitation). Finding this arm however, entails exploration - each arm pull provides incremental information about the payoff distribution for that particular arm. MyBehavior's suggestion making algorithm faces the same problem as the gambler. Initially, a user's most frequent calorie burning behaviors are unknown to the system. Over time these behaviors are revealed once food and physical activities are logged and clustered. However, potential changes in future behaviors cannot be known. Therefore the system also explores like the gambler by suggesting non-frequent behaviors to see if the user will frequently adopt them in the future.

Exploit-explore is common to all bandit algorithms but there are different strategies that are used [7]. In MyBehavior we use a EXP3 strategy [7]. In EXP3, most beneficial actions are frequently exploited with seldom exploration of less beneficial ones. In MyBehavior, 90% of the suggestions are exploited from frequent behaviors associated with high-calorie values and 10% of the time non-frequent behaviors are explored. One of the features of EXP3 is that it can adapt to changes in underlying payoff functions. This means if the user starts following new suggestions or the user's lifestyle changes (e.g., moving to a new location) then underlying caloric benefits of certain behavior will change. EXP3 strategy would tune to those changed circumstances quickly.

MyBehavior generates 10 food and 10 activity suggestions separately. The suggestion engine does not mix foods with exercise as the joint space of possibilities would make it a





**Figure 4: MyBehavior app screenshots (a) a set of activity suggestions for a user (b) a set of food suggestions for the same user (c) a set of suggestions at a different time for the same user (d) a set of activity suggestions for a different user**

combinatorially hard problem and lead to more complicated suggestions. For activity suggestions, changing stationary behavior is added to the mix of walking, running or manually input exercise suggestions. MyBehavior suggests users to change every hour of stationary event in a specific location with 3 minutes of walking. Such a mix often results into non-trivial changes in suggestions: for instance, Fig. 4(c) shows a ranking of MyBehavior suggestions where simply changing regular stationary episodes with 3 minutes of walk for every stationary hour can yield more calorie expenditure compared to the user's gym exercises. Regarding food suggestions, a separate bandit generates food suggestions that take into account intake frequency and calorie. MyBehavior makes a distinction between suggestions for meals and those for snacks, as the number of calories consumed can be different for these two food clusters.

Figure 4 shows different suggestions generated by MyBehavior. As seen in the screenshots, semantically meaningful messages are added with every suggestion. For suggestions generated by exploiting, MyBehavior asks the user to either continue positive activities (i.e., good calorie foods, walking, or exercise), make small changes in some situations (i.e., stationary activities), or avoid negative activities (i.e., frequent large meals). On the other hand, suggestions generated during exploration phase, the system asks the users to consider trying out the suggestions. All MyBehavior suggestions change overtime and are different for different users. Figure 4(a) and (c) are physical activity suggestions from the same user on different days. Figure 4(d) shows suggestions generated for a different user demonstrating the scalability of the system.

Finally, modeling MyBehavior as a MAB also has additional advantages. MAB is an online algorithm and incrementally makes decisions in a computationally efficient manner. This means MyBehavior can compute all suggestions inside the

phone which is an added privacy feature. MABs also have fewer parameters and are easy to learn.

#### *Deployment and lessons learned*

MyBehavior 1.0 was deployed in a 3-week pilot with 9 users (4 female). At the end of this study, we conducted semi-structured interviews with the participants about their experiences with MyBehavior. We also asked the participants to indicate whether they would be willing and able to follow each suggestion on an average day on a scale of 1-5 (5=Strongly Agrees that s/he can follow the suggestions; 1=Strongly Disagrees). Each participant rated 15 suggestions.

In the interviews, users reported MyBehavior suggestions to be actionable. In the suggestion rating survey, MyBehavior received an average of 3.4 out of 5 ( $\mu = 3.4$ ;  $\sigma = 1.4$ ). However, several areas of improvement are also identified. They are as follows:

1. **Difficulty in manual logging:** Users reported the manual food logging process to be self-reflective. However, they also found the searching and adding appropriate food items to be long and cumbersome. Furthermore, many manually added exercises were repeatedly done (e.g., gym or fitness classes). Users wanted quick ways to add repeated exercise rather than searching them every time.
2. **Lack of human control:** Although MyBehavior can dynamically adapt to lifestyle changes, on occasion MyBehavior was slow to adapt. For example, a user regularly played soccer with his friend but when his friend moved to a new location he could no longer do that activity. The user was frustrated that he could not remove the soccer suggestion. On the other hand, users are sometimes highly motivated about certain activities that they did not repeat much in the past. For example, several users wanted, "going to the gym" as a top suggestion even though they did not frequently go to gym in the past.

The initial 3-week pilot study was not long enough to show statistically significant changes in behavior. But, the study confirmed that automatically generated suggestions are indeed actionable and provided important usability feedback.

### MyBehavior 2.0: Easier logging and human-in-the-loop

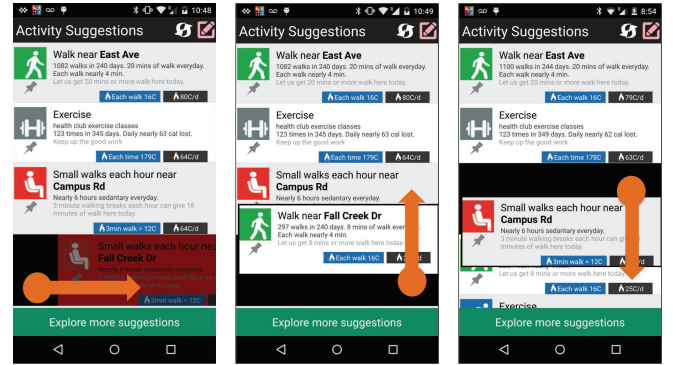
Based on insights and user feedback from the first pilot, we develop MyBehavior 2.0. We focus on easier ways to log food and exercise. In addition, we include the provision for user customization on MyBehavior generated suggestions. Below we describe these changes in detail.

