ORIGINAL RESEARCH

Usability testing of a monitoring and feedback tool to stimulate physical activity

Sanne van der Weegen[1]
Renée Verwey[1,2]
Huibert J Tange[3]
Marieke D Spreeuwenberg[1]
Luc P de Witte[1,2]

[1]Department of Health Services Research, CAPHRI School for Public Health and Primary Care, Faculty of Health Medicine and Life Sciences, Maastricht University, the Netherlands; [2]Research Centre Technology in Care, Zuyd University of Applied Sciences, Heerlen, the Netherlands; [3]Department of General Practice, CAPHRI School for Public Health and Primary Care, Faculty of Health Medicine and Life Sciences, Maastricht University, the Netherlands

Correspondence: Sanne van der Weegen
School for Public Health and Primary Care (CAPHRI), Maastricht University, PO Box 616, 6200 MD, Maastricht, the Netherlands
Tel +31 43 388 1479
Fax +31 43 388 4162
Email s.vanderweegen@maastrichtuniversity.nl

Introduction: A monitoring and feedback tool to stimulate physical activity, consisting of an activity sensor, smartphone application (app), and website for patients and their practice nurses, has been developed: the 'It's LiFe!' tool. In this study the usability of the tool was evaluated by technology experts and end users (people with chronic obstructive pulmonary disease or type 2 diabetes, with ages from 40–70 years), to improve the user interfaces and content of the tool.

Patients and methods: The study had four phases: 1) a heuristic evaluation with six technology experts; 2) a usability test in a laboratory by five patients; 3) a pilot in real life wherein 20 patients used the tool for 3 months; and 4) a final lab test by five patients. In both lab tests (phases 2 and 4) qualitative data were collected through a thinking-aloud procedure and video recordings, and quantitative data through questions about task complexity, text comprehensiveness, and readability. In addition, the post-study system usability questionnaire (PSSUQ) was completed for the app and the website. In the pilot test (phase 3), all patients were interviewed three times and the Software Usability Measurement Inventory (SUMI) was completed.

Results: After each phase, improvements were made, mainly to the layout and text. The main improvement was a refresh button for active data synchronization between activity sensor, app, and server, implemented after connectivity problems in the pilot test. The mean score on the PSSUQ for the website improved from 5.6 (standard deviation [SD] 1.3) to 6.5 (SD 0.5), and for the app from 5.4 (SD 1.5) to 6.2 (SD 1.1). Satisfaction in the pilot was not very high according to the SUMI.

Discussion: The use of laboratory versus real-life tests and expert-based versus user-based tests revealed a wide range of usability issues. The usability of the It's LiFe! tool improved considerably during the study.

Keywords: accelerometry, chronic obstructive pulmonary disease, diabetes mellitus type 2, heuristic evaluation, telemonitoring, thinking aloud

Introduction

Increased physical activity is associated with improvements in many health conditions, including cardiovascular diseases, obesity, insulin insensitivity, osteoporosis, and psychological conditions.[1,2] Therefore, guidelines recommend taking moderately intense aerobic physical activity for a minimum of 30 minutes on 5 days each week or vigorous-intensity aerobic activity for a minimum of 20 minutes on 3 days each week in order to maintain health.[3,4] However, many people do not meet these criteria, with percentages ranging from 41% in the Netherlands to 66% and 53% in the UK and USA.[4–7] It seems difficult to be sufficiently active, especially for people with a chronic disease.[8,9] In a Dutch sample, 66% of the people with chronic obstructive pulmonary disease (COPD) who were inactive agreed that sufficient exercise should be

part of their daily life. Of this group, however, 44% indicated that they needed help to achieve this.[10] Also, in people with type 2 diabetes, additional support seems to be needed to motivate and activate them.[11] That is why physical activity counseling in primary health care is recommended for people with chronic diseases. However, primary health care providers need strategies to improve their ability to counsel patients effectively.[12,13] New technologies can be applied to support health care interventions in all age groups.[14–16] In the It's LiFe! Study, we developed a tool[17] embedded in a self-management support program (SSP) that may support primary care professionals in their coaching role and patients with a chronic disease in improving their success in achieving an active lifestyle. The intervention helps to increase patients' awareness of the risks of inactivity behavior, in combination with self-monitoring of behavior, goal setting, action planning, discussing self-efficacy, and providing tailored feedback. The tool provides real-time feedback, on a smartphone application (app), about physical activity related to a personal goal. The tool also provides dialogue sessions about physical activity barriers and facilitators, a historic overview of activity behavior, and feedback messages about the results. Furthermore, the tool supports the primary care professional in accomplishing the coaching role by providing the activity data and results of dialogue sessions of their patients on a website.

The tool and SSP will only be a successful e-health intervention if they are adapted to the needs and preferences of the end users. This was achieved by following a user-centered development process.[18] An essential step in this process was a usability test. Testing for usability reduces errors, reduces the need for user training and user support, and improves acceptance by users,[19] which will probably lead to better compliance with the intervention.

The aim of the study reported in this paper was to test the usability of all parts of the It's LiFe! tool by end users (patients).

Methods

Usability is defined as, "The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use."[20] Indicators for effectiveness, efficiency, and satisfaction are error rate, task completion time, and a satisfaction rating questionnaire.[21] In this study, usability was tested in a mixed-method approach in the following four phases:

1. a heuristic evaluation by experienced technology users and developers;
2. a usability test in a laboratory (lab) setting with end users;
3. a real-life pilot test by end users;
4. a second usability test in the lab with end users.

The medical-ethical committee of Maastricht University Medical Centre+ approved the studies.

System description and use

The It's LiFe! tool consists of three elements:[17]

1. an activity sensor with Bluetooth connectivity worn on the hip, clipped on the belt;
2. a smartphone (Samsung Galaxy Ace; Samsung Electronics Co., Seoul, South Korea) with an app for mobile feedback;
3. a web client for comprehensive feedback and data entry for patient and practice nurse.

