Investing Time in the Public Good: A Web-based Threshold Public Good Experiment

Marco A Janssen
Arizona State University, USA

Allen Lee
Arizona State University, USA

Hari Sundaram
Arizona State University, USA

October 7, 2013
Investing Time in the Public Good: A Web-based Threshold Public Good Experiment

Marco A Janssen\textsuperscript{a}, Allen Lee\textsuperscript{b}, Hari Sundaram\textsuperscript{c},

\textsuperscript{a}Center for the Study of Institutional Diversity, School of Human Evolution and Social Change;
\textsuperscript{b}Center for the Study of Institutional Diversity, School of Human Evolution and Social Change;
\textsuperscript{c}School of Arts, Media and Engineering / School of Computing, Informatics, and Decision Systems
Engineering;

Corresponding author:
Marco A. Janssen
School of Human Evolution and Social Change
Center for the Study of Institutional Diversity
Arizona State University
PO Box 872402
Tempe, AZ 85287-2402, USA
Marco.Janssen@asu.edu

Abstract:
In traditional public good experiments participants receive an endowment from the experimenter that can be invested in a public good or kept in a private account. In this paper we present an experimental environment that uses participants time as their natural endowment, which can be invested in the public good experiment. The experiment runs for several days and participants can make contributions to the threshold public good by logging into a web application and performing virtual actions. Like traditional experiments we find an initial high level of contributions that decreases over subsequent days. We compared two treatments, with and without leaderboards, which compared the real-time performance of groups. We find that the leaderboard increases contributions during the first day, but in subsequent days no significant positive effect is found. Groups with low performances on the first day may have given up when confronted with the leaderboard rankings. These findings show the double-edged effect of information feedback.

Keywords:
Threshold public good, social influence, behavioral experiment.
Investing Time in the Public Good:  
A Web-based Threshold Public Good Experiment  

Marco A. Janssen  
School of Human Evolution and Social Change  
Arizona State University  

Allen Lee  
School of Human Evolution and Social Change  
Arizona State University  

Hari Sundaram  
School of Arts, Media and Engineering  
School of Computing, Informatics, and Decision Systems Engineering  
Arizona State University  

Introduction  
There is a substantial understanding of the conditions that lead to successful governance of the commons by small groups such as communities (Ostrom, 1990). Studies in small-scale communities and in controlled experiments (Poteete et al., 2010) show that the strength of groups in overcoming collective action problems lies in whether or not participants can communicate, whether they have input in the creation of the rules, whether there is group homogeneity, and whether institutional arrangements are monitored and enforced. In small-scale communities, participants have relatively low costs in deriving information to determine the trustworthiness of others. Small communities are characterized by low participant costs to monitor the behavior of others, as well as low costs for face-to-face meetings. The low costs of monitoring behavior and having face-to-face meetings are not generally possible at a large scale. Hence, one of the challenges is to scale up insights of success at the community level to larger scale collective action problems. Addressing a global scale problem like climate change requires actions at different levels of scales (Gibson et al., 2000), including bottom-up initiatives in a polycentric system (Ostrom, 2010).  

Computing has the potential to play an important role in addressing the challenges that appear at the large scale, including mitigating transaction costs, developing new mechanisms of trust, and addressing heterogeneity. First, computational infrastructures, including the increasingly popular social networking sites, reduce communication costs by enabling individuals at different spatial locations to message one another with minimal cost. Low-cost physical-world sensors such as smart meters enable individuals to understand energy use in real-time, as well as the activity breakdown (Froehlich et al, 2011; Larson et al., 2012). The low communication costs and the low sensing costs allow for rapid delivery of information, thus providing crucial real-time feedback on the consequences of our decisions and the decisions of other participants. Examples include smart energy meters (e.g., Mattern et al., 2010), smart water meters (e.g., Hauber-Davidson and Idris, 2006), tracking locations (e.g., Froehlich et al., 2009), and remote sensing of heat loss (e.g., Hay et al. 2011).
Second, with the increased participation in online networks, new forms of “trust” begin to appear, and there are computational mechanisms to extract smaller homogenous communities from large groups. In social networks, such as Twitter\(^1\) and Facebook\(^2\), people can passively, and asynchronously “follow” each other, creating a new form of “ambient intimacy” (Marlow et al., 2009) that is unavailable with inter-personal communication in the physical world, which are typically synchronous. It is possible with community detection algorithms (Lin et al. 2012) to address population heterogeneity by uniting similar individuals within the population into small homogenous groups or communities. Small homogenous groups will encourage the building of trusting relationships leading to greater cooperation. A community, in one operational definition, is a group of people interacting with each other in a consistent manner. Community detection algorithms identify the modular structure of a network, where nodes represent individuals and where links represent the interaction or similarity between individuals. These algorithms are closely related to the family of clustering algorithms. However, cross-validating the discovered communities is a challenge in large datasets since we frequently lack ground truth about community membership.

