"Good Enough!": Flexible Goal Achievement with Margin-based Outcome Evaluation

Gyuwon Jung
KAIST
Daejeon, South Korea
qonejung@kaist.ac.kr

Jio Oh
KAIST
Daejeon, South Korea
harryoh99@kaist.ac.kr

Youjin Jung
University of Washington
Seattle, Washington, United States
youjinj@uw.edu

Juho Sun
KAIST
Daejeon, South Korea
soju@kaist.ac.kr

Ha-Kyung Kong*
Seattle University
Seattle, Washington, United States
hkong@seattleu.edu

Uichin Lee†
KAIST
Daejeon, South Korea
uclee@kaist.edu

ABSTRACT

Traditional goal setting simply assumes a binary outcome for goal evaluation. This binary judgment does not consider a user's effort, which may demotivate the user. This work explores the possibility of mitigating this negative impact with a slight modification on the goal evaluation criterion, by introducing a 'margin' that is widely used for quality control in the manufacturing fields. A margin represents a range near the goal where the user's outcome will be regarded as 'good enough' even if the user fails to reach it. We explore users' perceptions and behaviors through a large-scale survey study and a small-scale field experiment using a coaching system to promote physical activity. Our results provide positive evidence on the margin, such as lowering the burden of goal achievement and increasing motivation to make attempts. We discuss practical design implications on margin-enabled goal setting and evaluation for behavioral change support systems.

CCS CONCEPTS

• **Human-centered computing** → **HCI theory, concepts and models**.

KEYWORDS

Behavior Change, User Experience Design, Interview, Service Design

ACM Reference Format:
Gyuwon Jung, Jio Oh, Youjin Jung, Juho Sun, Ha-Kyung Kong, and Uichin Lee. 2021. "Good Enough!": Flexible Goal Achievement with Margin-based Outcome Evaluation. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3411764.3445608

*Corresponding author for Study 1
†Corresponding author for Study 2

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445608

1 INTRODUCTION

Goal setting is one of the most widely used strategies in behavioral change [33]. Researchers have examined major goal-setting dimensions such as difficulty and proximity, in order to find out effective goal-setting strategies [50]. Goal setting design is an important research topic for Human-Computer Interaction (HCI) [13] in that due to the prevalent use of smart technologies, users set their goals via mobile applications that remind and help them achieve the goals. Traditional goal-setting systems have provided feedback based on goal-based binary evaluations. In this evaluation criteria, users will receive a success or failure message simply based on whether the goal is met or not.

However, the binary judgment may be less adequate for supporting users to continue their behavioral change. In the early stages of a behavioral change, the goals may seem difficult to users, considering they have to overcome their previous, dominant, and unhealthy behaviors [26], which in turn, the difficulty may result in cognitive anxiety [21] in achieving their goals. Moreover, if they have a fear of failure when striving for their goals, they may focus more on the evaluator's judgment in order to avoid being shamed [36] and would likely experience more stress and anxiety due to the expected negative feedback [32]. In addition, this evaluation criterion has a limitation in that it does not fully acknowledge how hard the user has tried to achieve their set goal. Even if the user almost reaches the goal (e.g., finishing 9,500 steps out of 10,000), her or his efforts will be counted as nothing and result in a lowered motivation toward the goal. As these negative experiences in goal-setting systems may lead the user to abandon her or his behavioral change plans altogether, HCI research should examine not only how to set an appropriate goal for the user, but also how to assess the user's performance in a positive and motivating manner.

In this paper, we propose a concept of "margin" to support the user's behavior change with a slight modification on the evaluation system. The margin is a range near the goal where the user's outcome will be counted as "good enough" even if the user fails to completely achieve the goal. With the inclusion of a margin, the evaluation becomes more flexible and relaxed, as it allows room for a little failure. To distinguish the assessment involving a margin from the traditional, goal-based assessment, we have titled this criteria as "mission." Thus, the condition for mission success becomes entering the margin area, and we could determine the user's

outcome as a mission success or failure. The concept of a margin is inspired by statistical process control (SPC) [35, 42] in the manufacturing field. In SPC, the quality of products within a certain level or boundary will be decided as 'in control', which means the products have a good-enough quality in terms of the statistical process. This study utilizes the boundary (or say, 'flexibility') concept in goal setting and investigates whether a similar concept could be applied in the field of behavioral change as well.

To explore the feasibility of margin-enabled behavioral coaching, we designed two experimental studies: (1) a large-scale, survey-based vignette experiment with 500 participants to understand how they perceive the margin-based assessment; and (2) a 10-day field experiment with 54 participants to help better understand how their behavior would be influenced by the margin in the evaluation criteria. For the field experiment, we designed FlexCoach, a margin-enabled coaching system that supports physical activity with goal-setting and feedback on the users' daily step counts.

Results from this study showed that participants rated their performances more positively in terms of goal achievement in scenarios with a margin, especially when their outcomes were within the margin area (i.e., outcomes that would be regarded as 'failure' in the traditional evaluation criteria, but as 'mission success' in our system). Field experiment results showed that the margin-based evaluation supported the users by reducing the negative emotions and psychological effects, although we could not find statistically significant differences in physical activity through the margin. Interestingly, most of the participants responded that their target remained the same as the original goal even if the standard of the mission success was more relaxed because of the margin.

Our results show that the flexibility in performance evaluation may help promote the users' behavior change. Specifically, we observe that the margin could be shown as a human-like feature of taking context into account when evaluating others' performance, and this could possibly allow users to become more engaged in the goal achievement process with less stress. Based on these findings, we propose several design implications for margin-enabled coaching systems, such as proper margin size, adaptive goals, and margin based on the user's accomplishment, feedback and rewards.

2 BACKGROUND AND RELATED WORK

This section offers an overview of the goal setting theory for behavior changes and statistical process control for data-driven clinical decision making. We then review how prior HCI studies considered goal setting and achievement evaluation in their system design.

2.1 Goal Setting Theories

Goal setting is one of the most widely used techniques in behavioral change systems [47] due to its positive effects in diverse domains [33], including diet and physical activity [45]. The major constituents of goal setting are: (1) properties (i.e., proximity, specificity, and difficulty), (2) components (i.e., progress feedback, and achievement rewards), and (3) sources (i.e., self-set, assigned, participatory, guided, or group-set) [50]. It is known that specific, challenging short-term goals are effective [33]. In addition, offering feedback on goal progress improves goal achievement, and internal (e.g., self-esteem) or external (e.g., recognition or money) rewards reinforce goal progress [40]. Goal sources (who set the goal) are less sensitive to behavioral performances [17], and expert-guided, patient-directed goal setting is most preferred among users [13]. These goal setting properties can be summarized as a S.T.A.R.T criteria for behavior changes [45]: Specificity, Timing, Acquisition (=type), Rewards and feedback, and Tools (for action planning and self-monitoring).

Goal achievement evaluation often results in a binary outcome, which could possibly demotivate users for behavioral engagement. In the rehabilitation domain, researchers have proposed using goal attainment scaling (GAS) [9, 53] that allows a flexible judgment of goal achievement for a given intervention. GAS offers flexibility by establishing "sub-goals" around the expected outcome as a baseline sub-goal: i.e., two higher sub-goals (+1 somewhat more; +2 much more) and two lower sub-goals (-1 somewhat lower; -2 much lower). GAS could mitigate the issue of binary judgment by recognizing a user's effort via sub-goals, but this approach faces a similar issue of binary judgment at the sub-goal levels. This approach is best suited for rehabilitation goals [9, 53], failing to offer practical guidelines on setting sub-goals in regular health behavior changes such as increasing physical activity (e.g., how to set sub-goals). Furthermore, GAS's scoring rubric has "negative scores" for lower sub-goals (-1 and -2), which may demotivate users—for example, regarding daily step goals, a user may have a series of "negatively scored" days. Prior studies reported that sub-goals are useful when expected goals are too complex and challenging as they offer tangible rewards [51], as is the case in rehabilitation scenarios.

2.2 Statistical Process Control for Clinical Decision Making

Statistical Process Control (SPC) is widely used in the manufacturing field to monitor the quality of a given manufacturing process [35, 42]. It considers statistical variations of quality metrics in the manufacturing process by setting control limits from a given dataset, and this technique is known as "control charts." Examining the lower and upper control limits in the control charts allows data-driven decision making on quality control because it helps us to systematically differentiate random variations from systematic errors. Prior studies on data-driven clinical decision making [46] leveraged SPC in several ways: detecting trend changes from baseline observations, judging the effectiveness of an intervention retrospectively (baseline vs. intervention comparison), and real-time tracking on the effects of interventions for adaptive treatment. Current applications of SPC are reactive in that it is similar to "behavioral analysis," as it aims to find out possible out-of-control events for early intervention. This concept can be extended so that behavioral coaching agents *proactively* guide users to set their behavioral change goals and to consider *permissible behavioral variations* for goal achievement evaluation (what we call "margin"). As a result, the agents can work with the users to set both behavioral goals and permissible variations. Our concept of *proactive process control* with behavioral coaching is in line with SPC in data-driven clinical decision making.

2.3 Persuasive Technology Design for Physical Activity Promotion

Prior studies examined diverse aspects of persuasive technology design such as goal setting, self-tracking tools [14], behavioral learning and personalizing [20, 49], behavioral tunneling [10], just-in-time reminding [11, 12], and interactive coaching [4, 24]. Consolvo et al. [13] used Ubifit Garden with self-tracking features [14] to evaluate diverse goal setting scenarios in terms of goal sources and periods. The advanced sensing features of interactive systems provide novel opportunities for persuasive interactions [30]. For example, mobile systems can learn users' behaviors and personalize behavioral interventions [49]. Physical activity sensing can alert users to be more active [8, 12, 34, 54]. Systems can offer structured interventions for behavioral guiding such as micro-breaks with physical activity [37] and gamified behavioral guiding when sensing physical activity transitions [10]. Natural language interactions with intelligent coaches help users to better reflect upon their physical activity history [24].