#### Easier logging

To reduce the burden of food logging, MyBehavior 2.0 contains a crowd-sourcing functionality that returns calorie information using photographs of the food taken by the user. This is similar to the implementation of the PlateMate framework by Noronha et al. [38] which coordinates Amazon Mechanical Turk [3] (AMT) labelers to ascertain nutritional information. Results from PlateMate deployment showed the nutritional information derived using the AMT was comparable to trained dietitians. In a pilot test, we found that AMT labelers often made mistakes if a very large database was used. This is overcome by using a smaller list of the 40 most frequently selected foods by nearly 50 million MyFitnessPal [33] mobile app users. For each image, 5 AMT labelers are asked for calorie information and the median is used as the final calorie number. In addition, we ask the AMT labelers to rate similarity of a food image with 16 other previous food images from the user. This similarity information is used for clustering similar foods. In a test image set of 50 images, we found an average of only  $\pm 70$  calories between ground truth values and the AMT labeled calorie estimate. In addition to this reduced burden in food logging, MyBehavior 2.0 also allows users to select from list of their past exercises for easier manual exercise logging.

#### Incorporating Human Customization in Suggestions

The second major modification in MyBehavior 2.0 is giving the user control to customize suggestion set. This is achieved by allowing the user to remove the suggestions the user does not want or is unable to follow due to a change in circumstances. In terms of interaction, users can swipe from left to right and remove suggestions (Figure 5(a)); a removed suggestion is never considered in the future. In addition, MyBehavior 2.0 allows the users to re-sort the suggestions in order of their preference. The user can long-press a suggestion and move the suggestion above or below another suggestion (Figure 5(b-c)). For instance, if a user prefers to go to the gym even though s/he did not do it often before, the user can simply move the gym suggestion to the top. This re-sorting creates a new ranking based on the user's preference in addition to the system generated suggestions. With this dual information, a final ranking is determined that considers both factors according to B.J. Fogg's [16] behavior model; where both preference and ability (i.e. perceived effort level) are important factors in how actionable a suggestion would be. We illustrate what Fogg's behavior model would suggest with an example. Let us assume that there are three suggestions for a user: walking near the office, walking near the



(a) (b) (c)  
Figure 5: Keeping human in the loop (a) dismissing a suggestion by removal (b) Moving a suggestion above (c) Moving a suggestions below

home, and going to gym. The user frequently walks near the office and prefers doing this. User also has a high preference for going to the gym, but is not good at gym work and goes infrequently. In addition, the user frequently walks near her house but is not keen on this activity. In this scenario, Fogg's behavior model would suggest that walking near the office is the most actionable. However, choosing between walking near home and going to gym would be a tie since one is easier to do while the other is more preferred.

Given this insight from Fogg's behavior model, we repurpose a principled technique in decision theory called *pareto-frontier* to balance between preference and low-effort [48]. Before moving into the details of the algorithm, we introduce some notations. We denote the set of suggestions as  $X$  where an element  $x_j \in X$  is a suggestion. For a suggestion  $x_j$ ,  $\nu_j$  refers to its rank from MAB algorithm whereas  $p_j$  refers to its rank after users finishes reordering the suggestions (Figure 5(a-c)). Thus a higher rank or value of  $\nu_j$  or  $p_j$  means the suggestion is more low effort or more preferred respectively. With this notation, the pareto algorithm works as follows. Let us assume that for two suggestions  $x_i, x_j$ , preferences and low-effort ranks are  $p_i, p_j$  and  $\nu_i, \nu_j$  respectively. If  $x_i$ 's both preference and low-effort ranks are higher than  $x_j$  then  $x_i$  ranks higher (or is more actionable) than  $x_j$  and we say that  $x_i$  *pareto-dominates*  $x_j$ . If  $x_i$ 's preference is higher than  $x_j$  while the low-effort rank is lower than  $x_j$  (i.e.,  $p_i > p_j$  and  $\nu_i < \nu_j$ ) or the other way around (i.e.,  $p_i < p_j$  and  $\nu_i > \nu_j$ ) then  $x_i$  and  $x_j$  receive the same rank and the more actionable suggestions can not be decided. Note here, that pareto-frontier makes no assumption about scale of  $p$  or  $\nu$  and can still balance between them. Finally, the ranking process works iteratively as shown in Algorithm 1. It starts with a set of all available suggestions  $X$ . At every iteration, a set of suggestions  $X_i$  are selected that pareto-dominates rest of the suggestions.  $X_i$  are then ranked higher than the rest and are removed from the set of  $X$ . The process then repeats.

Finally a specific case that needs special attention in the pareto ranking is when a *new* suggestion  $x$  arrives with low-effort rank  $\nu$  and unknown preference  $p$  since the user never ranked it. In this case, a fair policy is adopted that acts as follows: If  $x_1$  and  $x_2$  are two other suggestions such that  $\nu_1 > \nu > \nu_2$  and  $p_1 > p_2$  then no matter what the unknown

```

input : A set of suggestions  $X$  annotated with user
         preference and caloric benefit used in MAB

Initialize an index value  $i = 1$ ;

while  $X$  is non-empty do
    - find subset  $X_i$  in  $X$  that pareto dominates  $X - X_i$ ;
    - rank suggestion(s) in  $X_i$  with  $i$ ;
    - increment  $i$  by one and remove  $X_i$  from  $X$ ;
end

```

**Algorithm 1:** Ranking suggestion with pareto-frontier

value of  $x$ 's preference is,  $x$  would not be pareto dominated by  $x_2$  since  $x$  has a higher low-effort rank than  $x_2$ . Since the value of  $p$  is unknown, it is fairly assumed that this unknown value to be less than a known value  $p_2$ . This would assign  $x$  the same rank as  $x_2$  which is lower than  $x_1$ .

