How Gamification Affects Physical Activity: Large-scale Analysis of Walking Challenges in a Mobile Application

Ali Shameli
Stanford University
shameli@stanford.edu

Tim Althoff
Stanford University
althoff@cs.stanford.edu

Amin Saberi
Stanford University
saberi@stanford.edu

Jure Leskovec
Stanford University
jure@cs.stanford.edu

ABSTRACT

Gamification represents an effective way to incentivize user behavior across a number of computing applications. However, despite the fact that physical activity is essential for a healthy lifestyle, surprisingly little is known about how gamification and in particular competitions shape human physical activity.

Here we study how competitions affect physical activity. We focus on walking challenges in a mobile activity tracking application where multiple users compete over who takes the most steps over a predefined number of days. We synthesize our findings in a series of game and app design implications. In particular, we analyze nearly 2,500 physical activity competitions over a period of one year capturing more than 800,000 person days of activity tracking. We observe that during walking competitions, the average user increases physical activity by 23%. Furthermore, there are large increases in activity for both men and women across all ages, and weight status, and even for users that were previously fairly inactive. We also find that the composition of participants greatly affects the dynamics of the game. In particular, if highly unequal participants get matched to each other, then competition suffers and the overall effect on the physical activity drops significantly. Furthermore, competitions with an equal mix of both men and women are more effective in increasing the level of activities. We leverage these insights to develop a statistical model to predict whether or not a competition will be particularly engaging with significant accuracy. Our models can serve as a guideline to help design more engaging competitions that lead to most beneficial behavioral changes.

1. INTRODUCTION

Physical activity is critical to human health [45]. People who are physically active tend to live longer, have lower risk of several diseases including heart disease, stroke, Type 2 diabetes, depression, and some types of cancer, and are more likely to maintain a healthy weight (*e.g.*, [28, 37, 5]). However, only 21% of US adults meet official physical activity guidelines [11, 30] (at least 150 minutes a week of physical activity for adults), and less than 30% of US high school students get at least 60 minutes of physical activity every day [11]. It is estimated that physical inactivity contributes to 5.3 million deaths per year worldwide [24] and that it is responsible for a worldwide economic burden of $67.5 billion through health-care expenditure and productivity losses [15].

Given huge potential to improve public health, many interventions and small-scale studies have been designed towards increasing physical activity across the population (*e.g.*, [16, 17, 26, 31, 32, 33]). Unfortunately, many of these interventions are deemed either ineffective [17, 33] or are limited in that they only reach small populations [16, 26]. Recently, however, gamification techniques have become widely adopted and have been very impactful in obtaining behavioral outcomes [22]. Successful examples for incentivizing physical activity through so-called *exergames* [21, 35, 38] include in-game avatars [25] and geo-centric games such as Pokémon Go [6]. However, basic gamification mechanisms such as competitions and challenges have been relatively poorly explored and understood. While competitiveness is found to be associated with greater enjoyment [19], there have been no quantitative studies whether and how such competitions affect physical activity. Given the proliferation of mobile devices and health and activity tracking applications, effective and engaging competitions that increase physical activity have a huge potential to achieve population-wide improvements in public health and decrease in risk of various chronic diseases.

Here we study how various game design elements used by mobile health apps encourage exercise, fitness, and essentially weight loss. We analyze the effect of competitions on increasing the level of physical activity of participants. We study user physical behavior as captured by the Azumio Argus activity tracking app. The application allows users to create and engage in ad-hoc games that last from one to seven days and include an arbitrary number of participants. Participants then compete over who takes the highest total number of steps over the predefined duration of the competition. The dataset obtained from Argus contains nearly 2,500 physical activity competitions over a period of one year capturing over 800,000 person days of in-competition activity tracking. For each user we have a record of their daily physical activity, competition participation, as well as their demographic data (gender, age, height, and weight). This constitutes the largest studied dataset of physical activity competitions to date.

We analyze how participation in the game impacts the activity of participants during the time of the competition compared with their baseline activity levels. We find that during walking competitions, the average user increases their physical activity by 23%. Furthermore, we show that there are large increases in activity for both men and women across all ages, weight status, and baseline activity levels. In fact, we find the largest increases for users that were previously fairly inactive who exhibit an average increase of more than 2,500 steps per day throughout the challenge. Increases

WWW'17 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3054172

of this magnitude – if sustained over time – could lead to significant improvements in participants' health outcomes.

Then, we turn our attention to quantifying how much effort it takes to win a competition and find that winners increase their activity by 40-60% while the last-ranked users are on average less active than they were before. We also observe that the winner's effort increases in competitions with more participants.

We also find that the composition of participants greatly affects the dynamics of the competition leading to important design implications for exergames and mobile health applications:

1. Competitions lead to increases in physical activity and constitute a viable design element able to reach a broad user base across a wide variety of user demographics.
2. Competing participants should have similar pre-competition activity levels. Otherwise the effect of the competition on physical activity drops significantly.
3. Competitions should have a balanced mix of both men and women.
4. Competitions should ideally include some participants who have previously increased their activity in response to competitions to encourage the other participants.

We leverage these insights in a statistical model that predicts whether or not a competition will be particularly engaging to the participants. Our model can serve as a guideline to help group participants into competitions that are more competitive and thus lead to highest behavioral changes. Our approach can be potentially used across a variety of mobile health applications and games to recommend evenly-matched competitions to users which are most likely to be more active.

2. RELATED WORK

Next we survey related work and discuss our work in context.

Physical activity. The link between physical activity and improved health outcomes has been well-established (*e.g.*, [15, 24, 28, 37, 45, 5]). At the same time, only a small fraction of people in developed countries meet official physical activity guidelines [11, 30]. While, many interventions are aimed at increasing physical activity (*e.g.*, [16, 17, 26, 31, 32, 33]), many of them seem ineffective [17, 33] or were only reaching already active populations instead [16, 26].