Navigation through the smartphone works by swiping. For study phases 1, 2, and 4, dummy data were available on the phone and website. Three types of feedback are provided by the tool (Figure 1). The first feedback loop contains the real-time activity data compared to a personal goal on a widget and in the menu of the app, per hour, day, week, and month. The second feedback loop consists of dialogue sessions and feedback messages based upon the activity results. These are generated by the system and accessible from the app and the patient's website. Dialogue sessions consist of questions regarding the barriers facilitators and patients face in becoming active, preparatory questions for setting an activity goal, and advice for action planning. The third feedback loop is the feedback from the practice nurse during consultations. In these consultations, motivational interviewing, risk communication, and goal setting are used as counseling techniques. Figure 2 provides an overview of the entire intervention executed in the pilot.

Participants

For the heuristic evaluation (phase 1), six people were selected who were known for their experience with the development, evaluation, or extensive use of technology (technology experts). End users were people with COPD or type 2 diabetes, aged 40–70 years, and familiar with the Dutch language. In phases 2–4, these criteria were used to select participants. For the laboratory tests (phases 2 and 4), eleven patients were invited through an invitation letter. For the pilot in real life (phase 3), 20 patients were invited by practice nurses in two participating general practices.
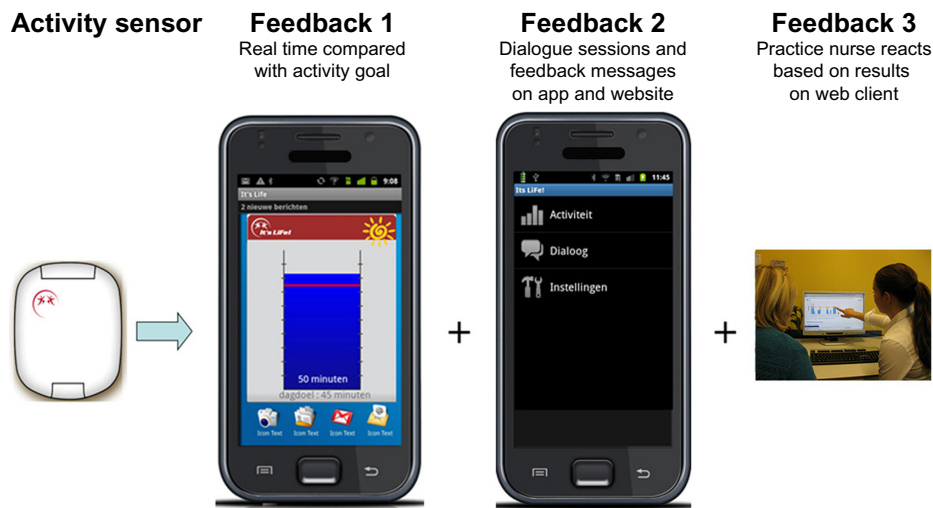
Figure 1 The It's LiFe! tool.
Abbreviation: app, application.

Study design
Below, the four study phases are described in detail.

Phase 1: Heuristic evaluation by experts
The heuristic evaluation was based on Nielsen's ten usability principles.[22] These principles concern, among other things, language use, error prevention, consistency, and efficiency of use. The test was performed by six technology experts. Each evaluator started by reading the manual. They were asked to write down any suggestions for improvement and thereafter to use the interface on the app twice – first to obtain a general idea about the app and then to go in depth for each screen – write down remarks, and score the ten usability heuristics on a scale from 1 (very bad) to 7 (very good). A heuristic was interpreted as violated by a score of 4 or lower. After completing the evaluation, the experts discussed their comments and scores with the researcher. The results of this phase were used to adjust the manual and develop a second prototype of the app.

Phase 2: Usability test in a laboratory setting by patients
In phase 2, a think-aloud procedure was performed by end users to evaluate the usability of the manual, app, and server.
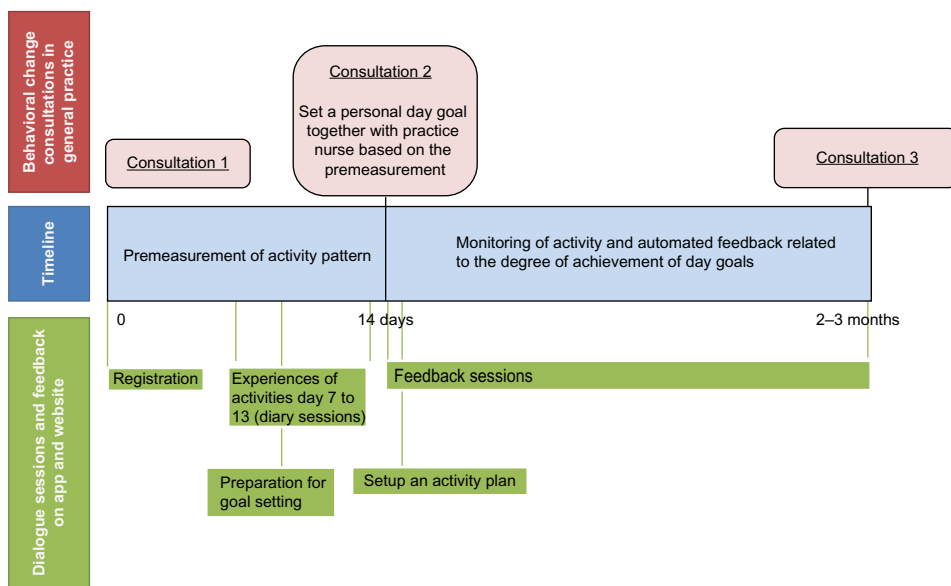

Figure 2 Timeline of the behavioral change consultations with practice nurse and dialogue sessions during the pilot.
Abbreviation: app, application.

