Evaluation of a personalized coaching system for physical activity: user appreciation and adherence

Julia S. Mollee
Vrije Universiteit Amsterdam,
the Netherlands
j.s.mollee@vu.nl

Anouk Middelweerd
VU Medical Center Amsterdam,
the Netherlands
a.middelweerd@vumc.nl

Saskia J. te Velde
VU Medical Center Amsterdam,
the Netherlands
saskia@tevelderesearch.com

Michel C.A. Klein
Vrije Universiteit Amsterdam,
the Netherlands
michel.klein@vu.nl

ABSTRACT

Physical inactivity is an increasingly serious global health problem, which implies a strong need for effective and engaging interventions. Smartphone technology offers new possibilities to address physical activity promotion. For app-based interventions to have an impact, both the effectiveness and user appreciation of the app are important. In this paper, we explore the user appreciation of the Active2Gether intervention, which offers personalized coaching to increase physical activity levels in daily life. The results are compared to the evaluation of a simplified version of the Active2Gether app (in which no coaching messages are sent) and the Fitbit app. Overall, the results reveal that users of a physical activity app appreciate a coaching feature to be included (on top of self-monitoring functionalities), but are also critical of how it is implemented (in terms of the number and content of the messages). The results also show that it is important to find a balance in the number of messages sent: too many messages seem to be perceived as annoying, but on the other hand, such system-initiated user interaction seems to reduce dropout.

Author Keywords

User experience; user evaluation; personalized coaching; behavior change; healthy lifestyle; physical activity.

ACM Classification Keywords

J.3. [Computer applications]: Life and medical sciences---health.

INTRODUCTION

Despite the well-known health and well-being benefits of physical activity [1, 2], about 50% of the adult population in western countries are less physically active than recommended by health authorities [3]. Moreover, engagement in moderate to vigorous physical activity decreases with age, in particular when transitioning from adolescence into (young) adulthood [4, 5]. Thus, effective interventions are needed to encourage young adults to become or remain physically active. Nowadays, smartphone technology offers new possibilities to address physical activity promotion, as smartphones are accessible at all times, convenient, accurate, can be used to (self-) monitor the levels of physical activity and can provide highly tailored and real-time feedback. The high adoption rate of smartphones (97% among adults aged 20-29 years) and the popularity of health and fitness apps and activity trackers in the Netherlands [6] suggest that young adults will appreciate and adopt a physical activity intervention that makes use of smartphone technology.

For app-based interventions to have an impact, it is not only important that they are effective, but also that they are accepted by users. After all, if people are unwilling to use a certain app, it won't be possible to prove its effectiveness. Research has shown that discontinuation of app use has several possible reasons, among which a lack of user friendliness and low engagement [7]. A recent systematic review reported that 17 out of 23 app-based health intervention studies found significant intervention effects on lifestyle behavior outcomes and related health outcomes [8]. Additionally, some studies demonstrated perceived effectiveness of such apps. For example, King et al. reported that 69% of participants mentioned that the apps motivated them to be more physically active and 71% reported that the apps helped them to exercise regularly [9]. Of the studies that found significant intervention effects, three studies examined associations between app usage and changes in the behavioral and health outcomes. All three showed that higher app usage was associated with improved physical activity and healthy eating [8].

The Active2Gether (A2G) intervention is an example of such an app-based physical activity intervention. This app-based intervention is linked with a Fitbit One activity tracker and aims to encourage young adults to adopt and maintain a physically active lifestyle, by focusing on the domains of active transport, stair walking and leisure time sports activities. To do so, it classifies users into one of three awareness categories (in need of education about a healthy level of physical activity, open for coaching and in need of positive feedback to maintain behavior), it helps users to select the most promising coaching domain (active

PervasiveHealth '17, May 23–26, 2017, Barcelona, Spain

ACM ISBN 978-1-4503-6363-1/17/05…$15.00
https://doi.org/10.1145/3154862.3154933

transport, stairs, sports), it helps users to set a goal and it sends motivational messages. Detailed information on the coaching system can be found in the subsequent section.

This study focuses on the *user evaluation* of the Active2Gether coaching system. We compare the evaluation of this app to two other related apps: a simplified variant of the same app (without coaching functionality), and a commercially available physical activity app (the Fitbit app). The objective of the study presented in this paper is threefold: (1) to evaluate how users used the app (adherence, interaction rates), (2) to assess how users evaluated the app with respect to perceived effectiveness, user friendliness etc., and (3) to evaluate the users' appreciation of the coaching messages sent. By evaluating different aspects of the apps, we form an idea of what is and is not appreciated by users of physical activity apps, which is vital information for developers of such systems.

ACTIVE2GETHER SYSTEM

The Active2Gether personalized coaching system evaluated in this paper aims to encourage young adults to adopt and maintain a physically active lifestyle, by focusing on the domains of active transport, stair walking and leisure time sports activities.

Initial assessment

New users start by filling out an online intake questionnaire, including questions about their daily life (e.g., occupation, significant locations) and about psychological factors underlying their physical activity behavior (e.g., self-efficacy, intentions, perceived barriers). Then, the users start a one-week assessment, to gauge their current physical activity level. The physical activity data is collected by means of a *Fitbit One* activity monitor, in combination with prompted daily user input about active transport and sports activities. The Fitbit One was chosen because of its relatively long battery life and possibility to synchronize the data continuously through Bluetooth LE.

Awareness classification

After the assessment week, the users are assigned to one of three awareness categories, namely *education*, c*oaching* or *feedback*. This classification is based on whether they meet the Dutch physical activity guidelines [1] and whether they think they should be more active. This classification is repeated every three weeks, in order to tailor the system to the user's latest awareness state.