#### *Pilot deployments with MyBehavior 2.0*

We conducted a 3-week pilot to study the improvement of MyBehavior 2.0. We recruited 7 users (3 female) for 21 days. Users were interested to see the crowd-source based calorie content of the foods and were satisfied with the accuracy of food labeling and clustering. Specifically we counted the number of foods logged per day over 3 weeks. We have found that the number of foods recorded per day per user with crowd-based approach ( $\mu = 4.2, \sigma = 2.5, q_{25} = 2.2, q_{50} = 4.1, q_{75} = 6.0$ ) is higher than the manual logging approach from MyBehavior 1.0 ( $\mu = 2.4, \sigma = 1.7, q_{25} = 1.4, q_{50} = 2.0, q_{75} = 2.9$ ). This increase is also statistically significant (Wilcoxon ranksum test,  $z = 2.5, p = 0.013$ ).

To measure the benefit of incorporating human preference, we showed MyBehavior 1.0 generated suggestions to the users after the study. They were asked to rate 8 food and 8 activity suggestions between a scale of 1 to 5. This rating represents whether users liked the suggestion and would act on it on an average day (1 = disagree and 5 = agree). After users finished rating the default set of suggestions (from 1.0), they were instructed on the use of the remove and reorder functions to incorporate their preferences. Users on the average changed 3.5 suggestions out of 16 suggestions. When users finished providing their preferences, we ran the pareto-frontier algorithm and then showed users the revised suggestions. We asked the users to rate again. Ratings without incorporating the human preference are similar to results from the MyBehavior 1.0 deployment ( $\mu = 3.5, \sigma = 1.2$ ). However, after incorporating human preference using pareto-frontier algorithm, there is a statistically significant increase of almost 19% ( $\mu = 4.2, \sigma = 1.1$ ).

## EVALUATION

After two iterations of improvements, we conducted a 14-week deployment and evaluation study of MyBehavior 2.0. The purpose of the evaluation is two fold: (1) to test whether MyBehavior has better efficacy compared to control condition, and (2) to assess if MyBehavior can enable change beyond the initial novelty period. In the rest of this section, we detail the user study design and report results.

## Study design considerations

Several pilot studies have been done to test early adoption and to make design improvements to MyBehavior, as have been argued by Klasanja et al. [23]. In order to quantitatively demonstrate early efficacy, Dallery et al. [12] and Onghena et al. [40] argue for small scale within subject trials, sometimes referred as "single case experiments". These experiments achieve sufficient statistical power with large number of repeated samples from a single individual. MyBehavior suits this requirement since enough repeated samples can be collected with automated sensing or daily manual logging [12].

In our study, we follow a single case experiment paradigm called *multiple baseline* design [12]. In a multiple baseline design, subjects are initially exposed to the control condition, which is followed by the experiment condition. However, the duration of the control condition before the experiment condition varies for different users. Such a variation is made as a *replication* strategy to show that the desired dependent variable consistently changes in the desired direction after the experiment condition starts. In our study, participants are exposed to either 2, 3 or 4 weeks of control condition before using MyBehavior as part of multiple baseline design. We also run the experiment condition for longer (7-9 weeks) than control condition. We do so to investigate MyBehavior's influence beyond initial novelty periods.

## Study procedure and participants

We sent an invitation for participating in MyBehavior's user study through Cornell University's Wellness Center's email list. Interested individuals emailed back an investigator and were requested to fill out a prescreening survey. The survey asked for age, gender, experience in using smartphones etc. We also asked *readiness* to act on healthier behavior as defined by the Trans-theoretical model (TTM) [46]<sup>1</sup>. We only included participants with (i) sound proficiency in using smartphones (ii) are either in ready or acting stages of TTM since in these stages people are willing or acting towards changing their behaviors [47]. The study investigators met with eligible participants and installed MyBehavior on their phones. In these meetings, we provided basic instructions to use MyBehavior. Participants also entered their gender, weight, height and weekly weight loss goals. Then the Harris-Benedict equation [18] is used to translate weight loss goals to daily calorie intake and expenditure goals.

The day after the face-to-face meeting, participants started the *baseline phase* of the study. In this phase, calorie goals were displayed in an on-screen widget in the phone's home screen. This widget also incorporated realtime updates of user's daily calorie intake and expenditure. We also added a daily chronological summary of physical activities and food intake. No suggestions were provided in this baseline phase. Note here

<sup>1</sup>TTM defines several stages to readiness: "Precontemplation" represents a stage of not feeling the need to change while in "Ready" stage there is intention to start eating well or doing exercise in near future but not taking actions. "Acting" stage on the other hand represents already taking actions but still need to strengthen commitment, or fight urges to slip. Finally, "Maintaining" stage means a lifestyle with regular health eating and exercise.



Variable	n(%)
<i>Gender</i>	
Male	7(43.7)
Female	9(56.3)
<i>Age</i>	
18 - 29	4(25.0)
30 - 39	6(37.5)
40 - 49	3(18.7)
> 50	3(18.7)
<i>Stage of behavior change before the study</i>	
Ready	7(43.7)
Acting	9(56.3)
<i>Previous experience with self-tracking</i>	
Maintained food diary	13(81.3)
Maintained exercise diary	11(68.7)

Table 1: User demographics in the long term study

that such widgets and daily logs are common for many modern health and fitness applications [15][33]. We ran the baseline phase for 3 weeks, since starting to use a health application often makes users more active temporarily even though no intervention is used. Such an effect is often referred to as “novelty effect” [49]. After the baseline phase, participants were exposed to the control condition of the study. Participants received generic prescriptive recommendations generated from a pool of 42 suggestions for healthy living, such as “walk for 30 minutes” and “eat fish for dinner”. A certified fitness professional created these generic suggestions after following National Institute of Health resources [36][35]. An external nutrition counselor also reviewed the suggestions to ensure that they were both healthy and achievable. We followed the multiple baseline design as described before and continue the control condition for different durations for different participants. The control condition ranged between 2-4 weeks depending on participants. Each day of the control phase, 8 physical activity and 8 food suggestions were randomly selected from the 42 prescriptive suggestions. These suggestions are shown in a list similar to MyBehavior suggestions are shown in Figure 4<sup>2</sup>. After the control phase, participants received MyBehavior suggestions for 7-9 weeks. Total participation period did not exceed 14 weeks for any participant. Participants were compensated \$120 for their regular participation in the study.