For the dialogue sessions, the researcher noted which tasks the participants completed: participants were free to complete the session from the server on the app or the website. Before the actual test, the participants completed a questionnaire that included questions regarding birth year, level of education, kind of mobile phone, and Internet use in hours per week. Participants rated the manual for comprehensiveness and readability on a scale from 1 (very bad) to 7 (very good). Then, after a short explanation, the participants individually performed seven predetermined tasks on the app and completed three to seven dialog sessions (depending on the time left). The users were asked to verbalize their thoughts while performing these tasks. Tasks were formulated in such a way as to guarantee that all functionalities of the interfaces were used and tested. All tasks are presented in Table 1. Participants were observed by a researcher throughout their task performance. The researcher registered all users' comments during task performance, including the need for assistance, the number of errors, and expressed suggestions for improvement. The researcher also registered relevant nonverbal communication (eg, confident or confused facial expressions). In addition, participants' facial expressions, audio, on-screen activity, and keyboard/mouse input were videotaped with Morae Recorder (version 3.1.1; TechSmith Corporation, Okemos, MI, USA).

Participants valued the complexity of each task on a scale from 1 (very difficult) to 7 (very easy). Dialogue sessions were also rated for comprehensiveness and readability on a scale from 1 (very bad) to 7 (very good).

All participants completed the translated post-study system usability questionnaire (PSSUQ). The PSSUQ consists of 19 items that are rated on a 7-point scale (strongly disagree [1] to strongly agree [7]).[23] The PSSUQ consists of an overall satisfaction scale and three subscales: system usefulness (items 1–8); information quality (items 9–15); and interface quality (items 16–18). Higher scores indicate better usability. Missing data were interpolated by averaging the remaining domain scores.[23] One item was not included as it was not applicable for the app or for the website. The PSSUQ was completed twice, once for the app and once for the website. The prototype was adapted on the basis of the results of this test.

Phase 3: Pilot test in real life

In phase 3, usability was tested by end users in real life, which implies that the tool was used in daily life (at home, at work, etc) and embedded in primary care. A practice nurse provided the tool to the patient in a first consultation that was aimed at behavioral change.[24] Subsequently,

the patients wore the tool for 2 weeks to get a baseline measurement of their physical activity. Patients received dialogue sessions on the app and website with questions about barriers and facilitators for physical activity. Every day, the results were automatically sent to the practice nurses' website.

After the baseline measurement, a consultation took place in which the patient and practice nurse set an activity goal in minutes per day. Thereafter, the patients continued wearing the tool for another 10 weeks. They composed an activity plan in a dialogue session and received feedback from the tool about their performance compared with their personal goal. In a final consultation, 3 months after the start of the intervention, the practice nurse reflected on the activity results in a final consultation. Instructions on using the tool were given in a written manual and instruction movies were available on YouTube.

After each consultation, participants were interviewed about their experiences. The interviews were audio taped and transcribed. At the end of the intervention period, participants completed the Software Usability Measurement Inventory (SUMI) questionnaire. The SUMI contains 50 items that have to be answered on a 3-point Likert scale (agree, undecided, disagree). The SUMI consists of a global scale and five subscales (efficiency, affect, helpfulness, control, and learnability).[25] The subscales are all linked to questions throughout the SUMI questionnaire. Software usability is considered reasonable with scores of 50 or more on each of the scales.[26] In addition, the number of errors, technical failures, defects, and causes of the defects were collected in logbooks kept by a helpdesk and the end users, including the practice nurses.

Phase 4: Usability test in a laboratory setting by patients

After the pilot test, the prototype was further improved. To ensure that the latest adaptations did not introduce new problems, the laboratory usability test was repeated with five new end users. To measure satisfaction, Microsoft's desirability toolkit (Microsoft Corporation, Redmond, WA, USA)[27] was added to the protocol. From a list of 118 words (60% with a positive and 40% with a negative meaning), participants were first asked to mark all words they found applicable to the system (app and website separately). Second, they were asked to choose from the selected words the five that most closely matched their personal reactions to the system, and to explain their choice. The desirability

Table 1 Results on task complexity, text comprehensiveness, and readability from tests in the laboratory setting

Manual	Phase 2		Phase 4	
	N	Mean (SD)	N	Mean (SD)
Comprehensiveness	4	4.5 (2.1)	5	6.2 (1.3)
Readability	4	4.8 (2.2)	5	6.0 (1.4)
It's LiFe! app				
Complexity of tasks on app	N	Mean (SD)	N	Mean (SD)
You see the widget, what is the activity goal?	5	6.2 (1.0)	5	6.6 (0.5)
Widget. Say out loud how many minutes of activity this person had today	5	7.0 (0.0)	5	7.0 (0.0)
Go to the It's LiFe! App	5	5.8 (0.1)	5	5.8 (1.8)
Press "activity," press "hour," and scroll through the days by swiping	5	5.8 (1.1)	5	6.8 (0.4)
Press "day," say out loud how many minutes this person performed moderately intense activities and very intense activities, 3 days ago	5	5.6 (1.1)	5	5.6 (1.7)
How many days did this person reach their goal in week 3	5	5.6 (2.1)	5	6.8 (0.4)
How many days did this person reach their goal in month X	5	6.2 (0.8)	5	6.6 (0.5)
Mean		6.0 (0.5)		6.5 (0.5)
It's LiFe! server				
Task complexity	N	Mean (SD)	N	Mean (SD)
Go to the website and log in	3w	6.3 (0.6)	5w	6.6 (0.5)
Fill out a reminder	3w	4.7 (2.1)	0	–
Registration session	2a	6.5 (0.7)	5w	5.4 (1.5)
	2w	6.5 (0.7)		
Diary sessions	2a	5.7 (1.2)	2a	7.0 (0.0)
	1w			
Preparation for goal setting	3a	4.0 (2.2)	5w	6.0 (0.7)
	2w			
Set up activity plan	3w	6.0 (1.0)	5w	6.0 (0.7)
Remarks of the day	1a	6.0 (1.4)	3a	6.0 (1.0)
	1w			
Mean		5.75 (0.9)		6.2 (0.6)
Comprehensibility of the text per task				
Register session	4	5.0 (2.0)	5	6.0 (0.7)
Diary sessions	3	6.3 (1.2)	2	7.0 (0.0)
Preparation for goal setting	4	4.5 (3.0)	5	6.6 (0.5)
Set up activity plan	3	6.3 (0.6)	5	6.4 (0.9)
Remarks of the day	2	5.5 (0.7)	3	6.3 (0.6)
Watch "the activity picture"	1	6.0 (0.0)	5	6.8 (0.4)
Mean		5.6 (0.7)		6.5 (0.4)
Readability of the text per task				
Register session	4	6.3 (0.5)	5	6.0 (0.7)
Diary session	3	6.3 (0.6)	2	7.0 (0.0)
Preparation for goal setting	4	4.3 (2.8)	5	6.6 (0.5)
Set up an activity plan	3	6.7 (0.6)	5	6.6 (0.5)
Remarks of the day	2	6.0 (0.0)	3	6.3 (0.6)
Watch "the activity picture"	1	6.0 (0.0)	5	6.8 (0.4)
Mean		5.9 (0.8)		6.6 (0.4)

