International Journal of Medical Informatics 128 (2019) 24–31

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Assessing usability of eHealth technology: A comparison of usability benchmarking instruments

Marijke Broekhuis[a,b,*], Lex van Velsen[a,b], Hermie Hermens[a,b]

[a] Roessingh Research and Development, Roessinghsbleekweg 33b, 7522AH, Enschede, the Netherlands
[b] Biomedical Signals and Systems, Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS), University of Twente, Enschede, the Netherlands

ARTICLE INFO

Keywords:
eHealth
Usability benchmarking
Think aloud
SUS
Usability task metrics
Evaluation

ABSTRACT

Background: It is generally assumed that usability benchmarking instruments are technology agnostic. The same methods for usability evaluations are used for digital commercial, educational, governmental and healthcare systems. However, eHealth technologies have unique characteristics. They need to support patients' health, provide treatment or monitor progress. Little research is done on the effectiveness of different benchmarks (qualitative and quantitative) within the eHealth context.

Objectives: In this study, we compared three usability benchmarking instruments (logging task performance, think aloud and the SUS, the System Usability Scale) to assess which metric is most indicative of usability in an eHealth technology. Also, we analyzed how these outcome variables (task completion, system usability score, serious and critical usability issues) interacted with the acceptance factors Perceived benefits, Usefulness and Intention to use.

Methods: A usability evaluation protocol was set up that incorporated all three benchmarking methods. This protocol was deployed among 36 Dutch participants and across three different eHealth technologies: a gamified application for older adults ($N = 19$), an online tele-rehabilitation portal for healthcare professionals ($N = 9$), and a mobile health app for adolescents ($N = 8$).

Results: The main finding was that task completion, compared to the SUS, had stronger correlations with usability benchmarks. Also, serious and critical issues were stronger correlated to task metrics than the SUS. With regard to acceptance factors, there were no significant differences between the three usability benchmarking instruments.

Conclusions: With this study, we took a first step in examining how to improve usability evaluations for eHealth. The results show that listing usability issues from think aloud protocols remains one of the most effective tools to explain the usability for eHealth. Using the SUS as a stand-alone usability metric for eHealth is not recommended. Preferably, the SUS should be combined with task metrics, especially task completion. We recommend to develop a usability benchmarking instrument specifically for eHealth.

1. Introduction

Usability is often named as one of the crucial requirements for an eHealth technology. Generally, usability is described as 'the extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use' [1]. This definition emphasizes how usability, and the perception of usability, can differ across products, target audiences and contexts. This is especially true when designing a usable system for the eHealth domain, because usability of eHealth differs from other domains on several aspects. First, user satisfaction with an eHealth system is difficult to establish. While e-commerce seduces customers with personal messages that fit perfectly with their needs, and thus attempt to increase user's satisfaction of the system, for eHealth the users need to be informed on both positive and negative effects of their health behaviour. This means that users sometimes need to hear advice they do not want to hear (*e.g.* taking a walk instead of watching television), which can influence their system satisfaction. Second, health communication needs to be tailored to the level of health literacy of individual users [2] to improve patient's health knowledge [3] and self-management of health [4]. Third, all of the above mentioned factors are further complicated since having a chronic illness can lead to heightened stress and anxiety [5], [6]. This hinders the uptake of information and learning skills for self-management.

* Corresponding author.
E-mail address: m.broekhuis@rrd.nl (M. Broekhuis).

https://doi.org/10.1016/j.ijmedinf.2019.05.001
Received 22 January 2019; Received in revised form 4 April 2019; Accepted 4 May 2019

When health care professionals also use the eHealth system, there are additional factors to consider, especially concerning information overload [7]. It is tempting to provide much information on a patient's health and progress, but care professionals can only digest a limited amount.

There are many methods for evaluating eHealth usability: Questionnaires are cost-friendly methods to quickly gather user feedback from large sample sizes [8]; thinking-aloud is very effective in identifying usability problems with only a small number of participants [9]; interviews and focus groups are great for collecting in-depth information on user perceptions of the system [10] and by applying usability task metrics one can assess how efficiently and satisfactorily participants perform tasks [11]. Klaasen and colleagues [12] found that questionnaires are the most preferred method (69%) for usability evaluations in eHealth. In 28.4% of the studies standardized questionnaires were applied, of which the System Usability Scale (SUS) [13] is most frequently used.

The popularity of the SUS for eHealth is understandable. Its method (questionnaire), length (10 items), easy score interpretation (range between 0 and 100), validity as established in non-eHealth domains [14]– [16] and availability (free of charge) make it a popular choice, also in the eHealth domain [17–,18,19,20,21]. However, although the scoring range goes from 0 to 100, few SUS scores drop below 50 [14], [22]. To overcome this problem, Sauro and Lewis [11] proposed a curved grading scale from A to F (A = excellent usability, F = clearly deficient), which is based on a normal distribution of the percentile range of average SUS scores [23]. However, this curved grading scale is based on a wide variety of technologies, such as commercial and financial websites [24], enterprise software applications, and landline telephones [14]. Because there are specific factors for eHealth that could affect the perceived usability (*e.g.* health literacy), it is unclear if the SUS still provides accurate results when compared to other benchmarks in the eHealth domain. Some studies compared the SUS with a seven-point adjective rating scale (worst imaginable – best imaginable) [16], [25] [26], and task metrics (such as completion rate and task completion time) [15], [27], but no comparisons have been made with the number of usability problems in a technology and their severity, that are derived from qualitative data collection methods. Since one wants the benchmark score to be predictive of actual usability (and hence, the (non)presence of usability problems), this is somewhat odd. After all, the list of actual usability problems and their effect on effective use of the system is the best indicator of a technology's usability. In short, it can be considered to be the 'golden standard'. Also, the validity of the SUS for eHealth has yet to be thoroughly examined. eHealth is often designed for specific patient groups with physical or cognitive impairments [28–30]. In its questioning and score calculation, the SUS does not take these factors into account.

