# NASA RTLX as a Novel Assessment Tool for Determining Cognitive Load and User Acceptance of Expert and User-based Usability Evaluation Methods

**M Georgsson**[1,2,3]*

[1]Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah, USA

[2]Faculty of Computing, Blekinge Institute of Technology, Karlskrona, Sweden

[3]Department of Health Sciences, University West, Trollhättan, Sweden

## Abstract

**Background:** Mobile health applications are frequently used to manage different health conditions. Diabetes is one disease where these are used in patients' personalized disease self-management, but where the usability is often deficient. Two methods to assess usability are the cognitive walkthrough (CW), which is expert-based and the think aloud (TA) which is user-based. Both offer advantages and disadvantages and detect many types of usability problems affecting the user. There is a lack of research, however, on how the usability evaluators themselves experience performing these methods and the method impact. This can be an important aspect to include due to its possible implications for the evaluation.

**Objectives:** In this article the focus was particularly on assessing the evaluators' cognitive load and method acceptance while performing the described methods.

**Methods:** In addition to the number of usability problems, and the system usability scores (SUS), the NASA RTLX instrument, novel for this purpose, was used together with in-depth interviews to assess the usability methods' cognitive impact.

**Results:** A total of 12 evaluators, six per method, detected 18 usability problems with the CW. Twenty were found with the TA. The SUS scores were 23.75 and 59.58 respectively. For both methods, users experienced a similarly high cognitive load with a RTLX score of 56.11 for CW and 53.47 for TA. According to the evaluators, both methods were cognitively demanding.

**Conclusion:** The results highlight the potential significance of these dimensions for inclusion in the usability evaluation and in decision-making purposes between different usability evaluation methods.

## Keywords

NASA RTLX; Cognitive load; Usability evaluation methods; Diabetes; Self-management

**Correspondence to:**

**Dr. M Georgsson, RN, PhD**
Applied Health Technology
Department of Health Sciences,
University West
Sweden
Tel: +46520223000
Email: mattias.georgsson@hv.se

**Citation:** Dr. M Georgsson. NASA RTLX as a Novel Assessment Tool for Determining Cognitive Load and User Acceptance of Expert and User-based Usability Evaluaion Methods. EJBI 2020; 16(2): 14-21.
**DOI:** 10.24105/ejbi.2020.16.2.4
Received: August 26, 2019
Accepted: February 10, 2020
Published: February 17, 2020

## 1. Introduction

This article is an extension of work originally presented for pHealth 2019 [1]. The extension focuses on expanding the information about in which areas of research the NASA RTLX has been used previously, on deepening the discussions of user characteristics related to usability and also on the applicability of NASA RTLX as an assessment tool when it comes to usability evaluation methods, as well as its utility in the area of health care and health technology assessments.

Mobile health applications are frequently used to manage different health conditions. Diabetes is one such disease requiring considerable self-management to monitor it. It is a demanding disease that puts a heavy burden on patients [2] due to that they are the largest contributors to their own care and often the most responsible for their own care and self-management [3]. Even if many patients utilize these mobile health applications for their personalized health management needs, studies emphasize that a large number of them have inadequate usability [4]. Usability is important in several regards for patients and especially to experience as they interact with the system and use it for their disease management needs and requirements [5].

Usability evaluations are usually part of the system development process or performed before the system is launched [6]. Two common evaluation methods are the cognitive walkthrough (CW) and think aloud protocol (TA). CW is an expert-based method, where usability experts perform all parts of the evaluation process while TA is a user-based evaluation method where the user, or in this case the patient, is the central person performing the evaluation [7]. The main intent of these usability evaluation methods is to determine the usability problems in the system so that these can be resolved. Both methods can also lead to the detection of usability problems of many types and offer advantages as well as disadvantages.

The CW can be used in the initial design phase, and also be utilized by users who are new to the system. In addition, it can generate results at low costs and rather quickly compared to other methods [8]. Its disadvantages include that it detects many minor or low-priority problems that are not always relevant to end-users [9]. The TA requires only a small user sample (5 to 8 users) to find 80-85% of the usability problems present in the system [10, 11]. The method provides detailed data instantly, and finds many recurring problems and serious problems**.** Its drawbacks are that it takes a lot of time to carry out and is also a costly and resource-demanding method [9].