Note: "a" indicates the sessions that are completed on the smartphone app and "w" indicates the sessions that are completed on the website. (set up activity plan was completed on the web only).
Abbreviations: app, application; SD, standard deviation.

toolkit was translated into Dutch by two independent researchers.

Statistics

For the quantitative measurements (baseline characteristics, Likert-scale questionnaires, PSSUQ questionnaire), means and standard deviations were calculated using SPSS software package 19 (IBM Corporation, Armonk, NY, USA). The SUMI data were analyzed by the SUMI-service using the proprietary software SUMISCO (Human Factors Research Group, University College, Cork, Ireland).

Missing values were scored as "undecided". Lists with more than four missing values were left out of the analysis. Qualitative data collected during the observations in the lab and in the real-life test were recorded, summarized, and analyzed with a directed content analysis.[28]

Results

Heuristic evaluation

The responses to Nielsen's heuristics indicated no major issues: all items scored on average 4 or higher. Help documentation could be improved by including information about the "back" and "on/off" buttons of the phone and not using the word "widget". According to some evaluators, the heuristic "visibility of the system status" was violated because it took too long for a session to open and there was no feedback about waiting time. Based on the results, the manual was rewritten and a new prototype was built with easier language, more consistency, other icons, extended swiping function to all screens, and an extended surface to the whole screen in the day view. The connectivity of the dialogue sessions was improved and indicators for progress and waiting time were added. There were several remarks about the term "sessions," but no better expression was found.

Usability test in lab

The new prototype was evaluated in a lab situation by four male patients and one female patient. The participants spent on average 21 hours a week on a computer. One participant had prior experience with a smartphone (Table 2).

The participants rated "comprehensiveness" and "readability" of the manual with an average of 4.5 and 4.8, respectively, on a scale ranging from 1 to 7 (Table 1). Based on the suggestions for improvement, the manual was extended with more information about the general use of the smartphone. One participant doubted whether the activity sensor was robust enough.

Although most participants had no previous experience with smartphones, none of the eight tasks on the app were rated as difficult (Table 1). Except in the case of opening the app for one person, no navigation errors were observed. The observers had to give only minor instructions about swiping and opening/returning to the app and widget. Results from the PSSUQ showed that the participants were, overall, satisfied with the usability of the app (details are presented in Figure 3).

Several suggestions for improvement were made during the thinking-aloud procedure. The most prominent suggestions included quicker response of the app on swiping, an always visible timescale in the hour view even when there is no activity, and better distinction between the bars in the month view. Those remarks were translated into improvements in the third prototype of the app.

The dialogue session "preparation for goal setting" was rated as the most difficult. Three out of four participants completed this task on the app. This task was a long session

Table 2 Demographic characteristics of the patient participants

	Phase 2 lab 1 (n=5)	Phase 3 pilot (n=20)	Phase 4 lab 2 (n=5)
Characteristics of the participants			
Age mean (SD)	61.4 (8.5)	60.2 (9.0)	58.6 (7.8)
Sex number (%)			
Male	4 (80%)	11 (55%)	3 (60%)
Female	1 (20%)	9 (45%)	2 (40%)
Disease number (%)			
COPD	2 (40%)	10 (50%)	1 (20%)
Diabetes type 2	3 (60%)	10 (50%)	4 (80%)
Education level number (%)			
≤ intermediate vocational education	3 (60%)	–	4 (80%)
> intermediate vocational education	2 (40%)	–	1 (20%)
Computer experience number (%)			
≤5 hours per week	1 (20%)	2 (10%)	1 (20%)
>5 hours per week	4 (80%)	18 (90%)	4 (80%)
Smartphone experience number (%)			
Yes	1 (20%)	6 (30%)	1 (20%)
No	4 (80%)	14 (70%)	4 (80%)

Abbreviations: COPD, chronic obstructive pulmonary disease; SD, standard deviation; lab, laboratory test.

with different input methods per question. Based on the results of the lab tests, a recommendation was added that long and complex dialogue sessions should be completed via the website rather than via the app. These sessions were: "registration", "preparation for goal setting", and "set up an activity plan".

In relation to completing sessions on the website, it was observed that the "home" button of the website should be made more prominent, all monitoring results (results per hour, day, and week) visible on the app should also be visible on the web interface, and the intention of the "reminder" function should be more evident. Two out of three participants made errors while using this function.

Furthermore, phrases like "you must" were perceived as paternalistic and should be changed to "you may".

According to the PSSUQ (see Figure 4), participants were satisfied with the usability of the website. For all components of the tool, it seemed that participants with a higher education level were more critical than people with a lower education level.

All suggestions were incorporated in the next prototype of the tool, except for the suggestion to present all monitoring results via the web interface. This was not technically feasible.