In this study, we examined the suitability of different usability benchmarking tools for the eHealth context. More specifically, we determined the relative value of the SUS and different usability task metrics: Task completion, time on task, task satisfaction, errors on task, and steps per task. The predictive value of these benchmarks were assessed, in relation to the number and severity of usability issues that were elicited from thinking-aloud sessions. For practitioners, this study defines which metrics they should choose for benchmarking eHealth usability.

2. Materials and methods

2.1. Case studies

We assessed the suitability of different usability benchmarking methods for the eHealth context via three case studies: A gamified application for training the physical condition of frail older adults, a tele-rehabilitation portal for rehabilitation professionals, and a mobile smoking cessation app.

Fig. 1. Screenshot of the gamified application 'Stranded'.

2.1.1. Case 1 – gamified application

The serious game 'Stranded' is developed to optimize the health of (pre)frail older adults (65+ years). In this game, players have to complete a physical training regimen in order to unlock pieces of a boat to escape an uninhabited island. Additionally, they can receive rewards such as mini-games and preparing meals in a virtual vegetable garden. It is connected to a web portal, where a physical therapist can create a personalized training regime, communicate with the patient, and provide health education (Fig. 1).

2.1.2. Case 2 – tele-rehabilitation portal

The tele-rehabilitation portal is an online tool for healthcare professionals, working in the children's department of a rehabilitation center. It supports monitoring the development of children, such as scheduling physical activities and setting new goals for them, and facilitates communication between parents and therapists (Fig. 2).

2.1.3. Case 3 – Mobile app

'Stopstone' is a smartphone app for motivating young adolescents to quit smoking. In the app, users can identify moments at which they find it difficult not to smoke and determine their strategies and motivations for dealing with these moments (Fig. 3).

2.2. Participants

Participants were recruited either via convenience or snowball sampling. For the gamified application, participants had to be 65 years or older and had to have basic computer skills, like for instance sending an e-mail. For the tele-rehabilitation portal, therapists of different domains (*e.g.*, physiotherapy, social care) were recruited. Adolescents between 19–25 years were recruited for the mobile app. All participants lived in the Netherlands and had no prior experience with the evaluated technology.

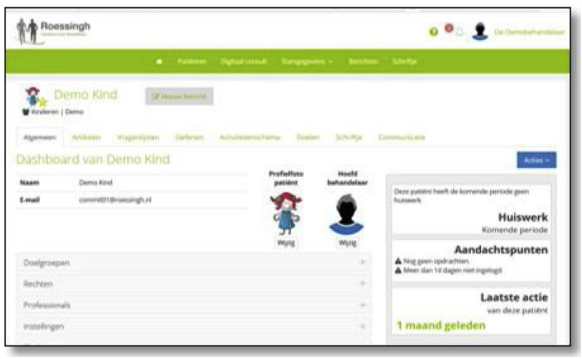

Fig. 2. Screenshot of the tele-rehabilitation portal.

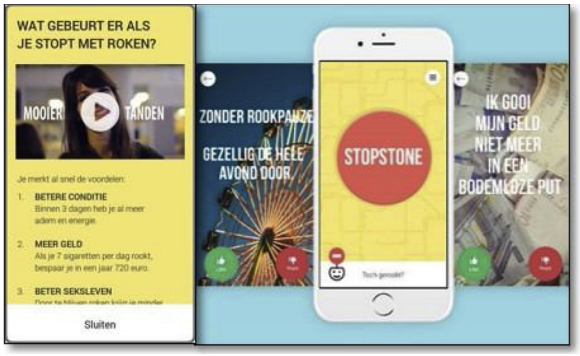

Fig. 3. Image of the mobile app 'Stopstone' (© 2016 Trimbos Institute. Reprinted with permission).

2.3. Study procedure

Each case used the same evaluation protocol. First, participants received a short demographics questionnaire (gender, age, education). Then, a concurrent think-aloud protocol was administered in which they were given several tasks to complete within the respective system while verbalizing their thoughts. This data was supplemented by researcher observations. At the same time, usability performance metrics (task completion, task completion time, satisfaction, steps, and errors) were assessed. Participants had five minutes to complete each task. If they did not complete the task within that time or did not want to proceed, they proceeded to the next task. The first task was to freely browse the eHealth technology for several minutes to simulate real-life usage of a new technology. The task metrics task completion, task completion time and task satisfaction were not measured for this explorative task. Then, the participants were given several specific tasks within the system. These tasks reflected central functionalities of the technology. For example, for the gamified application the participants had to perform a physical exercise (task 2) and find an e-mail from their therapist (task 3). For the tele-rehabilitation portal, the participants had to schedule a physical exercise for the patient (task 3) and write an e-mail to the parents of the patient (task 6). For the mobile app, participants had to add a stop-strategy (task 4) and calculate how much money they would save if they quit smoking (task 5). After each task, the participants were given the After-Scenario Questionnaire (ASQ) [31] to measure task satisfaction. After carrying out all tasks, they filled out the SUS. Last, a short interview was conducted to discuss participants' intentions to use the technology; we asked them about Perceived benefits, Usefulness and Intention to use [32–34].