If users do not meet the guidelines, but think they are sufficiently active, they receive *educational* messages to inform them about healthy levels and health benefits of physical activity. If users meet the guidelines and do not see the need to be more active, they receive affirmative *feedback* messages to maintain their current level of physical activity. If users don't meet the guidelines and understand that their physical activity level should increase, or if they meet the guidelines but still want to be more active, they enter the *coaching* phase.

Domain selection and goal setting

In the coaching phase, the system first suggests the user to select one of the three possible domains (i.e., active transport, stair walking or sports activities) to focus on for the next week. To do so, the user's behavior in each of the three domains is compared to what could be expected for this particular user based on personal context information. The domain with the lowest evaluation, and thus the largest potential for improvement, is suggested to the user, although they are free to select another domain instead.

After the domain selection, the user is prompted to set a domain-specific goal. If the user met his previous goal for this domain, the system suggests to increase it, and otherwise the user is suggested to keep the same goal.

Identification of promising coaching determinants

Then, the system runs simulations of a computational model to estimate what types of coaching messages are expected to be most effective for the user. To do so, the user receives a number of questions to assess the current state of personal determinants underlying physical activity behavior (e.g., self-efficacy, intentions). These states are translated into numerical values and inputted to the computational model, which describes the dynamics between those determinants and their effect on the behavior [10]. Using simulations, the system determines what the effect of improvement in each of the determinants on the behavior would be, and selects the three most promising determinants to be targeted in the coaching accordingly.

Coaching messages

Based on the selected domain and the identified most promising coaching determinants, the coaching messages are filtered to remove any messages that are irrelevant or not applicable to the user. At certain times (up to a maximum of three times per day), a message is selected from the remaining set of messages and sent to the user. Additionally, users may also receive messages to remind them to synchronize their data or to charge their Fitbit.

The coaching cycle (consisting of domain selection, goal setting, and identification of promising coaching determinants) is repeated weekly, in order to tailor the coaching to the user's current state and needs at all times.

Active2Gether app

The Active2Gether app shows a picture of a virtual coach with a welcome message that depends on the user's choice for a coaching domain, as well as the current daily number of steps and stairs and the user's progress towards the general weekly goal of 70,000 steps. Below that, the app shows an ordered graph with the user's total step count of the past seven days, among the data of up to six other users. These users are selected based on the user's preferred (upward or downward) direction of social comparison. Where possible, the app shows the data of Facebook connections, and if not available, the data of anonymized other users is shown. The same data is also accessible to the users by logging in to the Active2Gether website.

As explained above, users may receive different types of messages from the system. These pop up on the smartphone with a push notification, and are presented as overlay on top of the dashboard. As long as the app is not opened to read the message, the user receives a notification every 15 minutes.

More detailed information about the design of the Active2Gether system can be found in [11].

METHODS

This section describes the context in which the user evaluation was conducted, as well as the process of data collection and preprocessing. First, we describe the user study in which the data was collected. Then, we describe the conditions of this study in more detail. Finally, we describe the aim and content of the analyses.

User study

Participants were recruited at two university campuses in the Netherlands, as well as through referral of other participants. Interested participants were eligible if they were young adults (18 to 30 years old), healthy, and in possession of a smartphone running on Android or iOS.

Participants were assigned to one of three conditions, using a stratified randomization procedure based on their gender, type of smartphone and befriended participants. Each condition received (a variant of) a physical activity app: (1) Active2Gether Full, (2) Active2Gether Light, or (3) Fitbit. As the Active2Gether app was only available for Android smartphones, participants with an iPhone were automatically assigned to the Fitbit condition. The two other conditions were balanced on gender. Where possible, friends of participants were assigned to the same condition, in order to prevent them from comparing their apps during the study.

All participants were asked to fill out an online intake questionnaire, including questions about demographics, occupation, context, physical activity level and psychological constructs related to motivation to engage in physical activity. After the intake, the participants received a Fitbit One activity tracker, and were given instructions on how to install their assigned physical activity app and how to set up the synchronization. The participants used the app for a period of twelve weeks or longer, depending on their availability for the final appointment. After twelve weeks, the participants received a link to the final questionnaire, including questions about their experience with the app. At the final appointment with the researchers, they received €20 in gift vouchers as incentive for their participation.

Experimental conditions

As mentioned above, the participants were assigned to one of three conditions, each associated with (a variant of) a physical activity app. Figure 1 shows screenshots of the two different apps that were used.

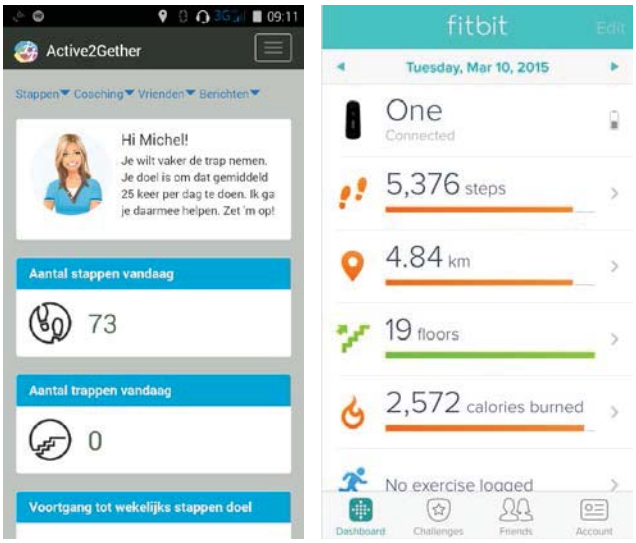

Figure 1. Screenshots of the Active2Gether app (left) and the Fitbit app (right).

