“Run Forrest Run!”: Measuring the Impact of App-Enabled Performance and Social Feedback on Running Performance

Completed Research Paper

Yash Babar
University of Minnesota
CSOM 19th Ave S
Minneapolis, MN 55455
babar008@umn.edu

Jason Chan
University of Minnesota
CSOM 19th Ave S
Minneapolis, MN 55455
jchancf@umn.edu

Ben Choi
Nanyang Technological University
50 Nanyang Ave, block s3,
Singapore 639798
benchoi@ntu.edu.sg

Abstract

Exercise tracking apps are a novel, scalable, and affordable tool for delivering personalized behavioral interventions. While thousands of fitness tracking solutions emerge in the market, there is a lack of systematic research that quantify their effectiveness on exercise outcomes, making it hard for for practitioners and users to know the true value of these apps. Drawing on the literature on motivation literature, this paper elucidates the effects of app-enabled motivation on running performance. Specifically, this study examines the two most common forms of feedback available to users of exercise tracking apps, namely performance feedback and social feedback. We conducted a 18-month long randomized field experiment, with 1,241 military servicemen, to assess the causal effect of these feedback types on actual exercise outcomes. Results from the experiment provided evidence that these two app features improve the running times of the service-men. We further discuss the temporality and heterogeneity of these effects.

Keywords: app effectiveness, field experiment, feature evaluation, usage impact, feedback

Introduction

With sustained improvements in processing power and continued success in miniaturizing electronic devices, we see a recent movement towards the increased usage and demand for activity tracking technologies and wearables. When designed and used effectively, tracking technologies can be novel and powerful tools for inducing behavior modifications that nudge casual users towards a more active lifestyle (Hebden et al. 2012). Termed broadly as mHealth, the industry comprising of lifestyle maintenance, health monitoring and community-based wellbeing apps, represented a non-trivial market of US$28 billion in 2017. According to PEW, approximately one in five smartphone users utilized at least one app to track their

1 This figure is expected to grow to US$100 billion by 2023 (ltd and Markets 2017)
progress towards some fitness goals, and 38% of health app users had downloaded at least one app for exercise tracking (Fox and Duggan 2010).

These trends have led to the entry of both large companies and start-ups in the fitness app market, with optimistic goals of extracting sizable revenues in the future. Over the last few years, mHealth apps have grown substantially, with more than 40,000 different app offerings available on the market (Aitken and Gauntlett 2013). While users are able to choose from thousands of fitness tracking solutions, the excessive diversity of apps available has made it difficult for clinicians and the public to discern which apps are most effective (Powell et al. 2014).

More worryingly, a lack of systematic research to evaluate and quantify the effectiveness of these apps has meant that the fitness apps are often developed without any grounding in scientific evidence or health behavior theory (Oowan et al. 2013). A study by Conroy et al. (2014) found that a majority of the top-ranked fitness apps in the market rarely have features that incorporate behavior change techniques. In addition, critics have cited the issue of low retention rate of apps as a barrier that undermines the effectiveness of fitness apps (Dennison et al. 2013). Currently, the evidence offered by existing studies examining app effectiveness is somewhat mixed (McCurdie et al. 2012). Through survey responses, Litman et al. (2015) found that individuals who use fitness apps were more likely to exercise during their leisure time, compared to those who did not use such apps. Yet, among others, Plow and Golding (2017) found that the usage of a mHealth app among adults did not materially improve users’ physical functions. The market of fitness apps has gravitated towards the development of ‘populist’ features instead of evidence-based features, which may not provide any real value to the fitness outcomes of users.

The mixed findings within the existing literature highlight the need for additional high-quality controlled trials to assess the effectiveness of fitness apps (Free et al. 2013). Furthermore, IS scholars have called for greater efforts towards the thorough evaluation of IT artifacts, so that IS research can provide more design-based guidelines to enhance the utility of various technologies (Bichler 2006). Responding to these needs, we conducted a randomized field experiment to assess the impact of fitness apps on physical outcomes. While past studies have reasoned that the lack of app features that incorporate behavioral change techniques (BCT) may severely limit the effectiveness of these apps, e.g., Conroy et al. (2014), no study has formally established the positive causal link between the individual BCT features and physical outcomes (Pagoto and Bennett 2013). We specifically focused on assessing two BCT features, namely performance feedback and social feedback.

We looked at the change in users’ running times for a fixed distance as a dependent variable over an extended period of time (i.e., 18 months), as medical studies have shown that running performance is an accurate indicator of cardiovascular health and physical fitness (Gupta et al. 2011, Berry et al. 2011). We were able to solicit the cooperation of a military organization in executing the field experiment. Through this collaboration, we observed the running times of its servicemen (before and after the prescription of experimental treatments), assessed through standardized physical fitness tests that were administered through the military organization. This unique partnership allowed us to have a direct measure of the physical outcomes of the users, thereby alleviating issues of measurement errors that are present in self-reported measures. The ability to randomly assign individuals to groups that utilized different app features further alleviated estimation challenges faced in previous work.

Our study makes both practical and theoretical contributions. First, our findings provide direct guidelines to several stakeholders. The assessment and verification of the impact of app features informs both users and app developers alike, of their efficacy. Second, our work contributes to the literature on the human-computer interaction and design aspects of IT artifacts. In this regard, our work provides a conceptual understanding of two common app-based feature based on the feedback type they offer, by which the relationships between these feedback types on athletic performance and behavioral outcomes are characterized using explanations derived from the theoretical underpinnings of user motivation. In addition, we provide finer theoretical insights into how various design features may influence and sustain usage behavior, which is currently not well understood given the low retention rates observed across the app market.
Background and Theory

Mobile Apps and Healthcare

With the increasing popularity of apps, IS research has embarked on the study of apps in a variety of ways. Specifically, past studies have examined the factors driving the sales and demand for apps (Wang et al. 2018), determinants of app success (Claussen et al. 2013), and usage patterns of apps (Kwon et al. 2016). There is also a small but growing stream of work that attempts to shed light on the design considerations of apps (Ghose et al. 2012, Gu et al. 2017). Specifically, with the increasing penetration and usage of mHealth apps and wearables, scholars in the management sciences have begun to assess the value of such apps and related interventions (Yan 2018).