Clearly, computing provides opportunities to scale up the strengths of self-governance observed in smaller communities. Unsurprisingly, there has been significant interest in developing websites and mobile applications for reducing an individual’s carbon footprint (stepgreen\(^3\)), energy use (OPOWER\(^4\), tendril energize\(^5\), energywiz\(^6\), peoplepower\(^7\), joulebug\(^8\)), transport (ubigreen\(^9\)), competitive sustainability challenges (ecochallenge\(^10\)), and water use (999 bottles\(^11\)) or sustainable behavior in general (practically green\(^12\), eEcosphere\(^13\), Rippl\(^14\)) (see also Dickinson et al. 2013).

To our knowledge, there has only been limited analysis of the effectiveness of these recent technologies. Mankoff et al. (2010) evaluated the effectiveness of a sample of users via self-reports. A more systematic analysis has been performed with OPOWER for a few hundred thousand households (Allcott, 2011). A significant reduction of energy use of around 2% has been found due to providing social feedback on energy bills. One of the challenges in measuring the effect of these apps and websites is the verification of a user’s actual behavior. Disaggregated energy use is possible with new smart meters (Froehlich et al, 2011; Larson et al., 2012) but it is much more difficult to monitor the many other behaviors (for example, eating vegetables instead of beef) affecting our resource use.

These new developments led us to develop a new type of experiment to test the effectiveness of different incentives on actual behaviors. To run experiments, we have developed

---

1 [http://twitter.com](http://twitter.com) accessed August 22, 2012
2 [http://facebook.com](http://facebook.com)
a web-based experimental environment capable of running public good experiments over several days with large groups of participants.

Traditional public good experiments typically have groups of two to ten participants who come to a designated laboratory and make decisions on how much of an experimenter-provided endowment should be invested in the public good, and how much to keep for themselves (Chaudhuri, 2011). The total investment in the public good is then multiplied and shared equally among the participants. The best outcome for the group would be for every participant to invest the whole endowment. The best material outcome for an individual would be to free ride on the actions of the others and not invest in the public good. We typically see an initial level of contributions around 50%, which declines over the subsequent rounds (Neugebauer et al., 2009). In recent work, researchers have run web-based experiments using Amazon’s Mechanical Turk (Paolacci et al., 2010; Rand et al., 2011; Mason and Suri, 2012). In these experiments participants must log in at the same time to participate, similar to traditional one-hour experiments run in economics labs.

We opted to develop a new experimental platform for several reasons. First, our approach mimics affordances of these websites while retaining the participatory challenges within web-based communities aimed at stimulating sustainable activities. Our platform allows us to run experiments over a number of days instead of one hour, which preserves the real-world challenge of participant retention and commitment. A second reason for not using Mechanical Turk is our future goal of conducting experiments with actions in the physical world. With the ability to develop our own infrastructure, we plan to introduce computational elements that can verify certain chosen tasks by examining the evidence submitted by the participant. Furthermore, such an infrastructure allows us to conduct longitudinal studies with varied experimental designs with the same set of participants.

While we are developing a specialized experimental infrastructure, we recognize the value of web sites that promote large-scale behavioral change. We are also actively collaborating with eEcosphere to run natural experiments within their web application to complement our controlled experiments. Some drawbacks of applied websites and apps include the difficulty of verifying actions, inability to test alternative designs, and working with constraints—the websites and the apps are often designed with specific business models in mind. While these drawbacks reduce the ability to test causal effects, they still present unique research opportunities at scale.