There is extensive research on that these methods can detect many types of usability problems, and also in what instances they provide the best results [9, 12]. There is, however, currently a lack of research on how the evaluators themselves experience performing these different methods and the impact the method itself may have on the evaluator. This may be an important aspect to consider and include due to that it can have implications for the overall evaluation and its outcomes.

## 2. Objectives

In this article the objective was particularly to assess the evaluators' cognitive load as well as method acceptance while performing the cognitive walkthrough (CW) and think aloud (TA) usability evaluation methods on a mHealth application intended for patients with diabetes. The usability and system satisfaction were determined as well as the participants' demographics, IT/computer, mobile phone knowledge, experience and use. The cognitive load as well as method acceptance of each of the methods was measured through a novel use of the NASA RTLX instrument along with in-depth interviews.

## 3. Methods

The study was performed at the University of Utah in Salt Lake City, Utah, USA. After approval was received from the University of Utah Institutional Review Board, a total of 12 evaluators were recruited; 6 experts from the university's Computer Science department and 6 diabetes patients from the university hospital's Diabetes and Endocrinology Center**.** The expert' tests were conducted at the department while the patients' usability tests were performed at the hospital. The mHealth application evaluated for usability in this study was a research application for diabetes self-management with which it is possible to keep track of and monitor blood glucose**,** insulin/medication, physical activity and food.

### *3.1 Inclusion and Exclusion Criteria*

The experts' inclusion criteria were: 1) 18 years of age or above, 2) at least 2 years of experience in human computer interaction, and 3) proficiency in the English language in terms of speaking and comprehension. The patients' inclusion criteria consisted of: 1) a diabetes diagnosis (either Type 1 or Type 2), 2) no cognitive impairment, 3) knowledge and familiarity with computers, the Internet and mobile phone, and 4) proficiency in the English language in terms of speaking and comprehension. None of the groups should have had any previous exposure to the mHealth application. The study participants all received 20 dollar gift cards for their participation in the study.

### *3.2 Study Procedures*

Each session began with taking participants' informed consent. Then, the pre-test questionnaires were distributed and completed. Pre-test questionnaires were then distributed to and completed by the participants. The demographic details collected included their gender, age, education, ethnicity, and occupation. They were also asked about their knowledge and frequency of use when it came to the computer, Internet and mobile phone (such as phone calls, text messaging, and app use). The mHealth application was then demonstrated to them and each evaluation performed in agreement with either the CW or TA method requirements**.** After, each participant completed the system usability scale (SUS), the NASA RTLX instrument and took part in a one-on-one digitally recorded interview on their acceptance of the methods.

### *3.3 The CW Evaluation*

The cognitive walkthrough (CW) by Polson and Lewis [13], based on theories of cognitive exploratory learning [14], consists of experts collectively going through the system while they try to imagine the problem solving process by the users to find the system usability problems. This method is particularly useful when it comes to systems that are new or unfamiliar to the user or about which the user has limited or no prior knowledge. The evaluation process itself starts with the performance of a task analysis. Here the experts specify the action sequences that a user needs to be able to perform to achieve a specific goal and what the system response would be. Then they evaluate the system by going through it together and answer 4 specific questions for each individual step. These are: 1) Will the user try to achieve the effect that the subtask has? (Does the user understand that this subtask is needed to reach the user's goal?) 2) Will the user notice that the correct action is available? e.g. is the button visible? 3) Will the user understand that the wanted subtask can be achieved by the action? e.g. the right button is visible but the user does not understand the text and therefore will not click on it. 4) Does the user get feedback? Will the user know that they have done the right thing after performing the action? [15]. The answer to these different questions for each step, determines the usability problems. Data are collected throughout the evaluation through

designated forms to be able to produce a usability problem list at the end [13-15].

### *3.4 The TA Evaluation*

The think aloud protocol (TA) is one of the most common methods for user tests and a method introduced by Lewis [16], then refined by Lewis and Rieman [17] and on work conducted by Ericsson and Simon [18-20].