Measurement of physical activity. Consumer wearable devices and smart phones are becoming more prevalent in the general population and could enable a better understanding of real-world health behaviors and physical activity and how to best support and encourage healthier behaviors [23, 34, 3]. However, few research studies to date have harnessed data obtained from consumer wearables to study physical activity [4, 6]. Medical studies have examined accelerometer-defined activity (*e.g.*, [40, 42]), but much of the social media research related to human health has relied on self-reports and proxy measures (*e.g.*, [13, 14, 27]), which have been found to be severely biased [41]. In contrast, we use objective physical activity measurements from smart phone accelerometers.

Incentivizing user behavior. Studies have found that use of pedometers and activity trackers for self-monitoring can help increase activity [39, 44] but other studies have reported mixed results [43]. Beyond enabling self-monitoring, encouraging additional activity through reminders led to increased activity only for the first week after the intervention and did not lead to any significant changes after six weeks in a randomized controlled trial [43]. However, in online domains gamification has been successful in changing user behavior [22, 36]. For example, badges increase engagement in question answering sites [7], online courses [8], and pro-social behaviors [2]. To encourage healthy behavior, researchers have studied the design of "exergames" [21, 35, 38], video games combined with exercise activity [25], and location-based games where game play progresses through the physical environment [6, 9]. Furthermore, social networks can also modify human behavior through peer influence. For example, researchers have highlighted the importance of facilitating social influence to encourage more exercise [12] and have found that sharing exercise activity through social networks has positive long-term effects on physical activity levels [4]. Furthermore, researchers have studied social interactions to better understand how people could be most supportive of others [1, 2].

Prior research also studied competitions and competitive behavior. Competitions can improve behavioral outcomes, for example in the running speed of children in short-distance races [20]. Men are more likely to embrace competitive formats than women [29], and while competitiveness is associated with greater enjoyment [19], there have been no quantitative studies of how competitions affect physical activity.

This work. Our work extends the existing literature on incentivizing healthy user behavior and exergame design implications by studying effects of competitions on physical activity. Unlike badges and simple activity tracking, competitions allow users to compete directly with each other. We use a large dataset of online competitions within a mobile activity tracking application in conjunction with objective measures of physical activity based on smart phone accelerometers. Our work has implication for the large number of mobile and web health applications using competitions to improve user engagement.

3. DATASET DESCRIPTION

We use a dataset of competitions within the Argus smartphone app by Azumio which allows users to track their daily activities. Competitions run for 1, 3, 5 or 7 days and can have any number of participants (who may or may not know each other outside the application). In this paper, we focus on the longest competitions running over seven days from Monday through Sunday. This ensures that each competition includes the exact same number of each weekday (one) and that any findings are not confounded by differences in weekday versus weekend activity. Furthermore, we restrict analysis to competitions with at least three participants. We use a dataset of 3,637 users in 2,432 competitions satisfying these constraints. These competitions occurred over a time period of 10 months and included a total of 535 million steps over 70,413 person days. All participants had used the activity tracking app prior to the start of their first competitions. Therefore, any increases in activity during competition periods are not an effect of self-monitoring. Table 1 further summarizes the dataset and shows that the distribution of age, gender, and weight is fairly representative of the overall population in many developed countries. For example, the median age is 34 years (official estimate in United States is 37 years) and the data is evenly split between males and females. Furthermore, over 34.4% of users are overweight and 18.8% of users are obese highlighting that not only healthy users participate in the walking competitions. Users that have participated in at least one competition take about 6164 daily steps on average outside of competitions. Compare this to 5926 for users that have never participated in a competition before. This shows that there is a slight selection effect; that is, competitions seem particularly attractive to those users with slightly elevated activity levels.

Physical activity levels are quantified using the number of steps in each day objectively measured through smartphone accelerom-

# 7 day competitions w. at least 3 particip.	2,432
# total users in competitions	3,637
Observation period	April 2015 – Jan. 2016
# days of steps tracking (within competition)	70,413
# days of steps tracking (outside competition)	817,666
# total steps tracked (inside competition)	535 million
Median age	34 years
% users female	51%
% underweight (BMI < 18.5)	3.1%
% normal weight (18.5 $\leq$ BMI < 25)	43.7%
% overweight (25 $\leq$ BMI < 30)	34.4%
% obese (30 $\leq$ BMI)	18.8%
Avg. daily steps outside competition for competition users	6,164
Avg. daily steps outside competition for non-competition users	5,926

Table 1: Dataset statistics. BMI refers to body mass index.

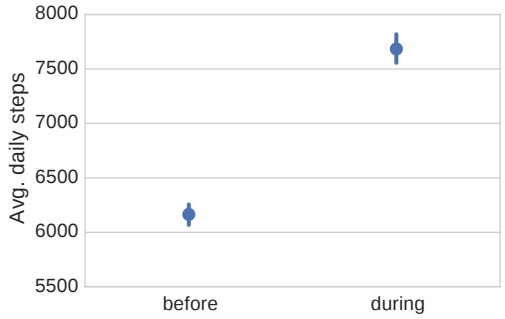


Figure 1: Average number of steps taken in the past when not participating in a competition vs. the average number of steps taken in a competition. We find that users take significantly more steps during competitions. Error bars in all figures correspond to 95% confidence intervals of the corresponding mean estimates.

eters (as done in many research studies [4, 6, 10, 42]). Objective measures are critical as commonly used self-reports of physical activity can be extremely biased with Tucker et al. reporting overestimates of up to 700% [41].

Our dataset is the largest dataset of physical activity competitions studied to date. It has three key properties: (1) It covers a large number of competitions covering a diverse population of participants in terms of age, gender, weight status, and activity level. (2) The activity levels of the users are measured objectively. (3) The activity levels of individuals are measured before and during the competition offering a baseline level for measuring the effect of competition. Therefore, this dataset uniquely enables the study of how online competitions affects offline physical activity.

4. DOES COMPETITION INCREASE ACTIVITY?

The first question we set to answer is whether competitions tend to increase user activity. Given the heterogeneity across users, we compare the physical activity of a user to his or her own baseline activity level. In other words, we measure the activity as the average number of steps per day over the duration of the competition, and compare that to the average daily activity of the user while not participating in a competition.