We now examine how existing persuasive technologies considered goal setting and achievement evaluations. Only a few HCI studies explored the major dimensions of goal setting, such as properties, components, and sources. Regarding goal properties (specificity and difficulty), researchers mostly considered trackable activities (e.g., step counts) [7, 12, 14, 31] and experimented with system-driven goal settings [7, 13, 23, 31]; e.g., setting weekly goals according to a user's baseline performance [31]. The Tracker Goal Evolution Model [41] highlighted that users also set qualitative goals such as a level/range of physical activity or a qualitative measure of perceived activities over different activities, which are closely related to their quantitative goals of physical activities.

A large-scale data analysis study on MyFitnessPal's weight loss goals revealed that users' behaviors in the first week are critical for goal achievement [18], and thus, it is very important to successfully maintain proximal goals (e.g., daily goals of physical activity). To better motivate users, Munson and Consolvo [39] experimented with dual-goal setting and achievement evaluation approaches where users set both primary and secondary goals as a main and a backup goal, and achievement evaluation was done by rewarding different badges for different goals (e.g., a gold badge for primary and a silver badge for secondary goals). Agapie et al. [2] proposed a system-driven lapse management method that allows users to use a fixed amount of cheat points per week for flexible goal achievement evaluation. However, offering user-driven flexibility (e.g., setting lower sub-goals) makes it difficult to achieve "behavioral process control." Our goal is to trade flexibility for process control by exploring "system-driven proactive process control."

3 MARGIN

Margin Definition: We propose a concept 'margin' that is a range of values near the goal where the user's outcome will be determined as "good enough." When the evaluation process includes a margin, outcomes that fall within the margin will be recognized even if they fall short of the goal. Figure 1 shows an example of tracking daily step-count goals, which aims to promote physical activities. A margin is given in the form of an area based on the goal that is either set by the users or assigned to them (or collaboratively set).

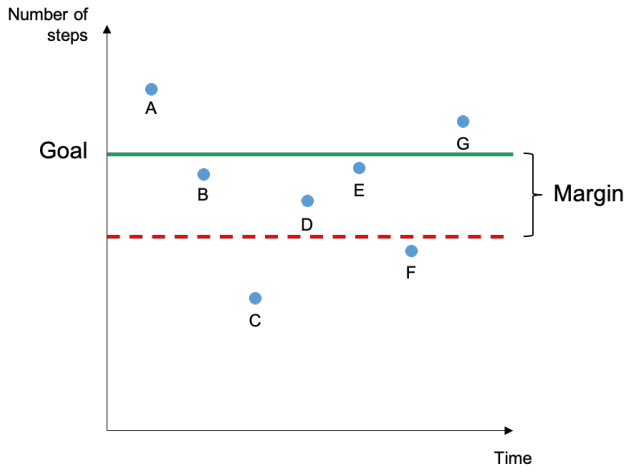


Figure 1: An example of daily step-count records for 7 days

If we apply the traditional binary evaluation criterion as described in Figure 1 (i.e., judging success or failure based on meeting the goal), there are five cases of failure out of seven. However, when a margin is included in the evaluation process, the assessment becomes relaxed, and the system recognizes points B, D, and E as being "good enough," not as failures. This concept is similar to permissible quality variation in the statistical process control chart (SPC) widely used in the manufacturing field. In the control chart, products within a certain boundary of quality metrics (i.e., an integer times standard deviation of the outcomes) will be recognized as quality in control despite its deviation from the central line (i.e., mean of the outcomes). In our scenario of behavioral goal setting, we provide a flexible and relaxed assessment by introducing a "good-enough zone" defined by a margin, which is similar to "permissible quality variation" in SPC. This allows us to recognize the user's effort and avoid small failures, which can possibly mitigate the negative effects of experiencing narrow failures. Thus, offering a margin for a given goal helps a coaching system to proactively control a user's behavioral process, which we call "system-driven proactive process control."

Margin Types: Depending on the type of the goal, margin placement may exist above, beneath, or in-between the goal. One of the well-known classifications of goals is approach versus avoidance by its motivation [15]. Goals for the approach motivation have a positive character in that people wish to achieve, whereas those for the avoidance motivation have a negative nature in that people try to avoid them if possible. In the approach goals, the user aims to achieve above the goal line, and thus, the margin will be placed beneath the goal (e.g., doing more exercise for weight loss or getting better grades in class). In contrast, avoidance goals aim not to pass the goal line, such as the maximum time allowance of smartphone use [22], and thus, the margin will be placed above the goal. There could be also maintenance goals; in this case, the margin should be placed in-between the goal line (e.g., weight maintenance: within +1 or -1 kg around the goal). Margin types can be further classified as fixed or variable in terms of margin size. So far we have discussed fixed margin scenarios, but margin values could vary over time. A

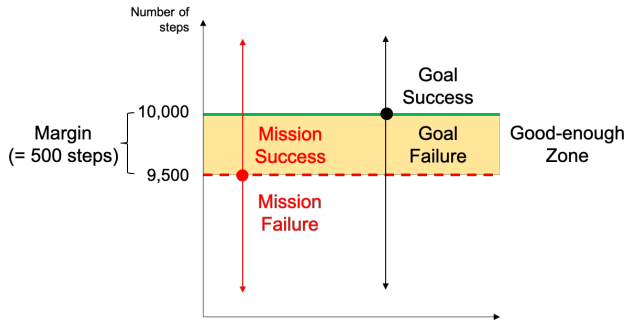


Figure 2: Possible states of outcomes from goal and mission

random margin can be drawn every day from a given distribution (e.g., uniform or Gaussian distribution). Fixed margin scenarios are comparable to random margin scenarios if the mean value of a margin distribution is equal to that of fixed margin scenarios.

Goal and Mission Separation: Relaxed assessment with a margin does not mean invalidating or changing a user's original goal. We ensure that the goal remains as it is, but the feedback provided by the evaluator is changed to prevent users from drawing negative conclusions from narrow failures. This can be done by introducing an additional criterion for relaxed assessment (called 'mission evaluation'), which is different from original 'goal evaluation.' For a given goal, we now have a mission to offer a flexible assessment of the user's outcome with respect to the margin around the goal. Figure 2 shows an example of the goal: walking more than 10,000 steps per day with a margin of 500 steps (5% of the goal). In this case, the goal is to reach 10,000 steps, and an associated mission is given as reaching more than 9,500 steps. Margin creates three possible outcome states as described in Figure 2: (1) mission success and goal achievement, (2) mission success and goal failure (i.e., within a good-enough zone), and (3) mission failure and goal failure. When a user sets a goal, the system offers an option for "margin" and this creates an associated mission for separate evaluation. We can differently treat two mission success states depending on how we recognize goal achievement (e.g., different messages and badges). Mission success is recognized as 'Well done!', but a system reminds users for goal anchoring, 'But your goal has not been achieved yet.'

Research Questions: As the first step toward understanding how a margin affects goal perception and achievement evaluation, we choose a health behavior goal of increasing physical activities, or daily step counts, which are widely supported in recent smartphones (e.g., Google Fitness or Apple Health). A lack of physical activity due to sedentary lifestyles is one of the major causes of various chronic diseases such as obesity, diabetes, and cardiovascular dysfunction [29]. To encourage physical activities, health insurance providers commonly use monetary incentive schemes that have positive effects on behavior changes [19, 44].

We first explore how users interpret the margin-enabled goal setting and achievement evaluation. For example, a margin may result in a positive assessment of goal achievement. We think that is critical because how users evaluate their progress may influence whether to keep trying to reach the goal or not. In addition, we examine what users are aiming for in the presence of a margin. A target could be either the original goal, or a reduced value by the margin; that is, achieving the goal vs. merely passing the mission (or reaching the good-enough zone). Thus, we set our first research question as follows: RQ1: How does the margin affect a user's perception of her or his goal and goal achievement?

Real-world user experiences of margin-enabled coaching may differ from the user's original perception. We design a margin-enabled coaching system to investigate how a margin influences the user's goal perception and her or his behaviors in reaching the goal and how the user experiences differ from a traditional binary goal-based assessment. Aiming to explore the design space of margin-enabled coaching, we set our second research question as follows: RQ2: How does the margin affect real-world user experiences of goal setting and achievement evaluation in margin-enabled behavioral coaching?

4 STUDY 1: SURVEY-BASED VIGNETTE EXPERIMENT

4.1 Methods

4.1.1 Survey Design. To evaluate how participants perceive the margin-based assessment, we conducted a large-scale survey-based vignette experiment. In a typical vignette study, researchers provide descriptions of certain situations or people to the participants and ask them some questions to see how they respond or make decisions about the settings [6].

Thus, we considered a situation of using a mobile application that supports users to walk more, providing (1) step-count goal, (2) margin, and (3) mission (i.e., margin-based evaluation criteria) every day. In the vignette, the coaching system provided a daily goal that was set as a 20% increment of the participants' baseline step counts. We asked the participants to select their typical daily step count values, among the options of 2,500, 5,000, 7,500, 10,000, and 12,500 steps. Participants saw three different scenarios with varying margin types: Scenario 1: no margin, Scenario 2: fixed margin (5% of the baseline steps), and Scenario 3: random margin (differing everyday with the average value of 5% of the baseline). We explained that the results would be evaluated based on the mission.

In addition, a hypothetical 14-day step counts result was given to the participants, with the supposition that these were their own step records. We made virtual step count records in 2.5% intervals, from 105% to 135% of the baseline and shuffled their sequence. The same set of records was given in all three scenarios. Table 1 shows an example of step counts and margin given for scenario 3 when the participant's baseline was 10,000 steps. In the case of scenario 1 and 2, we provided the same step counts as scenario 3, but the margin was given differently. (scenario 1: 0 step, scenario 2: 500 steps every day). There was no mock dashboard in this hypothetical setting, and we only presented step count and margin information to avoid confusion.