In this evaluation users perform tasks that are representative in the system while they express their thoughts out loud during the interaction. The goal of this process is to understand how users behave and think while interacting with the system through the tasks as well as their identification of the main usability problems in the system [21]. While the user is interacting with the system, the observer is only minimally involved to not interrupt their thought processes, except to remind them to keep on talking if they stop. To understand the users' decision making processes and how they experience the system is of key importance in this method. In the particular evaluation of this study the tasks consisted of setting blood glucose measurement units, entering readings for glucose, carbohydrate intake and insulin/medication. Participants also had to interpret entries in graphs, search and find entries and export entries. The interactions in the system were recorded digitally using Morae software.

The System Usability Scale (SUS) by Brooke was used for both the CW and the TA to determine the overall system usability as well as satisfaction with the system. The scores range from 0-100 [22]. SUS scores of 70 or above are considered acceptable, scores of 85 or above indicate a high level of usability, while scores of 50 or below are considered poor, or unacceptable [23].

### *3.5 The NASA RTLX Instrument and In-depth Interview*

The NASA RTLX instrument, by Byers et al. [24] is a simplified version of the NASA-Task Load Index developed by Hart and Staveland [25]. The instrument is used to measure the cognitive load regarding the performance of different tasks. The NASA RTLX has been used in several areas of research. These are for example in studies that focus on on-road assessments when it comes to cognitive distractions and their impact on drivers' visual behavior and braking performance where drivers performed demanding cognitive tasks while driving in city traffic [26], in studies on driver safety where the effect of Google Glass was measured on simulated lane keeping performance and tasks were performed during different conditions [27]. It has also been used to assess cognitive efficiency and effectiveness using sunlight as a cognitive stressor in visual display terminal work [28] as well as in the area of medicine to assess the cognitive load of different surgical techniques [29, 30].

In this study the NASA RTLX instrument was used to determine the method cognitive load for the two usability evaluation methods. This instrument consists of six dimensions that each have questions associated with them (Table 1).

The specific dimensions designate an activity's contribution to the cognitive workload from low to high on a scale with scores of 0-100. Scores are also weighted by the user through pairwise comparisons regarding their perceived importance. In the NASA RTLX this addition has been removed, however, but still allows for a high experimental validity [31]. The overall cognitive workload that the participant experiences is calculated through an addition of the scores and then division on the six different dimensions to get the average. The higher it is, the higher the experienced cognitive workload [25].

After filling out the NASA RTLX, the one-on-one audio recorded interview was conducted about how the method was experienced by the participants. The questions asked concerned: 1) the participants' thoughts about the performed method, 2) if it was easy or difficult to understand it, 3) if the method was easy or difficult to conduct and 4) their overall method experience.

### *3.6 Data Analysis*

Descriptive statistics were used to analyze the demographic characteristics, and technology experience and use questionnaires. The usability problems for the CW, was produced by the expert evaluators at the end of the evaluation as an actionable list as this is part of the method. The problems detected by TA were, on the other hand, determined by performing inductive coding on the imported, transcribed data in Nvivo 10.

System usability satisfaction (SUS) scores were determined by using Brooke's set of instructions [22]. These scores were

Table 1: Table 1 NASA Raw Task Load Index with six dimensions and questions (modified with a method focus) [24]

| Dimension | Question |
|---|---|
| Mental Demand | How much mental and perceptual activity was required? Was the method easy or demanding, simple or complex? |
| Physical Demand | How much physical activity was required? Was the method easy or demanding, slack or strenuous? |
| Temporal Demand | How much time pressure did you feel due to the pace at which the method or method elements occurred? Was the pace slow or rapid? |
| Performance | How successful were you in performing the method? How satisfied were you with your performance? |
| Effort | How hard did you have to work (mentally and physically) to accomplish your level of performance? |
| Frustration Level | How irritated, stressed, and annoyed versus content, relaxed, and complacent did you feel during the method steps? |

calculated by summing the scores on each of the 10 individual items that the instrument consists of. For items 1,3,5,7 and 9 one point was subtracted from the resulting score. For items 2,4,6,8 and 10, five points were subtracted from the resulting score. The final sum of all scores was then multiplied by 2.5 to get the overall usability satisfaction value [22]. To determine the NASA RTLX scores, the 6 dimensions were averaged for each participant for each method. The interviews were transcribed and analyzed by using thematic content analysis to determine what the users' experiences were of the methods [32].