Active2Gether-Full

Participants in the Active2Gether Full condition (A2G-Full) received the Active2Gether app, as described in the ACTIVE2GETHER SYSTEM section.

In order to facilitate timely data synchronization, the participants were instructed to install the Fitbit app as well, as this app enables synchronization of the Fitbit One activity tracker with the smartphone through Bluetooth LE. However, they were urged to only use the Active2Gether app, and not to view or use the Fitbit app instead.

Active2Gether-Light

Participants in the Active2Gether Light condition (A2G-Light) also received the Active2Gether app. However, in contrast to the participants in the Active2Gether Full condition, they were not sent any coaching messages. Apart from that, their app provided the same functionalities and layout as the full Active2Gether app.

Fitbit

Participants in the Fitbit condition were coached with the Fitbit app. The dashboard of the Fitbit app shows the users their current daily step and stairs data, as well as distance travelled, number of active minutes and calories burned. By clicking on any of these data tiles, the users can view graphs of their data on different levels of aggregation, varying from 5-minute epochs to yearly statistics. In addition, the Fitbit app allows users to log food and water intake, to log sports activities, and to monitor their sleep. The participants were neither encouraged nor discouraged to use these additional functionalities. Similar to the Active2Gether app, Fitbit offers the users a website with a dashboard of their activity data as well.

Data collection

The user study yielded different types of data that are of interest when evaluating the users' experience with the Active2Gether and Fitbit apps.

Intake questionnaire

First, the intake questionnaire provided information about the participants' demographics and baseline physical activity level. The demographics included information about gender, age, and height and weight.

The baseline physical activity level was obtained using a short version of the IPAQ [12], and interpreted with the Combi Norm. The Combi Norm states that people should meet at least one of two other norms, namely the Fit Norm [13] or the Dutch Norm for Healthy Exercise (Nederlandse Norm voor Gezond Bewegen, NNGB) [14]. In short, the Fit Norm requires to engage in vigorous-intensity physical activity for at least 20 minutes for at least three times per week. The NNGB states that adults should carry out at least 30 minutes of moderate to vigorous physical activity on a minimum of five days per week.

Final questionnaire

Second, a final questionnaire was used for information about the participants' subjective experiences. One question asked about prior experience with physical activity apps and activity trackers. An answer "yes" indicated that users are already using an app or tracker, "some" meant that they have tried an app or tracker before but were not currently using any, and "no" means that they had no prior experience. In addition, the final questionnaire contained Likert items about different aspects of the users' appreciation of the apps. For example, the participants were asked to evaluate the number of questions and messages sent by the app. The question items about user appreciation were based on [15] and [16], and included statements like "the app is easy to use" and "I would recommend this app to my friend". In addition, the questionnaire allowed the users to name their three most and least favorite features or aspects of the app they used during the study. Also, the questionnaire included questions for the users in the Active2Gether conditions about the number, content and tone of the messages and/or questions sent. Finally, the participants were also asked whether they experienced problems with the battery life of their phone (due to their assigned coaching app) or technical problems of any other kind.

Fitbit activity tracker data

Third, the Fitbit collects different types of physical activity data, such as steps, floors climbed, distance travelled and calories burned. The presence of step data was used as an indicator of dropouts: if no Fitbit data is synchronized, it indicates that the participant is no longer using the app.

Active2Gether app data

Finally, the Active2Gether app provided some information about the frequency of interaction with the users. The questions and messages sent to the user were logged, as well as whether they successfully reached the user's phone. First, this shows how much interaction the user had with the app. Also, logs of whether a message or question was successfully sent and received could indicate if the users experienced some technical problems or if they possibly removed the Active2Gether app from their smartphone.

Data analysis

In order to evaluate the Active2Gether app, we explored different aspects of the use of the intervention by the end user.

App use and dropouts

First, we investigated the dropout of participants based on their Fitbit data. This was done through a Kaplan-Meier survival analysis. The difference between the three groups in the survival curves was tested with a log-rank test. Also, the number of days that participants were using their app (based on their Fitbit data) was determined. As this data was not normally distributed, differences between the conditions were tested by comparing mean ranks with a Kruskal-Wallis test and Mann Whitney U post-hoc tests.

Interaction frequency through questions and messages

For users in the two Active2Gether conditions, we also investigated how much (system-initiated) interaction they had with the app in terms of received questions (A2G-Full, A2G-Light) and messages (A2G-Full).

User experience

The final questionnaire contained 20 Likert items about user appreciation of the apps. A factor analysis revealed that the data could be summarized in four factors. All four factors showed good to excellent internal consistency: α = [.942, .900, .906, .813]. Discussion between JM and StV resulted in the following labels of the four factors: (1) satisfaction, (2) user friendliness, (3) perceived effectiveness, and (4) professionality.

Examples of statements covered by each of the four factors are the following: (1) satisfaction: "the app meets my expectations", (2) user friendliness: "I can easily find the information I'm looking for", (3) perceived effectiveness: "the app motivates me to achieve my goals", and (4) professionality: "the app looks professional".

Differences between the user appreciation scores in the three conditions were assessed by means of a one-way Anova and Tukey post-hoc tests.

The questions about the experience of technical problems were answered on a 7-point Likert scale. It was considered as an occurrence of problems if the participant had selected one of the three answer options that reflected some extent of experiencing technical issues.