We recruited 16 participants. Table 1 shows the participant demographics. Our sample size was determined by following the literature of single case experiment design [12]. The literature argues that  $n \geq 4$  is sufficient for statistical power if enough repeated samples are collected per participant.

<sup>2</sup>Screenshots of suggestions during control phase and the list of 42 suggestions are added as supporting material of this paper.

### Daily phone survey

1. How many suggestions were you able to follow today?
2. How many suggestions did you want to follow?
3. How well did the suggestions relate to your life.
  - likert scale 1-7
  - 1- doesn't relate to your life
  - 7- relates to your life perfectly
4. Did you encounter any barrier to follow the suggestions today (e.g., weather or deadline)?
  - Yes/No
5. Rate your emotional state today
  - photographic affect meter (PAM) scale [44]

Table 2: Users answered the above 5 questions in a daily phone survey

### Outcome measures of the study

We utilize the food and exercise log data to measure changes in food calorie intake and calorie loss in exercise. During the study, we also used an in-phone survey that users filled out daily. The survey asks 5 questions as listed in Table 2. For the number of suggestions followed, we use self-report since it is hard to objectively judge whether an activity is done as part of regular actions or as a result of the suggestion. We ask how many suggestions users *wanted* to follow to measure user intentions or attitude [2]. Past literature shows that attitudes or intentions often indicate 19%-39% of future behavior [5]. A higher score in the 3rd question means the suggestions relate to a user's life and are potentially easy to implement. We ask the 4th and 5th questions because we want to investigate how MyBehavior suggestions perform against negative life circumstances as barriers and negative emotions have been shown to reduce chances of change [20].

Although weight loss is MyBehavior's main long term goal, calorie loss or user intentions to follow suggestions are important mediators to achieve weight loss. Recent work on adaptive interventions in clinical psychology (e.g., Behavior Intervention Technology [31]) and just-in-time adaptive interventions [34] argue that calorie loss or positive activities are essential subaims and are valid outcome measures for weight reduction applications.

### Analysis plan

We analyze the efficacy of MyBehavior against control condition by modeling our outcome measures (e.g., caloric loss or number of suggestions followed) as continuous variables using mixed effect models against time. We use mixed effect models [41] since they can handle imbalanced control vs. experiment conditions [10][43] and correlated data points from the same user [13]. In the models, we use intercept and time as random effects to respectively allow for inter-subject variations in initial starting points and growths over time [8][32]. Including these random effects significantly increased likelihood over fixed-effect-only models in likelihood

ratio tests [50]. Such an increase in model fit (i.e., likelihood) means inter-subject variability exists in our dataset and including random effect is necessary to properly isolate inter-subject variability from actual trends in fixed effects. As fixed effects, we use time and intervention type (i.e., control vs experiment). Intervention types are coded 0 for control and 1 for experiment phase. For time, the first week of the control phase is coded as 0 and incremented by 1 after each subsequent week. We observed non-linear changes in outcome measures over time, so we use non-linear time effect up to cubic polynomials [50]. In general, considering such polynomial time effects shows significant improvements in likelihood ratio tests compared to models without such polynomial time effects. On exception is for number of minutes walked where time or its polynomial forms as fixed effects did not improve the likelihood significantly. This approach of centering [50] time and intervention adjusts for time related effects (e.g., weather effect, or changes due to logging for longer periods) and isolates the change with MyBehavior over control as the co-efficient of intervention fixed effect ( $\beta_i$ ). In other words,  $\beta_i$ s reflect changes (e.g., number of minutes walked more) at the points of introducing MyBehavior. Finally, for the survey response of number of suggestions followed, we additionally include emotional state, barrier and their interaction with intervention types as fixed effects. We add these extra terms to explore interplay between MyBehavior and emotional states/barriers. Emotional states are coded as 0,1,2,3 respectively for negative high, negative low, positive low and positive high. Barriers are coded as 1 for presence of barrier and 0 for absence. Both barriers and emotional states are considered as categorical in the mixed model. The analyses are run using Matlab's statistical analysis toolbox with maximum likelihood.

Given significant intervention effects are achieved with mixed effect models, we explore the real-world end effect of MyBehavior in post-hoc analysis. We compare 2-4 weeks of using control condition to last 3 weeks of using MyBehavior. We consider the last 3 weeks of MyBehavior to measure change beyond initial novelty periods. Specifically, we describe the mean and standard deviations for these two conditions. We then use student t-tests and Cohen-d to measure the statistical significance and effect size. Similar pre-post analysis to measure real world end effect has been done in [8].

## Results

### Comparison with the control condition

Table 3 shows the results from the mixed model analyses for different outcome measures. Due to space limitations, we only include the relevant statistics. In 2nd column, we report the coefficient of intervention fixed effect ( $\beta_i$ ) and its significance. In third column, we also report the standard model fit statistics that underpin the values of  $\beta_i$ s. We include standard model fit statistics namely deviance, AIC and BIC scores [50]. We add significance of the fitted models (LR) against unconditional mean models (i.e., a baseline mixed model with only intercept as both fixed and random effects) [50] using a likelihood ratio test. From table 3, we observe that all the fitted mixed models for different outcome measures are significant improvements over the uncon-

ditional mean model. Furthermore, use of MyBehavior compared to control condition results in increased number of suggestions followed ( $\beta_i = 1.2, p < 0.0005$ ), walking minutes ( $\beta_i = 10.1, p < 0.005$ ) and calories burnt in non-walking exercises ( $\beta_i = 42.1, p < 0.05$ ) per day. Calorie consumption also decreased per meal ( $\beta_i = -56.1, p < 0.05$ ).