## 4. Results

### *4.1 Demographics, IT, Computer and Mobile Phone Knowledge, Experience and Use*

The CW participants consisted of mostly women and the average age was 28 years old. A majority of them were Caucasian and all were college or university educated. Half of the groups were students and the other half employed. The TA participants included an equal number of women and men with an average age of 52. A majority were Caucasian with college or university education. Most were also employed specific details in Table 2.

In terms of computer use and IT knowledge all CW participants had high computer/IT and Internet knowledge, and used the computer and Internet daily. When it came to mobile phones, all participants felt that they had high mobile phone knowledge and they used their mobile phone every day. A majority used it to make phone calls and all used it daily for e-mails, surfing and using apps.

Half of the TA group used the computer daily and the other half 3-6 times per week. One considered their computer/IT knowledge to be high and their Internet knowledge as high. Half of the group also used it 3-6 times a week. In terms of their mobile phone knowledge and use a minority felt that they had high mobile phone knowledge but the majority of participants used it daily. A large number of participants also used it to make phone calls daily, which was similar for text messaging, e-mails, surfing and apps specific details in Table 3.

### *4.2 Usability Problems and System Usability Satisfaction Scores*

The CW detected 18 usability problems while the TA found 20 problems. For both the CW and TA methods the problems centered on clarities in the interface such as what specific icons and functions meant. They also concerned difficulties in the system interaction, in the navigation between system views as well as the absence of system feedback on actions. The SUS scores showed a variation between the two groups; where the CW group felt the system had poor or unacceptable usability with a score of 23.75 and the TA group that it had an acceptable (between ok and good) usability with a score of 59.58.

### *4.3 NASA RTLX Scores and Method Perceptions*

The complete ICD-10-CM code can have 3-7 characters. We have 14,602 labels in total as prediction candidates in our dataset. Similarly, we use discharge diagnoses as our input to train this model, considering better chapter classification performance and the GPU hardware bottleneck. Our model has 0.625 on F1-score when we use 300 as our embedding dimension.

### *4.4 Performance in Department*

The NASA RTLX overall single workload scores showed high cognitive loads for the participants' respective methods. For the experts' and the CW, the experienced cognitive load was a mean of 56.11. For the patients and the TA, the mean was 53.47. The standard deviation scores (SD), showed a spread of 10.17 for the CW. The TA had a SD score of 14.16. The separate dimension subscales indicated that for both methods the mental demand and frustration placed the highest highlighting that this was a concern for participants in both methods. In addition, the TA also placed higher than the CW on the performance satisfaction dimension even if both had almost equally high scores on almost all dimensions. Both methods had the lowest scores, however, for physical demand, where the TA had the highest scores of the two (Figure 1) [33].

Their views of the methods, demonstrated that four experts considered the CW to be rigorous and in-depth. Five felt that the method was very time consuming. They also experienced it as difficult to get a broad picture of the application with it as well

Table 2: Participant demographics

| Participant demographics | Category | CW* | TA* |
|---|---|---|---|
| Gender | Male | 2 | 3 |
| | Female | 4 | 3 |
| Age | Years (Mean) | 28.2 | 52 |
| Ethnicity | White/Caucasian | 5 | 4 |
| | Black/African American | - | 1 |
| | Hispanic/Latino | - | 1 |
| | Asian/Pacific | 1 | - |
| Education | High School | - | 2 |
| | College/University | 6 | 4 |
| Occupation | Retired | - | 2 |
| | Employed | 3 | 4 |
| | Student | 3 | - |
| *CW = Cognitive Walkthrough, TA = Think Aloud usability test* | | | |

Table 3: Computer/IT, Internet, mobile phone knowledge, experience and use.