In this paper, we report on the first of a series of experiments with our infrastructure. In traditional one-hour experiments participants receive an endowment that they can invest in the public good or keep for themselves. In our experiment, participants do not receive an initial endowment. Instead, they must invest a small amount of their time repeatedly throughout the day to contribute to the public good by logging into the website and “performing” the virtual activities that are available to them at that time. In this sense, a participant’s time is their natural endowment and participation in the experiment competes with the participant’s other real-world activities. Since participants have to take time each day to make a contribution to the public good they have to remain motivated to participate.

Our experimental goal is to replicate a finding from a traditional public good experiment. Tan and Bolle (2007) found that if groups receive information that compares their performance to other groups—where the information does not affect their material rewards—the comparative information leads to an initial increase in contributions. Over the long term, the benefit from
comparative information disappears. In our experiment we will test the effects of information feedback on the level of participation.

We will now discuss the experimental design, before we discuss the results. The paper will close with a discussion of the possibilities of web-based experiments to test collective action for groups and social networks.

**Experimental design**

The experiment is based on a threshold public good problem. The theoretical formulation is as follows: Out of N individuals who can invest in the public good, if at least K units are contributed to the public good the reward for the public good is provided, otherwise it is not. In contrast to traditional threshold public good problems, our formulation does not include an endowment provided by the experimenter (Davis & Holt, 1993). As mentioned earlier, participants do not receive an endowment, but invest their time. If a participant’s group does not meet the threshold they will not receive a refund. From prior research on traditional threshold public good experiments, we know that participants invest less if there are no refund mechanisms compared to experiments where refund mechanisms exist (Isaac et al. 1989).

Participants may invest an amount equal to $x$ minutes to the public good, and may receive $y$ dollars if the threshold is received. Since $x$ is a variable that changes during the experiment, participants may decide to discontinue their contributions when the expected reward per hour invested falls below a certain threshold value. If the experiment is run over several days, we may expect a decrease in participation over time as the amount of effort / investment $x$ is increasing while the expected reward value $y$ is not.

Our experiment is framed as a carbon footprint reduction game where participants can perform virtual actions representing sustainable alternatives to common activities during a 5-day period. These sustainable alternatives are only available at certain time intervals throughout the day, coinciding loosely with when those activities are available (e.g., carpooling is available between 8-10AM and 4-6 PM). Participants can login to a website where they can view an update of their group’s progress (Figure 1) and the highlighted currently available activities (Figure 2). Each activity performed generates points for the group if the participant clicks on the button when the activity is available. This version of the experiment only requires participants to click on the actions. They don’t have to actually perform those actions in real life. In essence, this threshold public good game tests whether sufficient participants within a group can log in to the website at the right time to click on the Perform button for an available activity, contributing their time to the public good.

The reason we frame this public good experiment as a carbon footprint reduction game is to make it less abstract and more compelling for participants. We also found that the website design needed to be engaging and clear in order to maintain participant interest. This differs from experiments in laboratories where the participants are recruited for a certain time and the experimenter presumably has full attention from the participants. When Amazon Turk is used the commitment is somewhat lower than laboratory experiments, but still 90% of the participants remain in the experiment till the end (Suri and Watts, 2011). In our design, the experiment naturally competes with the various other activities a participant has going on in their daily life.
Figure 1: Screenshot of the Lighter Footprints experiment. Participants can view their group progress and the leaderboard comparing their group performance with other groups.

Figure 2: Screenshot of some of the actions the participants can take.
At the start of the experiment, registered participants are sent an email with their username and password to join the experiment. Once logged in they can view the currently available activities, select an activity to perform and earn points for their group. In order to make this experiment more engaging we divided the activities into three levels that progressively unlock if enough participants in a group contribute to the public good. Earnings are based on how quickly a group progresses through the levels during the week (Table 1).

At level 1, five activities are possible. At midnight if the average score per person in a group is 50 points, that group moves to the next level (max = 177). At level 2, five more activities are unlocked and the group average score needed to progress to level 3 increases to 125 green points (max = 323). Advancing to level 3 also unlocks five more activities and the group average score needed to complete the level is again raised to 225 (max = 434). If a group does not reach its target by midnight, the next day they start over with zero points, as points earned in previous days do not accumulate or count for the new day. It thus becomes more difficult over time to reach the group average score thresholds, respectively 28% (50 / 177), 39% (125 / 323) and 51% (225 / 434) of the maximum for levels 1, 2, and 3. From earlier pretests we learned to not be too strict on the first level so that groups can progress to higher levels and continue in the experiment.