Figure 6 shows different outcome measures (i.e., number of suggestions followed, minutes walked, calories burnt in exercise, calorie intake in meals) over time as commonly reported in multiple baseline designs [12][6]. All these values are predicted from the mixed models. For each outcome measure, we create three groups representing 2, 3, 4 weeks of using control conditions before exposing to MyBehavior. A dotted line shows the start time of using MyBehavior. Improvements in all outcome measures can be seen to occur in Figure 5 after the introduction of the MyBehavior phase irrespective of the start times. However, patterns over time differ for different outcome measures. Minutes walked did not change much over time. On the other hand, food calories consumption generally decreased over time although introduction of MyBehavior had some effect. Non-walking exercises generally decreased in control over time, but were sustained during MyBehavior usage.

Subjective responses namely number of suggestions participants wanted to follow ( $\beta_i = 2.9, p < 0.0005$ ) and relatedness of suggestions to life ( $\beta_i = 0.5, p < 0.0005$ ) were also higher for MyBehavior compared to control (Table 3). Including emotional state, barriers and their interactions with interventions significantly improved likelihood of predicting number of suggestions followed compared to excluding them in likelihood ratio tests ( $p = 0.05$ ). This means that there are significant interactions of MyBehavior vs. control with emotional state and barriers. Figure 7 visualizes these interactions as distributions of number of suggestions followed for different emotional states and barrier conditions.

### Pre-post real-world effect analysis

Pre-post analysis is summarized in Table 4. For all the outcome measures, values of Cohen-d indicate medium to large effects of MyBehavior. Although not shown in the table, all these changes are also statistically significant ( $p < 0.05$ ) in

Outcome measure	$\beta_i$	$-2\log L$	AIC	BIC	LR
# of sug. followed	1.2***	2491	2517	2576	***
# of sug. wanted	2.9***	2496	2518	2568	***
relatedness	0.5***	1551	1573	1623	**
walking/day (min) <sup>†</sup>	10.1**	4795	4809	4839	***
exercise/day (cal) <sup>a</sup>	42.1*	10959	10973	11006	**
each meal (cal)	-56.1*	16151	16165	16200	***

\*\*\* $p < 0.0005$ ; \*\* $p < 0.005$ ; \* $p < 0.05$ ; ~  $p > 0.1$

<sup>a</sup> non-walking exercises combined

<sup>†</sup> without time as fixed-effect

**Table 3: Summary of statistical differences between control and MyBehavior as collected from survey, physical activity and dietary logs**



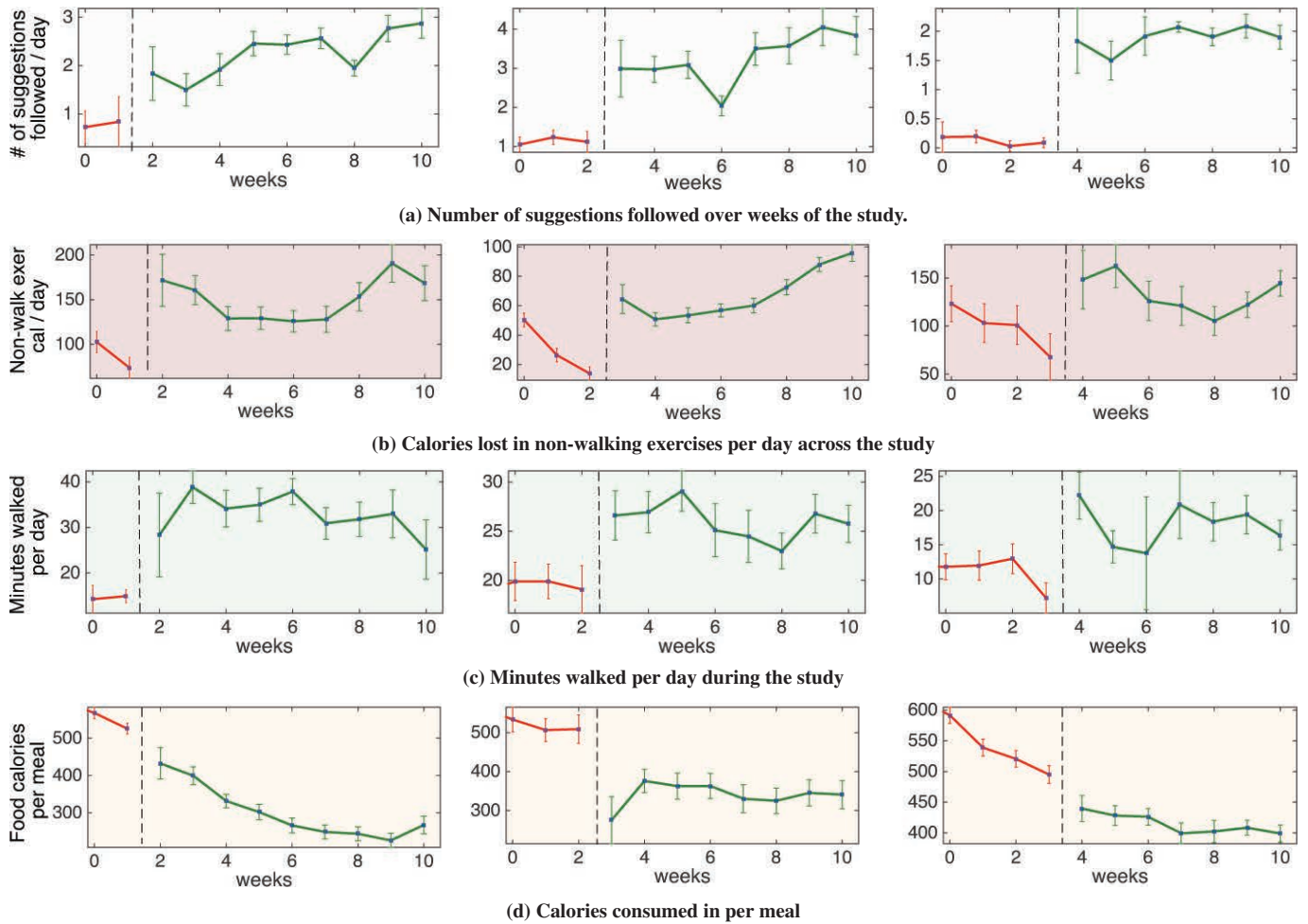


Figure 6: Changes in user behavior as predicted by the mixed model for multiple baseline design. The dotted lines represent the start of the intervention of MyBehavior. Left, middle, and right figures respectively show results from participants where intervention were started after 2, 3 and 4 weeks of using the control. Red color represents control phase where as green represents periods of using MyBehavior.