| Computer/IT, Internet, mobile phone experience and use | Time Period | CW* | TA* |
|---|---|---|---|
| Frequency of computer use | Every day | 6 | 3 |
| | 3-6 times/week | - | 3 |
| Computer/IT-knowledge | High | 6 | 1 |
| | Medium | - | 2 |
| | Small | - | 3 |
| Frequency of Internet use | Every day | 6 | 3 |
| | 3-6 times/week | - | 3 |
| Internet-knowledge | High | 6 | 1 |
| | Medium | - | 3 |
| | Small | - | 2 |
| Frequency of mobile phone use | Every day | 6 | 5 |
| | 3-6 times/week | - | 1 |
| Mobile phone-knowledge | High | 6 | 2 |
| | Medium | - | 2 |
| | Small | - | 2 |
| Mobile phone use for phone calls | Every day | 4 | 5 |
| | 3-6 times/week | - | - |
| | 1-2 times/week | 2 | 1 |
| Mobile phone use for text messaging | Every day | 6 | 5 |
| | 3-6 times/week | - | 1 |
| Mobile phone use for e-mails, surfing, apps | Every day | 6 | 4 |
| | 3-6 times/week | - | 1 |
| | Never | - | 1 |

** CW = Cognitive Walkthrough, TA = Think Aloud usability test*

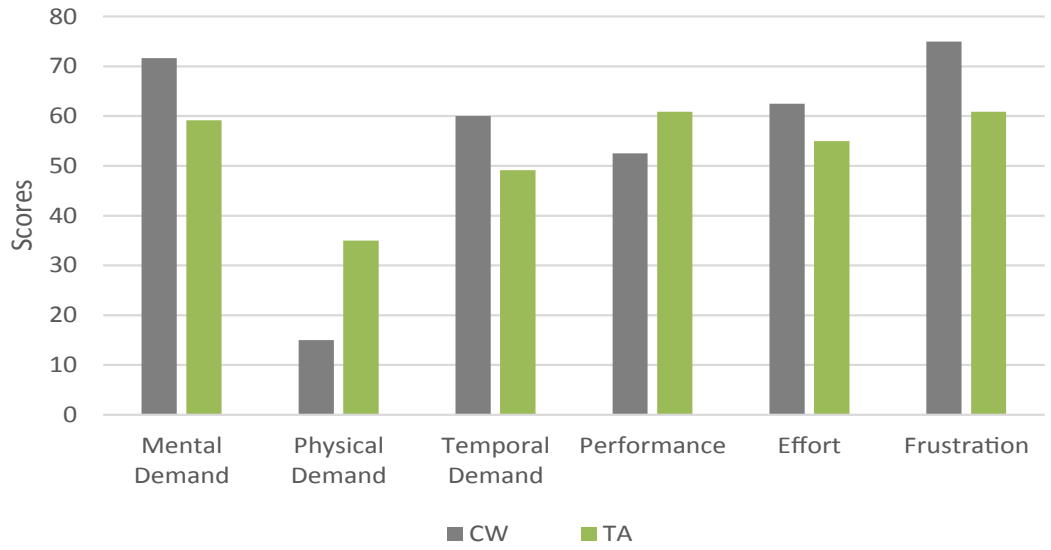


Figure 1: NASA RTLX dimensions.

as that the method was too granular, detailed, and redundant in some steps. One felt that there was a steep learning curve with the method, and three expressed that it took a while to figure out how to do it. Two considered it difficult to grasp how to perform it as well as to know if they performed it right. They also experienced the method to be rather difficult to carry out. One person expressed that encircling the user-perspective took some effort and two considered it difficult to put themselves in the position of the user of the system. For the TA, four patients experienced it as easy to understand. Two of them, however, thought it difficult to comprehend. One felt it wasn't hard to carry out, while half of the group experienced it as difficult to perform. Four considered it as an unusual method and rather uncommon. It was especially the think aloud component that they considered odd in this regard. Half felt that it was a rather awkward and stressful method. Four of the six participants also considered it to have a heavy cognitive load.

## 5. Discussion

Both the amount of usability problems that were found as well as the SUS scores indicate that participants in the CW as well as the TA experienced the mHealth application to be lacking in usability; the experts viewed it as poor and the patients as acceptable. In addition, the evaluations also demonstrated that despite differences in IT, computer, Internet and mobile phone knowledge, experience and use experts as well as diabetes patients experienced an almost equally heavy cognitive load for their methods as shown in their high RTLX scores and the different dimensions they highlighted as demanding. Also in their respective interviews, both evaluation groups expressed that it was rather difficult to understand and perform their different methods.