The expected reward for each participant is based on how quickly their group progresses to higher levels. Due to human subject requirements payments need to be processed in person. In order to minimize payment transaction costs at scale we decided to use a lottery system, where participants are selected to receive $100 with a probability dependent on how quickly their group completed each level of the experiment. There are 6 possible situations a group can experience. The most cooperative group can finish the third level on the third day. As we will see not all groups perform that well, with some groups not even completing level 1 after five days. The faster a group progresses through the levels, the higher the expected reward (ER). The maximum expected earning is similar to a traditional one-hour experiment. We defined the probability of earning $100 as

- 20% if the group completes level 3 by midnight on the third day (ER = $20),
- 14% if the group completes level 3 by midnight on the fourth day (ER = $14),
- 8% if the group completes level 3 by midnight on the fifth day (ER = $8),
- 4% if the group completes level 2 by midnight on the fifth day (ER = $4),
- 2% if the group completes level 1 by midnight on the fifth day (ER = $2),
- 1% if the group did not complete level 1 by midnight on the fifth day (ER = $1).

The probabilities for the group members are independent, which means that zero, one or more participants in a group can be drawn for a payment of $100.
Table 1: Activities for the different levels. Points for each activity are earned when a participant logs into the web application and presses the “Perform” button during the time slot when the given activity is available. One point roughly corresponds to 0.1 pound of CO$_2$ a day saved.

<table>
<thead>
<tr>
<th>Level 1</th>
<th>Activity</th>
<th>Points</th>
<th>Time Activity is Available</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Enable sleep function computer</td>
<td>10</td>
<td>All day</td>
</tr>
<tr>
<td></td>
<td>Eat local food lunch</td>
<td>15</td>
<td>Noon – 2pm</td>
</tr>
<tr>
<td></td>
<td>Carpool</td>
<td>94</td>
<td>8am – 10am and 4pm – 6pm</td>
</tr>
<tr>
<td></td>
<td>Adjust thermostat by 2 degrees</td>
<td>55</td>
<td>6am – 8am</td>
</tr>
<tr>
<td></td>
<td>Recycle</td>
<td>3</td>
<td>All day</td>
</tr>
<tr>
<td></td>
<td><strong>Total</strong></td>
<td><strong>177</strong></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level 2</th>
<th>Activity</th>
<th>Points</th>
<th>Time Activity is Available</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Turn off the water when you brush your teeth</td>
<td>8</td>
<td>7am-9am and 10pm-12pm</td>
</tr>
<tr>
<td></td>
<td>Bike or take public transport to go out</td>
<td>75</td>
<td>6pm-11pm</td>
</tr>
<tr>
<td></td>
<td>Recycle newspaper</td>
<td>6</td>
<td>All day</td>
</tr>
<tr>
<td></td>
<td>Turn off computer during the night</td>
<td>14</td>
<td>Midnight-8am</td>
</tr>
<tr>
<td></td>
<td>Eat no beef</td>
<td>43</td>
<td>6pm – 7pm</td>
</tr>
<tr>
<td></td>
<td><strong>Total</strong></td>
<td><strong>146</strong></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level 3</th>
<th>Activity</th>
<th>Points</th>
<th>Time Activity is Available</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Turn off the lights if you leave a room</td>
<td>23</td>
<td>6pm – 11pm</td>
</tr>
<tr>
<td></td>
<td>Green lunch</td>
<td>14</td>
<td>Noon-2pm</td>
</tr>
<tr>
<td></td>
<td>Wash with cold water</td>
<td>2</td>
<td>4pm-11pm</td>
</tr>
<tr>
<td></td>
<td>Air dry your clothes</td>
<td>20</td>
<td>Midnight – 6am</td>
</tr>
<tr>
<td></td>
<td>Vegan breakfast</td>
<td>44</td>
<td>7am-9am</td>
</tr>
<tr>
<td></td>
<td><strong>Total</strong></td>
<td><strong>103</strong></td>
<td></td>
</tr>
</tbody>
</table>
We developed two treatments based on Tan and Bolle (2007). Tan and Bolle showed that providing information about group performance compared to other groups temporarily increases the level of cooperation. Group comparison was implemented by adding a leaderboard on the front page of the experiment website (see Figure 1). We also sent a nightly email to each participant with the following text:

“Lighter Footprints Summary for DATE

You are a member of Group A. To participate in this experiment, please visit https://vcweb.asu.edu/lighterprints/participate and login with your experiment username and password.