All participants saw three hypothetical scenarios with different margin settings, and were asked to rate their perceived goal achievement levels. Based on the vignette, we asked three main questions to the participants. First, the participants were asked to rate daily step counts in each scenario on a 7–point Likert Scale ranging from completely unsuccessful (1) to moderate (4) to completely successful (7) in terms of goal achievement. Then, for the

	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7
Steps	12,500	11,750	12,250	11,250	12,000	12,750	11,500
Margin(S3)	400	1,000	300	700	0	600	500
	Day 8	**Day 9**	**Day 10**	**Day 11**	**Day 12**	**Day 13**	**Day 14**
Steps	13,250	11,000	13,000	10,500	12,000	13,500	10,750
Margin(S3)	400	1,000	300	700	0	600	500

Table 1: An example of step counts and margin of 14-day records in scenario 3 (i.e., random margin) when the baseline is 10,000 steps. Note that the goal is 20% increment from the baseline, which is 12,000 steps in this case.

two scenarios with a margin, we asked them whether their goal would be changed into the condition of mission success or remain as the goal provided by the coaching system. After that, we asked them to indicate the appropriate size of the margin in terms of supporting users to achieve the goal. Our study was approved by the university's Institutional Review Board (IRB).

4.1.2 Experiment Design and Data Analysis. We conducted an online survey-based experiment via a professional survey company with 500 participants (248 women; age: M = 43.7, SD = 13.4). In the survey, the order of scenario 2 (fixed margin) and scenario 3 (random margin) was randomized in order to minimize possible order effects. The participants were compensated with approximately 4.8 USD afterwards.

From the survey, we first analyzed how the participants rated the outcome across the three scenarios, particularly focusing on the difference in ratings when the outcome was within the good enough zone. Then, we counted the number of participants who thought their target goal had changed to reaching the good-enough zone and then examined what the appropriate size of margin to support the goal achievement would be.

4.2 Results

4.2.1 Goal Achievement Assessment. Figure 3 shows the self assessment results on the number of steps based on the goal. Here, we present our results in a ascending order with the lowest step count (i.e. 105% of the baseline steps) on the left side and the highest step count (i.e. 135% of the baseline steps) on the right. Participants rated all the days that were at or above the goal (i.e. 120% of their baseline steps) as 5 or more points out of 7. In each scenario, a descending trend was found in the participant's evaluation as the step count decreased. The minimum rating among the goal achievement cases was found on the days where the step count matched the goal *exactly*. We had two such "exactly matched" days (Day 5 and Day 12) among 14 days; the average assessment scores for Day 5 were 5.34, 5.32, and 5.31 for scenario 1, scenario 2, and scenario 3, respectively, whereas those for Day 12 were 5.34, 5.32, and 5.25, respectively. When the number of steps was below the goal, the participants' evaluation dropped drastically, and continued to decrease as the step counts became smaller.

We first analyzed how participants rated step counts within the different margin settings. The daily results of the hypothetical 14-day step counts were classified into one of the following categories; (1) goal achieved, (2) mission achieved, and (3) both not achieved. Then for each category, we conducted the one-way repeated measures ANOVA to see how participants' ratings differed across the margin settings (i.e., no, fixed, and random margin). From the analysis, there was a significant difference among conditions in the 'mission achieved' category ($F(2,1497) = 12.44$, $p < .001$) but not in the other two categories, which revealed weak or no difference (goal achieved: $F(2,1497) = 3.25$, $p = .04$, both not achieved: $F(2,1497) = 2.32$, $p = .10$). Moreover, the post-hoc pairwise Bonferroni correction method for the mission achieved category showed significant differences between when there was a margin and not (no vs. fixed margin: $p = .004$, no vs. random margin: $p < .001$).

We further analyzed whether there was a difference among conditions for *each* day. We found a statistically significant difference in assessment scores for each day in the 'mission achieved' category (Day 9, 7, 2, which are denoted as A, B, C in Figure 3, respectively) but not for any other day (i.e., days in the 'goal achieved' or 'both not achieved' conditions). As expected, the ratings for Days 9, 7, and 2 were higher in conditions with a margin. We hypothesize that this difference is due to the presence of the good-enough zone created by the margin, which led users to give a more positive rating for those days.

To explore the pattern found in margin area, we first examined two cases (Day 2 and Day 7) where the value fell within the margin range for both scenarios 2 and 3; the number of steps for these cases were Day 2 = 117.5% and Day 7 = 115% of the baseline behavior. In the case of 117.5%, participants rated their performance as 4.14 on average (SD = 1.39) when there was no margin given. The rating increased to 4.45 (SD = 1.29) and 4.43 (SD = 1.45) on average for the fixed and random margin, respectively. The one-way ANOVA test showed a statistically significant difference in the participants' assessment ($F(2,1497) = 7.77$, $p < .001$) across conditions. Moreover, the post-hoc pairwise Bonferroni correction method also confirmed that the differences between the scenario with no margin and both the fixed margin ($t(998) = 3.65$, $p < .001$) and random margin scenario ($t(998) = 3.16$, $p = .005$) resulted to be significant. There was no significant difference ($p = 1$) between the random margin and fixed margin scenario. A similar trend was found for Day 7 (115% of the baseline): the average rating in the scenario without a margin (M = 3.81, SD = 1.48) was significantly lower than the average ratings in the fixed margin (M = 4.29, SD = 1.43) and random margin (M = 4.34, SD = 1.49). Not only was there a noticeable increase in the ratings, but the change in average rating from below 4 points to over 4 points was especially meaningful. The cutoff point of success in our survey was 4 points, as ratings above 4 were considered as success and ratings below 4 as failure. This meant that the same step count could be perceived as success or failure based on the existence of a margin.

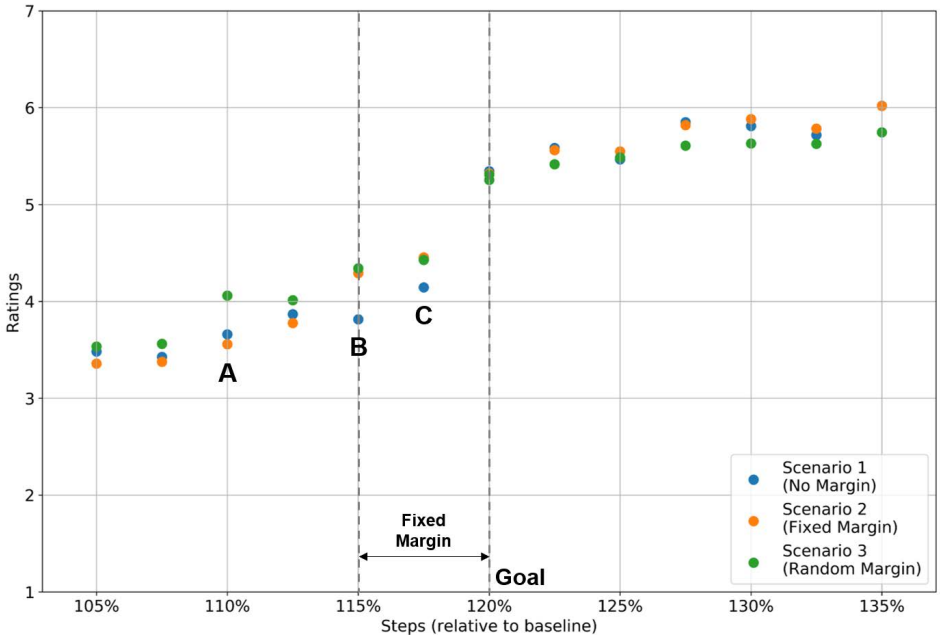


Figure 3: Self-assessment results on the step counts towards the goal. The three scenarios that showed statistically significant difference were labeled as A, B, and C in the figure. The step counts in the latter two cases (B and C) fell within the margin range for scenarios 2 and 3 whereas those in the first one (A) were within the good-enough zone for only scenario 3.

In addition, there was another interesting case (Day 9) where the step counts were within the good-enough zone for only the scenario with a random margin. This particular instance, which was 110% of the baseline, was considered as mission failure in the fixed margin scenario, since it did not fall in the 5% margin region. However, it was counted as mission success in the random margin scenario because a 10% margin was given for that day. Participants rated the outcome superiorly in the random margin scenario (M = 4.06, SD = 1.53) compared to in the no margin scenario (M = 3.66, SD = 1.54) and in the fixed margin scenario (M = 3.56, SD = 1.53). Consistent with the observations from Day 2 and Day 7 cases, participants tended to evaluate the outcome in a positive way when the steps were in the good-enough zone, and they even assessed the result as higher than 4 points. The difference was statistically significant through the ANOVA analysis ($F(2,1498) = 14.95$, $p < .001$), as well as the post-hoc Bonferroni correction method for both no margin ($t(998) = 5.75$, $p < .001$) and fixed margin scenario ($t(998) = 4.12$, $p < .001$) compared to the random margin scenario.

Regarding the result, we observed similar results when the data were normalized for each individual to adjust the individual differences. After the min-max normalization on the participants' ratings, the three days showed statistically significant differences in the participants' assessments when conducting the one-way ANOVA (Day 2: $F = 6.26$, $p = .002$, Day 7: $F = 21.27$, $p < .001$, Day 9: $F = 17.95$, $p < .001$). Also, there was a lack of explanation on why the random margin showed the lowest ratings on the rightmost 7 days from Figure 3. It appeared that the participants focused more on the good-enough days, and the goal-achieved days might not have stood out as much.