Extensive research has been performed on both the CW and TA. For example, have both methods been assessed on their performance, usability problem characteristics and on their positive and negative aspects [9, 34]. Authors have also compared them as stand-alone methods [35] or combined them with each other and or with other methods finding especially TA to be more effective as a combined method than CW with regards to its amount of detected usability problems as well as the severity of these problems [36]. Research has also been specifically performed on the usability of diabetes mHealth systems where usability has been an issue and user characteristics have influenced evaluation performance; where for example patients' with more IT knowledge and experience performed better compared to those with less experience [37, 38]. There is currently a lack in research, however, on the evaluator experiences regarding the usability method itself and also its cognitive impact and influence. This can be a dimension which can also be potentially important especially when it comes to the overall system experience and the result of the usability evaluation. If the evaluators consider the method to be difficult, as was partly the case here as shown in the high RTLX scores, these views can influence their overall system perception.

There is no doubt that evaluations for usability are important in figuring out how well the application satisfies the expectations that the patients have as well as their wants and needs and to safeguard patient outcomes and quality of care and for this reason have a significant purpose [6]. When performing these evaluations it may therefore also be especially important to consider the impact of the method for the patient evaluators due to the possibility that they can encounter added difficulties due to disease implications. As authors point out, the burden of the disease and treatment can be significant for patients with chronic conditions and here it is important that technological support does not add to this burden and to the cognitive load for patients [39]. Patients have such demands on themselves already when it comes to these aspects and if the burdens accumulate some patients can become overwhelmed, ultimately leading to discontinued use and poor healthcare outcomes [40]. Of importance is that while many patients voice their interest in using technologies for self-management, applications are not usable in a consistent manner [37], which is also the case here. Therefore, to design and test technology-enabled solutions that help lighten the work of chronic disease management while increasing the ability of the person with diabetes to engage in self-management and not add an additional burden to an already challenging situation is of the essence [40]. Using a usability evaluation method that has a cognitive load which is appropriate may be one aspect to consider if the long-term goal is to implement these types of applications and systems extensively within health care. As some authors state it is also important to adopt iterative and adaptive designs and evaluation processes for these types of technology-enabled self-management support systems and to be able to accommodate a wide variety of different users [41] as well as to facilitate more extensive usage for a larger number of users [38]. The NASA RTLX can also help in this regard as it can aid system designers to designate the origin of a workload or performance problem in the system [33] which in turn can assist them in developing systems that are better fitted to and more usable to the user.

## 6. Limitations

When it comes to the applicability of the evaluation results in this study, it might have been the case that the participants' characteristics could have influenced the groups' perceptions of both the methods as well as the application. The CW group, as the more experienced of the two, could have carried out their specific method more easily and provided a SUS score that was higher than the TA group. This seemed to not be the case; however, as it was clear that both groups felt that the cognitive load of their respective method was heavy and also that their methods were difficult. In addition, the experts' more critical view regarding the system's usability compared to the patients' demonstrates a result that is not too unusual for these different groups. Users' characteristics could also potentially have an impact on how the usability was experienced as noted in other research where researchers found out that those with less experience experienced lower usability and vice versa [37, 38]. Here there was a comparison between how experts and users experienced the cognitive load of two separate usability methods but future studies could also assess how the same group considered the different methods to determine the most preferable one for that specific group or how two similar groups considered each of the methods to select the most suitable one for a health technology assessment.

## 7. Conclusion

The study demonstrates that the assessment of users' cognitive load, as well as method acceptance as done here with the NASA RTLX instrument as an assessment tool and a short interview as a complement, can be useful due to that both groups experienced high cognitive loads and their evaluation methods to be rather demanding which could in the end also influence their system views. A method that puts on an additional cognitive demand can in particular have an impact on the chronic disease patient, adding to their disease burden which also is something to consider along with user characteristics in the development and evaluation process of these systems. The cognitive impact of the

method has, as of yet, not been explored to a great extent when it comes to usability evaluation methods. This may, however, be a significant dimension to take into account in future method selection processes.

## 8. Acknowledgements

The author would like to thank Eirik Arsand and the Norwegian Centre for E-health Research for access to the research application.