Evaluation of questions and messages (A2G)

The participants' evaluation of the number of questions and messages sent by the app was given on a 5-point Likert scale in the final questionnaire. The answer options 'too many' and 'far too many' were aggregated into one category, as well as the options 'too few' and 'far too few'. Then, the percentages of the answer selected were calculated for the two Active2Gether conditions as descriptive statistics.

For participants in the Active2Gether Full condition, the final questionnaire contained eight Likert items about the coaching messages. A factor analysis revealed that the data could be summarized in two factors. One negatively worded item was reversed. Both factors showed acceptable to good internal consistency: α = [.824, .760]. The factors were labeled by JM and AM as capturing (1) the tone of voice and (2) the content of the messages.

Examples of statements covered by the two factors are: (1) tone of voice: "the messages seem credible and trustworthy", and (2) content: "the messages are relevant to my personal situation".

Positive and negative aspects

In the question about the most positive and most negative aspects of the app, the participants could list up to three positive and three negative features in free text. To analyze the participants' feedback, two lists of categories were created while reading the responses (i.e., one for positive and one for negative aspects). The categories of positive aspects were (1) self-monitoring or insight, (2) social comparison, (3) coaching (messages), (4) goal setting, (5) clear, neat layout, (6) reminder, (7) perceived effect, (8) variety of data, and (9) other. The categories of negative aspects were (1) push notifications, (2) synchronization problems, (3) technical/battery problems, (4) inaccuracy of measurements, (5) lack of coaching, (6) excess or repetition of messages/questions, (7) irrelevance of coaching suggestions/messages, (8) missing functionalities, (9) unsatisfactory layout or user friendliness, (10) perceived demotivational effect, (11) use of activity tracker, and (12) other.

Then, all response items were classified using these lists and counted per category. This implies that one participant could mention more than one aspect in the same category, which would also be counted twice. Then, to compensate for the different numbers of participants in the three conditions, the counts were divided by the number of people in the corresponding condition to obtain a percentage.

All analyses were performed using SPSS 23.0 and Microsoft Excel 2010.

RESULTS

This section gives an overview of the participants who partook in the user study, and shows the results of the app evaluation, as outlined in the METHODS section.

Participant characteristics

Originally, 104 people signed up for participation. Eleven participants dropped out before the end of the study, for example because of technical problems (e.g., smartphone lacked storage or battery capacity to run the apps), because they strongly disliked wearing the activity tracker, or because participation in the study collided with other obligations.

Table 1 shows the remaining number of participants in each of the conditions, as well as the median age, the age range, the number and percentage of female participants, and the number of participants meeting the norm for physical activity.

	All	A2G-Full	A2G-Light	Fitbit
Number of participants	92	24	23	45
Age (median, range)	23 [18-31]	23 [19-30]	23 [18-30]	23 [18-31]
Number of females	72 (78%)	17 (71%)	19 (83%)	36 (80%)
Number meeting norm	50 (54%)	13 (54%)	9 (39%)	28 (62%)

Table 1. Participants' gender, age and baseline physical activity.

Table 2 shows how much prior experience the participants in each condition have with physical activity apps or trackers.

	All	A2G-Full	A2G-Light	Fitbit
Experience apps – yes	15 (16%)	6 (25%)	2 (9%)	7 (16%)
Experience apps – some	18 (20%)	2 (8%)	5 (22%)	11 (24%)
Experience apps – no	57 (62%)	15 (63%)	15 (65%)	27 (60%)
Experience trackers – yes	9 (10%)	4 (17%)	1 (4%)	4 (9%)
Experience trackers – some	8 (9%)	2 (8%)	3 (13%)	3 (7%)
Experience trackers – no	73 (79%)	17 (71%)	18 (78%)	38 (84%)

Table 2. Participants' prior experience with physical activity apps and trackers.

App use and dropouts

Table 3 shows the mean, median and range of the number of days that participants were using the app, how many participants dropped out per condition, and the percentage of participants that was still uploading Fitbit data after twelve weeks. Participants were marked as dropouts if they consecutively did not upload any Fitbit data for at least one day before the end of the experiment.

	All	A2G-Full	A2G-Light	Fitbit
Number of participants	92	24	23	45
Days using the app (mean, median, range)	70.5 84 [0-84]	79.0 84 [12-84]	81.0 84 [21-84]	60.6 84 [0-84]
Number of dropouts	35	8	3	24
Percentage using the app at 12 weeks	62.0%	66.7%	87.0%	46.7%

Table 3. Dropouts and participants that were still using the app after 12 weeks.

A Kruskal-Wallis test showed that the mean rank of the number of days that participants used the app (i.e., uploaded their Fitbit data) differed significantly between the three conditions, $\chi^2(2) = 10.671$, $p = .005$. Mann-Whitney U post-hoc tests revealed that differences that differences existed between the conditions Active2Gether Full and Fitbit ($U = 372.5$, $p = .022$) as well as between the Active2Gether Light and Fitbit groups ($U = 292$, $p = .001$), but not between the two Active2Gether conditions ($U = 219$, $p = .102$).

Figure 2 shows the Kaplan-Meier survival curves for the three conditions based on the availability of Fitbit data. The log-rank test revealed that the survival functions show statistically significant differences between the three conditions, $\chi^2(2) = 12.381$, $p = .002$.