Given the importance of this stream of work, researchers from fields outside IS have also attempted to assess the efficacy of mHealth app design in recent years. Researchers have examined the types of features implemented in fitness apps (Yang et al. 2015), the usability of these apps (Zapata et al. 2015), the impact of such apps on health and physical outcomes (Plow and Golding 2017), and the possible directions and frameworks for improving their design (Powell et al. 2014). Collectively, these research efforts reveal an important point. Based on the existing study results, it is unclear whether the currently available fitness apps are indeed effective in generating physical benefits. Even if certain apps were shown to influence exercise behaviors, no study till date has traced and attributed the specific app features that are responsible for inducing such effects, and for whom they are effective for.

The lack of consistent evidence on the positive impact of mHealth apps is no accident. Different fitness apps consist of different sets of features (Direito et al. 2014), and the joint effect of different features can lead to varied outcomes, making it possible for studies that examine different fitness apps to arrive at dissimilar results. Moreover, most existing studies on fitness apps do not have ready access to physiological measurements of the end users, but have instead relied on self reports to track users’ physical activity outcomes (Pagoto and Bennett 2013). This is problematic as these reports are subject to reporting biases and measurement errors. In addition, active individuals tend to self-select into the use of fitness apps. It is likely that those adopting mHealth apps, may also undertake other measures like changed diet and personalized exercise regimes. The selection bias herein compounds on the complicated task of attributing performance improvements to features within fitness apps. Further, past studies are often limited by short post-intervention observation periods (Bort-Roig et al. 2014), making it hard to tell whether the observed improvements in physical outcomes are indeed representative of the long-term effectiveness of these apps.

The evaluation of fitness apps is essentially a particular instance involving IT design and usage. Despite the call for more empirical studies to assess the efficacy of mHealth apps (Pagoto and Bennett 2013) and the practical fit of IS research to address such questions, efforts from the IS community to evaluate the design aspects of fitness apps have been scant. At their core, fitness apps are technological systems that provide users with different informational feedback. On top of helping to log the details of users’ run, fitness apps also convey specific information to users, with the aim of informing and motivating them to attain specific goals relating to physical activities.

In a review of fitness apps, Middelweerd et al. (2014) found that performance feedback is an informational feature that is available across all the apps that utilize behavioral change techniques (BCT). Similarly, features enabling social support are another commonly available BCT feature, with close to 80 percent of the top rated fitness apps having some form of social feature that allows users to communicate to other users (Yang et al. 2015). Conceptually, these two app features represent two disparate types of information with contrasting nature. Performance feedback is a set of structured, objective information, provided by the system when certain conditions are met. In contrast, social feedback refers to a set of unstructured, subjective information, provided by humans in a non-deterministic fashion. It is theoretically interesting to see and contrast how user behavior might be influenced by these two types of information, and the resultant effect on users’ physical outcomes.
**Performance Feedback**

Individuals can be motivated to tackle distant or difficult goals (e.g., fitness performance goals) by using intermediate goals that are smaller, proximal chunks of the final goal (Amari et al. 2011). In their seminal study, Amir and Ariely (2008) posit that any task can be characterized by the amount of progress information it embodies. In practice, personal trainers often employ this technique by verbally mentioning the number of completed "reps" of exercises in training sessions to make exercisers cognizant of their progress. Furthermore, Amir and Ariely (2008) postulate that progress information establishes natural intermediate goals, which are often considered more immediate and achievable, compared to the overarching distant final endpoint. Thus, by acknowledging users’ accomplishment of an intermediate milestone, an app feature which provides information on users’ exercise progress via their current performance outcomes can increase feelings of satisfaction and even strengthen the habit of engaging in training (Lally and Gardner 2013).

The satisfaction that accompanies the positive reinforcement of making progress in their performance can help maintain long-term behavior change. It does so by inducing the belief that the exerted effort is a good choice and is associated with cognitive rewards (Rothman 2000). Further, the HCI literature has widely recognized the importance of feedback designs that allow people to be aware of their intermediate achievement, which is essential for spurring greater subsequent effort and instilling goal persistence (Esakia et al. 2018). In addition to inducing positive reinforcement, by updating user perceptions of the covered distance and exercise duration, performance feedback can also alter perceived physical exertion (Tucker 2009). This feature is particularly helpful in motivating users to expend more effort when they realize that they have been exercising less than their expectations.

A unique aspect of performance feedback is the automated and deterministic nature in which it can be given. Such feedback is routinely and systematically given to users when they reach certain exercise milestones. While training sessions can be repetitive and monotonous, runners experience satisfaction when they learn that their timing and distance goals improve with each run. Acknowledgement of user progress can invoke persistent and sustained training behaviors. Thus, the routine and predictive nature of performance feedback can have the added effect of cultivating user commitment towards long-term goals that might seem cognitively distant or difficult to achieve at first glance (Dorris et al. 2012).

**Social Feedback**

In addition to computer-generated objective feedback, it is also useful to consider the use of social behavioral change strategies as an app feature. A common and simple way for implementing such feature is to allow users to receive social feedback. Social feedback can be implemented as an app feature that facilitates user-to-user interactions during app usage. By allowing for user-to-user interactions, fitness apps can induce effects of social facilitation, which can lead to performance improvement due to the mere presence of others. Individuals become motivated to work harder and improve their performance, when they are aware of the presence of others, who might be passive observers or active spectators of their efforts (Allport 1924). Past studies examining social facilitation have shown that the presence of others can enhance performance for a variety of cognitive tasks and motor activities (Guerin 1999).