Pilot test in real life

The third prototype in the pilot was evaluated in real life by eleven men and nine women, with a mean age of 60.2 years
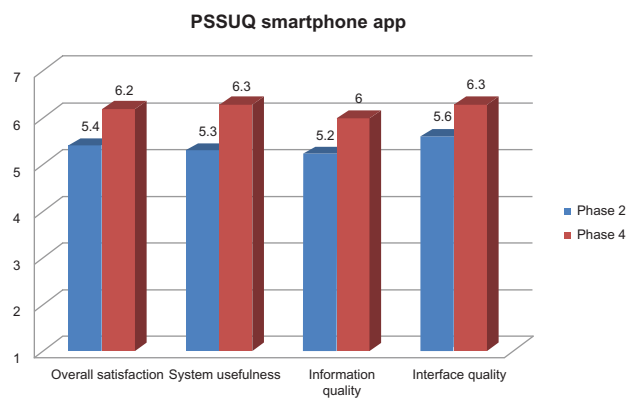
PSSUQ smartphone app

Figure 3 Post-study system usability questionnaire (PSSUQ) subscores for the smartphone app after phases 2 and 4.
Note: Higher scores indicate better usability.
Abbreviation: PSSUQ, post-study system usability questionnaire; app, application.

(standard deviation 9.0 [Table 2]). The data of the interviews and the log files were clustered into four themes: the sensor, data presentation, connection problems, and dialogue sessions.

Activity sensor

The participants had no difficulty wearing the activity sensor on a daily basis; however, most participants were afraid of losing the sensor. This problem was solved in the new prototype by adding a security clip with a thread that can be attached to belt loops. Two patients had to quit the pilot study because the hardware in the sensor broke, one of them

because the participant accidentally put the sensor in the washing machine.

Data presentation

Most participants were positive about the tool in general. They liked to see the distance to their target goal and the course of activities over the day in the hour view. However, almost all participants had the idea that the activity results were not consistent with their experienced activity. This inconsistency had two causes:

1. A delay or failure in transmitting the activity data from the sensor to the phone. Therefore, besides the automatic transmission of data every 15 minutes, synchronization occurs when opening the activity menu and a refresh button has been added in the new prototype to actively synchronize the app with the sensor and server.

2. The activity sensor starts counting if the average acceleration per minute is approximately ≥ 3.5 km/hour and upper body movements are not captured. This was better explained in a new version of the manual and in the instruction movies. In addition, the practice nurse will have the ability to lower the threshold to 2 or 3 km/hour if participants are not able to reach the threshold noted above.

There were almost no comments on the usability of the app. In the day view, the word "moderate" was changed to "active" and "intense" to "active plus", since the word
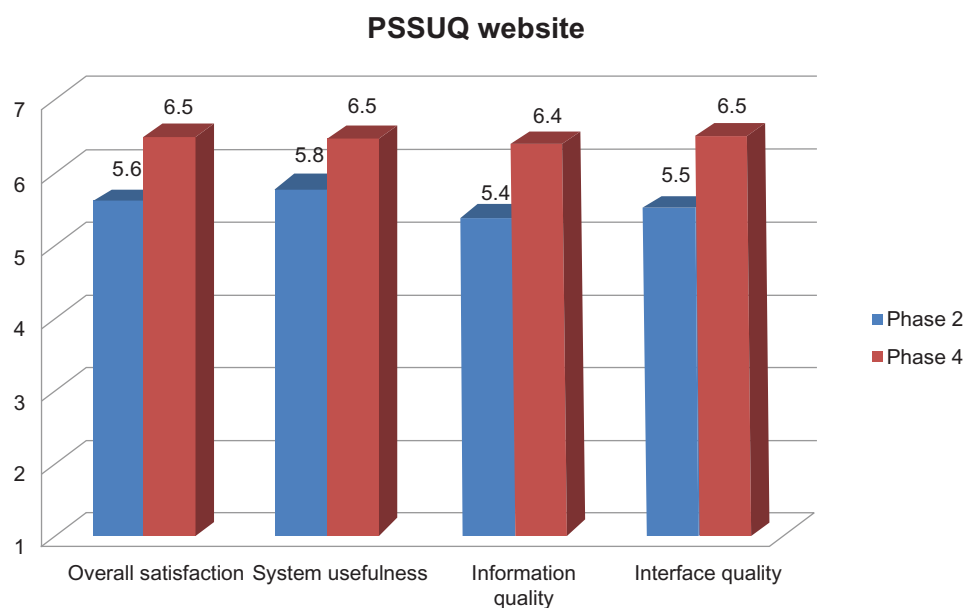
PSSUQ website

Figure 4 Post-study system usability questionnaire (PSSUQ) subscores for the website after phases 2 and 4.
Note: Higher scores indicate better usability.
Abbreviation: PSSUQ, post-study system usability questionnaire.

"moderate" was viewed as not encouraging. Furthermore, people were puzzled about the registered activity at 6 am, which seemed to be the summed activity and noise from midnight till 6 am. This has been solved by adding an "N" for night activity and raising the lowest threshold to separate noise from activity.

Connection problems

In order to make a connection between the phone and the server, participants had to log in on the phone once at the start of the intervention. Seven participants forgot to log in or did not manage to complete the task because they were not able to type in their correct user name and log-in on the phone. Based upon this, the registration session has been extended with a task to log in and an instruction on how to do this. Furthermore, the manual has been extended.

Seven patients complained that automatic data transfer from sensor to phone (which should occur every 15 minutes) did not work properly. It appeared that they had erroneously deactivated the smartphone's data connection. In addition, sometimes the Bluetooth connection failed because the sensor was out of range of the phone.

Due to the sleep mode of the phone and incorrect timings of data transmission, the connection between the phone and the server failed frequently, which meant that the patients and practice nurse did not see results on the website. This also meant that only two participants received more than one feedback message, since these messages depend on goal achievement and, therefore, the forwarded activity data.