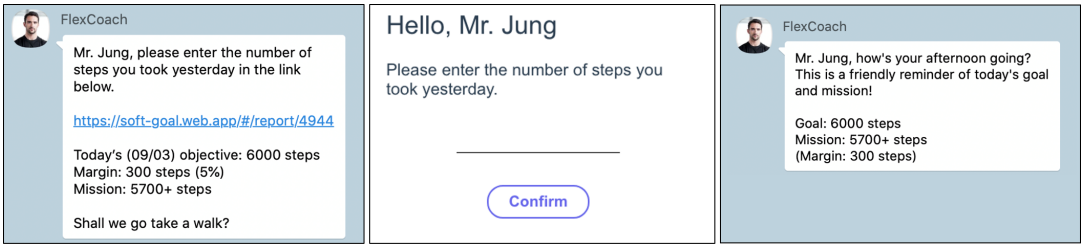
4.2.2 Goal Perception Changes. The majority of the participants responded that their target goal would still be the goal set by the coach even when a margin was given. Among the participants, 58.2% answered that their goals would not be changed from the original goals even though a fixed margin was provided. In the scenario with a random margin, 57.4% of the participants showed the same responses. We posit that participants' responses on goal perception could depend on the margin size. More than 70% of participants (n = 356 out of 500) said that an appropriate margin size would be 0–10% of their baseline steps. Specifically, 38.0% of participants chose 0–5% of their baseline as a proper margin and 33.2% of them chose 5–10%. The other options were chosen in order of 10–15% (13.2%), 15–20% (5.4%), and >20% (10.2%).

5 STUDY 2: FIELD EXPERIMENT

5.1 Methods

5.1.1 System Design. To understand real-world user experiences, we designed FlexCoach, a coaching system that aims to improve the physical activity of users by setting a goal for daily step counts along with a 'mission', a margin-based evaluation criteria. This mission distinguishes FlexCoach from existing pedometer applications that simply check whether the number of steps is above the goal.

FlexCoach is composed of two main parts: (1) activity data collection and (2) feedback on the user's outcome. Activity data collection is composed of a health application that collects step counts automatically and a web page through which users report their own steps every day. We used Samsung Health as a step counts collector since this application is widely used among Android users with more than a billion downloads and is easy to use for goal setting



(a) Morning call (a) Steps input page (c) Afternoon reminder

Figure 4: User Interfaces of FlexCoach: Activity data collection

as well as checking daily steps. In addition, the number of steps is shown in real-time in the notification drawer so that users can check their step counts by simply swiping the status bar.

We made a web page to collect the users' step counts from the previous day, and the collected records are then sent to the coaching system. To report the step counts, users check their previous day's steps in Samsung Health and manually enter the value on the web page. For the users' convenience, FlexCoach sends a message with a link to this web page through KakaoTalk, the most widely used messenger application in South Korea. The message is sent every morning to inform the user of their goal, margin, and mission for the day, and includes the link for entering the step counts (Figure 4.(a)). As the users input the step counts, the number is stored in a server with the date and user ID.

The user's activity is evaluated in the second part (i.e., feedback on the outcome) based on the mission, which applies a margin on the user's goal. After the users log their step counts from the day before, the server returns a daily report with (1) a badge indicating the mission success or failure, (2) a message about the result with short feedback, and (3) detailed information about the result (Figure 5). Since there are three possible results from the assessment (i.e., (1) mission success with goal achievement, (2) mission success within the good-enough zone, and (3) mission failure), we designed a different badge and message for each case.

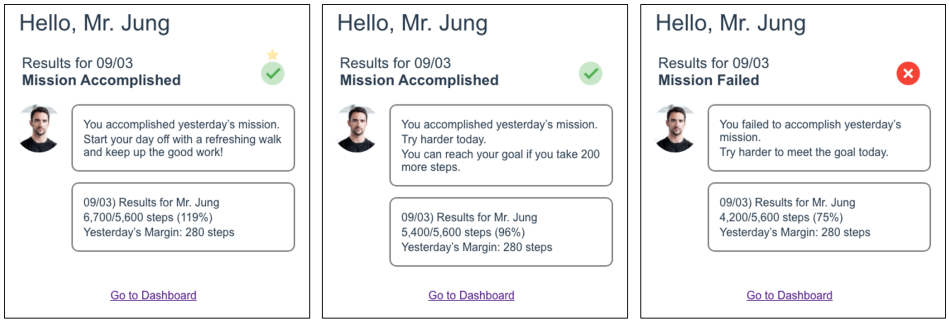
There are two types of badges: a checkmark on a green background for mission success and an X mark on a red background for mission failure. To remind the users of the original goal and encourage them to reach it, we added a small yellow star above the checkmark when they achieved the goal. Additionally, a message was sent to the users to remind them to check the result and what they should do to meet the goal. If the users succeeded in the mission *and* achieved the goal on the previous day, FlexCoach sends a message such as "You accomplished yesterday's mission. Start your day off with a refreshing walk and keep up the good work!" If the users failed to reach the goal, it says, "You failed to accomplish yesterday's mission. Try harder to meet the goal today." If the users reached the good-enough zone, then the FlexCoach says "You accomplished yesterday's mission. Try harder today. You can reach your goal if you take 200 more steps." For the consistency of interacting with FlexCoach, we designed the daily report page to resemble the instant messenger with the coach's profile picture and the placement of messages.

FlexCoach offers a dashboard web page showing the weekly progress as shown in Figure 6. The users can browse through different days to review their previous daily results. In particular, the dashboard contains a gauge chart, which is composed of two different background colors, dark and light orange, to indicate the margin area. For example, when the users complete the mission successfully, the progress bar is shown in green, and the end of the bar would be placed within the light orange area. When the users fail the mission, the progress bar is shown in red, and the result is within the dark orange area.

5.1.2 Participants. We recruited 57 participants (30 women; age: M = 23.14, SD = 7.41) for the field experiment by posting flyers on the online community and Facebook channel of a large university. The inclusion criteria were that the participant should be an Android user, within the range of 19 to 64 years old, and motivated enough to increase their physical activity during the experiment. To assess whether the participants had an intention to walk more, we utilized the Transtheoretical Model (TTM) of behavior change [48] and excluded those who had no interest at all (precontemplation stage) or already had their own exercise plans (maintenance stage).

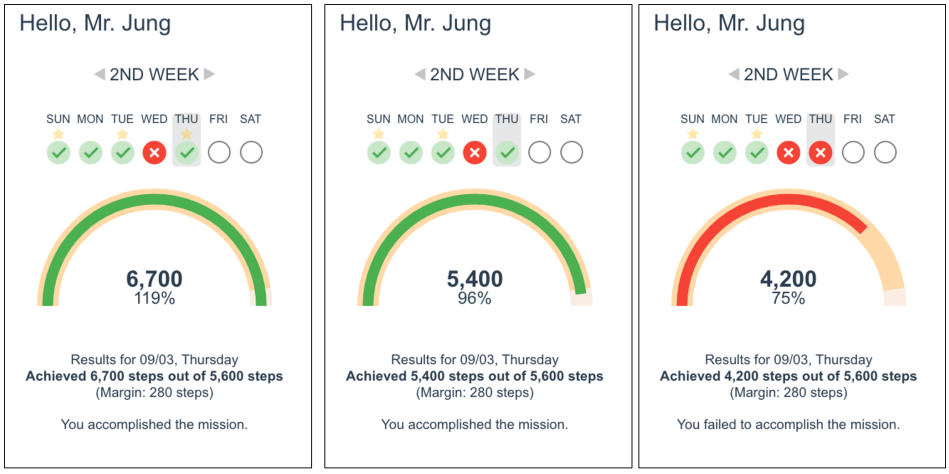
5.1.3 Study procedure. Before we began the experiment, we first collected the participants' baseline step count since FlexCoach set the goal and margin based on that value. For participants who used step count applications, their average number of steps over the last 20 days was used as their baseline. If a participant did not have this data, we collected 5-day step counts with a pedometer application after asking them to behave as usual. After the establishment of the baseline for all participants, we randomly assigned the participants into three groups (i.e. groups with no margin (control), fixed margin, and random margin). We conducted the experiment using a between-group design, while the mean values of age, TTM stage (2: contemplation, 3: preparation, 4: action), and baseline steps were similar from one another. In addition, each participant's goal was calculated by multiplying the baseline by 120% (i.e., 20% increment) then rounding the numbers to the nearest hundreds place.

We had online introduction sessions for each group where we explained how FlexCoach works. For the no margin (control) group, no information on margin was given, and the participants were only told that they would be evaluated based on whether they met the goal. They could also receive two types of badges; the red X mark and the green checkmark with a star above. To participants in the two margin groups, we explained how a margin works and



(a) Goal achieved (b) Good-enough zone (c) Mission failed

Figure 5: User Interfaces of FlexCoach: Daily report



(a) Goal achieved (b) Good-enough zone (c) Mission failed

Figure 6: User Interfaces of FlexCoach: Dashboard

how it would be given to them. For the random margin group, we added that the margin would be provided with an average value that is 5% of the baseline, but the exact value would be announced each morning.

The field experiment was conducted for 10 days. During this period, participants received a message from FlexCoach every morning regarding their goal, margin, and mission for that day, and a reminder with the same content in the afternoon. Participants repeated the following steps: reporting the previous day's step counts, and checking the feedback message and the dashboard. Regarding the step counts, we explained that the participants' records are stored at Samsung Health, and they will be checked to avoid possible issues from self-report. We did not find any participant who falsely reported the record. After the experiment period, we conducted a post-survey to see what their goal was during the experiment and what they viewed as an appropriate size of the margin. We conducted semi-structured online interviews with 19 participants from the two margin groups (10 from fixed margin, 9 from random margin), where we asked about their experience using FlexCoach.

5.1.4 Compensation. Participants were given approximately 30 USD for participating in the 10-day field experiment. They were also informed that they would receive 0.4 USD for each day they accomplished the daily mission. Prior studies documented the positive effects of financial incentives on encouraging physical activities [19, 44]. As recent HCI studies explored financial incentives in systems design [3, 52], we provided these micro incentives, with an amount that is comparable to that in prior studies [19, 44]. Regarding this additional compensation, the possibility of giving it to participants who were in the control group might be less than that of the other groups. However, the issue was minimized by limiting the amount of the incentive; the additional compensation for the mission accomplishment was only 1.3% of the baseline payment, and most participants responded that additional incentives did not influence their motivation. Additionally, each participant was paid 8.5 USD for the interview.