The usability tests had an average length of 60 min. The tests were conducted in a usability lab or on location. Each test was performed in a closed room to minimize distraction. Audio and screen capture recordings were made during the tests.

2.4. Ethics

All participants signed an informed consent form prior to the study. The nature of these general tests among healthy volunteers did not require formal medical ethical approval, according to Dutch law [35].

2.5. Qualitative analysis

Transcripts were used to identify usability issues using the following process:

1) One researcher (MB) identified all errors in the think-aloud transcripts and observational notes;
2) A second researcher (LvV) also examined this dataset. Discrepancies were solved and the first researcher (MB) re-analyzed the full data set with this final list.
3) The first researcher (MB) created an overview of usability issues by grouping similar errors into one usability issue (*e.g.*, recurring errors from clicking on non-clickable elements were grouped as 'the user has difficulty distinguishing clickable from non-clickable elements in the interface');
4) The second researcher (LvV) examined this usability issue overview. The researchers discussed discrepancies and created a final overview;
5) The first researcher awarded each usability issue with a severity score (minor, serious, or critical), following a procedure from [36]. The severity ratings were verified by the second researcher (LvV).

The answers to the interview questions were converted into binomial code (0 = negative, 1 = positive) to allow for statistical analyses. To ensure validity, the coding process was similar to that of the usability issues.

2.6. Statistical analyses

The data was analyzed using SPSS 19.0. Descriptive statistics were computed for demographic variables (means, percentages). Since normality tests indicated that normal distribution could not be assumed for most usability benchmarks, this data is presented non-parametrically. Binomial data (task completion, perceived intention-to-use, perceived benefits, and perceived usefulness) were analyzed with 95% binomial confidence intervals, using the Wilson Score method [37] from the episheet of Rothman and Boice [38]. A two-tailed Kendall Tau correlation was computed among the usability benchmark scores and number of usability issues (task completion, task completion time, task satisfaction, steps per task, errors per task, minor issues, serious issues, and critical usability issues). For this analysis, task completion scores were transposed to an ordinal scale (0 completions, 1 completion, *etc.*). Then, for the seemingly strong correlations significance tests were computed using the calculator of Lee and Preacher [39], which is based on the work of Steiger [40]. The variables Perceived benefits, Usefulness and Intention to use were each split into two categories: (1) perceiving benefits – not perceiving benefits, (2) useful – not useful and (3) intention to use, no intention to use. Mann-Whitney U-tests were conducted for each binomial variable to examine if the medians between the two categories were significantly different in relation to the SUS, task completion, serious and critical issues.

Table 1
Demographics (*N*, gender, age, education).

	N	Gender		Age		Education		
		Male	Female	M	SD	≥Lower vocational education	Vocational education	≤higher vocational education
Case 1	19	12 (63.2%)	7 (36.8%)	74.3	6.08	3 (15.8%)	12 (63.2%)	4 (21.1%)
Case 2	9	1 (11.1%)	8 (88.9%)	43.4	11.4	–	–	8 (100%)
Case 3	8	4 (50%)	4 (50%)	23.13	2.03	4 (50%)	3 (37.5%)	1 (12.5%)

Table 2
Binomial confidence intervals (task completion, perceived benefits, perceived usefulness, perceived intention-to-use).

		Case and tasks complete (percentage, 95% CI for percentage)		
		Case 1	Case 2	Case 3
Task completion	T1*	n.a.	n.a.	n.a.
	T2	8/19 (42.1, 23.1, 63.7)	9/9 (100, 70.1, 100)	8/8 (100, 67.6, 100)
	T3	9/19 (474 27.3, 68.3)	6/9 (66.7, 35.4, 87.8)	8/8 (100, 67.6, 100)
	T4	2/19 (10.5, 2.9, 31.4)	9/9 (100, 70.1, 100)	8/8 (100, 67.6, 100)
	T5	2/19 (10.5, 2.9, 31.4)	5/9 (56, 26.7, 81.1)	8/8 (100, 67.6, 100)
	T6	n.a.	9/9 (100, 70.1, 100)	n.a.
Perceived benefits		9/19 (47.4, 27.3, 68.3)	9/9 (100, 70.1, 100)	8/8 (100, 67.6, 100)
Perceived usefulness		2/19 (10.5, 2.9, 31.4)	8/9 (88.9, 56.5, 98)	4/8 (50, 21.5, 78.5)
Perceived intention-to-use		2/19 (10.5, 2.9, 31.4)	9/9 (100, 70.1, 100)	4/8 (50, 21.5, 78.5)

*Since task 1 was a free explore task, there is no completion rate.

3. Results

3.1. Demographics

In total 36 participants, nineteen older adults (case 1), nine therapists (case 2), and eight adolescents (case 3) participated in this study. Table 1 provides an overview of the demographics of participants per case. For case 1, ages ranged between 65 and 87 years, for case 2, between 32 and 60 years, and for case 3 between 19 and 25 years. Most participants had a vocational or higher vocational education. These educational backgrounds are typical for the end-user populations for each application.

