AUTHOR VERSION

Evaluation of the impact of extrinsic rewards on user engagement in a health promotion context

R.C.Y. Nuijten, P.M.E. Van Gorp, U. Kaymak, M. Simons, A.D.A.M. Kemperman, and P.E.W. Van den Berg

Abstract
Despite the many mHealth solutions available, it remains unclear what their success factors are. Specifically, there has been controversy on the effectiveness of extrinsic rewards. This study evaluates two design elements of an mHealth solution – i.e., social proof and tangible rewards – and their impact on user engagement. During a four-week campaign, a sample of 143 university staff members engaged in a health promotion campaign. Participants were randomly distributed over one of three treatment groups. It was found that the introduction of a sufficiently meaningful, unexpected, and customized extrinsic reward can engage participants significantly more in a health promotion context.

KEYWORDS
mHealth; health promotion; social proof principle; extrinsic rewards

1. Introduction

To date, many (mobile) lifestyle interventions exist, however their effective design elements remain unclear [11, 10]. Within this experiment, the effectiveness of two such elements – social proof and tangible rewards – are evaluated. In this scenario, social proof, or peer pressure, is an example of an intrinsic incentive, whereas the psychical reward is an example of an extrinsic incentive.

According to Deci, Koestner and Ryan, intrinsic incentives are more effective than extrinsic rewards [5]. Their work demonstrated that extrinsic rewards only provide motivation for *routine tasks*, and that for *creative tasks* (i.e., tasks that require some sort of problem-solving) they sometimes may even harm motivation [8, 5]. Additionally, sensitivity to extrinsic rewards is also a matter of personality; according to Depue and Collins particularly extroverts may be more sensitive to extrinsic rewards [6].

Despite the potential adverse impact of extrinsic rewards, it seems fair to assume that binding to internal goals and triggers requires more effort than providing an extrinsic incentive. Additionally, recent evidence suggests that "rewards seem to be an effective strategy for increasing physical activity, while there may be differential effects dependent on the reward type and framing" [13]. Therefore, this study also aims to evaluate whether a tangible reward can be made more powerful when its presentation (i.e., its frame) is altered. According to Kahneman and Tversky, Humans are *loss-averse* (i.e., our mind is programmed to avoid losses) [9]. Hence, an extrinsic reward may be perceived more as a driving force when framed as a loss, rather than when framed as a gain [9, 13]. As an additional advantage, a loss-framed reward might leverage the *rule for reciprocation*, which states that Humans are likely to return a benefit to those who have given them a benefit [3]. To optimize the return of this rule,

what is given first should be perceived as meaningful, unexpected, and customized [4].

Based on these findings, an experiment has been designed to evaluate the impact of different representations of tangible rewards when people are experiencing social proof in the context of a health promotion campaign. It was assumed that subjects would perceive the tasks they typically (are expected to) perform in a health promotion campaign more as creative tasks rather than as routine tasks, because these tasks likely require some adaption and planning, and therefore some sort of problem-solving. Hence, tangible rewards are not expected to be effective in motivating participants in this context, and may even have an adverse effect. Summarizing, it was hypothesized that subjects that only experienced social proof would be more engaged than subjects that experienced social proof *and* encountered a tangible reward (**H1**). Also, it was to be expected that the representational frame of the tangible reward would impact engagement, and that subjects encountering a loss-framed reward would therefore be more engaged than subjects that were introduced to a gain-framed reward (**H2**). It was hypothesized that this effect would be strongest for subjects that showed a tendency towards extroversion rather than to introversion (**H3**). Finally, it was expected that the most engaged subjects were seeing a larger impact on health measures and would be more satisfied with the campaign (**H4**).

2. Methodology

2.1. Participant recruitment

For this study, participants were recruited among staff members at Eindhoven University of Technology. Participants were allocated to a team corresponding to their affiliated (academic or support) departments, and therefore, potentially twenty teams of participants could have been recruited. Although the recruitment campaign was targeted at staff members only, students were also allowed to participate.