5.1.5 Data Analysis. For the quantitative analysis, we first set three metrics – goal achievement rate, goal success days, and mission success days – to assess the participant's physical activity and compare the three groups' results. Goal achievement rate was calculated per

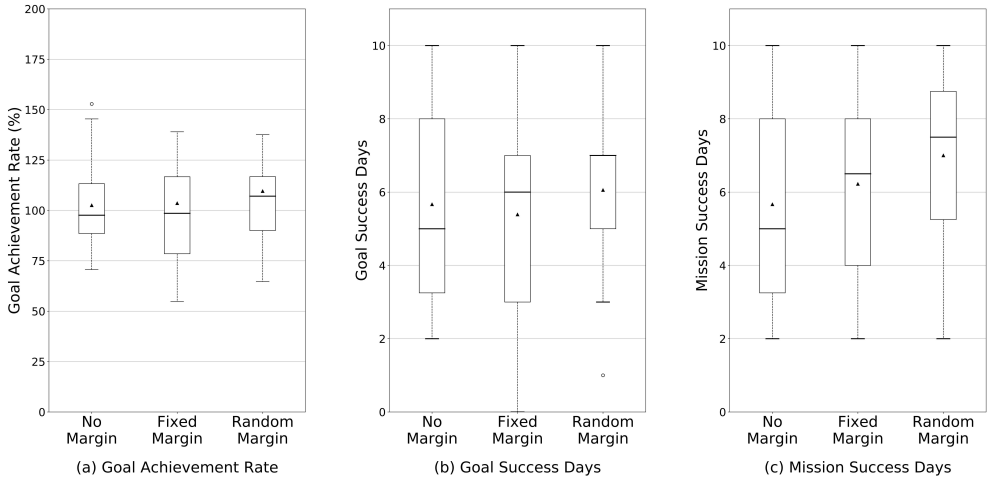


Figure 7: Physical activity statistics

day through dividing the total steps by the goal. The number of days the participants achieved their goals or completed the mission during the experiment period was defined as goal success days and mission success days, respectively. In addition, we conducted a post-hoc analysis on how many days each participant successfully achieved a certain level of steps. This evaluation method is commonly used in statistical process control. We analyzed the results for three levels - 95%, 90%, and 83% of the goal - to see whether the participants’ performance was ‘in control’ state when using a boundary of fixed margin (5%), maximum margin in the study (10%), and baseline steps (17%, which is (120-100)/120). This analysis aimed to see whether the margin may influence not only the goal achievement but also the consistent behavior near the goal, in terms of quality control.

Before analyzing the data, we excluded the data of three participants whose average daily steps were more than five times or less than one fifth of their goal value (i.e., the baseline did not represent their usual behavior well) or who did not check the messages from FlexCoach and reported their records too late. After the exclusion of outliers, 54 participants were equally assigned to three different groups (i.e. no margin, fixed margin, and random margin group) consisting of 7, 13, 9 women respectively. The participants were distributed to each group as the following: age (M = 22.22, 23.72, 23.89; SD = 3.49, 9.25, 8.95 respectively), TTM stage (M = 2.50, 2.39, 2.67; SD = 0.62, 0.70, 0.69 respectively), and baseline steps (M = 5998, 5744, 5891; SD = 3112.23, 3008.71, 3720.74 respectively). Regarding the interview data analysis, we transcribed the recorded data and conducted a content analysis on the following areas; (1) overall user experience on margin-based goal evaluation, (2) goal perception and behavioral engagement, (3) appropriate margin size, and (4) roles of badges. For each area, we reviewed the transcribed data a few times and collaboratively identified the major categories. The interview quotes of each participant were marked with the participant’s number (e.g., P07) and a letter indicating the margin condition (e.g., B=Baseline, F=Fixed, R=Random).

5.2 Results

5.2.1 Limited Influence of Margin on Physical Activity. We analyzed the goal achievement rate, goal success days, and mission success days across different groups (Figure 7). We found that there were no significant differences in the three metrics across different groups. Since the data were not normally distributed, we conducted Kruskal-Wallis tests. The p-value of each metric (i.e., goal achievement rate, goal success days, and mission success days) was given as .72, .78, and .35, respectively.

The goal achievement rate was slightly higher with a margin, but no significant difference was observed. Participants in the control group achieved 102.55% of the goal on average (SD = 21.93, Median = 97.62%) whereas those in the fixed and random margin group reached 103.56% (SD = 41.63, Median = 98.57%) and 109.58% (SD = 35.36, Median = 107.07%), respectively. Goal success days had similar tendencies, but no significant difference was observed. The number of days that participants achieved the goals during the 10-days experiment period was 6.06 days on average (SD = 2.29). However, participants in the fixed margin group reached the goal 5.39 days on average (SD = 2.91), which was slightly lower than the control group’s 5.67 days (SD = 2.72). Considering the sample size and the distribution of the samples in each group, we also compared the median of the goal success days. The result showed that the random margin group performed the best in goal success days, followed by the fixed margin and the control group. (7, 6, 5 days, respectively)

When it comes to the mission success days, there was only a slight difference (but not significantly different) among the groups owing to the participants who only reached the good-enough zone. For the fixed margin group, participants completed the mission

Group	Goal Count	95% Count	90% Count	83% Count
No Margin	5.67 (2.72)	5.89 (2.50)	6.28 (2.37)	6.61 (2.30)
Fixed Margin	5.39 (2.91)	6.22 (2.67)	6.44 (2.64)	6.83 (2.43)
Random Margin	6.06 (2.29)	6.94 (2.13)	7.22 (2.07)	7.44 (1.85)

Table 2: Days of performance 'in control'

6.22 days on average (SD = 2.67), and the random margin group succeeded for 7 days on average (SD = 2.28). Differences between the days of mission success and those of goal success showed the days when the participants completed the mission but failed to reach the goal. This value was 0.83 and 0.94 days for the fixed and the random margin group, respectively.

In terms of whether the outcome is 'in control', we compared the number of success days across different groups (Table 2). We did not find any statistically significant differences, but there was a tendency that groups with a margin had slightly more success days compared to the control group. When the criterion of success was to achieve 100% of the goal, the smallest value (5.4 days on average) was found in the fixed margin group as mentioned earlier. For evaluations with a lower standard of success (e.g., 95%, 90%, or 83% of the goal), we found that the random margin had the highest number of successful days on average, followed by fixed margin, then the control group. For instance, when we counted the outcomes of more than 90% of the goal as success, the participants in the random margin group succeeded for 7.22 days whereas the fixed margin group and the control group did for 6.44 and 6.28 days, respectively. In addition, this trend continued when we counted any record above the usual step counts (i.e. 83% of the goal) as a success. Random margin condition had an average of 7.44 days, Fixed margin had 6.83 days, and the control group had 6.61 days. Further studies with larger sample sizes are required to find whether margin-enabled evaluation results in a high number of 'in control' and 'success' cases.

5.2.2 Overall User Experiences on Margin-based Goal Evaluation. Participants named the reduced stress, the recognition of their effort, and the lower entry point after a relapse as the main benefits of having a margin as they strived to reach the goal. The primary effect of the margin was reducing stress and pressure during the goal achievement process as the margin provided a "safety net" for them. When only the goal was given, participants felt a lot of pressure to reach the given number of steps as P17-F reflected, "*Although I knew it was nothing more than the goal, I felt like I must achieve the goal, so it was kinda challenging for me.*" However, when a margin was given around the goal, participants were "*less worried about failing the mission. I could enjoy the activity itself rather than being obsessed with the number of steps*" (P01-R). P12-R similarly stated, "*I was less anxious about failing the mission because I thought of the margin as having a 'safe zone' not to be evaluated as a failure.*"

Next, they highly appreciated that they were recognized for their efforts in their interactions with the coach. P09-R shared, "*I liked how the coach evaluated my partial accomplishment as a success because I would have been demotivated by the fact that I failed to achieve the goal of the dichotomous evaluation.*" Thirdly, participants answered that they did not give up completely due to the existence of the margin, thinking 'Let's achieve at least the minimum steps of the mission success.' They believed that the margin would help them exercise steadily in the long term and "*help foster a behavior change because having the margin would decrease the amount of failure experienced when a person is demotivated or too busy to achieve the goal.*" (P13-F).

Two participants mentioned that the presence of the margin may interfere with the process of achieving the goal since they were content with the good-enough zone and perceived it as one of the goals. However for the majority of the participants, the lowered pressure to reach the goal did not lead to a lowered self-expectation or motivation: "*I still tried my best to achieve the goal even if I felt less pressure because of the margin*" (P10-R). Thus, participants continuously strived to reach the original goal contrasting our initial concern that participants' internalized goals could change to entering the good-enough zone instead of achieving the goal. The next section further explains the unchanged perception of the target goal.

5.2.3 Goal Perception and Behavioral Engagement. Throughout the experiment, participants mostly perceived the initial goal the coach assigned them (i.e., 120% of their usual steps) as their own goal, even though they were aware of how the margin-based evaluation works. In the post-experiment survey, participants who answered that they mainly focused on the assigned goal by the coach were 78.9% and 94.4% from the fixed and random margin, respectively. Participants mentioned their satisfaction when they reached the goal as the main reason why they targeted the goal given by the coach instead of the good-enough zone. Even though they were given a positive evaluation and additional compensation once within the margin, the majority of the participants "*didn't think of the good-enough zone as a complete success*" (P16-F). As such, they aimed to reach the goal once they were within the margin since only a little more effort was required: "*I only had a margin of 275 steps, so I would rather achieve the goal than achieving the good-enough zone*" (P07-F).