• You earned Y points while the average in your group was Z point(s).
• XX chat messages were posted by your group.
• Your group was ranked YY out of 15

Congratulations! Your group was promoted to Level 2. In order to be promoted to the next level your group must earn 125 points today.”

The bullet point listing the rank of the group compared to other groups was only sent to participants in the leaderboard treatment.

Results
The experiment was run in April 2013 with 150 participants recruited from a database of potential participants for economic experiments among undergraduates at Arizona State University. The participants signed up the week before the experiment and were informed they would receive instructions for the web-based experiment on a Sunday evening. The experiment ran from Monday, starting at midnight, and ended after 5 full days passed. Participants were informed about the length of the experiment when they were invited to participate.

The 150 participants were allocated into 30 groups of 5. Fifteen groups were placed in the leaderboard treatment, and 15 groups were placed in a control treatment without any information about their group performance compared with other groups.

Figure 3 shows the levels the groups reached during the days. Lighter colors refer to lower levels. Some groups remained at level 1 for the entire duration of the experiment while other groups quickly moved to finish level 3 at day 3. During the first day there is a statistically significant higher contribution by groups with leaderboards compared to groups without a leaderboard (64.24 vs. 45.96 points, which is a significant difference [p=0.089] using the Mann-Whitney test).

In the days that followed we do not see a significant difference. More groups finished level 1 on the first day with a leaderboard (10 vs. 7), but the groups without a leaderboard caught up. Figure 3 suggest that groups in the treatment with the leaderboard who fell behind in the first day do not catch up. It seems that they may have given up early. Participants see in real time how their group is doing compared to other groups in the leaderboard treatment. The ability to compare group performance in real-time requires participants to login. If members of a group do
not make many contributions and are therefore not logged in, they don’t see the level of their group performance. Participants who did not contribute or contributed very little receive information about the relative group performance via the email that is sent out at midnight. Participants in low performing groups who had contributed more than average will receive information about the low performance of their group compared with others and might be less eager to continue participation. This would be in line with conditional cooperation strategies often observed in public good experiments (Fischbacher et al., 2001). Participants who contributed less in a low performing group may not see a reason to increase their effort given the low performance of the whole group.

This kind of reasoning would be less prominent in groups that do not how they are doing compared to the other groups. Therefore it is no surprise that groups without the leaderboard catch up later in the week. On average the leaderboard treatment leads to an average expected earnings of $8.20 versus $6.90, but this difference is not statistically significant (p=0.652 with the Mann-Whitney test).

![Figure 3. Distributions of reached levels of the groups with (left) and without (right) leaderboard.](image)

Table 2 shows that groups with a leaderboard most frequently finish a level the first time they are on that level, or they do not finish the level at all. In groups without the leaderboard it is more common that groups take a few days to finish a level.
Table 2: Time it takes to finish a level for groups with and without Leaderboard.

**With Leaderboard**

<table>
<thead>
<tr>
<th>Group level</th>
<th>Number of days a group stays at that level</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>End level</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
</tbody>
</table>

**Without Leaderboard**

<table>
<thead>
<tr>
<th>Group level</th>
<th>Number of days a group stays at that level</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>End level</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>7</td>
<td>5</td>
<td>1</td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>9</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
</tbody>
</table>

Figure 4 shows the performance of the two types of groups relative to the threshold they need to reach to finish the level, and relative to the maximum number of points that can be earned at that level. In the first day the groups with a leaderboard have a significantly higher level than the groups without a leaderboard (p=0.089 Mann-Whitney test). In the ensuing days there is no significant difference between the two treatments (p > 0.1 for Mann-Whitney tests for each day after day 1). After day 3 fewer groups have to participate since they have already finished level 3 and completed the experiment. Although many participants of groups who had completed level 3 continued to participate, they did not do so at the same level. Removing groups that have finished the experiment leads to statistics indicating lower performing groups, so there is no surprise that the average performance is lower in days 4 and 5.
Figure 4: The two figures show two different measures of group performance: points relative to the maximum number of points possible and points relative to the threshold required to “level-up.” In each figure, the black solid line refers to the case without the leaderboard, and the solid gray line refers to the case with the leaderboard. At the end of day 3, some groups in both treatments have finished level 3. However, in both treatments some individuals from the groups finishing level 3 continue to participate. Thus we show additional lines for days 4 and 5. The black and gray dashed lines refer to the case when we include the performance of the groups who have already finished level 3. The solid lines for days 4 and 5 do not include the performance of groups who have finished level 3.