Competitions lead to increased physical activity of participants. We make several interesting observations (Figure 1). First, we notice that the average daily number of steps of a person that has ever participated in a competition is 6,164, which is about 200 steps more than an average user. This indicates that there is some selection effect and that more active users tend to participate in physical activity competitions. Second, we observe that during the competition the overall average activity increases to more than 7,500 steps, which is a 1,400 steps per day increase over the out-of-competition activity baseline. This means that participating in a competition leads to 23% average increase in the physical activity.

Discussion. To check for the robustness of our finding we also analyzed the differences between the group of users participaing in the competitions against those that do not participate in competitions. We found that both groups are very similar in terms of age, gender, and weight status. Furthermore, we also found that non-competition users were only slighly less active than competition users outside of competitions (5926 vs. 6164 average daily steps). On top of this we observed that the activity increase is about the same for users who start their own competitions and for users who accept an invitation to join an existing competition. All these findings suggest that insights derived from the above analyses are robust and generalize across user groups.

4.1 Who is changing how much?

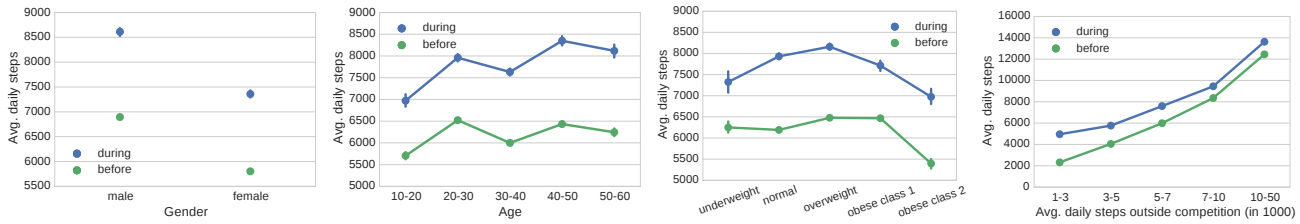
Our analysis showed that participants tend to increase their activity during competitions, which demonstrates that competitions may be an effective way for people to increase their physical activity. However, note that this increase is observed in average over all participants. The person who wins the competition may increase their activity significantly while the person who is last in the competition may not increase their activity at all. Furthermore, demographic indicators, like gender, age, and BMI (body mass index) affect physical activity and thus it is not clear how these factors modulate in-competition activity of individual participants.

In order to understand which users increase their activity the most during the competitions we measure the activity before and during the competition for different demographic groups. Figure 2 shows separate plots for gender, age, BMI, and baseline activity level. We make several observations.

Gender. Consistent with previous literature [10], Figure 2a shows that men tend to be more active than women on average. This is also true during the competitions where men increase their activity from just below 6,900 steps per day to about 8,500 steps per day. Similarly, women increase their activity from 5,800 steps per day to about 7,100 steps per day. Interestingly, in both cases we observe the same relative increase in activity. Both men and women tend to increase their activity by 23% during competitions.

Age. Examining physical activity levels as a function of age (Figure 2b), we find that the difference of activity levels across different age groups is rather minimal – people below age 20 are the least active group with 5,700 steps per day (outside competitions), while the age 20-30 group is the most active with 6,500 steps per day. Furthermore, we do not observe an expected decrease in physical activity as participants get older [10]. Surprisingly, even the group of 50-60 year old users take nearly 6,200 steps per day. We attribute this observation to a selection effect as a small but physically very active fraction of this age group may be participating in competitions. Perhaps more interestingly we observe that regardless of the age group, physical activity during competitions tends to increase for about 1,400 steps. The increase in physical activity is robust. For example, for the age group 10-20 the increase is 1,300 steps per day (22%), and then increases nearly linearly so that the age group 50-60 exhibits an increase of 1,800 steps per day (28%).

Body mass index (BMI). Next, we examine how competitions affect people with different body mass index (BMI). Using self-



(a) Gender **(b) Age** **(c) BMI** **(d) Previous Activity**

Figure 2: The average number of steps taken by users categorized by gender, age, BMI, and previous activity level. The green points represent the average number of steps when not participating in a competition and the blue points represent the average number of steps during competitions.

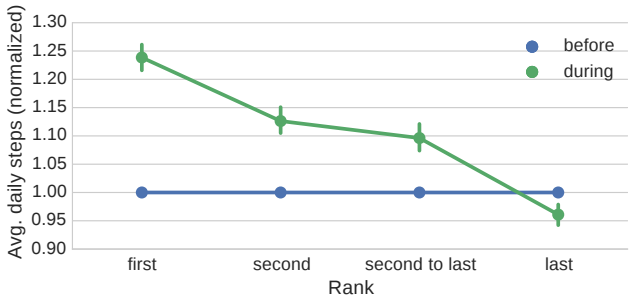


Figure 3: Avg. relative number of steps taken one week before and during a competition for the two top and bottom users.

reported height and weight we compute each participant's BMI and group them into five groups: underweight (15-18.5 BMI), normal (18.5-25 BMI), overweight (25-30 BMI), and then two groups of obese people (30-35 BMI and 35-40 BMI, respectively). In Figure 2c we observe that baseline physical activity is relatively constant for the first four groups and lower for only the most obese group (5,800 vs. 6,300 steps per day). Interestingly, however, the increase in physical activity is consistent across all the BMI groups. Underweight people increase their activity during competitions the least (1,000 steps, 18%), while the increase is the highest for normal-weight and overweight people (1,700 steps, 28%).

Baseline activity level. Last, we examine how previous activity modulates activity inside competitions. We find that competitions lead to increased physical activity regardless of baseline activity rate (Figure 2d). Moreover, low activity people tend to be most affected by the competitions—they increase their activity the most (to 4,700 steps per day; 103% increase) and both the total increase in activity as well as the relative increase tend to decay with the activity level before competitions. For example, the increase due to competitions is only 1,000 steps per day (9%) for people with over 10,000 average daily steps.