Interestingly, some participants intentionally set their goal high so that they could at least achieve the good-enough zone. P10-R explained that "*I always attempted to achieve the goal because, in that way, even if something came up, I could at least achieve the mission.*" In some cases, however, participants settled for falling within the good-enough zone rather than achieving the goal as they did not find the activity to be very important. P06-R hypothesized that "*I would have put more effort into goals if the goal was important to me like applying for colleges. I realized this goal of walking more was not that important to me, so I decided to achieve the mission, not the goal.*"

Most of the participants were eager to reach their goal and even tried many different ways such as reducing transportation usage, taking the stairs more, and taking a new, longer route to walk more to reach the goal. Then they checked their step counts in the evening and tried to achieve the goal when possible if they have

not yet. P13-F followed this pattern and said "*I usually checked up on my step counts after I got off from work. I went for a walk or run if I have not achieved the goal.*" However when participants realized there were too many steps remaining to achieve the goal by the end of the day, they began to take account of the margin and modified their target into entering the good-enough zone although their initial target was the coach-provided goal: "*I didn't have enough time to walk at the end of the day to achieve the goal, so I shot for the good-enough zone*" (P16-F). P09-R also "*looked again at the message the coach sent to check the margin when [they] figured it would be hard to achieve the goal.*" In this situation, participants started to take into account the provided margin, which they did not value high or important initially, and closely tracked the number of steps left. Some participants adjusted their goals due to the physical limitations they faced while achieving them, such as fatigue after work or pain while walking. P01-R explained, "*I wanted to do my best, but I had to give up achieving the goal and use the margin because my feet were so sore.*"

The main reason they still tried to achieve the good-enough zone even under unavoidable circumstances was that they at least wanted to feel gratified by their results and be acknowledged by the coach for what they have done. Though it was not the goal they had tried to achieve, they "*worked hard to get in the good-enough zone to feel a sense of accomplishment when it was nearly impossible to achieve the goal*" (P08-F). Participants also noticed that the coach evaluated their performances, so they wanted to be recognized by the system as P09-R stated, "*I didn't give up because I knew the coach was tracking my record.*" Similarly, P01-R said "*I tried to reach the good-enough zone because the coach recognized my effort even in the difficult situation.*"

Participants thought being in the good-enough area was a partial success ("*Though the coach evaluated my performance as a success, I did not feel like it was a complete success since the dashboard indicated 94% instead of 100%.*" - P02-R). However, they "*preferred the coach's evaluation criterion compared to that of other systems in that it acknowledged how hard they tried to reach the goal even in the tough situations*" (P12-R). P08-F shared their sentiment that "*I would feel bad about myself if I couldn't achieve the goal even with my best shot, but I felt like I was getting a pat on my back for my effort.*" Along the same line, P04-F said that "*If the system evaluated the good-enough zone as a failure, I would have felt discouraged and worried about failing on the following days.*" Thus when the participants' outcome was evaluated as 'mission success,' they also positively evaluated their results and maintained their motivation to reach the goal.

5.2.4 Appropriate Margin Size. According to the post-experiment survey data, most participants responded that 5–10% of their goal would be an appropriate size of the margin. Many participants in both margin conditions (42.1% in fixed margin, 55.6% in random margin) chose 5–10% of the goal to be the most appropriate size of a margin. They hypothesized that if the margin size was too large or small, it would hinder them from achieving the goal. More specifically, when the margin was too broad, they would be more likely to settle for the mission within the margin instead of reaching the goal as a significant amount of extra effort would be needed. P19-F explained that having a wider margin "*would certainly lower the burden of pressure, but I think there will be times when I start taking the margin for granted and become lazy.*" Similarly, P06-R said "*I may think I don't have to achieve the goal anymore when I am already in the good-enough zone.*" On the other hand, participants commented that "*if the given margin is too small, it wouldn't feel very different from the traditional binary evaluation of success or failure. So I think the pressure of achieving the goal would come back*" (P01-R). Thus, having an appropriate margin size was found to be an important factor in encouraging people to accomplish their goals.

Participants had a variety of different opinions on receiving random margins from the coach. Given a different margin every day, one participant said it "*felt like a real person was training me, and it was fun. It was like a real coach reducing the number of repetitions when I say it is too hard during the personal training*" (P09-R). Another participant "*enjoyed having a randomized margin because [they] paid more attention to what the coach had to say every day*" (P10-R). Thus, receiving a random margin gave participants the impression of interacting with a human coach and made their interactions with FlexCoach more entertaining. In addition, many participants reported that they focused mainly on their goal instead of the margin area because they could not keep track of the varying margins. P14-R elaborated that "*If I had a fixed margin, the value of the margin would have stuck in my head the whole time, which would have lowered my ultimate goal to the good-enough zone.*"

However, they felt it was important to inform users on how the coaching system might have decided the randomized margin. People even suggested different ways of deciding the margin size: "*I wasn't sure what determined the margin I had on that day. It would be more understandable if the coach adjusted the margin depending on physical abilities.*" (P02-R). Another participant suggested giving out "*a margin depending on what kind of days I was having. Maybe I had a hectic day or spent most of a day seated*" (P15-R).

5.2.5 Roles of Badges. Badges played a significant role in signaling success and achievement as participants checked on their results via daily reports and dashboards. The visually distinct badges for three outcomes (i.e. goal achievement, mission success, mission failure) provided a visual summary of their achievements while increasing their motivation to reach the goal. Participants tended to walk more to earn a green checkmark because they "*felt pretty bad about getting a [red] X mark*" when they failed to achieve the goal (P06-R). The design choice of indicating the good-enough zone with a checkmark also had an impact on how they perceived their outcome as P02-R remarked, "*I didn't think I did too bad on it because I technically did achieve the mission. The badges motivated me to walk more because I didn't want to end my success-streak.*"

While viewing mission success positively, the majority of participants commented that differentiating badges between achieving the goal and achieving the good-enough zone with a star was still effective. One person noted that "*I could easily recognize the days I achieved the goal by looking at my badges, and I was content with that*" (P01-R). Others even responded that they did not think of a checkmark (without a star) as a complete success, so they tried to get a badge with a star, if possible. They mentioned that having a star was "visually different" leading them to feel proud of themselves (P10-R), and that they enjoyed the process of collecting badges (e.g., "*I like collecting things in general, so I wanted to feel*

some type of satisfaction by achieving the goal and receiving the badge with a star" (P17-F)).

6 DISCUSSION

6.1 Summary of Key Findings

In this study, we introduced a 'margin-based goal evaluation framework' and explored how users would perceive and react to this concept. We aimed to anchor the user's goal without the confusion resulting from multiple goals, and this differentiates our approach from previous works using dual goals or a secondary goal for behavioral change. This seems in line with our result showing that most participants perceived the goals given by FlexCoach as their own regardless of the presence of a margin. If we use the traditional dual goal concept set by the users, the users still would not be free from binary judgment (i.e., a hard threshold dividing success and failure) in each goal and could set a secondary goal that is arbitrary or has a large gap from the primary one. However, our evaluation criterion recognizes a user's effort based on the closeness to each goal, and it offers a system-driven, 'constrained freedom' by controlling the flexibility of evaluation. In this sense, our approach could be interpreted as a digital commitment device for behavior changes [27, 28], where a user sets a behavior change goal, but its evaluation is delegated to the system.

Our results clearly showed the feasibility of margin-enabled coaching for behavioral changes. The key finding is that there was no significant difference in the participants' perception of the goal even with a margin. While the margin did not play a primary role in the goal achievement process, it served as an auxiliary component that lowers goal barriers, by provisioning a psychological buffer [55]. In Study 1, less than half of the participants reported the change in goal perception (41.8% for fixed margin, and 42.6% for random margin). In Study 2, the change in goal perception was significantly lower: 21.1% for fixed margin, and 5.6% for random margin. Our interview results showed that most participants did not lower their expectations, and they tried to reach the good-enough zone only as a contingency plan. In Study 2, the real-world usage of margin-based coaching for 10 days helped our participants to better understand the margin size. FlexCoach repeatedly explained their goal, margin, and mission. This repetitive explanation could have helped them to reinforce the distinction between the goal and the mission.

Margin-enabled coaching provided several positive user experiences to our participants: reducing the pressure on meeting the goal, preventing them from giving up completely, and motivating them to keep continuing by recognizing their efforts. Many participants mentioned that offering a margin is like a human coach who makes a flexible evaluation considering the participants' capability or situation. In other words, the 'human-like' characteristic was found from the existence of the margin itself, rather than something that needed to be added or implemented for the effectiveness of the margin-based evaluation. This human-likeness was one of the main reasons for stress reduction. A prior study reported that human-like features could possibly increase social engagement with an interactive system [38]. It would be beneficial to incorporate human-like flexibility in a goal setting process [50], which could possibly increase trust in a virtual coach as a social actor for behavioral changes. Furthermore, margin-enabled coaching lowered the burden of negative evaluations on goal failures. Negative feedback is known as a major demotivator of self-tracking tool usage [5, 16]. A margin can possibly lower the chances of goal failures (or prevent potential lapses), which helps to reduce the possibility of goal abandonment. Interestingly, positive feedback of success could induce another strive for seeking goal achievement.

6.2 Design Implications

As the first step toward examining the feasibility of margin-based evaluation, our work considered only one predefined margin rate (5%) associated with physical activity goals and two different margin allocation strategies (fixed and random margin). We leverage the existing goal setting dimensions [50] to further explore the design space of margin-enabled coaching: (1) goal properties (i.e., difficulty, specificity, and proximity), (2) process components (i.e., progress feedback, and achievement rewards), and (3) goal setting sources (i.e., self-set, assigned, guided, or group-set). In the following section, we discuss the design of margin-enabled coaching: (1) margin selection strategies, (2) margin adaptation opportunities, and (3) progress-centered coaching with margins.

6.2.1 Exploring Margin Selection Strategies. The first step of margin enabled coaching is the selection of a proper margin size. Considering the goal setting dimensions [50], we argue that margin selection should carefully consider two aspects of goal setting dimensions: goal setting properties and sources. A margin is mostly useful when a goal is specific (e.g., reaching 7,000 steps per day) and thus, we mainly consider the difficulty and proximity of goals.

Prior goal setting studies showed that too easy or too challenging goals are less effective [43]; for example, stretch goals may cause side-effects such as unethical or risky behaviors and psychological costs of goal failures. In our work, we only considered a moderate difficulty level (i.e., 20% increment from baseline step counts). We can show that difficulty level selection (e.g., easy, moderate, or difficult goal) is closely related to associated margin selection (e.g., narrow, moderate, or wide margin). Various combinations of goal-margin selection are feasible, and thus, existing goal setting findings may not be directly applicable to such settings. For example, a difficult goal could be considered as less challenging if a wide margin is given.