To recruit study participants, subjects have been invited via e-mail (e.g., internal corporate mailings) and several announcements were made via corporate and private social media channels (Facebook, Twitter and LinkedIn). Additionally, a professional advertorial video was distributed via several screens on the campus and roughly 250 flyers were handed out. Furthermore, in collaboration with the university's sports center, banners were distributed among the campus. Finally, a recruitment team has been present at the yearly employee health check that was organized by the university's sports center. At the same time, this event marked the start of the campaign. The active recruitment of participants was halted after the first week of the campaign. Nevertheless, participants were still allowed to join the experiment, up until the very last day of the campaign. In the statistical analysis however, only subjects that joined within the first week (and therefore before the treatment notification was sent) were evaluated.

On entering the study, informed consent was obtained from participants. Participants were free to opt-out at any time during the experiment. The study's experimental procedures involving human subjects are in agreement with the ethical principles as detailed in the declaration of Helsinki.

2.2. Study procedures

To test the impact of social proof and extrinsic rewards, a virtual competition was hosted among members of the departments of Eindhoven University of Technology. Participants were invited to demonstrate their own department as the healthiest of the entire organization. To prove the 'health status' of their departments, participants were requested to track their daily, healthy activities – such as active transport, sports, and fruit intake – using an mHealth solution named GameBus (see e.g., `www.gamebus.eu`). The mobile application would then award virtual points for each activity – see Appendix A for an overview of the number of virtual points awarded per activity – and the health status of a particular department would then be calculated as the average number of virtual points scored within that department. During a four-week campaign, participants were allowed to claim virtual points for their departments. To review their progress, the mHealth solution allowed participants to compare their team performance (i.e., performance of their departments) and individual performance within their team. The goal of the competition was to obtain the highest score as a team. During the campaign, the winning department was promised to receive a cup as proof of winning the 'TU/e Vitality Challenge 2018'. Note however that at participant recruitment time, nobody had been promised an opportunity to win a tangible reward of any kind.

To compare the motivational impact of the social proof principle and two differently framed tangible rewards, teams of participants were randomly distributed over three treatment groups: Group N, Group L, and Group G. Individuals in Group N did not have a change to win a tangible reward at all. The only (intrinsic) reward they could obtain was honors at the department level. On the contrary, participants in Group G were given a chance to win a physical reward – a mug – with their individual score exceeding 50 virtual points. Finally, to frame the tangible reward as a loss, participants in Group L received a mug at the beginning of the competition, which in turn they had to return (i.e., lose) whenever their individual score was lower than 50 virtual points. Note that participants in all treatment groups were (potentially) exposed to social proof, since the mobile application allowed all participants to compare their team as well as personal performance, thereby fostering social comparison and highlighting subject interdependence. Hence, based on the literature survey presented in the introductory section, it was expected that Group N would be more engaged than both Group L and Group G, because subjects of Group L and Group G were expected to be engaged just until they hit their personal goal, and thereby obtained their tangible reward (**H1**). Moreover, it was expected that some subjects in Group G would not even strive for the tangible reward, because the promoted behaviors were more creative than routine-based for the target group. Participants in Group L on the other hand were expected to actually do strive for the tangible reward, since the loss-aversion principle likely makes them want to hold onto it. Therefore, it was expected that Group L would be more engaged than Group G (**H2**).

Individuals within the different treatment groups were notified of their treatment on the tenth day since the start of the campaign via e-mail. Participants were unaware of the existence of other experimental groups. To minimize the chance of information exchange between members of different treatment groups, the assignment to treatment groups was made at the team level, rather than at an individual level.