Although different views of social facilitation have emerged, theorists agree that audience support is one of the main mechanisms that lead to increased performance. Audience support refers to the praise and encouragement given by the spectators who observe individuals’ performance (Kanouse et al. 1981). While praise focuses on acknowledging past performance, encouragement often is future-focused and emphasizes motivating sustained and/or improving current performance. Extant evidence suggests that praise can lead to performance improvement by enhancing individuals’ perceptions of self-competence and by making individuals feel good about themselves (Ryan and Deci 2000). Because of the positive experience of being praised, individuals are likely to continue to demonstrate the praised performance to sustain the attention of and approval by their audience. Encouragement is considered a form of social support, which can bolster individuals’ sense of self-esteem (Freeman et al. 2014). Further, the positive effects of encouragement on performance can be explained through the notion of distraction (Andreacci et al. 2002). In physical training,
individuals might focus on physical sensations, such as pain or exhaustion. Encouragement provides a dis-associative cognitive mechanism through which individuals’ focus on the physical sensations is redirected to the presence of others, thereby distracting them from physical discomfort and making the training feel less cumbersome (Scott et al. 1999). In sum, the presence of social feedback features within fitness apps can provide motivational effects that enhance users’ physical outcomes.

**Impacts of Performance and Social Feedback on Usage Behavior**

Ultimately, the efficacy of these feedback types on improving exercise outcomes is contingent on the actual usage behaviors. As such, it is worthwhile to contrast and discuss the theoretical characteristics underlying these two types of app features, to understand their potential effect on usage behaviors. The main difference between performance feedback and social feedback is the source by which information is given. While performance feedback is mechanically generated by a computer algorithm (i.e., machine-based feedback), social feedback is made available through other users’ (i.e., human-based feedback). The distinction in feedback source leads to secondary differences between the type of information provided, i.e., structured as opposed to unstructured, deterministic as opposed to non-deterministic, objective as opposed to subjective.

One of the defining aspects of computer-generated performance feedback is the reliability in which it is given. The deterministic nature of computer algorithms in generating feedback, repeatedly, whenever certain conditions are met meant that users can expect exercise progression information to be provided in a predictable and dependable fashion. Such regularity is helpful in fostering user trust and continued usage (Muir 1987). In contrast, social feedback given by human users is relatively non-deterministic. Physical and cognitive resources are depleted over time (Kaniasty and Norris 1993), by which the level of social support from human users would deteriorate with time (Lepore et al. 1991). As such, the reliability of social support can dissipate as soon as the audience are not as generous in providing approval (Henderlong and Lepper 2002). In this regard, it appears that performance feedback is more apt at inducing sustained usage compared to social feedback.

Although computer-generated performance feedback is more consistent in terms of its availability compared to human-based social feedback, this highly structured feedback can also be limiting in that it is less adaptive to the physical and emotive states of users. Most of the computer-generated feedback lack the ‘emotional intelligence’ to sense and respond appropriately to users’ effective states, and to skillfully adapt the delivery of message for constructive purpose (Pantic and Rothkrantz 2003). Thus, performance feedback can be perceived as unnatural/rigid (Nakatsu 1998) and consequently less efficacious and persuasive, compared to social feedback. Since human users are social beings, human-based social feedback are more superior in enhancing initial app adoption and usage, compared to computer-generated performance feedback. At the time, social feedback might be more suited for users who are inherently less motivated to participate in physical exercises, because social encouragement adapted on the basis of the user’s current performance can provide the appropriate emotional motivation to spur app usage and exercise behavior.

**Experimental Details**

**Study Context**

To study our research questions, we collaborated with a military organization to launch a field experiment in four regiments, spanning from October 2014 to mid of April 2016. In our study, a regiment on average comprises of 312 servicemen. Our study sample is made up of operationally ready male service personnel drafted at least 2 years prior to the experiment and were in the workforce. The process of mandatory conscription is advantageous to our research objectives, as any random selection of servicemen from military organization is likely to be a representative sample of the male population of the relevant age group of the country. Conscription has been widely used as a proxy for random selection in prior work (Meyer 1995) and

---

2 Under a non-disclosure agreement, we are unable to disclose the identity of the organization

---
its usage here enhances the generalizability of our results. Other advantages of using a military organization as a study context include: high treatment compliance, impartial periodic measurement of behavioral and physiological outcomes, and the ability to reduce cross-contamination between experimental groups (to be discussed next).

To assess their level of physical fitness, servicemen in the military organization undergo a compulsory physical fitness test on an periodic basis. The physical fitness test (abbreviated as PFT henceforth) systematically tracks exercise outcomes related to running and strength training. One of the assessments within the PFT is a 2.4km run, by which servicemen have to complete within a set amount of time. We observed test results (i.e., run times) of all subjects before and after the introduction of the experimental manipulations. Specifically, we observed the five latest test timings prior to the start of the experiment, and three test timings after the experimental conditions were initiated. Each subject within the study took the PFT within a six-month window during the experimental period. By observing the outcome of interest for each subject over different occasions, our study design mirrors a repeated measures experiment.

These standardized fitness tests allowed us to arrive at impartial measurements of a subject’s physical performances over time. It is well known that Hawthorne effect might hamper experiments, as the act of measuring and observing outcomes in these settings can influence the outcome itself (McCambridge et al. 2014). This was less of a concern in our study, as the PFT within the military organization existed long before the experiment was conducted, and is the standard test that all servicemen take during their military service. Moreover, it is a test system that utilizes an incentive structure that rewards servicemen such that they would perform as well as they possibly could when taking the physical test. Under military laws, servicemen who do not pass this compulsory fitness test have to undergo remedial training sessions, which takes additional time out of their daily lives. In addition, all servicemen (including those who are not in the study) who perform well in these fitness tests are given cash rewards. PFT tests have age determined passing criteria which must be met at least once annually. Also measured in these tests are physiological measures such as height, weight and BMI, which we use as controls. The standardized nature of executing the tests and time keeping also help to ensure that the run times of study subjects were measured in a consistent fashion. The recorded running time for the PFT test (measured in seconds) is our primary outcome of interest.