To conclude, we observe that competitions have significant and measurable effect on the physical activity of participants. Regardless of the gender, age group, or the body mass index, we observe a robust increase of about 1,400 steps (about 23%) per day in the physical activity of participants. If sustained, large increases like this would have a significant positive effect on the health of the participants [15, 18, 24, 28, 37, 45].

5. WHAT DOES IT TAKE TO WIN A COMPETITION?

In the previous section we observed remarkably strong effect of competitions on physical activity of participants. While the increase is stable across gender, age and weight groups, in the end there can only be one winner of the competition. So, in this section we investigate how much activity is needed in order to win a competition.

It takes 25% more steps per day to win. We analyze all competitions with duration of seven days that have at least three participants. This considers the following final placements: first, second, second to last, and last. We measure how much does a participant need to increase their physical activity (measured in the number of steps per day) compared to the baseline activity which is their average number of steps one week before the competion. Figure 3 shows the relative change in the number of steps as a function of the final placement of the participant.

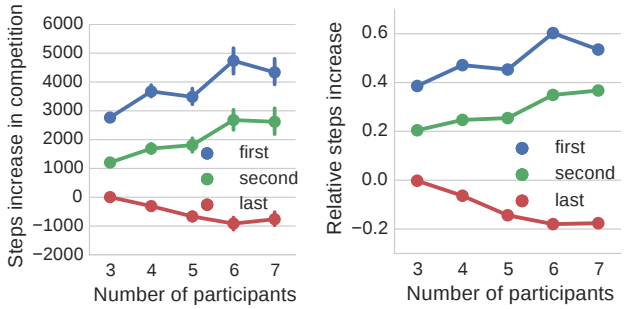
We observe that the winner of the competition increases his or her activity for 25% over their baseline activity, while the second person increases it for only 13%, second to last person still increases their activity over the baseline for about 10%, while the last person actually drops their activity below their baseline activity (4% drop). We conclude that winners increase their activity the most, followed by the mid-placed people, while the last person slightly decrease their activity. Regardless of this, the overall average activity across all participants in the competition is still higher than outside the competition.

Winner's effort increases as there are more competitors. We also examine how does the winner's effort increase as there are more competitors in a competition. Figure 4a plots the absolute number of steps per day of the winner as a function of participants in the competition, while Figure 4b plots the increase in the number of steps when compared to the out-of-the-competition baseline. Each separate curve plots the activity of the top, second, and last placed participants.

In Figure 4a we observe that the absolute activity of the winner and the second-placed participant increase with the competition size while the activity of the last-placed person actually decreases. Examining data in Figure 4b we find that with each additional participant in the competition the final winner of the competition increases their activity for 420 steps, while the effect on the second-placed participant is about 100 steps smaller (320 steps per addit. participant).

Activity over time. Last, we also examine how participant activity changes over time. We examine only competitions with exactly 5 participants to control for competition size. After every day of the competition we compute the current position of every participant and ask: given the participant was ranked k "yesterday", how many steps are they going to do today?

Figure 5 plots the average daily steps "today" for a person who was "yesterday" at current position k (blue line). We compare the activity level with the average activity of the participant in competitions regardless of his/her position (green line). The difference



(a) Steps per day **(b) Steps delta**

Figure 4: (a) Average number of steps taken by users by the number of participants in the competition. (b) Average increase in steps compared to previous outside competition activity by the number of participants in the competition.

between the lines can be interpreted as the "boost" participant at current rank k gets because of their current position.

We observe that yesterday's leader always increases their activity on the next day beyond their average in-competition activity. Similarly, second placed person only slightly increases their activity. Participants at current positions 3 and 4 maintain their level of activity to be the same as their average in-competition activity. However, participant at rank 5 (last position) performs worse than expected. This is interesting as it seems to suggest that while today's leader keeps his/her next day activity above the baseline, the person at the last position drops their activity to the level even below their baseline.

6. WHAT MAKES A COMPETITION ENGAGING?

The focus of the last section was on the winner of the competition. In this section, we look at the dynamics of the competition more broadly. In particular, we focus on more engaging competitions in which there is a close race for the top position with participants changing rank position multiple times. We will observe that these competitions are more successful in increasing the overall activity of the participants. We will also identify the impact of other important factors such as the gender composition of the group as well as the inherent competitiveness of the participants.

Probability of winning. We start with the simplest possible model of competition as our null hypothesis. Consider the model in which all participants increase their activity level uniformly and at the same rate. In other words, they start with their baseline levels of activity, then increase it to the in-competition level, and then effectively keep that same activity level throughout the competition. If that were the case, then we would expect that the person who takes the most steps after the first day of the competition would finally also win the competition.

We examine this hypothesis in Figure 6a, where we quantify the probability that the leader on a given day finally wins the competition. Again, we examine 7 day competitions which all start on a Monday and end on a Sunday. We measure how often the current leader wins the overall competition. We observe that the leader after day 1 of the competition tends to win 58% of the cases. The probability then linearly increases up to 1.0, meaning that the current leader on day 7 always wins the competition (because the competition is finished).

So overall, in slightly more than half of the competitions, the early leader can maintain the lead throughout the competition. Next, we will look at the dynamics of the more interesting competitions in which there are multiple changes in the leaderboards position and the participants truly compete with each other.

The dynamics of close competitions. One way to quantify the competitiveness of a race is to measure the number of leaderboard changes (or swaps) as the competition unfolds. Here, Figure 6b plots the number of leaderboard changes as a function of the absolute difference between the most and the least active participant in the period before the competition. In other words, the figure shows how inequality in the baseline level of activity of participants affects the number of leaderboard swaps. We observe that when the difference in baseline activity is small, there are about 4 leaderboard changes in a competition. However, when this difference increases to 7-15 thousand steps per day, then the number of leaderboard changes drops but still remains non-trivial at around 2.5.