Let us say that a baseline state is denoted as B, and a user's goal as G with α% improvement: i.e., $G = B*(1+\alpha)$. This equation tells us that the gap between the goal and the baseline condition is simply given as $G - B = B * \alpha$. Let M denote the margin. We assume that this gap should be greater than an associated margin; otherwise, a user can go below the baseline condition, which may nullify goal setting. This permissible margin assumption can be represented as $B-G = B*\alpha \geq M$. This equation shows that the target improvement factor α from the baseline B is the key determinant of permissible margin size. We can further define an additional factor β to control a possible range of M as in $M = \beta * B\alpha$—the higher the β the larger is the size of a margin. Thus, we can control two parameters, i.e., α (improvement factor) and β (margin factor) for goal setting and associated margin selection, respectively.

Along with these variable selections, we can additionally consider temporal proximity in goal properties and goal setting sources (e.g., self, others, and guided). In terms of temporal proximity, it is possible to consider setting different periods for goal achievement (e.g., achieving 5 minutes of physical activity per hour; or achieving 5 hours of physical activity per week). Again, we can easily show that our margin selection model can be directly applied to diverse temporal proximity cases. For goal setting sources, it is interesting to show that our model allows us to systematically explore possible goal-margin ranges, by varying α and β values. Given that users prefer collaborative goal setting [13], this can be extended to margin selection as well. A margin-enabled coaching system can recommend users a set of possible goals and permissible margins associated with a chosen goal, and a user can choose a preferred goal and margin.

6.2.2 Seeking Margin Adaptation Opportunities. As shown earlier, a less than optimal goal (e.g., too easy or too difficult) may be demotivating. A possible approach of dealing with this issue is to adaptively change goals based on a user's performance [1, 25]. For example, Konrad et al. [25] showed that for stress intervention, dynamically adjusting a user's goal based on a previous day's performance (i.e., increasing difficulty upon success, or decreasing difficulty upon failure) helps to maintain compliance with app usage, and this resulted in better stress reduction over time. In the random margin case, a different margin was assigned to the participants every day. Interestingly, some participants wondered whether the margin size was set based on their performance. In addition, most participants attributed human-like coaching to varying margin assignments. With margin-enabled coaching, it is possible to consider "margin adaptation" along with the goal adaptation. For example, a challenging goal or stretch goal is initially set, and a wide margin is permitted to encourage participants to try. After several rounds of "mission success" we then can decrement the margin, and this will help to gradually improve a user's overall performance. Furthermore, as in adaptive goal setting [1, 25], we can increment the margin upon successive mission failures. Unlike traditional approaches of goal adaptation where a goal is a moving target, in our case, a goal is fixed, but the margin is adaptively changing. Margin adaptation can be incorporated into goal adaptation while using different time scales. For example, a daily goal is adjusted on a weekly basis, and a daily margin can be adjusted daily (by considering a user's performance on a previous day). Instead of frequently changing goal targets, margin adaptation allows users to maintain the same goal for a fixed period of time, and yet offers flexibility in performance evaluation.

6.2.3 Progress-Centered Coaching with Margins. Margin-enabled coaching has two evaluation mechanisms: mission evaluation and goal achievement. Our experimental results showed that this separation did not influence a user's goal perception, thanks to "progress-centered coaching." We basically leveraged the anchoring effect in that despite mission success, participants were informed that the goal has not been achieved yet. In addition, goal achievement is presented differently via using an additional badge (a checkmark with a star). This difference in the presentation and framing of goal and mission distinguishes a goal of 10,000 steps with a margin of 500 steps from a goal of 9,500 steps. A goal process model [50] identified two critical components related to goal achievement: i.e., progress feedback and achievement rewards. We can configure progress feedback and reward assignments to promote progress-centered coaching. In our scenario, progress feedback was given, when we informed the current step counts and evaluated mission success and goal achievement. We further offered badges and financial incentives as achievement rewards. Our results clearly showed that it is important to anchor target goals when a system provides progress feedback. In addition, we can consider distributing a set of rewards over two evaluation criteria: mission success and goal achievement. For example, monetary rewards and a default badge can be given upon mission success. Additional recognition, such as a star badge and praising messages, can be provided.

6.3 Limitations

Our experiment results clearly showed the feasibility of margin-enabled coaching in that adding a margin did not harm the goal achievement. Despite a lack of a significant effect, we could find a possibility of managing the lapses from the behavior change with a permissible area when evaluating the outcome. The interview responses showed that the finding was not limited to the users' emotional change but also affected their strategies to reach the goal. This prompts for future work with a larger sample size and longer study period to systematically investigate the effect on margin-enabled coaching compared to other methods.

In addition, it is required to validate the effect of margin-based evaluation in comparison with different goal setting strategies for behavioral change. Our aim for this work was to gain a basic understanding of the new 'margin-based goal evaluation framework,' its effectiveness, and its user experience. Thus, we only used the traditional static, single-goal as the baseline for this study so that we could control confounding factors such as the number of goals and the agent of evaluation. However, tuning the various parameters as described in our discussion would be an interesting future work. We may also consider who sets the goal and margin (e.g., based on the experts' guidelines, by system with certain criteria, or by the users' own decision) to extend this study. After that, we could establish its benefits, and further examine its effectiveness compared to other superior approaches in the future.

7 CONCLUSION

We introduced a flexible evaluation system that allows some margin where a user's outcome would be considered as "good-enough" even though the user fails to reach the goal. Our goal was to explore how the margin-based evaluation affects the user's perception of her or his goal and goal achievement and how it influences the user experience of goal setting and achievement evaluation in the real world. Results from our study showed that the margin-base evaluation makes users evaluate themselves more positively. In addition, we found that the margin supports goal achievement by reducing negative emotions and psychological effects while anchoring them to strive toward their goal even if the evaluation becomes relaxed. In this sense, our approach demonstrated the potential benefits from separating the goal and the evaluation criteria, by helping users continue their effort even when they faced lapses and small failures during their behavior change. However, further studies are

required to investigate whether this concept influences the user behavior.

Margin-enabled coaching made the first step toward progress-centered, human-like coaching that can guide behavioral change more flexibly. We expect our evaluation system to be extended with several practical design implications such as margin selection and adaptation, and it could be studied further in the HCI field to better support the user's behavior change.

ACKNOWLEDGMENTS

This research was supported by the KAIST-KU Joint Research Center, KAIST, and by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the the Korea government (MSIT) (2020R1A4A1018774)

REFERENCES

[1] Marc A Adams, Jane C Hurley, Michael Todd, Nishat Bhuiyan, Catherine L Jarrett, Wesley J Tucker, Kevin E Hollingshead, and Siddhartha S Angadi. 2017. Adaptive goal setting and financial incentives: a 2× 2 factorial randomized controlled trial to increase adults' physical activity. *BMC public health* 17, 1 (2017), 1–16.

[2] Elena Agapie, Daniel Avrahami, and Jennifer Marlow. 2016. Staying the course: System-driven lapse management for supporting behavior change. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1072–1083.

[3] Nabil Alshurafa, Jayalakshmi Jain, Rawan Alharbi, Gleb Iakovlev, Bonnie Spring, and Angela Pfammatter. 2018. Is More Always Better? Discovering Incentivized mHealth Intervention Engagement Related to Health Behavior Trends. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 2, 4 (2018), 1–26.

[4] Daniel Aranki, Gao Xian Peh, Gregorij Kurillo, and Ruzena Bajcsy. 2018. The feasibility and usability of runningcoach: A remote coaching system for long-distance runners. *Sensors* 18, 1 (2018), 175.

[5] Christiane Attig and Thomas Franke. 2020. Abandonment of personal quantification: A review and empirical study investigating reasons for wearable activity tracking attrition. *Computers in Human Behavior* 102 (2020), 223–237.

[6] Christiane Atzmüller and Peter M Steiner. 2010. Experimental vignette studies in survey research. *Methodology European journal of research methods for the behavioral & social sciences.* 6, 3 (2010), 128–138.

[7] Timothy W Bickmore, Lisa Caruso, and Kerri Clough-Gorr. 2005. Acceptance and usability of a relational agent interface by urban older adults. In *CHI'05 extended abstracts on Human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 1212–1215.

[8] Dale S Bond, J Graham Thomas, Hollie A Raynor, Jon Moon, Jared Sieling, Jennifer Trautvetter, Tiffany Leblond, and Rena R Wing. 2014. B-MOBILE-A smartphone-based intervention to reduce sedentary time in overweight/obese individuals: a within-subjects experimental trial. *PloS one* 9, 6 (2014), e100821.

[9] Thamar JH Bovend'Eerdt, Rachel E Botell, and Derick T Wade. 2009. Writing SMART rehabilitation goals and achieving goal attainment scaling: a practical guide. *Clinical rehabilitation* 23, 4 (2009), 352–361.

[10] Scott A Cambo, Daniel Avrahami, and Matthew L Lee. 2017. BreakSense: Combining physiological and location sensing to promote mobility during work-breaks. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 3595–3607.

[11] Jessica R Cauchard, Jeremy Frey, Octavia Zahrt, Krister Johnson, Alia Crum, and James A Landay. 2019. The Positive Impact of Push vs Pull Progress Feedback: A 6-week Activity Tracking Study in the Wild. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–23.

[12] Woohyeok Choi, Sangkeun Park, Duyeon Kim, Youn-kyung Lim, and Uichin Lee. 2019. Multi-stage receptivity model for mobile just-in-time health intervention. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 2 (2019), 1–26.

[13] Sunny Consolvo, Predrag Klasnja, David W McDonald, and James A Landay. 2009. Goal-setting considerations for persuasive technologies that encourage physical activity. In *Proceedings of the 4th international Conference on Persuasive Technology*. Association for Computing Machinery, New York, NY, USA, 1–8.