We used the popular application Endomondo to deliver treatment interventions to our subjects. There are several reasons for choosing this particular app. First, it is a widely used exercising app that which has features that are common across most running apps. Second, Endomondo allows for the delivery of progress and social feedback during exercise through its Audio Coach and Pep Talk features, respectively. Furthermore, it is possible to selectively manipulate the availability of these features by turning on and off these features in Endomondo, which allowed us to create different treatment groups that only differed in terms of the selected feature. Third, Endomondo delivers audio interventions in addition to plain text/image based interventions. This is another advantage as spoken feedback during exercise has been shown in the past to have strong behavioral impacts (McNair et al. 1996). Finally, the application also allowed a LiveTracking feature which allowed the live broadcasting of a focal user’s run to his/her social network. This enables peers to be aware of the user’s training efforts in real-time, thereby providing opportunities for synchronous social feedback. Friends of exercising individuals would be alerted under the live tracking feature. Such a feature design encourages social feedback to be solicited in time windows close to or during each individual’s exercises, which is a helpful experimental design because the app features scopes the occurrence of both feedback types under study for better comparison.

**Experiment Design**

Four main experimental conditions were deployed for the experiment. Each regiment was randomly assigned to one of the four experimental groups: Control, Performance Feedback (PF), Social Feedback

---

3 For ease of reference, we refer to each period between two adjacent tests as an experimental wave.

4 Note that our study subjects are not full-time soldiers. They have regular jobs and have regular lives outside their reservist duties. Servicemen who spent time on additional remedial sessions are not financially compensated for their time.
The purpose of randomly assigning each regiment to one type of experimental treatment was to minimize the effects of contamination between experimental groups, due to between-subject communications. In each of these experimental groups, the servicemen within the regiment were briefed twice, once before experiment Wave 1 and once before experiment Wave 2. In the first briefing, the servicemen were asked to download the running app, Endomondo, and were instructed to turn on specific features and to use the app in their personal training. To encourage usage and adherence to the treatment, servicemen were given exercise armbands to hold their mobile devices while exercising. In the Control group, subjects were asked to use the standard features of Endomondo, excluding in-exercise feedback that was prescribed to the treated groups. Subjects in the Performance Feedback (PF) condition received in-exercise auditory feedback in addition to the basic features of Endomondo used by the control group. Specifically, users were prompted of their running progress whenever they completed one kilometer of running/walking in single session. The message delivered included total distance covered, total time taken and the time taken for the last kilometer covered (known in-app as a lap). The subjects in the Social Feedback (SF) condition could follow each others’ progress through the app and provide text-based comments on a peer’s activity in real time. These user-generated comments were read out to the focal user during his exercise. Based on our data, we found that most comments were inspirational, encouraging, sportingly competitive, or appreciative. Some examples include: “Run like someone’s chasing you”, “The end is right in front”, “Kudos! I can see that U have been running a lot!”. In the Performance & Social Feedback condition, subjects received auditory feedback on their performance as well as comments generated by peers.

As training programs, housing and socialization were confined by organizational structures within a regiment, the addition of peers from the same platoon was an intentional design consideration to minimize cross-contamination across treatment groups and capitalize on the existing offline social networks. Besides the four main experimental groups, we also randomly assigned subjects within each regiment to alternative baseline conditions: a group without any apps Baseline, and a group with a random non-exercise app BaselineDummyApp. The latter was created to contrast and account for potential endowment effects of a social mobile application.

**Manipulation Checks** At the end of each wave, officers randomly checked the apps of servicemen, to see if they followed the instructions in turning on and using the assigned app feature(s). This check was helpful in reducing non-compliant behavior. Subjects were free to use the app as frequently/infrequently as they saw fit. We also administered a post experiment survey a week after the final wave finished. The survey was delivered over two weeks in three batches to randomly selected subjects from each treatment condition with the Endomondo app. This was done to minimize collusion behaviors in the responses within each military sub-organization. Each subject was surveyed only once and by the end of the two weeks, and all 825 participants who had used Endomondo were surveyed. The questions measured awareness of features available in the subject’s condition.

**Power Analysis** We checked that our experimental design delivered adequate statistical power as suggested in (Enderlein, 1985). We used the run times of subjects one year prior to treatment as a proxy for the mean of our primary outcome and estimated a medium sized effect of 10% (Cohen’s $d = 0.45$). The breakdown of the sample size for each experimental group was as follows: Baseline (N=212), BaselineDummyApp (N=204), Control (N=219), PF (N=202), SF (N=203), PF&SF (N=201).

**Data**

The average serviceman was about 30 years old, with an annual income of 33,010 US dollars and held a diploma degree. We compared the activity levels of our sample to that of the country’s population at large, by age groups (Win et al., 2015). The study sample was found to be highly comparable to the overall population, suggesting that the pre-treatment behavior and physiology of our sample were representative of the nation’s population of young men.

---

On top of these five friends, subjects were free to add other friends to their app.
Randomization Checks We tested for differences in outcome variables and covariates between groups in the pre-treatment period using t-tests, to assess whether the groups are properly balanced through our randomization assignment. All covariates were balanced across groups, except for BMI (averaged over last 3 years for each subject). We also performed analysis of variance tests with Tukey’s ‘Honest significant difference’ method. These too showed no differences between conditions, other than pre-treatment BMI values. While the differences in BMI values were statistically different, the magnitude of these differences across groups were practically small. For instance, the control condition had an average BMI 1.2% higher than the performance feedback condition. While the magnitude of this difference was small, BMI is an important predictor of running speeds (Campbell 1985). As such, we took extensive measures in our analyses to account for any potential impact of this difference on our outcomes. A summary of the descriptive statistics of the in-experiment variables used in our analysis are presented in Table 1.

Analysis and Results

Our primary outcome of interest is RunTime\(_{i,t}\), which was the measured PFT runtime of subject \(i\) at time \(t\). The subjects’ running performances manifested as a right skewed distribution with few slow runners populating the long tail. To account for this skewed distribution, we considered the logarithm of RunTime\(_{i,t}\) as our primary dependent variable.

Main Results

We formally tested the treatment effects using regression analyses. To account for temporal variations, we included experimental wave fixed effects \(\gamma_t\) in our specification. To ensure that the differences in BMI did not influence our results, we controlled for pre-experimental BMI\(_i\) of each subject. Recall that each subject received the same experimental treatment throughout the experiment. As such, in lieu of individual fixed effects, we accounted for individual heterogeneity in running performance by including each subject’s latest pre-experimental logged PFT runtime: PastRunTime\(_i\), as a control.