Throughout the campaign, the organizing committee has communicated with the subjects via e-mail. Several e-mails have been sent to inform and engage participants. At the start of the campaign, an e-mail was sent to inform everybody on how to get

started. After one week, an informative message was sent, explaining how to register activities automatically. On the tenth day, the treatment notification was sent, which consisted of a general and personalized message. The general part of the message highlighted the actual interim score at that time, in order to foster peer pressure. The personalized part of the message depended on a subject's allocated treatment group. Subjects in Group L were notified that they would receive a mug in the upcoming week, which they were obliged to return when their personal goal of obtaining 50 virtual points was not met. Similarly, subjects in Group G were notified about the opportunity they had to win a mug, whenever they hit their personal goal of obtaining 50 virtual points. Subjects in Group N did not receive any additional personalized message. On the 22^{nd} day another informative message was sent, informing subjects on how to obtain bonus points by performing additional tasks (i.e., Deskercises). On the 25^{th} day, another personalized notification was sent to foster peer pressure among subjects. On the 29^{th} day a closing message was sent, and on the 30^{th} day subjects of Group L and Group G were informed on the procedure to either return or collect their prizes.

2.3. Measurements

To perform quantitative analysis and test the hypothesis, engagement measurements were derived from raw data of the mHealth solution, as well as from an online post-test questionnaire. Finally, the engagement of participants was approximated as: 1) the number of activities an individual had performed (i.e., both self-reported and automatically recorded), and 2) the number of days a participant had checked on the mHealth application (i.e., the number of distinct days a participant had been online, as objectively derived from the mHealth solution).

Participants' self-reported degree of extroversion was included in the analysis as a control variable for subjects that filled out the additional questionnaire. Extroversion was measured using a subset of The International English Big-Five Mini-Markers [14]. Additionally, the post-test included questions to assess the subjective impact of the campaign on participants' 1) level of physical activity, 2) fruit intake, and 3) interaction with colleagues, as well as questions to measure the satisfaction of participants with the campaign.

2.4. Statistical analysis

To assess the degree of similarity (i.e., independence) of treatment groups at baseline (i.e., in the first week, before the treatment notification), one-way ANOVA tests have been performed in R [12] on continuous variables (i.e., engagement metrics) and a chi-square analysis has been performed for the categorical variable gender.

Subsequently, using *lme4* [2], a mixed-effect regression model with random intercepts and random slopes was used to test the primary and secondary hypothesis of a difference in the rate of change of the number of activities performed and the number of days online between the different treatment groups. Random intercepts were fit per subject and team (i.e., department) and random slopes were allowed for subjects over different time periods. Both hypothesis **H1** and **H2** were assessed by comparing the slope of change in the number of activities and number of days online over the 4-week treatment period between the three groups, with a positive slope reflecting increased engagement. To assess hypothesis **H3**, subjects that filled out the online post-test

questionnaire were evaluated separately. In this analysis, the impact of incorporating a subject's degree of extroversion as a co-variate in the model was evaluated. Finally, again on this subset, one-way ANOVA tests have been performed to assess hypothesis **H4** and thereby determine whether different treatment groups perceived the campaign differently in terms of impact on their 1) level of physical activity, 2) fruit intake, and 3) interaction with colleagues, as well as on 4) campaign satisfaction measures.

3. Results

3.1. Study participants

A total of 143 participants (i.e., with 46, 53, and 45 participants in respectively Group N, Group L, and Group G) engaged in the campaign and all teams had participants subscribed (average number of participants per team = 8.35, maximum number of participants per team = 36, minimum number of participants per team = 1), although only 16 out of 20 teams had active participants involved (i.e., participants scoring more than zero virtual points). Two participants – one in Group N and one in group G – chose to leave the experiment prematurely. Of the participants subscribed, 43% were males whereas 57% were females. Three participants were students, while the other 140 participants were university staff members. One-Way ANOVA tests and a chi-square analysis revealed that treatment groups were not statistically different from each other in terms of engagement metrics and gender distribution in the week before the treatment notification was sent.

3.2. Outcomes

Statistical analysis was performed on a subset of participants that were subscribed within the first week of the campaign (n = 122, with 42, 49, and 31 participants in respectively Group N, Group L, and Group G). The number of activities performed per week ranged from 0 to 88, whereas the number of distinct days participants were online per week varied from 0 to 7. Before the analysis, the square root of the number of activities performed was taken to assure valid interpretation of the regression model. Visual inspection of Figure 1 suggests differences in mean engagement between Group N and both the extrinsically rewarded groups (i.e., Group L and Group G), but no differences between Group L and Group G.