Dialogue sessions

The participants did not give detailed feedback on the content of the sessions; however, the following suggestions for improvements were revealed. Participants were confused about the difference between the diary sessions and the "remarks of today" and, in their view, there were too many sessions. In response, the sessions were renamed, the session "remarks of today" was no longer announced by email, the diary sessions were offered less often, and some text fields were enlarged. Based upon these results, all errors were solved in a new prototype and the help documentation was extended, the manual and instruction movies were adapted, and all comments about ambiguities were merged in a "frequently asked questions" file.

The SUMI questionnaires of 14 participants were analyzed (four did not fill out the questionnaire and two had more than four missing values). The results are presented in Figure 5. The score of 50 on the global scale indicates that satisfaction with the It's LiFe! tool is reasonable. The efficiency, helpfulness, control, and learnability could be improved. The only sub-score above average is "affect", which indicates that the users liked the interfaces and the idea of the tool.

Usability test in lab

The fourth prototype was evaluated by three men and two women, with a mean age of 58.6 years (standard deviation 7.8), in a lab situation. Only one participant had experience with a smartphone.
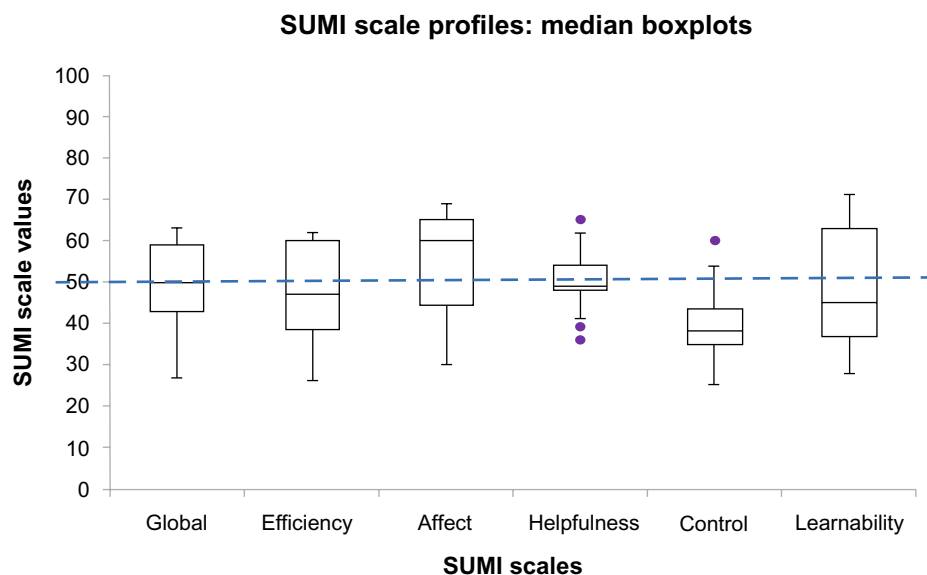

SUMI scale profiles: median boxplots

Figure 5 Results of the SUMI questionnaire, completed by 14 participants after the pilot in real life.
Note: Usability is considered reasonable at 50.
Abbreviation: SUMI, Software Usability Measurement Inventory.

Table 3 Desirability of It's LiFe! App

Participant number	Desirability app				
6	Advanced	Fascinating	Innovative	**Motivating**	**Usable**
7	Attractive	**Suitable**	**Easy to use**	**Motivating**	**Understandable**
8	**Accessible**	Clear	**Easy to use**	Enthusiastic	**Useful**
9	**Accessible**	Comprehensive	**Easy to use**	Familiar	**Understandable**
10	Business-like	**Suitable**	Straightforward	**Usable**	**Useful**

Note: The words in bold were chosen by more than one participant.
Abbreviation: App, application.

The comprehensiveness and readability of the manual had clearly improved compared to the first usability test in the lab (phase 1), as is shown in Table 1. During the thinking-aloud procedure, the main suggestion made was to replace technical or English terms in the manual with easier terms or Dutch language.

All tasks on the app were scored as less or equally complex compared to the first lab test, few errors were observed, and difficulties were only faced with opening the app menu and the understanding of the terms "active" and "active +" for moderate and intense activities, respectively. According to the scores on the PSSUQ, user satisfaction with the usability of the app had also improved, as shown in Figure 3. During the thinking-aloud procedure, no suggestions for improvement were made regarding the app.

The "registration session," which was rated as the easiest session in the first lab test, was now rated the most difficult session (Table 1). This could be explained by the fact that this session was extended with a procedure to prevent people from not logging in (one of the major issues in the real-life test). As a result, the "registration session" had become more complex: due to the small keyboard on the phone, typing errors were made and the backspace and symbol buttons were hard to find. In the final prototype, the procedure for logging in on the phone is explained extensively in the registration session and an instruction movie for this session has been made available.

Based upon the results of the thinking-aloud procedure during completion of the dialogue sessions on the website, some text was adapted, the activity plan could be filled in

twice, the structure of the "compose activity plan" was changed, and the number of questions was lowered. Results from the PSSUQ (Figure 4) show that satisfaction with the usability of the website improved compared to the earlier version of the prototype used in phase 1.

The participants rated the desirability of the app and website positively. To describe the app, three participants chose the phrase "easy to use" and two participants chose the words "motivating", "usable", "understandable", "useful", "suitable", and "accessible". The only chosen word that could be interpreted as negative was "business-like". Concerning the website, "accessible" was chosen three times and "clear", "interesting", "understandable", and "stimulating" were chosen by two participants (Tables 3 and 4). Again, people with higher education levels tended to be more critical. Figure 6 shows the final interfaces of the app.

Discussion

In response to the heuristic evaluation and tests in the laboratory and in real life, a new prototype of the It's LiFe! tool was developed. The usability of the tool improved during the study. The interface of the app needed relatively small adaptations. Most adaptations were made to the dialogue sessions on the phone involving the keyboard. In addition, connectivity problems were identified and solved.