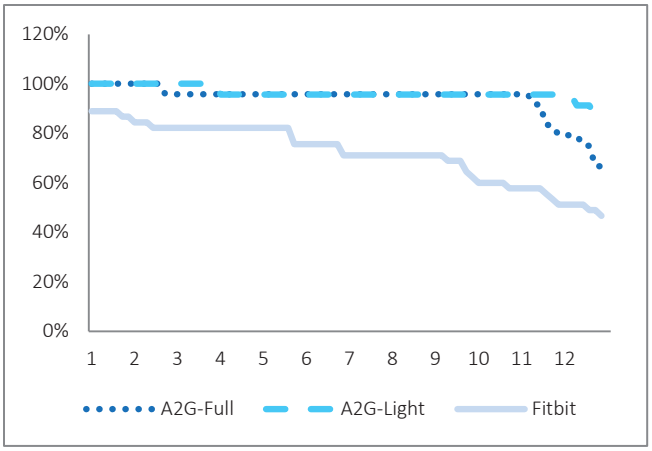

Figure 2. Percentage of participants using the app (based on synchronization of Fitbit data) over period of 12 weeks.

Interaction frequency through questions and messages

In total, 8,556 questions were successfully received by the 47 participants in the Active2Gether Full and Active2Gether Light conditions, over the course of twelve weeks. All questions that were derived were also sent, and 112 questions were sent but not received. The 24 participants in the A2G-Full condition received 1,324 messages successfully. In contrast, 48 were derived but not sent and 57 were sent but not received.

Further analysis shows that Active2Gether (Full or Light) participants received on average 182 questions during the twelve-week period. All derived questions were successfully sent, but for 22 out of 47 users, a question was not received by the phone at some point. For one user, no questions were derived and therefore this user did not receive any questions. This suggests some technical problems or unsuccessful installation of the app.

Similarly, the logs show that participants in the Active2Gether Full condition received an average of 55 messages in twelve weeks. For five out of 24 users, a derived message was not sent at some point, which indicates that the app was removed before the end of the study. For nine users, a sent message was not received by the phone, and one user did not receive any messages at all.

User experience

Of the 93 participants that completed the study, 90 filled out the complete final questionnaire. Two participants did not finish the questionnaire, and one participant did not do the final questionnaire at all.

As described in the METHODS section, a factor analysis was performed that revealed four factors. Table 4 and Figure 3 show the average scores on those factors for the three conditions. A one-way Anova showed that the user experience ratings for all four factors differed between the three conditions; satisfaction: $F(2,87) = 21.991$, $p < .001$; user friendliness: $F(2,87) = 5.065$, $p = .008$; perceived effectiveness: $F(2,87) = 6.303$, $p = .003$; professionality: $F(2,87) = 15.437$, $p < .001$. Tukey post-hoc tests revealed that differences existed between the conditions Active2Gether Full and Fitbit ($p < .001$; $p = 0.022$; $p = .006$; $p = .006$), and between the Active2Gether Light and Fitbit groups ($p < .001$; $p = .039$; $p = .030$; $p < .001$), but not between the two Active2Gether conditions ($p = .892$; $p = .987$; $p = .891$; $p = .129$).

	All	A2G-Full	A2G-Light	Fitbit
(1) Satisfaction (mean, sd)	3.71 (1.56)	2.90 (1.46)	2.73 (1.38)	4.61 (1.24)
(2) User friendliness (mean, sd)	5.10 (1.30)	4.65 (1.65)	4.71 (1.69)	5.51 (1.07)
(3) Perc. effectiveness (mean, sd)	4.21 (1.60)	3.54 (1.85)	3.75 (1.64)	4.77 (1.40)
(4) Professionality (mean, sd)	4.44 (1.36)	4.13 (1.56)	3.44 (1.44)	5.09 (1.04)
Overall (mean, sd)	4.36 (1.23)	3.81 (1.43)	3.66 (1.24)	4.99 (1.00)

Table 4. Average scores on user appreciation (range [1,7]).

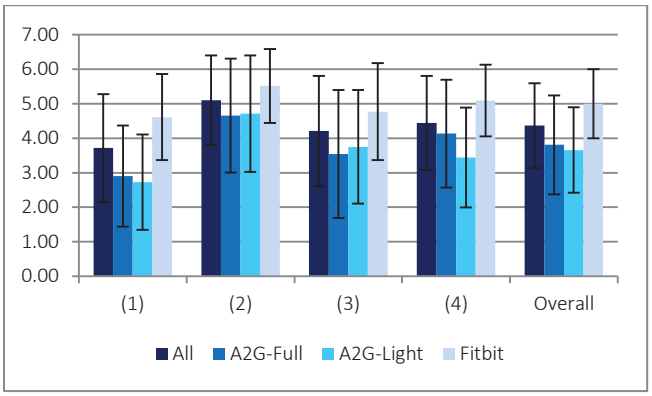

Figure 3. Average scores on four factors of user appreciation: (1) satisfaction, (2) user friendliness, (3) perceived effectiveness, and (4) professionality.

The differences between the three conditions are also apparent in the overall user experience rating, $F(2,87) = 15.929$, $p < .001$. A Tukey post-hoc test showed that the difference between the Active2Gether Full and Fitbit groups is significant ($p < .001$), as well as between the Active2Gether Light and Fitbit condition ($p < .001$), but not between the two Active2Gether conditions ($p = .883$).

Figure 4 shows the percentage of participants that expressed battery problems or other technical issues.

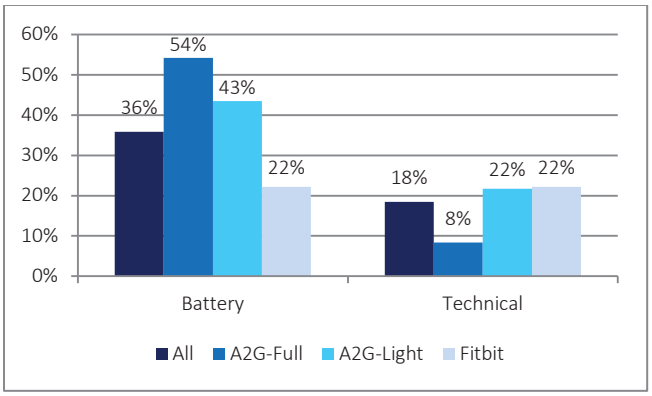

Figure 4. Percentage of participants with battery problems and/or other technical problems.