Another way to quantify how competitive a competition is to measure the final total difference in the number of steps between the top and the bottom placed participants (Figure 6c). We plot the relationship between the final step difference between the winner and the loser of the competition as a function of the absolute difference between the most and the least active participants in the period before the competition (blue line). The green line provides a null-model that quantifies the expected final difference. Here we simply take the baseline difference and multiply it with the duration of the competition. Our reasoning is that if, for example, in the baseline period the daily activity difference between the most and the least active person is say 1,000 steps, then the expected final difference after a 7 day competition would be 7,000 steps.

Examining Figure 6c we make two observations. First, as the inequality of the baseline physical activity of the participants increases, the final-step difference also increases. Second, we observe that for inequalities of less than 5,000 steps per day, the final difference is in fact larger than what would be expected under the null-model. This means that in tight competitions the winner strongly increases their activity level and the difference in activity grows larger. We also observe that when participants with very different baseline levels of activity compete, the final difference is in fact smaller than the baseline. This means that when uneven people are matched to compete, their level of activity actually gets closer to each other and the effect of the competition on the physical activity is smaller.

Competitiveness of participants. Lastly, we also examine how the past increases in the activity of participants determines their overall increase in the current competition. What happens if a competition is comprised of competitive participants – the ones who had increased their activity levels significantly in previous competitions? Do they increase the level of activity of the whole group? Do these competitive tendencies have a compounding effect and raise the overall activity level significantly?

We perform the following experiment. For every competition we compute the relative increase in activity in past competitions averaged over all participants. This gives us a sense of how competitive the participants are in the competition. We then also compute the average activity of participants in the current competition to understand how the competitiveness of participants affects the overall activity.

Figure 6d shows the results. We observe that as the average historic competitiveness of participants increases so does the overall competitiveness of the competition. In other words, if we put together people who tend to increase their activity by a lot, then they will also increase it in the current competition. Not surprisingly,

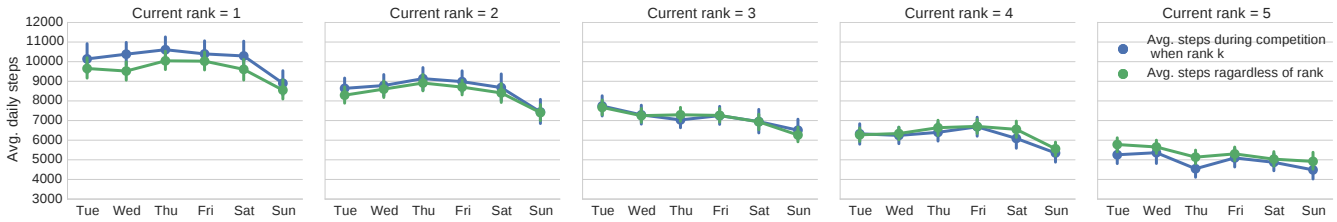
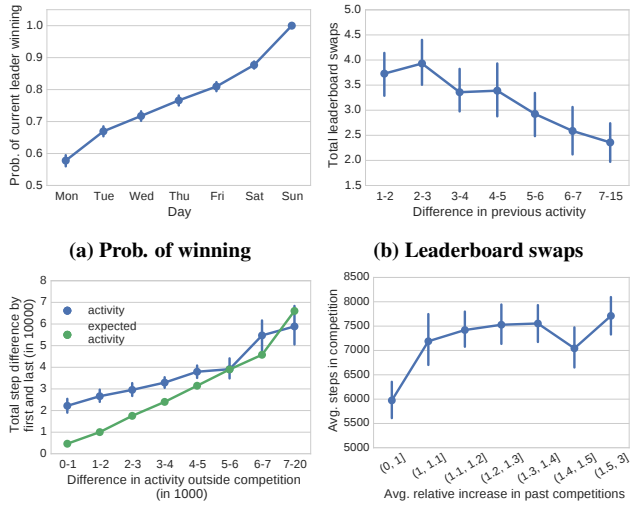


Figure 5: Average number of steps taken by users with specific current ranks in the competition over different days of the week for 7-day competitions with 5 participants.



(a) Prob. of winning **(b) Leaderboard swaps**

(c) Activity spread **(d) Activity increase**

Figure 6: Competition dynamics. (a) Probability of current leader winning as a function of time; (b) Number of leaderboard swaps as a function of the difference between the most and the least active participant; (c) Total step difference between the top and the bottom ranked participant at the end of the competition as a function of the difference between the most and the least active participant before the competition. (d) Average participant's steps per day in a competition vs. average relative increase in past competitions.

we also observe that when competitions are comprised of people who do not tend to increase their activity in competitions then the average activity is low.

More interestingly, the effect of the competitiveness of the participants on the average level of activity is sharp but then it quickly levels off. In particular, the overall activity tends to stabilize after the average past increase of activity (of participants when in competition) is over 10%. This means that as soon as competitions are comprised of participants who generally tend to increase their activity during competitions then the average steps per participant steps will reach up to 7,500 per day. If the competition is among extremely competitive participants, the average activity only slightly increases above that level.

The effect of gender diversity. Figure 7 plots the average number of steps taken by participants in the competition by across varying compositions of male and female participants in the competition. Note that competitions with a balanced number of female and male participants have the highest level of activity. This is surprising because men are on average more active than women (*i.e.*, they take more steps per day) both in our dataset as well as published estimates [10]. Nevertheless, competitions with all male participants have a smaller average than competitions in which the number of

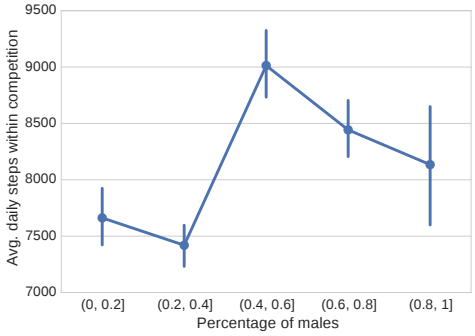


Figure 7: The average number of daily steps taken by participants as a function of the fraction of male competitors.

men and women is close to each other. This observation suggests that gender diversity can lead to more active and beneficial competitions.