The combination of laboratory and real-life tests, and the combination of "expert-based" and "user-based" usability tests, revealed a wide range of usability issues.[19] The heuristic evaluation and the thinking-aloud procedure in the laboratory

Table 4 Desirability of It's LiFe! Website

Participant number	Desirability website				
6	Helpful	**Interesting**	Motivating	Personal	**Stimulating**
7	**Accessible**	Calm	**Clear**	Comprehensive	**Understandable**
8	**Accessible**	Connected	Comfortable	Enthusiastic	Reliable
9	**Interesting**	New	**Stimulating**	**Understandable**	Useful
10	**Accessible**	Approachable	**Clear**	Innovative	Suitable

Note: The words in bold were chosen by more than one participant.
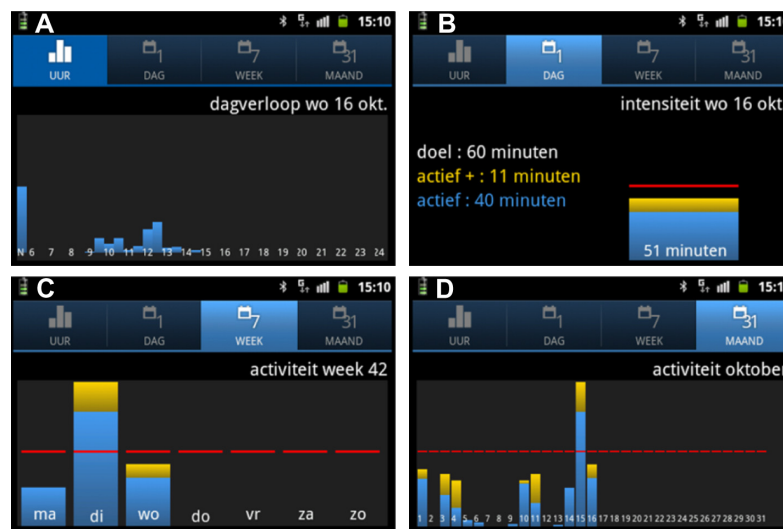
Figure 6 Screenshots of the final prototype of the smartphone app.
Notes: (**A**) Day view per hour from 6 am to 12 pm – "N" indicates summed activity between midnight and 6 am. (**B**) Day view – The yellow parts indicate high intense activities, blue indicates moderate intense activities, the red line "doel" is "goal", and "minuten" is "minutes". (**C**) Week view. (**D**) Month view.
Abbreviation: app, application.

tests revealed the most valuable and detailed feedback on the interfaces and texts. The pilot in real life revealed practical issues such as connectivity problems and overall usability.

After the first test in the lab, usability was considered as good. The real-life test, however, revealed a whole different range of usability problems, and satisfaction with the It's LiFe! tool was low according to the SUMI results. This low appraisal was a logical consequence of the connectivity problems that occurred during the pilot. The high score on "affect" indicates that satisfaction with the interfaces and the idea of the tool was high, which is most likely due to the involvement of end users in the development process of the tool, the prior usability test in the laboratory, or because the participants liked the concept of the tool very much. It is possible that usability was rated more positive in the lab tests and in the interviews during the real-life test compared to the SUMI questionnaire because of social desirability bias. In other fields, it is observed that socially desirable answers are given more often in face-to-face interviews.[29,30] People in our lab tests may have wanted to prove that they had the capabilities necessary to use the system or wanted to satisfy the researcher.[31]

All results should be considered with caution because of the small sample size. Nevertheless, it is known that tests with five participants are able to uncover 85% of usability issues. This number of evaluators is stated to be a good tradeoff between completeness and investment.[32,33] Therefore, we think most usability issues have been revealed.

This study shows the importance of a mixed-method approach, since different issues were revealed in the lab compared to the real-life test. The interfaces can be very effective, efficient, and desirable in a lab situation, but if communication fails between different components of the tool in a real-life situation, satisfaction will be low.

Almost all technical errors and suggestions for improvement have been incorporated in the newest version of the It's LiFe! tool. A crucial aspect that could not be handled is the need to log in on the phone in order to make a connection between the phone and the server. This is because privacy must be respected in all cases. Hopefully, the guidance provided by the instruction movie added to the registration session will be sufficient in further use. Furthermore, a hip-worn activity sensor has well-known restrictions, such as not capturing upper body movements. During the development process, the addition of another physiological measure was considered; however, this does not significantly improve the assessment of energy expenditure and reduces wearing comfort.[34]

The effectiveness of the tool in combination with the SSP on physical activity level (exercise) will be tested in a randomized controlled trial.

Acknowledgments

and especially the patient representatives, Jos Donkers and Ina van Opstal, for their remarks during the research meetings. Thanks to Trudy van der Weijden for her advice during the development process and Science Vision for the adaptations of the pictures.