Evaluation of questions and messages (A2G)

Figure 5 shows the percentage of participants that perceived the number of questions received as 'too many', 'just right' or 'too few'.

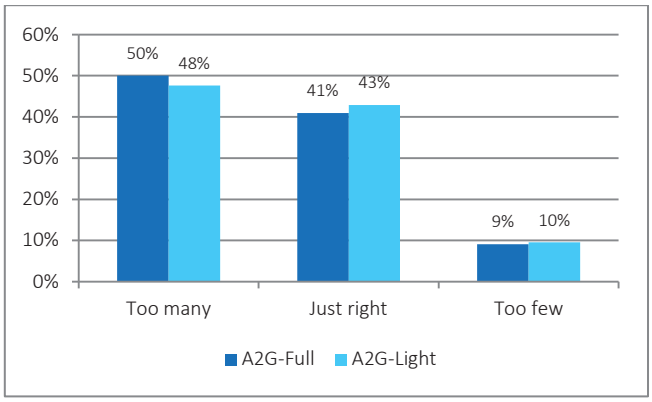

Figure 5. User evaluation of number of questions in Active2Gether conditions.

In addition to the questions, the users received messages through the app. The Active2Gether Light participants did not receive motivational coaching messages, but were only sent messages about their Fitbit's low battery life or overdue data synchronization. Figure 6 shows the percentage of participants that selected certain answers.

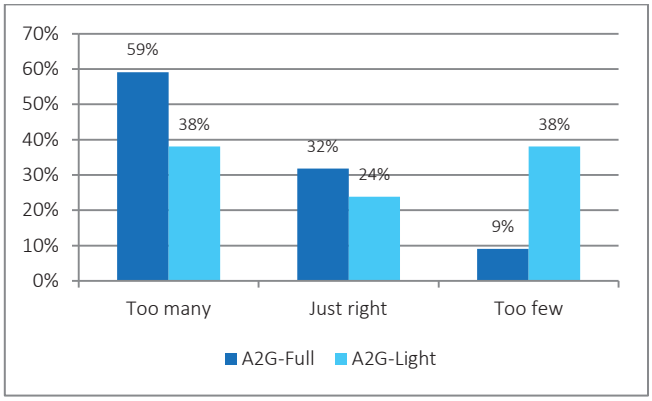

Figure 6. User evaluation of number of messages in Active2Gether conditions.

Table 5 shows the average scores on the two factors on which the messages in the Active2Gether Full condition were evaluated.

	Mean	St. dev.
Tone of voice	3.02	0.865
Content	2.30	0.816
Overall	2.66	0.737

Table 5. Average scores on eight statements about evaluation of messages (range [1,5]).

Positive and negative aspects

As explained in the METHODS section, the reported positive and negative aspects of the apps were classified into overarching categories.

Table 6 shows the categories of positive aspects that were mentioned most often in the Active2Gether Full condition, and Table 7 lists the categories of negative aspects.

#	Aspect	Count	Perc.
1	Self-monitoring or insight	13	54%
2	Coaching (messages)	10	42%
3	Social comparison	7	29%
	Layout	7	29%

Table 6. Most often reported positive aspects in the Active2Gether Full condition, with count and percentage.

#	Aspect	Count	Perc.
1	Technical/battery problems	15	63%
2	Excess or repetition of messages/ questions	12	50%
3	Irrelevance of coaching suggestions/ messages	7	29%

Table 7. Most often reported negative aspects in the Active2Gether Full condition, with count and percentage.

Table 8 lists the categories of positive aspects that were mentioned most often in the Active2Gether Light condition, and Table 9 enumerates the categories of negative aspects.

#	Aspect	Count	Perc.
1	Self-monitoring or insight	17	77%
2	Layout	14	61%
3	Social comparison	7	32%

Table 8. Most often reported positive aspects in the Active2Gether Light condition, with count and percentage.

#	Aspect	Count	Perc.
1	Missing functionalities	14	61%
2	Synchronization problems	11	48%
3	Technical/battery problems	9	39%

Table 9. Most often reported negative aspects in the Active2Gether Light condition, with count and percentage.

Table 10 shows the categories of positive aspects that were mentioned most often in the Fitbit condition, and Table 11 lists the most common categories of negative aspects.

#	Aspect	Count	Perc.
1	Self-monitoring or insight	41	91%
2	Layout	23	51%
3	Variety of data	19	42%

Table 10. Most often reported positive aspects in the Fitbit condition, with count and percentage.

#	Aspect	Count	Perc.
1	Inaccuracy of measurements	26	58%
2	Synchronization problems	11	24%
3	Missing functionalities	8	18%

Table 11. Most often reported negative aspects in the Fitbit condition, with count and percentage.

DISCUSSION AND CONCLUSIONS

This section provides an interpretation of the results. The most important findings are discussed, as well as their implications for the design of physical activity apps or interventions.

Participants of the study were young adults (18 to 31 years old), and a majority was female (78%). Approximately half of the participants were sufficiently physically active according to health recommendations, which is in line with overall findings about the adult population in western countries [3]. The majority of the participants had no prior experience with physical activity apps (62%) or activity trackers (79%).