student t-tests. An additional result we point to is the changes in number of suggestions followed for barriers and emotional states. Users followed more MyBehavior suggestions where there was no barrier ( $p < 0.001, d = 0.84$ ) such as bad weather. Similar significant increase is also found for positive emotion ( $p < 0.001, d = 0.82$ ). Furthermore, MyBehavior suggestions were still followed more than control suggestions even when there were barriers ( $p < 0.001, d = 0.44$ ) or when the user experienced negative emotion ( $p < 0.001, d = 0.55$ ). However, effect sizes are smaller for barrier and negative emotions.

## DISCUSSION AND RELATED WORK

### Primary findings

To the best of our knowledge, MyBehavior is the first recommendation system to automatically generate health feedback from physical activity and food log data. It utilizes concepts of low-effort [16] and self-efficacy [4] from behavior change theory literature and operationalized them in machine learning optimization functions. Through several deployments, we created a usable MyBehavior app that utilize the benefits of algorithmic computation in usable form.

In a 14-week study, participants subjectively reported MyBehavior suggestions to be more related to their life and they wanted to follow the suggestions in higher numbers. We believe such higher actionability and relatedness result from MyBehavior's prioritization of low effort suggestions. The higher actionability and relatedness also translated to actual behavior with increased walking, exercise and decreased food calorie intake. These favorable results are replicated as part of multiple baseline design as shown in Figure 6. This adoption may result from low-effort suggestions that should enable actual adoption according several behavior change theories [16][21][4][20][2]. Finally, in the pre-post real-world effect analysis, MyBehavior suggestions were followed more during no-barrier or positive emotions states compared to barriers or negative emotional states. We believe this happens because low effort suggestions similar to MyBehavior are adopted in higher numbers during high motivation states like no-barrier or positive emotions [16]. Nonetheless, some MyBehavior suggestions were followed during barrier or negative emotional states. According to Fogg [16], low-effort suggestions similar to MyBehavior may still stay actionable

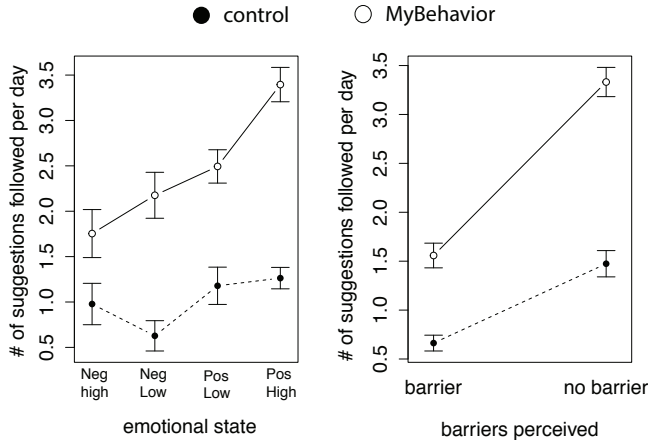


Figure 7: Number of suggestions followed for control and experiment conditions with respect to barriers and emotional states

in low motivation states like with barriers and negative emotions.

### Related work and MyBehavior

MyBehavior’s automated personalization scheme differs from prior pervasive health literature. Ubifit [11] or Be-Well [25] relied on overall statistics (e.g., total amount of activity) for providing feedback without personalized actionable suggestions. On the other hand, previous literature on life-logging [56] relied on visualizing the whole personal data that users have to interpret and find actionable information by themselves. MyBehavior, in comparison, breaks down each user’s behavior and finds personalized actionable suggestions. Such an approach is only remotely similar to tailored health communication [24] approaches where suggestions are tailored for groups of users with similar age, gender or stages of behavior change. This also means MyBehavior personalization approach is the first automated  $N$ -of-1 [17][53] or small data [14] system that treats each user differently in creating its suggestions.

In addition to personalization, MyBehavior also explored the space of generating low-effort suggestions automatically for the first time. According to B.J. Fogg’s behavior model, low-effort is similarly important as motivation [16]. However, earlier literature on gamification [30], goal seeking [54] or self-regulation [51] relied on increasing user motivation. To the best of our knowledge, easiness or low-effort suggestions were only provided through health coaches in earlier work. Nonetheless the work on increasing user motivation is orthogonal to our work on low-effort. Higher user motivation can increase probability of executing low-effort suggestions [16].

Using Multi-armed Bandit (MAB) also solves a few practical issues of generating suggestions. MAB is an online learning algorithm that learns, adapts and decides simultaneously. All of these learning and adaption are done with relatively less data since MABs are not heavily parameterized. This is crucial at the early stages when less data is available from users. MAB’s online nature also means model update needs

Outcome measure	Control	MyBehavior	Cohen- $d$
# of sug. followed	1.1 (1.1)	3.1 (2.7)	0.76
# of sug. wanted	2.1 (1.2)	4.4 (2.4)	1.07
relatedness	3.8 (1.1)	4.5 (1.2)	0.54
walking/day (min)	14.5 (5.9)	24.9 (7.4)	1.41
exercise/day (cal) <sup>a</sup>	83.5 (33.1)	126.7 (35.3)	1.23
each meal (cal)	540 (137.2)	362 (134.1)	1.30
# of sug. followed <sup>b</sup>	1.3 (2.2)	3.4 (2.8)	0.84
# of sug. followed <sup>c</sup>	0.6 (2.1)	1.6 (2.5)	0.44
# of sug. followed <sup>d</sup>	1.2 (1.9)	3.2 (2.6)	0.82
# of sug. followed <sup>e</sup>	0.7 (1.5)	1.9 (2.1)	0.55