3.2. Case 1 – gamified application

3.2.1. Usability benchmarks

The participants evaluated the overall system usability (SUS) with a score of Mdn = 27.5 (95% CI: 10–42.5)). This score falls far below the acceptability baseline of the SUS. When looking at the task completion rates (see Table 2), it shows that participants had difficulty executing the tasks. Tasks 4 and 5 were considered most difficult for the participants, with a 10.5% (95% CI: 2.9, 31.4%) completion rate. Tasks 2 and 3 were relatively easier. 42.1% (95% CI: 23.1, 63.7%) of the participants completed task 2 and 47.4% (95% CI: 27.3, 68.3%) of the participants completed task 3. Table 3 provides an overview of the usability task metrics task completion time, satisfaction, errors, and steps of case 1, the gamified application. It shows that task 3 had the quickest task completion time with an Mdn of 102.4 s (95% CI: 31, 189). Task 5 had the lowest task satisfaction, with an Mdn of 1 (95% CI: 1, 2).

3.2.2. Usability issues

The think-aloud method elicited 287 usability issues. Almost half of these issues (48.8%) were serious issues, with an average Mdn of 8 issues (95% CI: 6, 9) per participant. There were 80 (27.9%) critical issues (Mdn = 4, 95% CI: 3, 5) and 67 (23.3%) minor issues (Mdn = 3, 95% CI: 2, 5) on average. Critical issues consisted of problems such as: 'The user wants to exit the system because s/he cannot find what s/he is looking for in the gamified application'. Examples of serious issues were 'Users with color blindness have difficulty distinguishing elements in the interface'. Minor issues were problems such as 'The user does not like the introduction movie'.

3.2.3. Perceived benefits, usefulness, and intention-to-use

The interviews revealed that 47.4% (95% CI: 27.3–68.3%) of the participants did see some benefits of the gamified application (see Table 2). However, most participants mentioned the system could support their cognitive skills instead of physical activity. Just two participants (10.5%, 95% CI: 2.9–31.4%) thought the system would be useful to support their physical exercises and believed they would use the system.

Table 3
Usability task metrics of the gamified application stranded.

		Task completion time (sec.)	Satisfaction	Errors	Steps
T1	N	n.a.	n.a.	18 **	18**
	Mdn	n.a.	n.a.	8.5	31.5
	95% CI	n.a.	n.a.	4, 11	21, 45
T2	N	8	19	18 **	18 **
	Mdn	166.5	2.3	6	14.5
	95% CI	85, 280	1, 6	2, 13	8, 23
T3	N	9	19	19	19
	Mdn	100	2.7	8	16
	95% CI	31, 189	1, 4.3	3, 15	9, 25
T4	N	2	19	19	19
	Mdn	157	2.3	16	27
	95% CI	154, 160	1.3, 6	10, 33	19, 42
T5	N	2	19	19	19
	Mdn	196*	1	22	29
	95% CI	94, 298	1, 2	6, 27	19, 42
Av.	N	13	19	19	19
	Mdn	142	2.4	13.4	26.6
	95% CI	83, 228	1.4, 4.2	10.8, 16	23.6, 29.8

Table 4
Usability task metrics (task completion time, satisfaction, errors, and steps) of the tele-rehabilitation portal.

		Task completion time (sec.)	Satisfaction	Errors	Steps
T1	N	n.a.	n.a.	9	9
	Mdn	n.a.	n.a.	.0	17
	95% CI	n.a.	n.a.	.0, 1	11, 24
T2	N	9	9	9	9
	Mdn	21	5.3	.0	2
	95% CI	12, 32	4.7, 5.7	.0, .0	2, 3
T3	N	6	9	9	9
	Mdn	127	4.7	4	16
	95% CI	59, 234	3.3, 6.3	.0, 15	7, 28
T4	N	9	9	9	9
	Mdn	70	6.7	.0	7
	95% CI	37, 122	6, 7	.0, 4	7, 17
T5	N	5	8	8	8
	Mdn	99	4.2	3	13
	95% CI	62, 131	1, 7	.0, 22	8, 25
T6	N	9	9	9	9
	Mdn	120	6	.0	9
	95% CI	37, 229	2.3, 7	.0, 7	6, 24
Av.	N	9	9	9	9
	Mdn	81.6	5.3	3.3	12.8
	95% CI	69.5, 146.8	4.8, 5.5	.3, 5.5	9.2, 16

3.3. Case 2 – Tele-rehabilitation portal

3.3.1. Usability benchmarks

The tele-rehabilitation portal had a SUS score of Mdn = 77.5 (95% CI: 60–85), which means the usability of the system is considered good but could be further improved [11]. All participants completed tasks 2, 4, and 6. Task 5 had the lowest task completion rate of 55.6% (95% CI: 26.7–81.1%), see Table 2. Tasks 3 and 5 were considered more difficult to execute and had higher numbers of errors and steps, see Table 4. Task satisfaction was positively rated, with an average of Mdn = 5.3 (95% CI: 4.8, 5.5).