The Fitbit data showed that the dropout of the intervention was lower in the Active2Gether conditions (both variants) than in the Fitbit condition. The percentage of users after 12 weeks was highest in the Active2Gether Light condition (87.0%) and lowest in the Fitbit condition (46.7%). This pattern was already visible in earlier weeks of the intervention. Interestingly, this cannot be explained by the experience of technical or battery problems (see Figure 4), as those factors were comparable or even higher in the Active2Gether Light condition.

Research has shown that dropout numbers in health interventions are very diverse, e.g. from 6% in an 8-week physical activity intervention [17] to 73% in a 14-week healthy lifestyle intervention [18]. This makes it difficult to compare the results, but it is interesting to see that the dropout between the Active2Gether conditions and the Fitbit condition differed so strongly, even though the setup of the study was otherwise exactly the same. This could suggest that the Active2Gether conditions offer something that retains the users' interest for a longer period of time. Other research has shown that adherence in health apps is generally quite low: 26% of health apps is only used once after downloading, and 74% of health app users indicated to have stopped using the app within ten times of using it [7]. In light of these findings, the adherence in the current study was very acceptable.

The systems logs showed that the participants in the two Active2Gether conditions received approximately 182 questions during the twelve-week period. Almost 99% of the questions were received successfully. Approximately half of the users perceived the number of questions as too high, which could be resolved by replacing some user input by automated registration (e.g., of sports activities and transport options).

Similarly, over 92% of the derived coaching messages were received successfully by the users in the Active2Gether Full condition. Over the twelve-week period, these participants received an average of 55 coaching messages. This is less than the system allows (i.e., up to three messages per day), which indicates that there were not always relevant messages available for the user, and the set of messages should be extended to cover more combinations of context variables. However, the participants in the two Active2Gether conditions also indicated that the number of messages was too high (59% and 38%). Since 38% of the users in the Active2Gether Light condition also perceived the number of messages as too high, even though they only

received messages about the status of their Fitbit battery and data synchronization, it is possible that the coaching messages were not the main contributor to these sentiments. Also, it is possible that participants did not clearly distinguish between questions and messages, and perceived the overall number of app-initiated interactions as too high.

Over all four factors of user appreciation, the Fitbit app was rated higher than the two Active2Gether conditions. Generally, the full Active2Gether app scored slightly better than the simplified version, although these differences were not significant. Reasons for the relatively low scores could be explained by the feedback on the apps' negative aspects. Both Active2Gether conditions reported quite some technical problems (63% and 39%, respectively), for example with respect to their smartphone's reduced battery life. For the full Active2Gether app, the repetition in the questions and messages was disliked (50%) and the messages were perceived as not very personal or relevant (29%). The main criticism on the Active2Gether Light app was its simplicity (61%). The participants in the Fitbit condition complained most often about its inability to reflect certain activities (58%), as well as delays in of problems with synchronization (24%) and lacking functionalities (18%).

On the other hand, the feedback on the positive aspects shows that participants in all three conditions highly value the possibility to review their behavior (54%, 74% and 91%, respectively). In the Fitbit condition, the percentage is probably higher because of the option to view activity data in more detail (i.e., per 5 minutes) and in different types of parameters (i.e., active minutes, calories burned, etc.). These aspects are mentioned by 42% of users in the Fitbit condition. In addition, participants in both Active2Gether conditions appreciated the comparison to other users (29% and 30%), and the clean layout of the app (29% and 61%). Finally, users of the full Active2Gether app praised the coaching aspect (42%).

One of the key strengths of this study is that the user evaluation was based on considerable use of the app, as the participants were asked to use their app for at least twelve weeks. This allows for a substantiated evaluation. In addition, since a variety of different aspects of the apps were considered, the evaluation in this paper gives a rather complete picture of the likes and dislikes of the participants. While the focus of the evaluation is on one of the apps, the full Active2Gether app, the comparison to its simplified version and a commercially available app provides more insight in the aspects that are appreciated by users.

A limitation of the present study is that the results might not be easily transferable to the general population. It covered only young adults (18 to 31 years old), and the majority of the participants was female (78%). Also, all participants signed up voluntarily, so they probably were already intrinsically motivated to improve their physical activity levels through an app-based intervention. Moreover, it is possible that participants who use a physical activity app in context of an experiment perceive their experience differently from users who download the app solely for their own use. In addition, although different aspects of the apps were evaluated, it is difficult to say which aspects or features contributed to specific scores on their satisfaction, user friendliness, perceived effectiveness and professionality. Further (qualitative) research should reveal exactly which aspects were liked and disliked by the users. Also, since the Active2Gether app was only available for Android smartphones, the assignment of participants to conditions was not completely random, which in theory could have influenced the results. Finally, although the subjective user evaluation is very important for the user experience and adherence, it does not necessarily imply the apps' effectiveness as well. In order to develop and offer successful physical activity interventions, both the user experience and effectiveness should be ensured.

Overall, we can conclude that users of a physical activity app want a coaching feature to be included (on top of self-monitoring functionalities), but are also critical of how it is implemented (in terms of number and content of the messages). It is important that the coaching is perceived as personal and relevant, and it should be sufficiently diverse in order not to become too repetitive. Thus, it is important to find a (personal) balance in the number of messages: too many messages seem to be annoying, but on the other hand, such system-initiated user interaction seems to reduce dropout. Further research should reveal how this perfect balance can be achieved.

ACKNOWLEDGMENTS

This research is supported by Philips and Technology Foundation STW, Nationaal Initiatief Hersenen en Cognitie NIHC under the partnership program Healthy Lifestyle Solutions. The authors would like to thank prof. dr. Johannes Brug for his feedback on the manuscript.