Disclosure

References

1. Haapanen N, Miilunpalo S, Vuori I, Oja P, Pasanen M. Association of leisure time physical activity with the risk of coronary heart disease, hypertension and diabetes in middle-aged men and women. *Int J Epidemiol*. 1997;26(4):739–747.
2. Fentem PH. ABC of sports medicine. Benefits of exercise in health and disease. *BMJ*. 1994;308(6939):1291–1295.
3. World Health Organisation. Global recommendations on physical activity for health. In: World Health Organisation, editor. WHO Press Switzerland. Geneva; 2010.
4. Haskell WL, Lee IM, Pate RR, et al. Physical activity and public health: updated recommendation for adults from the American College of Sports Medicine and the American Heart Association. *Circulation*. 2007;116(9):1081–1093.
5. Schiller JS, Lucas JW, Ward BW, Peregoy JA. Summary health statistics for US adults: National Health Interview Survey, 2010. *Vital Health Stat 10*. 2012(252):1–207.
6. Craig R, Mindell J, Hirani V. *Health Survey for England 2008. Volume 1: Physical Activity and Fitness*. London: NHS Information Centre; 2009.
7. Hildebrandt VH, Bernaards CM, Stubbe JH. Trend Report Exercise and Health 2010/2011. TNO Quality of Life, Physical Activity and Health; 2013. Available from: http://www.webcitation.org/6MxYEo8d6. Accessed September 28, 2012.
8. Lung Alliance the Netherlands. Zorgstandaard COPD [Care standard COPD]. Amersfoort: Long Alliantie Nederland [Host]; 2010. Dutch.
9. Dutch Diabetes Federation. NDF Zorgstandaard: transparantie en kwaliteit van diabeteszorg voor mensen met diabetes type 2 [NDF Care Standard: Transparency and Quality Diabetes are for people with type 2 Diabetes]. Amersfoort: Nederlandse Diabetes Federatie; 2007.
10. Baan D, Heijmans M. Mensen met COPD in beweging. Factsheet. 2012.
11. Albright A, Franz M, Hornsby G, et al. American College of Sports Medicine position stand. Exercise and type 2 diabetes. *Med Sci Sports Exerc*. 2000;32(7):1345–1360.
12. Carroll JK, Fiscella K, Epstein RM, et al. Physical activity counseling intervention at a federally qualified health center: improves autonomy-supportiveness, but not patients' perceived competence. *Patient Education and Counseling*. 2013;92(3):432–436.
13. Hebert ET, Caughy MO, Shuval K. Primary care providers' perceptions of physical activity counselling in a clinical setting: a systematic review. *Br J Sports Med*. 2012;46(9):625–631.
14. Bardsley M, Steventon A, Doll H. Impact of telehealth on general practice contacts: findings from the whole systems demonstrator cluster randomised trial. *BMC Health Services Research*. 2013;13:395.
15. Price M, Yuen EK, Goetter EM, et al. mHealth: a mechanism to deliver more accessible, more effective mental health care. *Clin Psychol Psychother*. Epub August 5, 2013.
16. Esposito M, Ruberto M, Gimigliano F, et al. Effectiveness and safety of Nintendo Wii Fit Plus™ training in children with migraine without aura: a preliminary study. *Neuropsychiatr Dis Treat*. 2013;9:1803–1810.
17. van der Weegen S, Verwey R, Spreeuwenberg M, Tange H, van der Weijden T, de Witte L. The Development of a Mobile Monitoring and Feedback Tool to Stimulate Physical Activity of People With a Chronic Disease in Primary Care: A User-Centered Design. *JMIR Mhealth and Uhealth*. 2013;1(2):e8.
18. Shah SG, Robinson I. Benefits of and barriers to involving users in medical device technology development and evaluation. *Int J Technol Assess Health Care*. 2007;23(1):131–137.
19. Jaspers MW. A comparison of usability methods for testing interactive health technologies: methodological aspects and empirical evidence. *Int J Med Inform*. 2009;78(5):340–353.
20. The international organization for standardization. 9241-11. Ergonomic requirements for office work with visual display terminals (VDTs). *Part 11: Guidance on Usability* 1998.
21. Frøkjær E, Hertzum M, Hornbæk K. Measuring usability: are effectiveness, efficiency, and satisfaction really correlated? Proceedings of the SIGCHI conference on Human Factors in Computing Systems. April 1–6, 2000. The Hague, the Netherlands. ACM Press. New York, NY, USA. 2000.
22. Nielsen J, Mack RL. *Usability Inspection Methods*. New York: Wiley; 1994.
23. Lewis JR. IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. *Int J Hum Comput Interact*. 1995;7(1):57–78.
24. Verwey R, Weegen Svd, Spreeuwenberg M, Tange H, Weijden Tvd, Witte Ld. A pilot study of a tool to stimulate physical activity in patients with COPD or type 2 diabetes in primary care. *J Telemed Telecare*. 2014;20(1):29-34.
25. Kirakowski J, Corbett M. SUMI: the Software Usability Measurement Inventory. *Br J Educ Technol*. 1993;24(3):210–212.
26. Kirakowski J. The Software Usability Measurement Inventory: Background and Usage. P Jordan, B Thomas, B Weerthmeester (Editors). Usability Evaluation in Industry. Taylor and Frances, London, UK.
27. Benedek J, Miner T. Measuring Desirability: New methods for evaluating desirability in a usability lab setting. *Proceedings of Usability Professionals Association*. July 8–12. Orlando 2002.
28. Hsieh H-F, Shannon SE. Three approaches to qualitative content analysis. *Qual Health Res*. 2005;15(9):1277–1288.
29. Waruru AK, Nduati R, Tylleskar T. Audio computer-assisted self-interviewing (ACASI) may avert socially desirable responses about infant feeding in the context of HIV. *BMC Med Inform Decis Mak*. 2005;5:24.
30. Luke N, Clark S, Zulu EM. The relationship history calendar: improving the scope and quality of data on youth sexual behavior. *Demography*. 2011;48(4):1151–1176.
31. Sauro J. [serial on the internet] 9 Biases in Usability Testing. [updated August 21, 2012]. Available from: http://www.measuringusability.com/blog/ut-bias.php. Accessed November 18, 2013.
32. Why You Only Need to Test with 5 Users. Evidence-Based User Experience Research, Training, and Consulting [serial on the internet]. Nielsen J; 2000 [updated March 19, 2000]. Available from: http://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/. Accessed November 18, 2013.
33. Nielsen J, Landauer TK. A mathematical model of the finding of usability problems. Paper presented at: Proceedings of the INTERACT'93 and CHI'93 conference on Human Factors in Computing Systems; April 24–29, 1993. Amsterdam, the Netherlands. ACM Press. New York, NY, USA. 1993.
34. Plasqui G, Bonomi AG, Westerterp KR. Daily physical activity assessment with accelerometers: new insights and validation studies. *Obes Rev*. 2013;14(6):451–462.