3.3.2. Usability issues

We identified 51 usability issues, of which 23 serious (45.1%), 22 minor (43.1%), and 6 (11.8%) critical. On average, participants had an Mdn of 3 serious issues (95% CI: 1, 4), and a Mdn of 3 minor issues (95% CI: 1, 3). Critical issues (Mdn = .0, 95% CI: 0.0, 2) were only found with the scheduling of exercises for patients: 'The user does not know how to schedule an exercise for the patient in the exercise-interface'. Serious issues were problems like 'The system does not clearly stipulate that the parents, not the children, are the contact persons'. By sending a message to the patient, the therapist is actually sending a message to the parents. Minor issues were issues such as 'The tele-rehabilitation portal does not have a navigational aid, such as a bread crumb trail, for users to keep track of their location within the system'.

3.3.3. Perceived benefits, usefulness, and intention-to-use

All participants perceived the benefits of the tele-rehabilitation portal, see Table 2. The therapists believed the online portal provides a better overview on the progress and activities of the patient, which could improve the patient and parent involvement. All therapists indicated they would use this system because it prevents having to use different systems both for patients as therapists.

3.4. Case 3 – Mobile app

3.4.1. Usability benchmarks

The mobile app had a SUS score of Mdn = 71.3 (95% CI: 45–87.5)). Table 2 shows that the participants had little difficulty completing the tasks in the mobile app. The participants gave tasks 2 (Mdn = 6.2, 95%

Table 5
Usability task metrics (task completion time, satisfaction, errors, and steps) of the mobile app.

		Task completion time (sec.)	Satisfaction	Errors	Steps
T1	N	n.a.	n.a.	8	8
	Mdn	n.a.	n.a.	1.5	40
	95% CI	n.a.	n.a.	.0, 14	27, 62
T2	N	8	8	8	8
	Mdn	83.5	6.2	.0	18.5
	95% CI	65, 223	4.3, 7	.0, 17	10, 28
T3	N	8	8	8	8
	Mdn	67.5	6	.0	14
	95% CI	41, 128	4.7, 7	.0, .0	10, 22
T4	N	8	8	8	8
	Mdn	69	5.3	.0	17.5
	95% CI	41, 187	2.7, 6.7	0, 14	6, 38
T5	N	4	4	4	4
	Mdn	56	4.7	.0	16.5
	95% CI	35, 138	3, 6.7	.0, 2	9, 38
Av.	N	8	8	8	8
	Mdn	87.4	5.7	2.5	21.9
	95% CI	54.3, 132	4.7, 6.6	.0, 4.6	17.4, 29.5

Table 6
Correlation table for the usability metrics.

		SUS	Task completion	Av. time on task	Av. task satisfaction	Av. steps on task	Av. error on task	Minor issues	Serious issues	Critical issues
SUS	*R*	–	–	–	–	–	–	–	–	–
	95% CI	–	–	–	–	–	–	–	–	–
Task completion	*R*	.61**	–	–	–	–	–	–	–	–
	95% CI	.36, .78	–	–	–	–	–	–	–	–
Av. time on task	*R*	−.282*	−.45**	–	–	–	–	–	–	–
	95% CI	−.56, .05	−.68, -.14	–	–	–	–	–	–	–
Av. task satisfaction	*R*	.54**	−.65**	−.16	–	–	–	–	–	–
	95% CI	.26, .74	−.80, -.41	−.46, .18	–	–	–	–	–	–
Av. steps on task	*R*	−.27*	−.31*	.40**	−.14	–	–	–	–	–
	95% CI	−.55, .06	−.58, .02	.08, .64	−.45, .19	–	–	–	–	–
Av. error on task	*R*	−.52**	−.58**	.39**	−.41**	.51**	–	–	–	–
	95% CI	−.72, -.23	−.76, -.30	.07, .63	−.65, -.01	.22, .72	–	–	–	–
Minor issues	*R*	−.27*	−.35**	.17	−.36**	.17	.32*	–	–	–
	95% CI	−.55, .06	−.61, -.02	−.16, .47	−.61, -.03	−.16, .47	−.01, .58	–	–	–
Serious issues	*R*	−.4**	−.64**	.41**	−.50**	.35**	.57**	.43**	–	–
	95% CI	−.64, -.08	−.80, -.40	.09, .65	−.71, -.21	.03, .61	.29, .75	.12, .67	–	–
Critical issues	*R*	−.47**	−.75**	.44**	−.53**	.36**	.69**	.33*	.69**	–
	95% CI	−.69, -.17	−.87, -.56	.13, .67	−.73, -.24	.04, .62	.47, .83	.0, .59	.46, .83	–

*$p \leq 0.05$, **$p \leq 0.01$

CI: 4.3, 7) and 3 (Mdn = 6, 95% CI: 4.7, 7) a high task satisfaction score, see Table 5. Interesting is that while participants needed to follow more steps to complete the tasks, the number of errors is quite low, with an average of Mdn = 2.5 (95% CI: 0.0, 4.6).

3.4.2. Usability issues

A total of 29 usability issues were identified, of which 14 (48.3%) were minor and 15 (51.7%) were serious issues. On average, participants had an Mdn of 2 serious issues (95% CI: 0.0, 5), and a Mdn of 3 minor issues (95% CI: 0.0, 4). No critical issues came up. Serious issues were problems such as 'The user has difficulty finding the location where a cessation strategy can be added for a difficult moment'. Minor issues consisted of problems such as 'The interface does not explain what type of notifications the app can send you'.