REFERENCES

[1] W. H. O. "Global Recommendations on Physical Activity for Health," 2010. [Online]. Available: http://www.webcitation.org/6ITBUAKj4. [Accessed March 2017].

[2] I. M. Lee, E. J. Shiroma, F. Lobelo, P. Puska, S. N. Blair, P. T. Katzmarzyk and L. P. A. S. W. G. , "Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy," *The Lancet,* vol. 380, no. 9838, pp. 219-229, 2012.

[3] W. H. O. "Physical activity Fact Sheet No.385," February 2014. [Online]. Available: http://www.webcitation.org/6SMZsQsXZ. [Accessed March 2017].

[4] M. Y. Kwan, J. Cairney, G. E. Faulkner and E. E. Pullenayegum, "Physical activity and other health-risk behaviors during the transition into early adulthood: a longitudinal cohort study," *American Journal of Preventive Medicine,* vol. 42, no. 1, pp. 14-20, 2012.

[5] S. Bell and C. Lee, "Emerging adulthood and patterns of physical activity among young Australian women," *International Journal of Behavioral Medicine,* vol. 12, no. 4, pp. 227-235, 2005.

[6] TelecomNieuwsNet, "Onderzoek: Smartphonepenetratie vlakt in Nederland af," 2016. [Online]. Available: https://telecomnieuwsnet.wordpress.com/2016/03/08/onderzoek-smartphonepenetratie-vlakt-in-nederland-af/. [Accessed March 2017].

[7] "Motivating Patients to Use Smartphone Health Apps," Consumer Health Information Corporation (CHIC), 2011. [Online]. Available: http://www.consumer-health.com/motivating-patients-to-use-smartphone-health-apps/. [Accessed March 2017].

[8] S. Schoeppe, S. Alley, W. V. Lippevelde, N. A. Bray, S. L. Williams, M. J. Duncan and C. Vandelanotte, "Efficacy of interventions that use apps to improve diet, physical activity and sedentary behaviour: a systematic review," *International Journal of Behavioral Nutrition and Physical Activity,* vol. 13, no. 1, p. 127, 2016.

[9] A. C. King, E. B. Hekler, L. A. Grieco, S. J. Winter, J. L. Sheats, M. P. Buman, B. Banerjee, T. N. Robinson and J. Cirimele, "Harnessing different motivational frames via mobile phones to promote daily physical activity and reduce sedentary behavior in aging adults," *PLoS One,* vol. 8, no. 4, p. e62613, 2013.

[10] J. S. Mollee and C. N. van der Wal, "A Computational Agent Model of Influences on Physical Activity Based on the Social Cognitive Theory," in *Boella G., Elkind E., Savarimuthu B.T.R., Dignum F., Purvis M.K. (eds) PRIMA 2013: Principles and Practice of Multi-Agent Systems. PRIMA 2013. Lecture Notes in Computer Science, vol 8291. Springer, Berlin, Heidelberg*, 2013.

[11] M. C. A. Klein, A. Manzoor, A. Middelweerd, J. S. Mollee and S. J. te Velde, "Encouraging Physical Activity via a Personalized Mobile System," *IEEE Internet Computing,* vol. 19, no. 4, pp. 20-27, 2015.

[12] "International Physical Activity Questionnaire," [Online]. Available: https://sites.google.com/site/theipaq/home. [Accessed March 2017].

[13] American College of Sports Medicine position stand, "The recommended quantity and quality of exercise for developing and maintaining cardiorespiratory and muscular fitness, and flexibility in healthy adults," *Medicine & Science in Sports & Exercise,* vol. 30, no. 6, pp. 975-991, 1998.

[14] H. G. C. Kemper, W. T. M. Ooijendijk and M. Stiggelbout, "Consensus over de Nederlandse Norm voor Gezond Bewegen [Consensus on the Dutch standard for healthy exercise]," *Tijdschrift voor Sociale Gezondheidszorg,* vol. 78, pp. 180-183, 2000.

[15] A. M. Lund, "Measuring Usability with the USE Questionnaire," *Usability Interfae,* vol. 8, no. 2, pp. 3-6, 2001.

[16] J. Sauro, "SUPR-Q: A Comprehensive Measure of the Quality of the Website User Experience," *Journal of Usability Studies,* vol. 10, no. 2, pp. 68-86, 2015.

[17] A. C. King, E. B. Hekler, L. A. Grieco, S. J. Winter, J. L. Sheats, M. P. Buman, B. Banerjee, T. N. Robinson and J. Cirimele, "Effects of Three Motivationally Targeted Mobile Device Applications on Initial Physical Activity and Sedentary Behavior Change in Midlife and Older Adults," *PLoS One,* vol. 11, no. 6, p. e0156370, 2016.

[18] N. J. Safran, Z. Madar and D. R. Shahar, "The impact of a Web-based app (eBalance) in promoting healthy lifestyles: randomized controlled trial," *Journal of Medical Internet Research,* vol. 17, no. 3, p. e56, 2015.

[19] L. Hebden, A. Cook, H. P. van der Ploeg, L. King, A. Bauman and M. Allman-Farinelli, "A mobile health intervention for weight management among young adults: a pilot randomised controlled trial," *Journal of Human Nutrition and Dietetics,* vol. 27, no. 4, pp. 322-332, 2014.

[20] World Health Organization, "Physical activity," World Health Organization, 2016. [Online]. Available: http://www.who.int/topics/physical_activity/en/. [Accessed October 2016].