<sup>a</sup>non-walking exercises combined

<sup>b</sup> for no barrier, <sup>c</sup> with barrier

<sup>d</sup> for positive emotion, <sup>e</sup> for negative emotion

Table 4: Pre-post analysis for the control condition and last 3 weeks of experiment condition. Means and standard deviations (within bracket) are shown along with effect size measures.

only processing the latest data with less computation. A competing technique to MAB is the Markov Decision Processes (MDP) [45], the most used reinforcement learning algorithm for decision making. In comparison to MABs, MDPs are highly parameterized and often require large amount of data to train. Because of MAB’s low computational requirement, MyBehavior can generate all suggestions inside the phone without significantly lowering the battery. This also means location and activity traces do not need to leave the phone, which is an added privacy feature [22]. Finally, MABs are used before for personalized recommendation in other domains. For example, Yahoo on their front page uses MABs to suggest personalized news articles [27]. Google also uses MABs to dynamically serve their advertisements [29].

### CONCLUSION

In this paper, we present the design, implementation, and evaluation of the MyBehavior smartphone app that provides personalized health suggestions automatically. We build the underlying automatic suggestion generation system using two different decision theory techniques, namely, multi-arm bandit and pareto-frontier algorithm. The combination of these techniques provides a novel way to tailor feedback without requiring expensive and difficult-to-scale interventions from health coaches. We present the results from a 14-week study that shows significant improvement over an appropriately chosen control condition that lasted beyond the initial novelty phase. As more and more people use automated technologies to track their health, we believe MyBehaviors ability to auto-personalize suggestions holds great promise for providing actionable feedback at scale.

### ACKNOWLEDGMENT

This work has been partially supported by Intel Science and Technology Center for Pervasive Computing. We thank Julie Kientz for her feedback on early manuscript and for suggesting the multiple baseline design.

## REFERENCES

1. B. E. Ainsworth, W. L. Haskell, S. D. Herrmann, N. Meckes, D. R. Bassett, C. Tudor-Locke, J. L. Greer, J. Vezina, M. C. Whitt-Glover, and A. S. Leon. 2011 compendium of physical activities: a second update of codes and met values. *Medicine and science in sports and exercise*, 43(8):1575–1581, 2011.
2. I. Ajzen. Theory of planned behavior. *Handb Theor Soc Psychol Vol One*, 1:438, 2011.
3. Amazon Mechanical Turk. <http://www.mturk.com/>, 2013. [Online; accessed 19 March 2013].
4. A. Bandura and D. C. McClelland. Social learning theory. 1977.
5. Behavioral Intentions. <http://chirr.nlm.nih.gov/behavioral-intention.php>, 2013. [Online; accessed 26 February 2015].
6. A. Biglan, D. Ary, and A. C. Wagenaar. The value of interrupted time-series experiments for community intervention research. *Prevention Science*, 1(1):31–49, 2000.
7. S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*, 2012.
8. M. N. Burns, M. Begale, J. Duffecy, D. Gergle, C. J. Karr, E. Giangrande, and D. C. Mohr. Harnessing context sensing to develop a mobile intervention for depression. *Journal of medical Internet research*, 13(3), 2011.
9. B. Caballero. The global epidemic of obesity: an overview. *Epidemiologic reviews*, 29(1):1–5, 2007.
10. A. Cnaan, N. Laird, and P. Slasor. Tutorial in biostatistics: Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Stat Med*, 16:2349–2380, 1997.
11. S. Consolvo, D. W. McDonald, T. Toscos, M. Y. Chen, J. Froehlich, B. Harrison, P. Klasnja, A. LaMarca, L. LeGrand, R. Libby, et al. Activity sensing in the wild: a field trial of ubifit garden. In *Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 1797–1806. ACM, 2008.
12. J. Dallery, R. N. Cassidy, and B. R. Raiff. Single-case experimental designs to evaluate novel technology-based health interventions. *Journal of medical Internet research*, 15(2), 2013.
13. P. Diggle, P. Heagerty, K.-Y. Liang, and S. Zeger. *Analysis of longitudinal data*. Oxford University Press, 2002.
14. D. Estrin. Small data, where  $n = me$ . *Commun. ACM*, 57(4):32–34, Apr. 2014.
15. Fitbit, Inc. <http://www.fitbit.com/>, 2013. [Online; accessed 19 March 2013].
16. B. Fogg. A behavior model for persuasive design. In *Proceedings of the 4th international Conference on Persuasive Technology*, page 40. ACM, 2009.
17. W. M. Grove. Thinking clearly about psychology.
18. J. Harris and F. Benedict. Biometric studies of basal metabolism. *Washington, DC: Carnegie Institution*, 1919.
19. D. Haytowitz, L. Lemar, P. Pehrsson, J. Exler, K. Patterson, R. Thomas, M. Nickle, J. Williams, B. Showell, M. Khan, et al. Usda national nutrient database for standard reference, release 24, 2011.
20. G. Hochbaum, I. Rosenstock, and S. Kegels. Health belief model. *United States Public Health Service*, 1952.
21. D. Kahneman. *Thinking, fast and slow*. Macmillan, 2011.
22. P. Klasnja, S. Consolvo, T. Choudhury, R. Beckwith, and J. Hightower. Exploring privacy concerns about personal sensing. In *Pervasive Computing*, pages 176–183. Springer, 2009.
23. P. Klasnja, S. Consolvo, and W. Pratt. How to evaluate technologies for health behavior change in hci research. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3063–3072. ACM, 2011.
24. R. Kukafka. Tailored health communication. *Consumer Health Informatics: Informing Consumers and Improving Health Care*, pages 22–33, 2005.
25. N. D. Lane, M. Mohammad, M. Lin, X. Yang, H. Lu, S. Ali, A. Doryab, E. Berke, T. Choudhury, and A. T. Campbell. Bewell: A smartphone application to monitor, model and promote wellbeing. In *5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth2011)*, 2011.
26. J. Lester, C. Hartung, L. Pina, R. Libby, G. Borriello, and G. Duncan. Validated caloric expenditure estimation using a single body-worn sensor. In *Proceedings of the 11th international conference on Ubiquitous computing, Ubicomp '09*, pages 225–234, New York, NY, USA, 2009. ACM.
27. L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010.
28. H. Lu, J. Yang, Z. Liu, N. D. Lane, T. Choudhury, and A. T. Campbell. The jigsaw continuous sensing engine for mobile phone applications. In *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems*, pages 71–84. ACM, 2010.
29. T. Lu, D. Pi, and M. Pi. Showing relevant ads via lipschitz context multi-armed bandits. In *Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010.
30. S. McCallum. Gamification and serious games for personalized health. *Stud Health Technol Inform*, 177:85–96, 2012.