3.4.3. Perceived benefits, usefulness, and intention-to-use

All eight adolescents thought the mobile app 'Stopstone' had some benefits, see Table 2. They liked the app because it is easy to use and because it has multiple options that confronts users with smoking habits, especially the 'budget option', an option in which you can calculate how much money you save by not buying cigarettes. Although they all saw the advantages of the system, only four adolescents (50%, 95% CI: 21.5–78.5%) perceived the system to be useful for themselves. One of the reasons being that some participants believed that the motivation to quit smoking should stem from the user, not from an app. Four adolescents (50%, 95% CI: 21.5–78.5) thought they would use the mobile app because it would provide them insights into how smoking affects their life, such as identifying moments they find it difficult not to smoke.

3.5. Correspondence among SUS, usability task metrics, and usability issues

We analyzed the relationships between the usability benchmarks and the number and severity of the usability issues. These correlations were computed across the three case studies. The correlation matrix can be viewed in Table 6. The table shows that (1) task completion has stronger correlations with task metrics and usability issues than the SUS, and (2) serious and critical issues have stronger correlations with task metrics, except for task satisfaction, than the SUS. The correlation matrix shows considerable disparities between the SUS and task completion on serious issues ($r\tau = -0.397$ vs $r\tau = -.644$), critical issues ($r\tau = -.470$ vs $r\tau = -.753$) and task completion time ($r\tau = -.282$vs $r\tau = -.447$,). Two-tailed significance tests of the correlations [40], show that there are only significant differences found between the SUS and task completion on critical issues ($z = 2.62$, $p = 0.01$) and serious issues ($z = 2.02$, $p = 0.04$), not for task completion time ($z = 1.18$, $p = .236$).

3.6. Correspondence between SUS, task completion, serious and critical issues on perceived benefits, usefulness, and intention-to-use

The Kendall-Tau correlation and significance tests of the correlations revealed that there are significant disparities between the correlations of the SUS and task completion on critical and serious issues. As a final step, additional Mann-Whitney U tests were conducted between the SUS, task completion, serious and critical issues and the binomial variables Perceived benefits, Usefulness and Intention-to-use. The results showed that for all variables there were significant differences ($p < .001$), as can be seen in Table 7. The medians of the SUS and task completion were significantly higher among the participants that did

Table 7
Mann-Whitney *U* test for usability benchmarks and intention to use indicators.

		SUS (Mdn,95% CI)	Task completion (Mdn, 95% CI)	Serious Issues (Mdn, 95% CI)	Critical issues (Mdn, 95% CI)
Perceived benefits	Yes	68.8 (45, 75)	4 (2, 4)	3.5 (2, 5)	0 (.0, 3)
	No	13,8 (2.5, 30)	.5 (.0, 2)	8 (6, 10)	5 (3, 5)
	U	U = 21,5, p ≤ .001	U = 22, p ≤ .001	U = 34 p ≤ .001	U = 34.5 p ≤ .001
Perceived usefulness	Yes	76.3 (67.5-85)	4 (3, 5)	3 (.0, 5)	.0 (.0, 2)
	No	30 (12.5, 45)	1.5 (1, 3)	6.5 (4, 8)	3.5 (3, 5)
	U	U = 13, p ≤ .001	U = 47, p ≤ .001	U = 72,5, p ≤ .01	U = 64.5 p ≤ .01
Perceived intention to use	Yes	72.5 (67.5, 77.5)	4 (3, 5)	2 (1, 4)	.0 (.0, 2)
	No	30 (12.5, 45)	1 (.0, 2)	7 (5, 8)	4 (3, 5)
	U	U = 27, p ≤ .001	U = 43, p ≤ .001	U = 50,5, p ≤ .001	U = 60, p ≤ .001

perceive benefits and usefulness of the system and intended to use it in comparison to those participants that did not. Likewise, the medians of serious and critical issues were significantly lower among the Yes-group in contrast to the No-group for each of the three acceptance factors.

4. Discussion

Our results suggest that the SUS is inadequate as a stand-alone usability benchmark for eHealth technology, as it is a weaker indicator of the presence of critical and serious usability issues than the task completion rates. These results are in line with recent studies on the SUS, in particular the research of Harrati et al. [27], who also found that for usability evaluations of eLearning systems, the SUS in itself is not sufficient. So, at the very least, evaluators should report these task completion rates alongside SUS scores in their usability reports or articles. With regard to predicting intention to use, we found that the usability benchmarks are interchangeable.

This lack of predictive power of the SUS can have several reasons. First, the SUS is a subjective evaluation instrument. Therefore, the estimation of usability, as measured by the SUS, might be mixed with other perceptions about the technology (*e.g.*, usefulness, fun). Second, the SUS only provides a general score of the usability. Participants who evaluated the gamified application had more difficulty completing the tasks than participants who evaluated the other systems. This was reflected in the average SUS scores. The gamified application had a low SUS score of 27.5, while the tele-rehabilitation portal and the mobile app had much higher scores (respectively 77.5 and 71.3). However, when comparing the tele-rehabilitation portal and the mobile app, the SUS scores found in this study do not accurately reflect the actual performance of the users. While participants had more difficulty in completing tasks in the tele-rehabilitation portal (between 55.7 and 100%) than in the mobile app (100%), the average SUS score of the tele-rehabilitation portal was higher. These results suggest that task completion is a stronger predictor than the SUS for the presence or absence of usability issues (and their severity), which we consider to be the golden standard.