## References

1. Georgsson M. NASA RTLX as a Novel Assessment for Determining Cognitive Load and User Acceptance of Expert and User-Based Evaluation Methods Exemplified Through a mHealth Diabetes Self-Management Application Evaluation. Stud Health Technol Inform. 2019; 261: 185-190.
2. Powers MA, Bardsley J, Cypress M, Duker P, Funnell MM, Fischl AH, et al. Diabetes Self-management Education and Support in Type 2 Diabetes. Diabetes Educ. 2017; 43(1): 40-53.
3. Collins MM, Bradley CP, O'Sullivan T, Perry IJ. Self-care coping strategies in people with diabetes: a qualitative exploratory study. BMC Endocr Disorders. 2009; 9: 6.
4. Fu H, McMahon SK, Gross CR, Adam TJ, Wyman JF. Usability and clinical efficacy of diabetes mobile applications for adults with type 2 diabetes: A systematic review. Diabetes Res Clin Pract. 2017; 131: 70-81.
5. Veazie S, Winchell K, Gilbert J, Paynter R, Ivlev I, Eden K, et al. Mobile Applications for Self-Management of Diabetes. AHRQ Comparative Effectiveness Technical Briefs. Rockville (MD): Agency for Healthcare Research and Quality. 2018.
6. Goldwater JC. Human Factors and Usability in Mobile Health Design-Factors for Sustained Patient Engagement in Diabetes Care. Proc of the International Symposium of Human Factors and Ergonomics in Healthcare. 2014; 3(1): 63-70.
7. Kushniruk AW, Patel VL. Cognitive and usability engineering methods for the evaluation of clinical information systems. J Biomed Inform. 2004; 37(1): 56-76.
8. Wang E, Caldwell B. An Empirical Study of Usability Testing: Heuristic Evaluation Vs. User Testing. Proc of the Human Factors and Ergonomics Society Annual Meeting. 2002; 46(8): 774-778.
9. Jeffries R, Miller JR, Wharton C, Uyeda K. User interface evaluation in the real world: A comparison of four techniques. Proc of the SIGCHI Conference on Human Factors in Computing Systems; New Orleans, Louisiana, USA. ACM 1991; 119-124.
10. Nielsen J, Landauer TK. A mathematical model of the finding of usability problems. Proceedings of the INTERACT ,93 and CHI ,93 conference on Human factors in computing systems; Amsterdam, The Netherlands. ACM 1993; 206-213.
11. Virzi RA. Refining the test phase of usability evaluation: How many subjects is enough? Hum Factors. 1992; 34(4): 457-468.
12. Gupta S. A Comparative study of Usability Evaluation Methods. IJCTT. 2015; 22(3): 103-106.
13. Polson PG, Lewis C, Rieman J, Wharton C. Cognitive Walkthroughs-A Method for Theory-Based Evaluation of User Interfaces. Int J Man-Machine Studies. 1992; 36(5):741-773.
14. Polson PG, Lewis CH. Theory-Based Design for Easily Learned Interfaces. Human-Computer Interaction. 1990; 5(2/3): 191.
15. Wharton C, Rieman J, Lewis C, Polson P. The cognitive walkthrough method: a practitioner's guide. In: Nielsen J, Mack RL, editors. Usability inspection methods. New York: John Wiley & Sons, Inc. 1994: 105-140.
16. Lewis C. Using the „thinking-aloud" method in cognitive interface design. IBM Watson Research Center, Yorktown Heights, NY, USA; 1982.
17. Lewis C, Rieman J. Task-centered User Interface Design: A Practical Introduction. Boulder: University of Colorado. 1993.
18. Ericsson KA, Simon HA. Verbal Reports as Data. Psychol Rev. 1980; 87(3): 215-251.
19. Ericsson KA, Simon HA. Verbal reports on thinking. In: Færch C, Kasper G, (eds). Introspection in second language research. Clevedon, Avon: Multilingual Matters 1987; 24-54.
20. Ericsson KA, Simon HA. Protocol analysis: verbal reports as data. Cambridge, Mass. MIT Press 1984; 426.
21. Nielsen J, Mack RL. Usability inspection methods. New York: John Wiley & Sons, Inc. 1994: 413.
22. Brooke J. SUS: a „quick and dirty" usability scale. In: Jordan PW, Thomas B, Weerdmeester BA, McClelland IL, editors. Usability evaluation in industry. London: Taylor & Francis. 1996: 189-194.
23. Bangor A, Kortum P, Miller J. Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale. Journal of Usability Studies. 2009; 4(3): 114-123.
24. Byers JC, Bittner AC, Hill SG. Traditional and raw task load index (TLX) correlations: are paired comparisons necessary? In: Mitai A, editor. Advances in Industrial Ergonomics and Safety I. New York: Taylor & Francis. 1989; 481-485.
25. Hart SG, Staveland LE. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In: Peter AH, Najmedin M, editors. Advances in Psychology. North-Holland.1988; 52: 139-183.
26. Harbluk JL, Noy YI, Trbovich PL, Eizenman M. An on-road assessment of cognitive distraction: Impacts on drivers' visual behavior and braking performance. Accid Anal Prev. 2007; 39(2): 372-379.