Our first set of analyses examined the impact of treatment on run times by modelling each experimental condition as independent dummies Cond\(_i\). First, we compared each condition against BaselineNoApp and BaselineDummyApp to assess if the use of different variants of the app would lead to an improvement in run times. Next, to understand whether the app features of interest lead to incremental effects of using the app, we compared the different feedback conditions to the control group that utilized the same app but without the feedback features: Control. All three regressions were of the form:

\[
\log(\text{RunTime}_{i,t}) = \gamma_t + \delta_1 \cdot \text{BMI}_i + \delta_2 \cdot \log(\text{PastRunTime}_i) + \delta_3 \cdot \text{Cond}_i + \epsilon_{i,t}
\]  

(1)

The results are presented in Table 2. In Model I the BaselineNoApp group was treated as the reference group for comparison. It should be noted that these log linear regressions with standard errors clustered at the individual subject level are essentially a multi-factor analysis of co-variance (ANCOVA), often used as the primary test in repeated measures experimental analysis (Unwin 2013). We found that in general the use

---

**Table 1: Descriptive Statistics for Key Variables**

<table>
<thead>
<tr>
<th>Statistic</th>
<th>N</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>PerformanceFeedback(PF)</td>
<td>2.475</td>
<td>0.488</td>
<td>0.500</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>SocialFeedback(SF)</td>
<td>2.475</td>
<td>0.490</td>
<td>0.500</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>SFFrequency(SFfrq)</td>
<td>2.475</td>
<td>5.750</td>
<td>7.102</td>
<td>0</td>
<td>26</td>
</tr>
<tr>
<td>RunTime(Secs)</td>
<td>3.721</td>
<td>834.071</td>
<td>197.286</td>
<td>476</td>
<td>1,436</td>
</tr>
<tr>
<td>UsageFrequency(UF)</td>
<td>2.475</td>
<td>38.876</td>
<td>23.519</td>
<td>1</td>
<td>159</td>
</tr>
<tr>
<td>TimePeriod</td>
<td>3.723</td>
<td>2.000</td>
<td>0.817</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>AppFriends</td>
<td>2.475</td>
<td>4.364</td>
<td>5.270</td>
<td>0</td>
<td>21</td>
</tr>
</tbody>
</table>

Note: All variables reported are aggregated by user per experimental wave. We codified the treatment conditions (i.e., PF and SF) into binary indicators and reported their respective statistics.
of the Endomondo app leads to a significant decrease in overall individual runtimes (i.e., an improvement in running performance). For the subjects in the Control condition, a decrease of 1.3% (i.e., a 11.4 seconds improvement compared to the average run times of the Baseline NoApp group) was observed. Those in performance feedback group experienced the greatest improvement in runtimes (4.3%, 37.7 seconds), followed by those in the joint condition (3.9%, 34.19 seconds). Social feedback was found to reduce run times by 3.2% (28.05 seconds). While the improvements in run time in terms of percentages might seem modest at first glance, the actual reduction in run times (i.e., seconds) are practically noteworthy. To see this more concretely, every grade for the 2.4km running test in the military organization is separated by a 40-second difference. That is, servicemen would improve their running grade by one level, should they reduce their running times by 40 seconds. In general, a time difference of 30 seconds or more for semi-long runs, such as that of a 5km run - a standard Olympic track and field event, could distinguish good runners from mediocre runners. Thus, the magnitude of the effect brought by PF, SF and joint treatment are practically significant.

As expected, past BMI and pre-experiment test runtimes are correlated with the current running performance of the participants. It might be possible that the mere presence of a mobile phone on runs (as needed in all treatment conditions) could drive these basic results. To account for this, we re-estimated the effects with the Baseline Dummy App condition as the comparison group in Model 2. These results were found to be consistent with the earlier results. To tease out the incremental effects of the aforementioned app features over the use of basic tracking features in Endomondo, we restricted our next analysis to include only subjects in the groups: Control, PF, SF and SF&PF. The results of this analysis are reported in Model 3. In subsequent analyses, we adopt the same procedure of restricting our sample to these four main experimental groups, with the Control group serving as the comparison group. Using Control as the comparison group, we found that PF is the most effective in reducing average running times (across all waves) followed by SF&PF and finally by SF (see Model 3).

By specifying the model with dummies denoting experimental conditions, Equation 2 estimates the effect of each treatment group with respect to the control group. A limitation for this approach is that it cannot estimate the marginal effect of a feedback at different levels of the second feedback, i.e., the interaction effect of PF and SF. Thus, we repeated our analysis using binary indicators to denote each type of treatment. PF_i and SF_i were marked 1 or 0 depending upon whether the subject i received performance and social feedback, respectively. The resultant model estimated is shown in Equation 2.

\[
\log(\text{RunTime}_{i,t}) = \gamma_i + \delta_1 \text{BMI}_i + \delta_2 \log(\text{Past RunTime}_i) + \beta_1 \text{PF}_i + \beta_2 \text{SF}_i + \beta_3 (\text{PF}_i \times \text{SF}_i) + \epsilon_{i,t} \quad (2)
\]