Another explanation for the relatively low predictive power of the SUS is that it does not take eHealth specific factors into account that affect usability (such as information overload, accessibility for the visually or cognitively impaired, *etc.*). In the literature on serious games for health, there is a growing awareness that there is a need for a standardized framework for usability evaluations [41,42]. Future studies should lead to an exhaustive overview of eHealth-specific factors that affect eHealth usability. Then, using this list, one can fine-tune usability testing and benchmarking methods for the eHealth context, ultimately leading to an easy to use usability benchmarking tool for eHealth, with high predictive power.

4.1. Study limitations

In this study, we chose to let participants get familiarized with the system before starting with the tasks. This was done to resemble real-life experience with a new technology. In the tele-rehabilitation portal and the gamified application, there were many options and areas to explore besides the locations and activities set in the research protocol. Using a technology with fewer functionalities, like the 'Stopstone' mobile app, there was more overlap between the free exploration task and the subsequent tasks which had a specific goal the participants had to complete. This could have affected their performances. A second limitation was that we did not measure logging on to the system. In the free exploration task, we saw that older adults had difficulty with the entry field for the e-mail address, more specifically creating special reading characters, like the 'at' sign (@). Contrary, in the mobile app participant had to fill out a long list of demographics and smoking habits before entering the main screen. These differences in system accessibility could have influenced participant's perceptions on usability and user-friendliness. However, this study's results provide usability ratings of the system in general, including system access, to examine differences in usability benchmarks. When analyzing the usability of eHealth for further optimization and development, it would be beneficiary to examine and compare the usability of different elements of a system.

4.2. Conclusions

In the field of eHealth, new innovations are produced very rapidly. However, the way in which we test the usability of these applications, or their prototypes, has been the same for decades. The results in our study indicate that we might also need to innovate the usability testing toolkit for eHealth, as we showed that the System Usability Scale (SUS) might not be the best instrument to benchmark the usability of an eHealth technology. We hope that this study will inspire other researchers and usability practitioners to closely look at the tools they use during their eHealth usability tests and to fine-tune these tools for this particular context.

Conflicts of interest

The authors declare they have no conflict of interest.

Summary table

What was already known on this topic?

- The System Usability Scale (SUS) is one of the most popular usability benchmarking instruments for eHealth.
- Context-specific factors can influence the perceived usability of eHealth, such as health literacy and physical/cognitive impairments of patient groups.
- Eliciting usability issues via qualitative methods is considered the 'golden standard' for assessing usability.

What this study added to our knowledge?

- The SUS is insufficient as a stand-alone usability benchmark for eHealth.
- Critical usability issues and task completion have more predictive value for the actual usability of eHealth than the SUS.
- This study emphasizes the need to develop a new usability benchmarking instrument, specifically designed for eHealth.

References

[1] International Organisation for Standardisation (ISO), Ergonomics of Human-System Interaction - Part 11: Usability: Definitions and Concepts, (2018).
[2] B.E. Chew, Bradley LD, KA, "Brief questions to identify patients with\rinadequate health literacy, Fam. Med. 36 (8) (2004) 585–594.
[3] K.M. Åkesson, B.I. Saveman, G. Nilsson, Health care consumers' experiences of information communication technology-a summary of literature, Int. J. Med. Inform. 76 (9) (2007) 633–645.
[4] L.M. MacKey, C. Doody, E.L. Werner, B. Fullen, Self-management skills in chronic disease management: what role does health literacy have? Med. Decis. Mak. 36 (6) (2016) 741–759.
[5] B.S. McEwen, Central effects of stress hormones in health and disease: understanding the protective and damaging effects of stress and stress mediators, Eur. J. Pharmacol. 583 (2–3) (2008) 174–185.
[6] T.J. Shors, Stressful experience and learning across the lifespan, Annu. Rev. Psychol. 57 (1) (2006) 55–85.
[7] M. Zwaanswijk, R.A. Verheij, F.J. Wiesman, R.D. Friele, Benefits and problems of electronic information exchange as perceived by health care professionals: an interview study, BMC Health Serv. Res. 11 (2011) 1–10.
[8] A.W. Kushniruk, V.L. Patel, Cognitive and usability engineering methods for the evaluation of clinical information systems, J. Biomed. Inform. 37 (1) (2004) 56–76.
[9] M.W.M. Jaspers, A comparison of usability methods for testing interactive health technologies: methodological aspects and empirical evidence, Int. J. Med. Inform. 78 (5) (2009) 340–353.
[10] D. Rubin, J. Chisnell, Handbook of Usability Testing [Electronic Resource] : How to