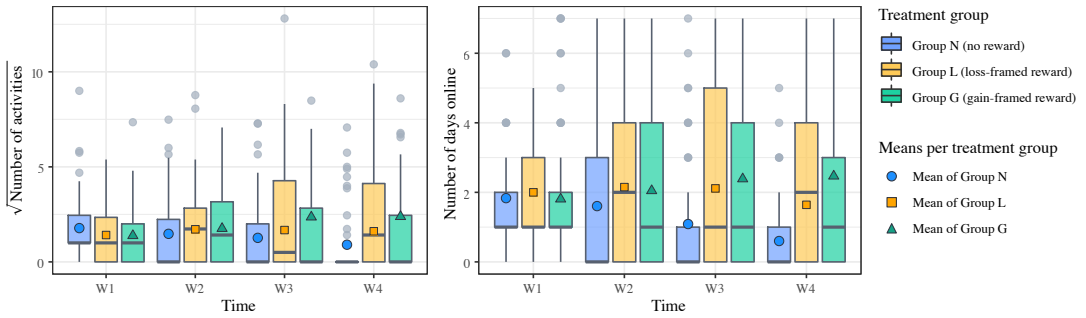

Figure 1. Box plots with highlighted means of engagement metrics over time per treatment group (n = 122).

The square root of the number of activities that a participant in Group N performed

is reduced by 0.279 (95%-CI = -0.475; -0.0785) weekly. The square root of the number of activities that a participant in Group L and Group G performed were weekly increased by 0.0673 (95%-CI = -0.118; 0.253) and 0.145 (95%-CI = -0.0876; 0.378) respectively. Both the difference between Group N and Group L and the difference between Group N and Group G are significant (p = 0.0354, and p = 0.0196, respectively), while between Group L and Group G, no statistically significant difference is found (p = 0.862).

The number of days a participant was online in Group N was reduced by 0.414 (95%-CI = -0.623; 0.144) weekly. The number of days a participant was online in Group L and Group G were weekly decreased by 0.069 (95%-CI = -0.531; 0.162) and increased by 0.106 (95%-CI = -0.137; 0.350) respectively. Again, the difference between Group N and Group G is significant (p = 0.0048), but the difference between Group N and Group L is not (p = 0.0858). Between Group L and Group G, no statistically significant difference is found (p = 0.369).

Of the post-tested subjects (n = 48), 18.8% reported to have engaged in physical activity more often than usual, 14.6% reported to have eaten more pieces of fruit than usual, and 35.5% reported to have interacted with colleagues more often, see Figure 2. In an analysis of post-tested subjects that were subscribed within the first week (n = 42), a subject's degree of extroversion was not a statistically significant term. Additionally, none of the one-way ANOVA tests reported statistically significant differences in ratings on health impact and campaign satisfaction between different treatment groups.

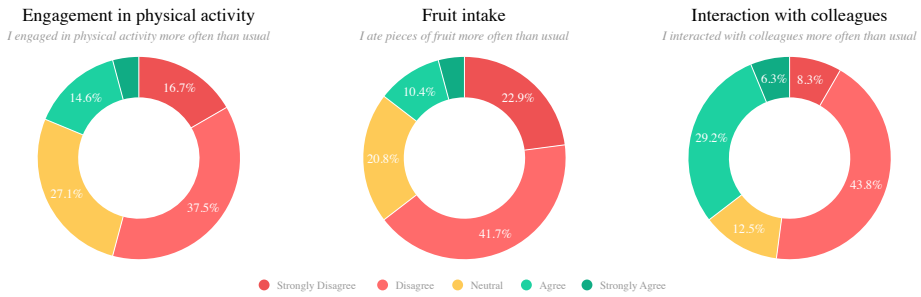

Figure 2. Donut charts of post-test responses on the intervention's health impact measures (n = 48).