31. C. D. Mohr, M. S. Schueller, E. Montague, N. M. Burns, and P. Rashidi. The behavioral intervention technology model: An integrated conceptual and technological framework for ehealth and mhealth interventions. *J Med Internet Res*, 16(6):e146, Jun 2014.
32. L. G. Morrison, C. Hargood, S. X. Lin, L. Dennison, J. Joseph, S. Hughes, D. T. Michaelides, D. Johnston, M. Johnston, S. Michie, et al. Understanding usage of a hybrid website and smartphone app for weight management: A mixed-methods study. *Journal of medical Internet research*, 16(10), 2014.
33. MyFitnessPal, LLC. <http://www.myfitnesspal.com/>, 2013. [Online; accessed 19 March 2014].
34. I. Nahum-Shani, S. N. Smith, A. Tewari, K. Witkiewitz, L. M. Collins, B. Spring, and S. Murphy. Just in time adaptive interventions (jitais): An organizing framework for ongoing health behavior support. *Methodology Center technical report*, (14-126), 2014.
35. National Heart, Lung and Blood Institute, National Institutes of Health. Getting started and staying active. [http://www.nhlbi.nih.gov/health/public/heart/obesity/lose\\_wt/calories.htm](http://www.nhlbi.nih.gov/health/public/heart/obesity/lose_wt/calories.htm), 2011. [Online; accessed 19 March 2013].
36. National Heart, Lung and Blood Institute, National Institutes of Health. Healthy eating plan. [http://www.nhlbi.nih.gov/health/public/heart/obesity/lose\\_wt/calories.htm](http://www.nhlbi.nih.gov/health/public/heart/obesity/lose_wt/calories.htm), 2013. [Online; accessed 19 March 2013].
37. Netflix. <http://www.netflix.com/>, 2013. [Online; accessed 19 March 2014].
38. J. Noronha, E. Hysen, H. Zhang, and K. Z. Gajos. Platemate: crowdsourcing nutritional analysis from food photographs. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 1–12. ACM, 2011.
39. C. L. Ogden, M. D. Carroll, B. K. Kit, and K. M. Flegal. Prevalence of obesity in the united states, 2009–2010, 2012.
40. P. Onghena and E. S. Edgington. Customization of pain treatments: Single-case design and analysis. *The Clinical journal of pain*, 21(1):56–68, 2005.
41. R. L. Ott and M. T. Longnecker. *An Introduction to Statistical Methods and Data analysis*, 4th. New York: Duxbury Press, 1993.
42. T. Philipson and R. Posner. Is the obesity epidemic a public health problem? a decade of research on the economics of obesity. Technical report, National Bureau of Economic Research, 2008.
43. J. Pinheiro and D. Bates. Nlme: Software for mixed-effects models, 2000.
44. J. P. Pollak, P. Adams, and G. Gay. Pam: a photographic affect meter for frequent, in situ measurement of affect. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 725–734. ACM, 2011.
45. W. B. Powell. *Approximate Dynamic Programming: Solving the curses of dimensionality*, volume 703. John Wiley & Sons, 2007.
46. J. O. Prochaska and W. F. Velicer. The transtheoretical model of health behavior change. *American journal of health promotion*, 12(1):38–48, 1997.
47. M. Rabbi, A. Pfammatter, M. Zhang, B. Spring, and T. Choudhury. Automated personalized feedback for physical activity and dietary behavior change with mobile phones: A randomized controlled trial on adults. *JMIR mHealth uHealth*, 3(2):e42, May 2015.
48. M. C. Roberts, A. S. Dizier, and J. Vaughan. Multiobjective optimization: Portfolio optimization based on goal programming methods.
49. E. M. Rogers. *Diffusion of innovations*. Simon and Schuster, 2010.
50. J. D. Singer and J. B. Willett. *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford university press, 2003.
51. L. Springvloet, L. Lechner, H. de Vries, M. J. Candel, and A. Oenema. Short-and medium-term efficacy of a web-based computer-tailored nutrition education intervention for adults including cognitive and environmental feedback: Randomized controlled trial. *Journal of medical Internet research*, 17(1):e23, 2015.
52. R. Sriraghavendra, K. Karthik, and C. Bhattacharyya. Fréchet distance based approach for searching online handwritten documents. In *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, volume 1, pages 461–465. IEEE, 2007.
53. A. Tashakkori and C. Teddlie. *Mixed methodology: Combining qualitative and quantitative approaches*, volume 46. SAGE Publications, Incorporated, 1998.
54. B. E. Wilde, C. L. Sidman, and C. B. Corbin. A 10,000-step count as a physical activity target for sedentary women. *Research Quarterly for Exercise and Sport*, 72(4):411–414, 2001.
55. T. Zhang, R. Ramakrishnan, and M. Livny. Birch: an efficient data clustering method for very large databases. In *ACM SIGMOD Record*, volume 25, pages 103–114. ACM, 1996.
56. L. M. Zhou and C. Gurrin. A survey on life logging data capturing. *SenseCam 2012*, 2012.