27. Young KL, Stephens AN, Stephan KL, Stuart G. An Examination of the Effect of Google Glass on Simulated Lane Keeping Performance. Procedia Manufacturing. 2015; 3: 3184-3191.

28. Roberto G. Rodriguez, Andrea Pattini. Effects of a Large Area Glare Source in Cognitive Efficiency and Effectiveness in Visual Display Terminal Work. LEUKOS. 2012; 8(4): 283-299.

29. Corker HP, Singh P, Sodergren MH, Balaji S, Kwasnicki RM, Darzi AW, et al. A randomized controlled study to establish the effect of articulating instruments on performance in single-incision laparoscopic surgery. Journal of surgical education. 2015;7 2(1): 1-7.

30. Shiber LJ, Ginn DN, Jan A, Gaskins JT, Biscette SM, Pasic R. Comparison of Industry-Leading Energy Devices for Use in Gynecologic Laparoscopy: Articulating ENSEAL versus LigaSure Energy Devices. Journal of Minimally Invasive Gynecology. 2018; 25(3): 467-473.

31. Bustamante EA, Spain RD. Measurement Invariance of the Nasa TLX. Proc of the Human Factors and Ergonomics Society Annual Meeting. 2008; 52(19): 1522-1526.

32. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res. 2005; 15(9): 1277-1288.

33. Hart SG. Nasa-Task Load Index (NASA-TLX); 20 Years Later. Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 2006; 50(9): 904-908.

34. Fu L, Salvendy G, Turley L. Effectiveness of user testing and heuristic evaluation as a function of performance classification. 2002; 21(2): 137-143.

35. Karat CM, Campbell R, Fiegel T, editors. Comparison of Empirical Testing and Walkthrough Methods in User Interface Evaluation 1992. Acm Press, United States. 1992.

36. Frökjær E, Larusdottir MK. Prediction of usability: Comparing method combinations. Managing Information Technology Resources in Organizations in the Next Millennium, Hershey, USA. 1999.

37. Sarkar U, Gourley GI, Lyles CR, Tieu L, Clarity C, Newmark L, et al. Usability of Commercially Available Mobile Applications for Diverse Patients. J Gen Intern Med. 2016; 31(12): 1417-1426.

38. Georgsson M, Staggers N. Quantifying usability: an evaluation of a diabetes mHealth system on effectiveness, efficiency, and satisfaction metrics with associated user characteristics. J Am Med Inform Assoc. 2016; 23(1): 5-11.

39. Tran VT, Barnes C, Montori VM, Falissard B, Ravaud P. Taxonomy of the burden of treatment: a multi-country web-based qualitative study of patients with chronic conditions. BMC Med. 2015; 13: 115.

40. May CR, Eton DT, Boehmer K, Gallacher K, Hunt K, MacDonald S, et al. Rethinking the patient: using Burden of Treatment Theory to understand the changing dynamics of illness. BMC health services research. 2014; 14: 281.

41. Greenwood DA, Gee PM, Fatkin KJ, Peeples M. A Systematic Review of Reviews Evaluating Technology-Enabled Diabetes Self-Management Education and Support. J Diabetes Sci Technol. 2017; 11(5): 1015-1027.