### Table 2. Impact of treatment conditions on 6 monthly PFT Run times from Equation 2.

<table>
<thead>
<tr>
<th>Ref Group =</th>
<th>NoApp (1)</th>
<th>DummyApp (2)</th>
<th>Control (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>cond BaselineDummyApp</td>
<td>0.0002 (0.0034)</td>
<td>0.0002 (0.0034)</td>
<td>0.0002 (0.0034)</td>
</tr>
<tr>
<td>cond BaselineNoApp</td>
<td>-0.0131*** (0.0031)</td>
<td>-0.0002 (0.0034)</td>
<td>-0.0129*** (0.0034)</td>
</tr>
<tr>
<td>condControl</td>
<td>-0.0433*** (0.0036)</td>
<td>-0.0331*** (0.0036)</td>
<td>-0.0291*** (0.0036)</td>
</tr>
<tr>
<td>condPF</td>
<td>-0.0322*** (0.0036)</td>
<td>-0.0320*** (0.0036)</td>
<td>-0.0181*** (0.0036)</td>
</tr>
<tr>
<td>condSF</td>
<td>-0.0391*** (0.0037)</td>
<td>-0.0389*** (0.0036)</td>
<td>-0.0255*** (0.0036)</td>
</tr>
<tr>
<td>condSF&amp;PF</td>
<td>-0.0063*** (0.0009)</td>
<td>0.0063*** (0.0009)</td>
<td>0.0114*** (0.0015)</td>
</tr>
<tr>
<td>BMI</td>
<td>1.0032*** (0.0062)</td>
<td>1.0032*** (0.0062)</td>
<td>0.9863*** (0.0100)</td>
</tr>
<tr>
<td>RunT ime</td>
<td>3721</td>
<td>3721</td>
<td>2473</td>
</tr>
<tr>
<td>Adj. R^2 (full model)</td>
<td>0.9550</td>
<td>0.9550</td>
<td>0.9554</td>
</tr>
<tr>
<td>Adj. R^2 (proj model)</td>
<td>0.9546</td>
<td>0.9546</td>
<td>0.9548</td>
</tr>
<tr>
<td>Wave FE</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Note: ***p < 0.01, **p < 0.05, *p < 0.10
SE are clustered at the individual level.
We also note that the strength of the social feedback treatment was contingent on the frequency by which each subject received messages from their peers on the app. To account for treatment intensity in our analysis, we replaced $SF_i$ with a continuous measure of the frequency of social feedback received: $SfF_{req}$.

![Predicted RunTimes vs Social Feedback Frequency](image)

**Figure A.** Change of test Runtimes with social feedback frequency

When modeling the joint feedback condition as an interaction of individual binary dummies (as per Equation 2), we found that Performance Feedback ($PF, \beta = -0.0291, P \text{ val} < 0.01$) and Social Feedback ($SF, \beta = -0.0238, P \text{ val} < 0.01$) continued to exhibit significant reductions in run times. Interestingly, when used in conjunction, their effect on run times was non-additive ($PF*SF, \beta = 0.0218, P \text{ val} < 0.01$). The average impact of PF was found to be dampened by the co-presence of SF. We relied on the frequency of social feedback received to capture treatment intensity of social feedback. Through the coefficient of $SfF_{req}$ ($\beta = -0.0014, P \text{ val} < 0.01$), we saw that the effect of social feedback did improve run times with the increase in the number of social interactions. We note that being in the social feedback group did not necessarily mean that a user would indeed receive social feedback. Such cases comprise of 0.6% of our sample. After removing these rows from our analysis, the results remained qualitatively similar.

The joint impact of performance and social feedback should be understood through the following sum: $\beta_1 + SfF_{req} \times (\beta_2 + \beta_3)$, which equates to $-0.0238 - 0.0002 \times SfF_{req}$, which is negative in effect. In essence, when we considered the frequency of feedback received, the joint condition outperformed both independent conditions in improving average run times. Figure A illustrates this interaction effect, by showing the variation in logged runtimes over the different social feedback intensity. The impact of the social only condition (SF) increased in $SfF_{req}$ but were less that of (PF) alone. The same figure also reveals another interesting trend: while the overall improvement to runtimes in the joint condition is larger, the slopes indicate that the marginal gains from social feedback volume are much lower in the joint condition compared to that of the SF only condition. Such a result suggests that in the presence of multiple app features, the additional gain from social features may be limited by individuals’ physical ability to improve their exercise outcomes.

**Robustness Checks**

**Inclusion of Random Effects** Our estimated effects could possibly be influenced by inter-regimental differences, as each regiment was assigned to only one treatment condition. While this ensured that subjects did not learn of alternative treatment schemes (which would lead to a more serious issue of cross-contamination), it could mean that systematic dissimilarities in training programs, social dynamics and organizational leadership could affect the estimation results. To account for this, we included random effects for each regiment. In addition, to account for individual heterogeneity beyond the observable covariates, we introduced individual random effects to all of our models. Individual specific effects are also effective
for controlling regimental differences, as each individual throughout our time window belongs to only one regiment.\footnote{Regimental or individual fixed effects cannot be applied in the between-subject experimental setting, as each regiment or individual is in the same experimental condition throughout the study.}

In specifications that include regiment and individual specific random effects, we found that the results for Equation 2 are quite similar. Moreover, they were qualitatively similar to the main results presented earlier. Users in the $PF$ condition showed a decline in run times of $-2.9\%$, in $SF$ of $-1.8\%$ and in the $SF\&PF$ condition of $-2.6\%$ (compared to the control condition). The results of Equation 2 with individual random effects are also qualitatively as well as quantitatively similar to those from the basic models without random effects.

**Matching Subjects across Conditions** Since it is plausible that individual differences may have persisted after randomization, we took the further step of applying matching to the treated participants, to ensure that our main results are robust. The approach of matching across experimental conditions is a recommended step even for randomized trials (Rubin [1973]). To this end, we used coarsened exact matching (CEM) with multiple treatment conditions as proposed in (Iacus et al. 2012) to match individuals in treatment conditions to those in the control group. The matching was performed on pre-experiment $BMI$ and run times $PastRunTimes$ with the $Control$ condition as the baseline. We replicated models from Equation 2 with matching. The controls were replaced with strata fixed effects indicating which strata the matched subjects belonged to. Each observation was weighed within strata as per CEM’s weighting procedure. Overall we found our results to be largely robust to matching. For Equation 2, a $2.1\%$ reduction in PFT Runtimes was found for $PF$, $1.8\%$ for social feedback ($SF$) and that of $2.3\%$ for the joint feedback condition. The absolute magnitudes of treatment effect were diminished (although qualitatively similar to the unmatched results) in the matched sample.