4. Discussion

Despite the social proof being present, engagement seems to diminish over time in participants in Group N (negative effect sizes for both engagement metrics, although the negative direction of the effect cannot be verified for the number of days online, since its confidence interval includes zero). Engagement in Group L seems to increase over time in terms of activities (positive effect size) and decrease in terms of number of days online (negative effect size), however additional data has to be collected to determine whether these effects truly exist, since both confidence intervals include zero. Finally, engagement in Group G seems to increase over time (positive effect sizes for both outcome measures), although again additional data has to be collected to determine whether these effects truly exist, since again both confidence intervals include zero.

The difference in the number of activities performed between Group N and both Group L and Group G (**H1**) is statistically significant, meaning that the introduction of a tangible reward could have engaged participants significantly. Contradictory, it was expected that participants in Group N would be more engaged than participants in both Group L and Group G. This controversy may be explained by the fact that effort was taken to optimize the potential impact of the tangible reward by introducing a prize that was likely perceived as meaningful, unexpected, and customized [4]. Additionally, this controversy may have emerged from subjects experiencing the rewarded tasks as routine tasks, rather than as creative tasks, which may have constituted an environment in which extrinsic incentives actually are powerful [5, 8].

Within this experiment, no statistical differences in engagement were found between Group L and Group G (**H2**), meaning that the frame of the reward has not influenced engagement. The supposed absence of an effect may be explained by the fact that participants in Group L did not perceive the mug as sufficiently important to actively prevent losing it, or by the fact that these participants were not expecting the organizing committee to actually reclaim the prize in case their personal goal was not met. At the same time, this effect could not be explained by participant's degree of extroversion (**H3**), since inclusion of a subject's degree of extroversion did not significantly alter the model. Hence, it may be concluded that more extrovert subjects in Group L and Group G were not engaged more. However, this result may also be explained by the fact that this analysis was performed on a small subset only (n = 42; post-tested subjects that had subscribed within the first week).

Finally, the higher engagement of participants in Group L and Group G did not result in a statistically more positive evaluation of the campaign's impact on health measures, or campaign satisfaction (**H4**).

4.1. Study limitations

Study limitations included a sampling bias, which is a potential threat to the study's external validity. In particular, since participants were recruited at convenience, the recruited participants were expected to be generally more motivated to demonstrate their team (i.e., department) as the healthiest of the entire organization, or to participate in an health promotion campaign. Additionally, this study has been prone to a potential threat to internal validity, since subjects in different treatment groups have been communicated with differently. Particularly, subjects of Group L have been in physical contact with a member of the organizing committee during the distribution of tangible rewards, whereas subjects of Group N and Group G have not. Furthermore, since this study focused on assessing the impact of different treatments, baseline measurements on participants' level of physical activity, fruit intake, and interaction with colleagues were not recorded. Finally, this study relied mostly on self-reported measures, which may affect validity of statistical results.

4.2. Future research

Future research should focus on collecting more data in order to estimate true effect sizes and to better estimate the impact of extroversion and sensitivity to extrinsic rewards. Collecting more data seems particularly difficult for obtaining post-test responses, since only 48 participants from a sample of 143 have completed the online questionnaire (34%). Therefore, a beneficial line of research is to explore methods to

approximate survey responses (i.e., personality traits) from application event data and user performance metrics. Additionally, future research experiments may include baseline measurements on participants' health status, such that the intervention's health impact can be estimated objectively. Furthermore, a baseline measurement may enable tailoring the intervention towards a subject's personal needs. Finally, future research should focus on conducting an experiment where the impact of unexpected rewards is compared with the impact of expected rewards.

5. Conclusions

This study aimed to simultaneously analyze the effectiveness of different mHealth design elements (i.e., social proof and tangible rewards), using a realistic dataset. It was found that – despite peer pressure being present – engagement of participants turned out to diminish over time. Nevertheless, the introduction of a sufficiently meaningful, unexpected, and customized tangible reward can engage participants significantly in a health promotion context.