Plan, Design, and Conduct Effective Tests, 2nd ed., (2008).
[11] J. Sauro, J.R. Lewis, Quantifying The User Experience. Practical Statistics for User Research. (2012).
[12] B. Klaassen, B.J.F. van Beijnum, H.J. Hermens, Usability in telemedicine systems—A literature survey, Int. J. Med. Inform. 93 (2016) 57–69.
[13] J. Brooke, SUS - A quick and dirty usability scale, in: P.W. Jordan, B. Thomas, B.A. Weerdmeester, I.L. McClelland (Eds.), Usability evaluation in industry, Taylor & Francis, London, 1996, pp. 189–194.
[14] A. Bangor, P.T. Kortum, J.T. Miller, An empirical evaluation of the system usability scale, Int. J. Hum. Comput. Interact. 24 (6) (2008) 574–594.
[15] B.A. Campbell, C.C. Tossell, M.D. Byrne, P. Kortum, Voting on a smartphone: evaluating the usability of an optimized voting system for handheld mobile devices, Proc. Hum. Factors Ergon. Soc. (2011) 1100–1104.
[16] P.T. Kortum, A. Bangor, usability ratings for everyday products measured with the system usability scale, Int. J. Hum. Comput. Interact. 29 (2) (2013) 67–76.
[17] E.I. Konstantinidis, G. Bamparopoulos, P.D. Bamidis, Moving Real exergaming engines on the web: the webFitForAll case study in an active and healthy ageing living lab environment, IEEE J. Biomed. Heal. Informatics 21 (3) (2017) 859–866.
[18] F.W. Simor, M.R. Brum, J.D.E. Schmidt, R. Rieder, A.C.B. De Marchi, Usability evaluation methods for gesture-based games: a systematic review, JMIR Serious Games 4 (2) (2016) e17.
[19] S.M. Jansen-Kosterink, R.M.H.A. Huis in' t Veld, C. Schönauer, H. Kaufmann, H.J. Hermens, M.M.R. Vollenbroek-Hutten, a serious exergame for patients suffering from chronic musculoskeletal Back and neck pain: a pilot study, Games Health J. 2 (5) (2013) 299–307.
[20] M. Georgsson, N. Staggers, Quantifying usability: an evaluation of a diabetes mHealth system on effectiveness, efficiency, and satisfaction metrics with associated user characteristics, J. Am. Med. Informatics Assoc. 23 (1) (2016) 5–11.
[21] L. Wozney, et al., Usability, learnability and performance evaluation of intelligent research and intervention software: a delivery platform for eHealth interventions, Health Informatics J. 22 (3) (2016) 730–743.
[22] P. Kortum, C.Z. Acemyan, How low can you go? Is the system usability scale range restricted? J. Usability Stud. 9 (1) (2013) 14–24.
[23] J.R. Lewis, The system usability scale: past, present, and future, Int. J. Hum. Comput. Interact. 34 (7) (2018) 577–590.
[24] T. Tullis, B. Albert, Measuring the user Experience: Collecting, Analyzing, and Presenting Usability Metrics, 2nd ed., (2013).
[25] K. Orfanou, N. Tselios, C. Katsanos, Perceived usability evaluation of learning management systems: empirical evaluation of the system usability scale, Int. Rev. Res. Open Distrib. Learn. 16 (2) (2015) 227–246.
[26] A. Bangor, P. Kortum, J. Miller, Determining what individual SUS scores mean: adding an adjective rating scale, J. Usability Stud. 4 (3) (2009) 114–123.
[27] N. Harrati, I. Bouchrika, A. Tari, A. Ladjailia, Exploring user satisfaction for e-learning systems via usage-based metrics and system usability scale analysis, Comput. Human Behav. 61 (2016) 463–471.
[28] Y.Y. Hoogendam, et al., Older age relates to worsening of fine motor skills: a population-based study of Middle-aged and elderly persons, Front. Aging Neurosci. 6 (2014) 259.
[29] K.M. Gerling, F.P. Schulte, J. Smeddinck, M. Masuch, Game design for older adults: effects of age-related changes on structural elements of digital games, Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) vol. 7522 LNCS, (2012), pp. 235–242.
[30] E. Flores, G. Tobon, E. Cavallaro, F.I. Cavallaro, J.C. Perry, T. Keller, Improving patient motivation in game development for motor deficit rehabilitation, Proceedings of the 2008 International Conference in Advances on Computer Entertainment Technology - ACE' 08, (2008), p. 381.
[31] J.R. Lewis, Psychometric evaluation of an after-scenario questionnaire for computer usability studies, ACM SIGCHI Bull. 23 (1) (1991) 78–81.
[32] F.D. Davis, perceived usefulness, perceived ease of use, and user acceptance of information technology, MIS Q. 13 (3) (1989) 319–340.
[33] F.D. Davis, R.P. Bagozzi, P.R. Warshaw, User acceptance of computer technology: a comparison of Two theoretical models, Manage. Sci. (1989).
[34] H.R. Hartson, T.S. Andre, R.C. Williges, Usability evaluation methods, Society 13 (February) (2015) 373–410 2001.
[35] CCMO, Dutch Law on Medical-Scientific Research, [Online]. Available: (2018) www.ccmo.nl.
[36] L. van Velsen, T. van der Geest, R. Klaassen, Identifying usability issues for personalization during formative evaluations: a comparison of three methods, Int. J. Hum. Comput. Interact. 27 (7) (2011) 670–698.
[37] E.B. Wilson, Probable inference, the law of succession, and statistical inference, J. Am. Stat. Assoc. 22 (158) (1927) 209-121.
[38] K.J. Rothman, J.D. Boice, Epidemiologic Analysis With a Programmable Calculator, (1979).
[39] Calculation for the Test of the Difference between Two Dependent Correlations With One variable in Common, (2013) [Computer software].,Available from http://quantpsy.org.
[40] J.H. Steiger, Tests for comparing elements of a correlation matrix, Psychol. Bull. 87 (2) (1980) 245–251.
[41] L. Sardi, A. Idri, J.L. Fernández-Alemán, A systematic review of gamification in e-health, J. Biomed. Inform. 71 (2017) 31–48.
[42] S. McCallum, Gamification and serious games for personalized health, Stud. Health Technol. Inform. 177 (2012) 85–96.