Summary. We observe that the composition of participants greatly affects the dynamics of the competition. In particular, if highly unequal participants get matched to compete, then the competition suffers and the overall effect of the competition on the physical activity drops significantly. Second, competitions with a balanced gender ratio are more effective in increasing the level of activities. Lastly, the effect of competitiveness of participants as defined by their performance in the previous competitions has a sharp impact on the level of activity but its impact quickly levels off. With these insights in mind, the next section takes on the task of predicting such dynamics to serve as a guideline for the formation of interesting and engaging competitions.

7. PREDICTING ENGAGING CHALLENGES

In the previous sections we studied the factors of engaging competitions – those that lead to large increases of activity, close races between the first and the last ranked user, and competitions with many changes in the leaderboard ranking. This section builds on those insights and builds a series of models to predict which competition will be particularly engaging given only information available before the competition begins. We demonstrate that while there is large variability in the engagement of competitions, the factors described in this work allow to predict competition engagement with significant accuracy.

Predicted outcomes. We formulate the prediction task as a binary prediction of whether or not a competition will be engaging based on three different outcomes:

- ΔACTIVITY: The average relative increase of activity during the competition compared to activity before the competition across all participating users. For the baseline activity before the competition we exclude any activity that happened during previous competitions. Ideal engaging competitions would lead to large increases in activity across all users.

Model	ΔACTIVITY	FIRST-LAST	RANK SWAPS
Random	0.500	0.500	0.500
Logistic Regression	0.698	0.749	0.643
Gradient Boosted Trees	0.719	0.742	0.611

Table 2: Prediction performance of several models predicting three outcomes of engaging competitions. Performance values correspond using the area under the ROC curve.

	Model	ΔACTIVITY	FIRST-LAST	RANK SWAPS
1	None	0.500	0.500	0.500
2	# participants	0.507	0.682	0.500
3	User demographics	0.645	0.668	0.622
4	User experience	0.584	0.689	0.599
5	Prev. activity (outside)	0.668	0.707	0.609
6	Prev. activity (during)	0.576	0.751	0.606
7	Prev. increases (during)	0.717	0.691	0.641
8	All features	0.719	0.742	0.611

Table 3: Prediction performance of Gradient Boosted Tree models using different feature sets predicting three outcomes of engaging competitions. Performance values correspond using the area under the ROC curve.

- FIRST-LAST: The absolute difference in total number of steps between the first and the last ranked user over the 7 day competition. Engaging competitions are close races where the first and last user are not too far apart.
- RANK SWAPS: The number of total changes in the leaderboard over the time of the competition measured as the minimum number of inversions in the day-to-day rankings. Engaging competitions are competitions where participants are competing in a tight race leading to many changes in the leaderboard over the course of the competition.

Each of these outcomes is a continuous variable which we transform into a (approximately) balanced binary prediction problem by splitting at the median value (ΔACTIVITY 1.199 factor of increase, FIRST-LAST 37,421 total steps, RANK SWAPS 4 swaps).

Data and methods. We use the dataset of 7 day competitions and at least three participants (N=2,432).We use 75% for training and 25% for testing at random. Area under the ROC curve is used as a measure of predictive performance on the test set. We report performance for Gradient Boosted Tree and L_1 penalized logistic regression models and optimize number of trees, tree depth, learning rate, and regularization parameter through 5-fold cross-validation on the training data.

Features used for learning. We featurize the state before the competition as follows. In all cases we only use data from before the start of the competition:

- **Participants:** Number of participants in competition.
- **User demographics:** We use the number and fraction of male users and the number and fraction of obese users (BMI>30) as well as the median age of all participants.
- **User experience:** We use number and fraction of users who have participated in a competition before and the number and fraction of users who have won a competition before. Furtermore we use summary statistics (further explained below) of the distribution over how many times users have participated in a competition or won a competition.
- **Previous outside-competition activity:** We use summary statistics of the distribution over number of steps taken in previously outside of any competitions by all the participants.
- **Previous in-competition activity:** We use summary statistics of the distribution over number of steps taken in previous competitions by all the participants.
- **Previous increases in competition activity:** We measure how much each user has increased their daily number of steps in previous competitions (relatively and absolutely). We then use summary statistics over this distribution of increases across all participants.

As summary statistics we use mean, standard deviation, max, min, second largest, second smallest, difference between max and min, difference between max and second largest, difference between second smallest and min, and the number/fraction of users within 0.5, 1.0, and 2.0 standard deviations of the mean.

Predictive performance. Prediction accuracies using the ROC AUC measure are shown in Table 2. Overall, we observe encouraging predictive performance. We find that Logistic Regression models and Gradient Boosted Trees perform similarly well after optimization of parameters through cross validation. Predicting ΔACTIVITY yields a performance of 0.72 ROC AUC. The difference between the first and last user in total steps yields 0.75 ROC AUC. Predicting whether the competition will see many leaderboard swaps is more challenging and the model achieves 0.64 ROC AUC.

Overall, we show that it is able to predict which competition will be engaging before it start with ROC AUC scores between 0.64-0.75.

Factors affecting predictability. Next, we study which factors are particularly helpful in predicting the three outcomes. This is relevant to application settings where not all features might be available (*e.g.*, missing knowledge about participants previous activity level, previous in-competition activity, or missing information on user demographics). We investigate performance of the individual feature sets described above. We report prediction accuracies for Gradient Boosted Tree models as they performed slightly better than than Logistic Regression models, particularly on small sets of features. The predictive performance for the individual feature sets is reported in Table 3.

We observe that knowing the number of participants (Model 2) is only useful for predicting the final difference between the first and last ranked users but not for the other two measures. This is likely due to the effects observed in Figure 4 which showed that the difference between the first and last ranked users increases in larger competitions. Knowing just basic user demographics (age, gender, and obesity status) allows to predict all three outcomes with significant accuracy (Model 3). In particular, the number of RANK SWAPS can be predicted with ROC AUC of 0.622 which is as good as using all features (Table 2).

Past experience with competitions of all users (Model 4) is also predictive of engaging competitions. In particular, it allows good predictive accuracies for predicting FIRST-LAST (0.689 ROC AUC) and relatively good performance for predicting RANK SWAPS (0.599 ROC AUC) close to the the best performing model.