References

[1] B. E. Ainsworth, W. L. Haskell, S. D. Herrmann, N. Meckes, D. R. Bassett Jr, C. Tudor-Locke, J. L. Greer, J. Vezina, M. C. Whitt-Glover, and A. S. Leon. 2011 compendium of physical activities: a second update of codes and met values. *Medicine & science in sports & exercise*, 43(8):1575–1581, 2011.

[2] D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. .

[3] R. B. Cialdini. Harnessing the science of persuasion. *Harvard Business Review*, 79(9):72–81, 2001.

[4] R. B. Cialdini. *Pre-suasion: A Revolutionary Way to Influence and Persuade*, pages 153–157. Random House Books, 2016.

[5] E. L. Deci, R. Koestner, and R. M. Ryan. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological bulletin*, 125(6):627, 1999.

[6] R. A. Depue and P. F. Collins. Neurobiology of the structure of personality: Dopamine, facilitation of incentive motivation, and extraversion. *Behavioral and brain sciences*, 22(3):491–517, 1999.

[7] Gezondheidsraad. *Beweegrichtlijnen 2017*. Gezondheidsraad, The Hague, The Netherlands, 2017. URL `https://www.gezondheidsraad.nl/documenten/adviezen/2017/08/22/beweegrichtlijnen-2017`.

[8] S. Glucksberg. Problem solving: Response competition and the influence of drive. *Psychological Reports*, 15(3):939–942, 1964.

[9] D. Kahneman and A. Tversky. Choices, values, and frames. *American Psychologist*, 39(4):341, 1984.

[10] Z. H. Lewis, M. C. Swartz, and E. J. Lyons. What's the point?: A review of reward systems implemented in gamification interventions. *Games for health journal*, 5 (2):93–99, 2016.

[11] S. Michie, M. Richardson, M. Johnston, C. Abraham, J. Francis, W. Hardeman, M. P. Eccles, J. Cane, and C. E. Wood. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus

for the reporting of behavior change interventions. *Annals of behavioral medicine*, 46(1):81–95, 2013.

[12] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL `https://www.R-project.org/`.

[13] A. N. Sullivan and M. E. Lachman. Behavior change with fitness technology in sedentary adults: a review of the evidence for increasing physical activity. *Frontiers in public health*, 4:289, 2017.

[14] E. R. Thompson. Development and validation of an international english big-five mini-markers. *Personality and individual differences*, 45(6):542–548, 2008.

Appendix A. Number of virtual points awarded per activity

During the campaign, participants were rewarded virtual points for their daily, healthy activities. These activities were rewarded relative to the effort they would take to perform (based on their MET value [1]), see table A1. Activities that classify as lightly intensive (MET between 1.5 and 3) are rewarded 1 to 3 virtual points; activities that classify as moderately intensive (MET between 3 and 6) are rewarded 3 to 5 virtual points; activities that classify as highly intensive (MET larger than 6) are rewarded 5 to 10 virtual points. The prescribed minimum duration of activities has been based on Dutch guidelines [7]. Note that participants were not awarded a multiple of virtual points (e.g., 750 meters of walking was not awarded with 3 virtual points, but rather just with 1 virtual point) to stimulate performing more sessions, rather than longer sessions. According to post-tested subjects (n = 48), more than half (52%) of the participants agreed that this allocation of virtual points was fair.

Table A1. Number of virtual points awarded per activity

Rule	MET	Virtual points
Walking (250 m)	3	+1
Running or jogging (15 min)	6	+5
Biking (15 min)	5	+5
Swimming (15 min)	6	+5
Fitness, yoga or dance (25 min)	7	+10
Any ball sports (25 min)	7	+10
Ice or roller skating (25 min)	7	+5
Make a Work Walk with colleagues	3	+3
Join a sports lesson at SSC/e	3–6	+3
Eat an apple (or any other fruit)	1.5	+2
Visit a TU/e Vitality Week event	?	+3
BONUS: Deskercise (15 reps)	?	+3