Participants invest their time in the public good. Even if their actions cost them only a few minutes a day, they have to remember to do their activities. From the comments in the group discussion boards we know that some participants set their alarms for the times when they could get the most points and even created Google Docs spreadsheets listing the optimal times to perform activities.

Participants also received nightly digests summarizing their group’s activity that day to remind them about the experiment. The distribution of contributions is unequal. Figure 5 shows the percentage of participants who have participated during a day, taking into account the groups who had not finished the third level. We see that less than 80% of the participants actually participate in the game during a day. In a traditional experiment in the laboratory this percentage will be 100% since the recruited participants are focused for about an hour on making decisions in an experiment. In Mechanical Turk experiments this percentage was found to be around 90% (Suri and Watts, 2011). Since we provided many elements in the design of the experiment to explicitly keep participants involved (participants signed up by themselves with full knowledge of the duration of the experiment, a $100 possible reward, a more engaging interface and game-like characteristics) this percentage shows the challenge of retention in online activities.

Using the Fisher’s exact test we find that the participation is not different among the two treatments, except for day 3 (p=0.028) and day 5 (p=0.064) when groups without leaderboard have more participation. It is interesting that the participation on day one is similar between the two treatments, but that groups with a leaderboard had a significantly higher number of points. This shows that individuals in those groups returned more frequently to the website to earn the points only available during limited timeslots.
When we look at the distribution of individual contributions we do not see a significant difference between the treatments (Figure 6). There is a bit more inequality in the distribution of contributions among the individuals who had a leaderboard treatment compared to those who did not see a leaderboard (gini coefficient 0.544 versus 0.488). Indeed about 16% of the participants did not participate at all despite signing up for the experiment the week before the experiment started, and receiving a daily digest of their group’s activities.

When do participants login during the day? A high level of points can be earned during particular time slots. Will participants coordinate their logins closely with the availability of points? Figure 7 shows when and how many points can be earned. High valued time slots are between 8 and 9 am and between 6pm and 7pm. The high level of points can be earned during times coinciding with commuting and during meals.
If we look at when participants log in, we don’t see a difference between the two different treatments (Figure 8). What is interesting is that participants largely login just after midnight to start earning points and follow closely during the day for the first opportunities to earn more points. In the communication data it has been mentioned that some participants set their alarms to remind themselves to visit the website.

Figure 7. The amounts of points available at certain times of day. Points for performing a specific activity can only be earned once per day so if a participant performed the “Carpooling” activity at 8 AM they would not be able to perform it again at 5 PM.

We close the result section with an econometric analysis (Table 3). As the dependent variable we used the number of points relative to the threshold and the maximum level the group is playing for. This relative metric will normalize the absolute level of points. We find that the leaderboard has no significant effect. Groups who have finished level 3 invest significantly less points in days 4 and 5. In later days we also find that groups earn less points relative to the threshold or maximum amount of points.
On average there are 0.1 likes and 1 chat message per day per group. We find that there is a positive effect of chat messages during the previous days on the relative performance of the groups. The nightly digest email would also mention the number of messages sent during the past day.

Table 3: Multi-level linear regression on daily group level data. The number between brackets denotes the standard deviation. In columns 1 and 3 the data of all groups are used for the 5 days of the experiments. In columns 2 and 4 the dependent variable does not include the first day due to including lags (t-1) in the analysis.