Knowing the exact previous activity levels of the users – both outside (Model 5) and inside of competitions (Model 6) – performs even better. Outside of competition activity is a good indicator for the final difference between the first and last ranked users (Figure 6c) with a prediction performance of 70.7% ROC AUC. Prediction of FIRST-LAST improves even more when knowing previous inside competition activity levels with 75.1% ROC AUC (best performing model for FIRST-LAST).

Baseline activity levels outside of competitions and activity levels during previous competitions allows to calculate how much each

user increased (or decreased) their activity in response to previous competitions (Model 7). This allows for high predictive performance for both ΔACTIVITY (0.717 ROC AUC; close to best performing model) and RANK SWAPS (0.641 ROC AUC; best performing model).

As expected, using all combined features allow for high predictive performance (Model 8). However, it is interesting to not that for FIRST-LAST and RANK-SWAPS the model performs slightly worse compared to just using the best subset of features.

Implications. The prediction results are encouraging in multiple ways. First, we can predict key outcomes capturing various aspects of engaging competitions with significant accuracy. The proposed models could be used when recommending which people to add to a competition to optimize the likelihood of high engagement. Second, we demonstrate that having access to only small subsets of the features studied in this work still allows for models achieving similarly high accuracy. This has direct implications for the use of competition features in practice. Many mobile apps and websites use, but not yet optimize, competitions to engage with their users. Given our results, they could use for example basic demographic information or outside competition activity levels to create particularly engaging competitions. We also note that during competition behavior is not necessary for high predictive performance (*e.g.*, Models 3 and 5 in Table 3). This partially alleviates potential cold start problems when creating new competition-based applications.

8. CONCLUSION

The focus of this paper is on how competition incentivizes users to increase their physical activity levels. We analyzed a large dataset gathered by a smartphone activity tracking app measuring the total number of daily steps taken by a user. We studied games where multiple users compete over several days with the goal of achieving the highest total number of steps.

We measured the effect of competitions on user behavior and showed that in general, users increase their activity level by 23% when participating in a competition. We also studied the design elements contributing to an engaging competition and identified the importance of factors like age, gender, BMI, diversity, and prior activity levels. We found that the group compositions strongly affects the dynamics of the competition, which leads to important design implications for exergames and mobile health applications:

1. Competitions lead to increases in physical activity and constitute a viable design element able to reach a broad user base across a wide variety of user demographics.
2. Competing participants should have similar pre-competition activity levels. Otherwise the effect of the competition on physical activity drops significantly.
3. Competitions should have a balanced mix of both men and women.
4. Competitions should ideally include some participants who have previously increased their activity in response to competitions to encourage the other participants.

Finally, we leveraged these insights in a statistical model that predicts how effective and engaging a given competition is going to be. Such models can lead to designing more evenly-matched competitions that are most likely to help users increase their daily activity. In summary, this work enhances the understanding of effective mechanisms that can potentially be used to engage people in healthier behaviors.

Acknowledgments. We thank Azumio for providing us with anonymized activity data that enabled this study. This research has been supported in part by NSF IIS-1149837, NIH BD2K Mobilize Center, ARO MURI, DARPA SIMPLEX and NGS2, Stanford Data Science Initiative, Lightspeed, SAP, Chan Zuckerberg Biohub, and Tencent.

9. REFERENCES

[1] T. Althoff, K. Clark, and J. Leskovec. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *TACL*, 2016.

[2] T. Althoff, C. Danescu-Niculescu-Mizil, and D. Jurafsky. How to ask for a favor: A case study on the success of altruistic requests. In *ICWSM*, 2014.

[3] T. Althoff, E. Horvitz, R. W. White, and J. Zeitzer. Harnessing the web for population-scale physiological sensing: A case study of sleep and performance. In *WWW*, 2017.

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

[5] T. Althoff, R. Sosic, J. L. Hicks, A. C. King, S. L. Delp, and J. Leskovec. Quantifying dose response relationships between physical activity and health using propensity scores. In *NIPS ML4H*, 2016.

[6] T. Althoff, R. W. White, and E. Horvitz. Influence of Pokémon Go on physical activity: Study and implications. *J Med Internet Res*, 18(12):e315, Dec 2016.

[7] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec. Steering user behavior with badges. In *WWW*, 2013.

[8] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec. Engaging with massive online courses. In *WWW*, 2014.

[9] N. M. Avouris and N. Yiannoutsou. A review of mobile location-based games for learning across physical and virtual spaces. *J. UCS*, 18(15):2120–2142, 2012.

[10] D. R. Bassett, H. R. Wyatt, H. Thompson, J. C. Peters, and J. O. Hill. Pedometer-measured physical activity and health behaviors in U.S. adults. *Med. Sci. Sport. Exer.*, 2010.

[11] Centers for Disease Control and Prevention. Facts about physical activity. May 23, 2014. `http://www.cdc.gov/physicalactivity/data/facts.htm`. Accessed on September 1, 2016.

[12] S. Consolvo, K. Everitt, I. Smith, and J. A. Landay. Design requirements for technologies that encourage physical activity. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 457–466. ACM, 2006.

[13] M. De Choudhury, E. Kiciman, M. Dredze, G. Coppersmith, and M. Kumar. Discovering shifts to suicidal ideation from mental health content in social media. In *CHI*, 2016.

[14] M. De Choudhury, S. S. Sharma, and E. Kiciman. Characterizing dietary choices, nutrition, and language in food deserts via social media. In *CSCW*, 2016.

[15] D. Ding, K. D. Lawson, T. L. Kolbe-Alexander, E. A. Finkelstein, P. T. Katzmarzyk, W. van Mechelen, M. Pratt, and Lancet Physical Activity Series 2 Executive Committee and others. The economic burden of physical inactivity: A global analysis of major non-communicable diseases. *Lancet*, 2016.

[16] R. K. Dishman, J. F. Sallis, and D. R. Orenstein. The determinants of physical activity and exercise. *Public Health Reports*, 100(2):158, 1985.