**Mechanisms and Heterogeneous Impacts**

*Usage Frequency as a Mediator*

<table>
<thead>
<tr>
<th></th>
<th>$log(RunTime_{i,t})$</th>
<th>$UF_{i,t}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2SLS Stage 1</td>
<td>2SLS Stage 2</td>
</tr>
<tr>
<td>$PF$</td>
<td>$15.083^{***}$</td>
<td>$-1.908^{**}$</td>
</tr>
<tr>
<td></td>
<td>(1.192)</td>
<td>(1.19)</td>
</tr>
<tr>
<td>$SF$</td>
<td>$7.415^{**}$</td>
<td>$-9.455^{***}$</td>
</tr>
<tr>
<td></td>
<td>(1.19)</td>
<td>(1.702)</td>
</tr>
<tr>
<td>$PF \times SF$</td>
<td>$0.0095^{***}$</td>
<td>$-1.019^{**}$</td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
<td>(0.534)</td>
</tr>
<tr>
<td>$BMI$</td>
<td>$0.9536^{***}$</td>
<td>$-16.984^{***}$</td>
</tr>
<tr>
<td></td>
<td>(0.01)</td>
<td>(3.214)</td>
</tr>
<tr>
<td>$log(PastRunTime)$</td>
<td>$-0.0019^{**}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td></td>
</tr>
<tr>
<td>Num. obs.</td>
<td>2473</td>
<td>2473</td>
</tr>
<tr>
<td>Adj. $R^2$ (full model)</td>
<td>0.9999</td>
<td>0.7833</td>
</tr>
<tr>
<td>Adj. $R^2$ (proj model)</td>
<td>0.9289</td>
<td>0.9302</td>
</tr>
<tr>
<td>Wave FE’s</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Anderson-Rubin Wald Test</td>
<td>43.29***</td>
<td>64.556</td>
</tr>
<tr>
<td>Cragg Donald F-Stat</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stock-Yogo Critical Value(5%)</td>
<td>13.91</td>
<td></td>
</tr>
<tr>
<td>Sargan Statistic</td>
<td>1.648, $PVal = 0.43$</td>
<td></td>
</tr>
</tbody>
</table>

Note: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$  

**Table 3. The impact of different forms of feedback on PFT runtimes**
Performance outcomes $\text{RunTime}_{i,t}$ are expected to be driven by the frequency of app usage in each experimental wave $t$. To assess this mediating link, we conducted a mediation analysis. As we had multiple treatments, we split the analysis considering $\text{PF}$ as the primary treatment and $\text{SF}$ as a covariate. This resulted in two models of mediation analysis (one with $\text{SF} = 0$ and another with $\text{SF} = 1$), with usage as the mediator of the treatment effect on $\log(\text{RunTime}_{i,t})$. We performed a parametric analysis with standard errors clustered by individual under Tingley et al.'s (2014) proposed implementation. We found that $\text{UF}$ (Usage Frequency) significantly mediates the treatment effect of $\text{SF}$ and $\text{PF}$.

Besides this test, we also performed an alternative analysis using an instrumental variable (IV) approach. Our IV approach relies on the fact that the random assignment of treatments would exogenously induce varying levels of app usage, which in turn affects running performance. While not perfect, the IV test offered some defense against the endogenous nature of app usage frequency when we were inferring the intermediate role that app usage plays in affecting running performance. The results of the IV analysis are reported in Table 3. Based on the Stock-Yogo cutoff values and the Sargan statistics, our model does not seem suffer from weak instrument issues (Stock and Yogo 2002) nor from over-identification (Sargan 1958). We found that all three treatment conditions significantly increased app usage behavior: $\text{PF}$ was the most effective in this regard followed by $\text{SF} \times \text{PF}$ and finally $\text{SF}$. In the second stage regression, we found that app usage was significantly correlated with run times, in that subjects with higher app usage ran faster than their counterparts, thereby augmenting the case for the mediating role played by app usage frequency.

**Temporal Effects on Usage**

Having seen that usage frequency is a mediating mechanism that underlies the treatment effect on running outcomes, it is of value to understand how app usage might be change over time. In particular, we were interested in understanding whether different app features influence usage outcomes over time. To do so, we utilized the models in Equations 2 replacing the DV with $\text{UF}_{i,t}$, the weekly frequency of app usage to report exercise behavior. We modeled weekly usage frequency $\text{UF}_{i,t}$ as a function of treatment dummies, covariates, and user’s tenure of app usage (in weeks) $t$ along with its interactions with treatment indicators. The quadratic tenure term $t^2$ and its interactions are also included to examine nonlinear treatment effects over time. The model was ran as a poisson regression.

We simulated predicted values under different treatments, using sample average values of pre-experimental PFT run times and BMI. We used our model to predict weekly usage outcomes for tenure values ranging from 0 to 74 weeks. The simulated results are plotted in Figure 4. The plot shows the rise of app usage across different treatment conditions at differing rates followed by an eventual decay. It should be noted that the $\text{SF}$ condition users peaked fastest (36 weeks) in terms of app usage but were also the fastest to decline. On the other hand, the $\text{PF}$ condition exhibited a reduction in slope but continue to increase through 70 weeks. This implied that performance feedback was the most effective in leading to sustained app usage, while $\text{SF}$ was least effective in sustaining long term usage (though it is helpful in boosting initial usage intensities).

**Usage Heterogeneity** Next, we explored how app usage frequency might vary across users. Knowledge of this aspect would help in the customization and refinement of app design. The vast variance in $\text{RunningTime}$ as seen in our study sample indicates that users have different levels of inherent physical fitness and athletic capabilities for running. At the same time, various app features might appeal differently to various individuals based on their running aptitude. For instance, a seasoned runner may appreciate social feedback and find it useful, while a weak runner can potentially be over-conscious of his/her poor ability when observed by peers. For such individuals, performance feedback may prove less daunting as it does not put the user under scrutiny of peers.