[14] Sunny Consolvo, David W McDonald, Tammy Toscos, Mike Y Chen, Jon Froehlich, Beverly Harrison, Predrag Klasnja, Anthony LaMarca, Louis LeGrand, Ryan Libby, et al. 2008. Activity sensing in the wild: a field trial of ubifit garden. In *Proceedings of the SIGCHI conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 1797–1806.

[15] Andrew J Elliot. 1999. Approach and avoidance motivation and achievement goals. *Educational psychologist* 34, 3 (1999), 169–189.

[16] Daniel A Epstein, Monica Caraway, Chuck Johnston, An Ping, James Fogarty, and Sean A Munson. 2016. Beyond abandonment to next steps: understanding and designing for life after personal informatics tool use. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1109–1113.

[17] Paul A Estabrooks, Candace C Nelson, Stanley Xu, Diane King, Elizabeth A Bayliss, Bridget Gaglio, Paul A Nutting, and Russell E Glasgow. 2005. The frequency and behavioral outcomes of goal choices in the self-management of diabetes. *The Diabetes Educator* 31, 3 (2005), 391–400.

[18] Mitchell Gordon, Tim Althoff, and Jure Leskovec. 2019. Goal-setting and achievement in activity tracking apps: A case study Of MyFitnessPal. In *The World Wide Web Conference*. Association for Computing Machinery, New York, NY, USA, 571–582.

[19] Kristin A Harkins, Jeffrey T Kullgren, Scarlett L Bellamy, Jason Karlawish, and Karen Glanz. 2017. A trial of financial and social incentives to increase older adults' walking. *American journal of preventive medicine* 52, 5 (2017), e123–e130.

[20] Katja Herrmanny, Jürgen Ziegler, and Aysegül Dogangün. 2016. Supporting users in setting effective goals in activity tracking. In *International Conference on Persuasive Technology*. Springer, Springer International Publishing, Cham, 15–26.

[21] Graham Jones and Andrew Cale. 1997. Goal difficulty, anxiety and performance. *Ergonomics* 40, 3 (1997), 319–333.

[22] Inyeop Kim, Hwarang Goh, Nematjon Narziev, Youngtae Noh, and Uichin Lee. 2020. Understanding User Contexts and Coping Strategies for Context-Aware Phone Distraction Management System Design. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 4, Article 134 (Dec. 2020), 33 pages. https://doi.org/10.1145/3432213

[23] Jaejeung Kim, Hayoung Jung, Minsam Ko, and Uichin Lee. 2019. GoalKeeper: Exploring Interaction Lockout Mechanisms for Regulating Smartphone Use. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 1 (2019), 1–29.

[24] Rafal Kocielnik, Lillian Xiao, Daniel Avrahami, and Gary Hsieh. 2018. Reflection companion: A conversational system for engaging users in reflection on physical activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 1–26.

[25] Artie Konrad, Victoria Bellotti, Nicole Crenshaw, Simon Tucker, Les Nelson, Honglu Du, Peter Pirolli, and Steve Whittaker. 2015. Finding the adaptive sweet spot: Balancing compliance and achievement in automated stress reduction. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 3829–3838.

[26] Dominika Kwasnicka, Stephan U Dombrowski, Martin White, and Falko Sniehotta. 2016. Theoretical explanations for maintenance of behaviour change: a systematic review of behaviour theories. *Health psychology review* 10, 3 (2016), 277–296.

[27] Hyunsoo Lee, Auk Kim, Hwajung Hong, and Uichin Lee. 2021. Sticky Goals: Understanding Goal Commitments for Behavioral Changes in the Wild. In *In CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. ACM, New York, NY, USA.

[28] Hyunsoo Lee, Uichin Lee, and Hwajung Hong. 2019. Commitment devices in online behavior change support systems. In *Proceedings of Asian CHI Symposium 2019: Emerging HCI Research Collection*. Association for Computing Machinery, New York, NY, USA, 105–113.

[29] I-Min Lee, Eric J Shiroma, Felipe Lobelo, Pekka Puska, Steven N Blair, Peter T Katzmarzyk, Lancet Physical Activity Series Working Group, et al. 2012. Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. *The lancet* 380, 9838 (2012), 219–229.

[30] Uichin Lee, Kyungsik Han, Hyunsung Cho, Kyong-Mee Chung, Hwajung Hong, Sung-Ju Lee, Youngtae Noh, Sooyoung Park, and John M. Carroll. 2019. Intelligent positive computing with mobile, wearable, and IoT devices: Literature review and research directions. *Ad Hoc Networks* 83 (2019), 8 – 24.

[31] James J Lin, Lena Mamykina, Silvia Lindtner, Gregory Delajoux, and Henry B Strub. 2006. Fish'n'Steps: Encouraging physical activity with an interactive computer game. In *International conference on ubiquitous computing*. Springer, Springer-Verlag, Berlin, Heidelberg, 261–278.

[32] Edwin A Locke. 1996. Motivation through conscious goal setting. *Applied and preventive psychology* 5, 2 (1996), 117–124.

[33] Edwin A Locke and Gary P Latham. 2006. New directions in goal-setting theory. *Current directions in psychological science* 15, 5 (2006), 265–268.

[34] Yuhan Luo, Bongshin Lee, Donghee Yvette Wohn, Amanda L. Rebar, David E. Conroy, and Eun Kyoung Choe. 2018. Time for Break: Understanding Information Workers' Sedentary Behavior Through a Break Prompting System. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14.

[35] John F MacGregor and Theodora Kourti. 1995. Statistical process control of multivariate processes. *Control Engineering Practice* 3, 3 (1995), 403–414.

[36] Holly A McGregor and Andrew J Elliot. 2005. The shame of failure: Examining the link between fear of failure and shame. *Personality and social psychology bulletin* 31, 2 (2005), 218–231.

[37] Dan Morris, AJ Bernheim Brush, and Brian R Meyers. 2008. SuperBreak: using interactivity to enhance ergonomic typing breaks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1817–1826.

[38] Lilia Moshkina, Susan Trickett, and J Gregory Trafton. 2014. Social engagement in public places: a tale of one robot. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. Association for Computing Machinery, New York, NY, USA, 382–389.

[39] Sean A Munson and Sunny Consolvo. 2012. Exploring goal-setting, rewards, self-monitoring, and sharing to motivate physical activity. In *2012 6th international conference on pervasive computing technologies for healthcare (PervasiveHealth) and workshops*. IEEE, IEEE, 25–32.

[40] Mitchell J Neubert. 1998. The value of feedback and goal setting over goal setting alone and potential moderators of this effect: A meta-analysis. *Human Performance* 11, 4 (1998), 321–335.

[41] Jasmin Niess and Paweł W Woźniak. 2018. Supporting meaningful personal fitness: the tracker goal evolution model. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–12.

[42] Robert James Oakland and John S Oakland. 2018. *Statistical process control*. Routledge, Abingdon-on-Thames, England, UK.

[43] Lisa D Ordóñez, Maurice E Schweitzer, Adam D Galinsky, and Max H Bazerman. 2009. Goals gone wild: The systematic side effects of overprescribing goal setting. *Academy of Management Perspectives* 23, 1 (2009), 6–16.

[44] Mitesh S Patel, David A Asch, Roy Rosin, Dylan S Small, Scarlett L Bellamy, Jack Heuer, Susan Sproat, Chris Hyson, Nancy Haff, Samantha M Lee, et al. 2016. Framing financial incentives to increase physical activity among overweight and obese adults: a randomized, controlled trial. *Annals of internal medicine* 164, 6 (2016), 385–394.

[45] Erin S Pearson. 2012. Goal setting as a health behavior change strategy in overweight and obese adults: a systematic literature review examining intervention components. *Patient education and counseling* 87, 1 (2012), 32–42.

[46] Al Pfadt and Donald J Wheeler. 1995. Using statistical process control to make data-based clinical decisions. *Journal of applied behavior analysis* 28, 3 (1995), 349–370.

[47] Charlie Pinder, Jo Vermeulen, Benjamin R Cowan, and Russell Beale. 2018. Digital behaviour change interventions to break and form habits. *ACM Transactions on Computer-Human Interaction (TOCHI)* 25, 3 (2018), 1–66.

[48] James O Prochaska, Sara Johnson, and Patricia Lee. 2009. The transtheoretical model of behavior change. (2009), 59–83.

[49] Mashfiqui Rabbi, Min Hane Aung, Mi Zhang, and Tanzeem Choudhury. 2015. MyBehavior: automatic personalized health feedback from user behaviors and preferences using smartphones. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. Association for Computing Machinery, New York, NY, USA, 707–718.

[50] Mical Kay Shilts, Marcel Horowitz, and Marilyn S Townsend. 2004. Goal setting as a strategy for dietary and physical activity behavior change: a review of the literature. *American Journal of Health Promotion* 19, 2 (2004), 81–93.

[51] Victor J Strecher, Gerard H Seijts, Gerjo J Kok, Gary P Latham, Russell Glasgow, Brenda DeVellis, Ree M Meertens, and David W Bulger. 1995. Goal setting as a strategy for health behavior change. *Health education quarterly* 22, 2 (1995), 190–200.

[52] Yuan-Chi Tseng, Hui-Yen Chang, and Shih-Wei Yen. 2018. The different effects of motivational messages and monetary incentives on fostering walking behavior. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–6.

[53] Lynne Turner-Stokes. 2009. Goal attainment scaling (GAS) in rehabilitation: a practical guide. *Clinical rehabilitation* 23, 4 (2009), 362–370.

[54] Saskia Van Dantzig, Gijs Geleijnse, and Aart Tijmen Van Halteren. 2013. Toward a persuasive mobile application to reduce sedentary behavior. *Personal and ubiquitous computing* 17, 6 (2013), 1237–1246.

[55] Xinyue Zhou and Ding-Guo Gao. 2008. Social support and money as pain management mechanisms. *Psychological Inquiry* 19, 3-4 (2008), 127–144.