[17] M. Dobbins, K. DeCorby, P. Robeson, H. Husson, and D. Tirilis. School-based physical activity programs for promoting physical activity and fitness in children and adolescents aged 6-18. *The Cochrane Library*, 2009.

[18] T. Dwyer, A. Pezic, C. Sun, J. Cochrane, A. Venn, V. Srikanth, G. Jones, R. Shook, X. Sui, A. Ortaglia, et al. Objectively measured daily steps and subsequent long term all-cause mortality: The tasped prospective cohort study. *PloS one*, 10(11):e0141274, 2015.

[19] C. M. Frederick-Recascino and H. Schuster-Smith. Competition and intrinsic motivation in physical activity: A comparison of two groups. *Journal of Sport Behavior*, 26(3):240, 2003.

[20] U. Gneezy and A. Rustichini. Gender and competition at a young age. *The American Economic Review*, 94(2):377–381, 2004.

[21] S. Göbel, S. Hardy, V. Wendel, F. Mehm, and R. Steinmetz. Serious games for health: Personalized exergames. In *MM*, 2010.

[22] J. Hamari, J. Koivisto, and H. Sarsa. Does gamification work?–a literature review of empirical studies on gamification. In *HICSS*, 2014.

[23] E. C. Hayden. Mobile-phone health apps deliver data bounty. *Nature*, 531(7595):422–423, 2016.

[24] I.-M. Lee, E. J. Shiroma, F. Lobelo, P. Puska, S. N. Blair, P. T. Katzmarzyk, and Lancet Physical Activity Series Working Group and others. Effect of physical inactivity on major non-communicable diseases worldwide: An analysis of burden of disease and life expectancy. *Lancet*, 380(9838):219–229, 2012.

[25] J. J. Lin, L. Mamykina, S. Lindtner, G. Delajoux, and H. B. Strub. Fish'n'steps: Encouraging physical activity with an interactive computer game. In *UbiComp*, 2006.

[26] A. Marshall. Challenges and opportunities for promoting physical activity in the workplace. *Journal of Science and Medicine in Sport*, 7(1):60–66, 2004.

[27] Y. Mejova, H. Haddadi, A. Noulas, and I. Weber. # foodporn: Obesity patterns in culinary interactions. In *ACM DH*, 2015.

[28] L. Miles. Physical activity and health. *Nutrition Bulletin*, 32(4):314–363, 2007.

[29] M. Niederle and L. Vesterlund. Do women shy away from competition? Do men compete too much? *The Quarterly Journal of Economics*, pages 1067–1101, 2007.

[30] U. D. of Health and H. Services. Physical activity guidelines for Americans, 2008.

[31] R. S. Reis, D. Salvo, D. Ogilvie, E. V. Lambert, S. Goenka, R. C. Brownson, and Lancet Physical Activity Series 2 Executive Committee and others. Scaling up physical activity interventions worldwide: Stepping up to larger and smarter approaches to get people moving. *Lancet*, 2016.

[32] J. F. Sallis, A. Bauman, and M. Pratt. Environmental and policy interventions to promote physical activity. *American Journal of Preventive Medicine*, 15(4):379–397, 1998.

[33] J. Salmon, M. L. Booth, P. Phongsavan, N. Murphy, and A. Timperio. Promoting physical activity participation among children and adolescents. *Epidemiologic Reviews*, 29(1):144–159, 2007.

[34] K. Servick. Mind the phone. *Science*, 350(6266):1306–1309, 2015.

[35] J. Sinclair, P. Hingston, and M. Masek. Considerations for the design of exergames. In *GRAPHITE*, 2007.

[36] M. Souza-Júnior, L. Queiroz, J. Correia-Neto, and G. Vilar. Evaluating the use of gamification in m-health lifestyle-related applications. In *New Advances in Information Systems and Technologies*, pages 63–72. Springer, 2016.

[37] P. B. Sparling, N. Owen, E. V. Lambert, and W. L. Haskell. Promoting physical activity: The new imperative for public health. *Health Education Research*, 15(3):367–376, 2000.

[38] A. E. Staiano and S. L. Calvert. Exergames for physical education courses: Physical, social, and cognitive benefits. *Child Development Perspectives*, 5(2):93–98, 2011.

[39] C. Thorup, J. Hansen, M. Grønkjær, J. J. Andreasen, G. Nielsen, E. E. Sørensen, and B. I. Dinesen. Cardiac patients' walking activity determined by a step counter in cardiac telerehabilitation: Data from the intervention arm of a randomized controlled trial. *Journal of Medical Internet Research*, 18(4), 2016.

[40] R. P. Troiano, D. Berrigan, K. W. Dodd, L. C. Masse, T. Tilert, M. McDowell, et al. Physical activity in the United States measured by accelerometer. *Medicine and Science in Sports and Exercise*, 40(1):181, 2008.

[41] J. M. Tucker, G. J. Welk, and N. K. Beyler. Physical activity in U.S. adults: Compliance with the physical activity guidelines for Americans. *AJPM*, 2011.

[42] C. Tudor-Locke and D. R. Bassett Jr. How many steps/day are enough? *Sports Medicine*, 34(1):1–8, 2004.

[43] J. B. Wang, L. A. Cadmus-Bertram, L. Natarajan, M. M. White, H. Madanat, J. F. Nichols, G. X. Ayala, and J. P. Pierce. Wearable sensor/device (Fitbit One) and sms text-messaging prompts to increase physical activity in overweight and obese adults: A randomized controlled trial. *Telemedicine and e-Health*, 21(10):782–792, 2015.

[44] J. B. Wang, J. K. Cataldo, G. X. Ayala, L. Natarajan, L. A. Cadmus-Bertram, M. M. White, H. Madanat, J. F. Nichols, and J. P. Pierce. Mobile and wearable device features that matter in promoting physical activity. *Journal of Mobile Technology in Medicine*, 5(2):2–11, 2016.

[45] WHO. *Global recommendations on physical activity for health*. World Health Organization, Geneva, 2010.