<table>
<thead>
<tr>
<th></th>
<th>Relative to target</th>
<th>Relative to maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>1.323*** (0.117)</td>
<td>0.287*** (0.040)</td>
</tr>
<tr>
<td></td>
<td>1.459*** (0.163)</td>
<td>0.323*** (0.052)</td>
</tr>
<tr>
<td>Level</td>
<td>-0.068 (0.060)</td>
<td>0.087*** (0.021)</td>
</tr>
<tr>
<td></td>
<td>-0.122* (0.069)</td>
<td>0.024 (0.028)</td>
</tr>
<tr>
<td>Finished</td>
<td>-0.208* (0.120)</td>
<td>-0.106** (0.042)</td>
</tr>
<tr>
<td></td>
<td>-0.155 (0.113)</td>
<td>-0.123*** (0.043)</td>
</tr>
<tr>
<td>Leaderboard</td>
<td>0.005 (0.117)</td>
<td>0.004 (0.039)</td>
</tr>
<tr>
<td></td>
<td>-0.126 (0.110)</td>
<td>-0.039 (0.034)</td>
</tr>
<tr>
<td>Day</td>
<td>-0.133*** (0.031)</td>
<td>-0.0653*** (0.011)</td>
</tr>
<tr>
<td></td>
<td>-0.147*** (0.039)</td>
<td>-0.048*** (0.014)</td>
</tr>
<tr>
<td>#Chat events</td>
<td>0.037 (0.023)</td>
<td>0.022*** (0.008)</td>
</tr>
<tr>
<td></td>
<td>0.041** (0.024)</td>
<td>0.020** (0.009)</td>
</tr>
<tr>
<td>#Likes</td>
<td>0.147 (0.100)</td>
<td>0.038 (0.035)</td>
</tr>
<tr>
<td></td>
<td>0.099 (0.111)</td>
<td>0.040 (0.041)</td>
</tr>
<tr>
<td>Relative to target (t-1)</td>
<td>0.147 (0.100)</td>
<td>0.038 (0.035)</td>
</tr>
<tr>
<td>Relative to max (t-1)</td>
<td></td>
<td>0.315*** (0.106)</td>
</tr>
<tr>
<td># chat events (t-1)</td>
<td>0.044*** (0.021)</td>
<td>0.021*** (0.008)</td>
</tr>
<tr>
<td># Likes (t-1)</td>
<td>-0.096 (0.094)</td>
<td>-0.049 (0.034)</td>
</tr>
</tbody>
</table>

N 150 120 150 120
Wald $\chi^2$ 104.24 (p<0.01) 115.91 (p<0.01) 71.49 (p<0.01) 124.38 (p<0.01)
Log likelihood 74.195 42.807 84.682 80.814
$\chi^2$ 13.15 (p<0.01) 3.67 (p>0.028) 14.66 (p<0.01) 0.93 (p=168)

Discussion

This paper presented the first results of a new experimental environment where participants invest time in the public good during a period of days. We find the major inequality to be the amount of participation among the participants, even though they signed up for the experiment just days before and received a reminder digest email every evening. When participants have to decide to invest their time to contribute to the public good, this investment of time faces competition with alternative activities. This is not the case when recruited experiments are signed up for and performed in the laboratory or on Mechanical Turk.

Adding a leaderboard to the experiment has a positive effect during the first day, but this benefit disappears in subsequent days. This finding replicated the observed effect of Tan and Bolle (2007). On average the treatments generated the same level of points, but there is more inequality of contributions in the groups with a leaderboard. The results suggest that the leaderboard stimulated a rapid start for motivated groups of individuals but may have reduced the motivation of groups who did not perform well during the first day.

Our findings confirm earlier studies that show participants in public goods act as conditional cooperators. When participants see others cooperating or expect others to do so, they may be more likely to contribute to the public good. In a world of conditional cooperators information feedback is important. Although the leaderboard initially increases cooperation...
during the first day, it also gave a negative signal to low performing groups. This observation is in line with the negative effect found in low performing neighbors in public good experiments in a social network (Suri and Watts, 2011). The negative effect of information feedback to conditional cooperators reduces the overall initial benefit of the leaderboard.

Our experimental environment uses participants’ time as their natural endowment, which can be invested in the public good experiment or in other activities. As such, the experimental environment may be used to perform controlled experiments for the increasing use of websites and mobile apps to stimulate sustainable behavior.

Acknowledgements
We would like to thank Rolee Sinha and Shelby Manney, Steven Elias, Shawn FitzPatrick and Brian Tang for the help in implementing and designing the experiment in vcweb. We acknowledge financial support for this work from the National Science Foundation, grant number 1210856.

References