To test such potential differences in treatment effects by inherent running ability, we added an indicator $\text{FailedPastTest}$ to the model in Section to denote whether a subject failed or passed his latest PFT test prior

---

8 For instance, health conscious or athletic individuals may have still self select to use the app more even under the random assignment into various treatment groups
Social and Performance Feedback: Impact on Running

Figure 2. Change of App Usage with tenure across conditions

the experiment. This variable was also interacted with the treatments and time vectors. While the overall trends are similar, users that had poorer physical fitness exhibited greater usage frequency over time in experimental groups that only have PF feature, compared to those with social features. Interestingly, the joint condition worked best in increasing app usage for the strong runners. We believe that this might be driven by a difference in the amount of social affirmation and cues received across fast and slow runners. To investigate this possibility, we contrasted the change in social feedback received over a wave between these two user groups. We found that fast runners (i.e., users who passed the test previously) solicited a significantly greater increase in social feedback, relative to the slow runners (i.e., users who failed the test previously) \((t = -2.83, P_{\text{val}} < 0.01, \Delta \mu = 0.992)\).

Discussion

To our knowledge, this is one of the first studies that assessed the causal impact of app features on actual exercise performance, over an extended period of time (18 months). Using a field experiment executed in the military setting, we explored the impact of different app features, in the form of different feedback types, on run times. Our analysis revealed that performance feedback and social feedback positively influenced both running performance and exercise frequency, affirming the incremental effect of app-based feedback in enhancing running performance, over and beyond the basic tracking features of running apps. The degree of improvements brought about by the social feedback is found to be a function of the quantity of feedback received. It was also revealed that the greatest improvements to running performance come under joint feedback conditions with high social feedback frequency. In view of this result, we further performed a post-hoc survey, in which we gauged the servicemen’s perceptions of social presence across different experimental groups. We adopted three items from Gefen and Straub (2004) to measure subject’s perceived social presence. Individuals in social conditions displayed higher levels of perceived social presence compared to participants who were not in social conditions. Responses from this survey offer psychometric support that social facilitation effects are at work in the social feedback based conditions.

On top of these main effects, we also attempt to shed light on the underlying drivers of the treatment effects. App usage for exercise tracking appears to be a major pathway by which these design features influence running performance. These app features were also found to affect the usage behaviors directly, and these effects are rather nuanced. While performance feedback generates sustained advantages in usage frequency, social feedback has a curvilinear impact on app usage over time. However, we note that social feedback exhibits the fastest usage improvements. There are multiple possible reasons for the curvilinear impact, the chief among which is the decline in social feedback given by peers over time. We explore this possibility by comparing the average number of messages received in each experimental wave, and found that users in the SF condition indeed received lesser feedback with each passing wave.

Finally, we explored heterogeneous effects among users based on their innate physical fitness. We found that performance feedback is most helpful in enhancing the usage behaviors of weak runners, while the
combination of social and performance feedback is most ideal in inducing usage behavior for strong runners. This finding along with the discovery of nuanced relationship bears several theoretical and practical implications, which we discuss next.

**Theoretical Implications**

Our work contributes to an emerging body of IS literature that documents the effectiveness of app features, as well the broader set of work that examines the efficacy of mHealth apps in improving physical outcomes. We theorized that a fitness app can be effective in improving exercise performance through features that offer objective feedback that reinforces goal achievement, and features that provide social feedback that encourages user effort. The finding of a complementary effect also suggests that user motivation can be further enhanced when disparate, yet supportive modes of feedback are provided. This adds to the overarching understanding of human motivation (Atkinson 1964), especially to that of motivation in exercise and sports (Frederick and Ryan 1993).

We should also note that social-based features can exert opposite effects on different users. As proposed by Zajonc (1965), social facilitation effects can either heighten or dampen user motivation and performance, depending on the task type. Arousal due to social presence would enhance one’s confidence and induce the individual to perform better, provided that the task at hand is something that the individual is relatively competent at. However, the presence of others can cause individuals to experience heightened anxiety and evaluation apprehension if he/she struggles with the focal task, leading to reduced performance (Zajonc 1965). Our study results appear to agree with the applied predictions of this theory.

**Practical Implications**

Our study has several takeaways for application companies, app designers, users, fitness coaches, and sports clinicians. First, a direct implication of our causal-based finding is that app companies are now able to advertise fitness apps with performance and social feedback as being effective digital tools for enhancing physical performance. With experimental evidence of the efficacy of such apps, fitness coaches and sports clinicians can now prescribe the use of fitness apps with such features to users as a means of improving their training outcomes. Second, our findings also validate the effectiveness of social features in these apps. Despite social based features being effective in spurring initial adoption and usage, app designers need to be cautious in utilizing such features, as our results reveal that social features alone do not necessarily lead to sustainable usage improvements. Social features should be complimented with goal based features such as performance feedback, so that usage levels can be sustained over time. Linking other social networks, recruitment of friends, and utilizing strategies to combat the drop in social feedback are solutions that app designers should consider in their work.

Social features are also shown to have varying effects across users. In particular, the finding of a detrimental effect on weak runners highlights to app companies and designers that the design or inclusion of social features need to be reconsidered, such that negative effects of evaluation apprehension do not set in. A possible strategy, for instance, could be to design the social features such that each user is paired with peers who are on par with their fitness level, so that weak performers would not need to worry about judgment. Another strategy would be to allow users to decide how much information about their exercise to reveal on the app. To illustrate, a user can choose to simply reveal that he/she has gone for a run on a certain day, and not show details of the speed and/or the distance covered. Under such a presentation formation, weak runners would be less concerned with possible negative evaluation from peers. In such cases, it is also important that the recommendation of social features are not delayed for too long, as they seem to deliver fastest gains in the early stage usage.

Finally, our study finds that only the strong runners appreciated both the performance and social feedback features, in that these users exhibited the greatest app usage frequency when placed in the joint condition. The effectiveness of the joint condition for strong runners is telling of the users’ needs and preferences for a fitness app. The presence of advanced features tends to spur greater app usage among strong runners.
From a business perspective, while it might seem reasonable for app companies to incorporate new app features that other developers are introducing, such a strategy might bring value to only a subset of users, i.e., the highly invested users. Thus, an app developing company should assess the relative size of its “pro” users to its regular users, before deciding whether to invest in the development and introduction of new app features.

References


Plow, M., and Golding, M. 2017. “Using mHealth Technology in a Self-Management Intervention to Promote Physical Activity Among Adults With Chronic Disabling Conditions: Randomized Controlled Trial,” *JMIR mHealth and uHealth* (5:12).


