TOWARD PATIENT-CENTERED, STANDARDIZED, AND REPRODUCIBLE APPROACHES OF EVALUATING THE USABILITY OF mHEALTH CHRONIC DISEASE SELF-MANAGEMENT SYSTEMS FOR DIABETES

Mattias Georgsson

Blekinge Institute of Technology
Doctoral Dissertation Series No. 2018:03
Department of Creative Technologies
Toward Patient-Centered, Standardized, and Reproducible Approaches of Evaluating the Usability of mHealth Chronic Disease Self-Management Systems for Diabetes

Mattias Georgsson
Toward Patient-Centered, Standardized, and Reproducible Approaches of Evaluating the Usability of mHealth Chronic Disease Self-Management Systems for Diabetes

Mattias Georgsson

Doctoral Dissertation in
Applied Health Technology

Department of Creative Technologies
Faculty of Computing
Blekinge Institute of Technology
SWEDEN
To my family

Success is not final, failure is not fatal: it is the courage to continue that counts

- Sir Winston Churchill
Abstract

Diabetes is a chronic disease affecting 422 million patients worldwide according to World Health Organization data with 30.3 million in the United States and 64 million in Europe. The prevalence speaks to the need for improved ways to support patients in disease self-management. mHealth solutions are increasingly used for this; however, usability is a current challenge affecting patients’ mHealth use. Recent literature emphasizes an increased focus on patient-centeredness in diabetes care, user-centeredness in chronic disease mHealth system design and standardized, systematic approaches for usability evaluation. The aim of this thesis and its individual studies was to incorporate these foci into the evaluation of two mobile health self-management systems for diabetes.

Study I used ISO standard 9241-11 to examine the relationship between selected group characteristics of diabetes patients on specific interaction outcomes to quantitatively identify needed system modifications. Study II utilized a multi-method design to assess diabetes patients’ mHealth usage and combined two novel analytic methods to structure and analyze results. Study III used a modified, user-oriented heuristic evaluation (HE) method, validated tasks and in-depth severity factor ratings to identify critical problems from patients’ point of view. By developing and employing a modified, user-centered cognitive walkthrough method (UC-CW), study IV assessed its effectiveness and efficiency in finding relevant usability problems for users as well as patients’ acceptance. The modified CW was validated against the golden-standard user test with Think Aloud (TA).

Study I emphasized the importance of considering user characteristics in mHealth performance as these influenced interaction outcomes. All patients had difficulties with multiple-step tasks. Patients more recently diagnosed were able to perform tasks more successfully, with fewer errors and at faster times and had higher satisfaction scores; similar outcomes to the more experienced users. Educational level did not, however, seem to influence performance. In study II, the usability test with Think-Aloud, in-depth interviews and questionnaires contributed to 19 consolidated issues, and triangulated on 5 critical usability problems for users. The combined analysis methods resulted in structured, categorized descriptions to aid in problem-solving. In Study III, the disease-related, critical information deficiencies found by expert evaluators using the modified, structured method also converged on and highlighted potentially adverse user concerns. Study IV demonstrated that the UC-CW found more critical user problems compared to the user test with TA despite both methods producing similar major average severity ratings and violations of heuristic categories. The modified method was more efficient per detected problem and experienced as less cognitively demanding and with a higher ease of use.

These studies offer different approaches that include patient-centered, efficient and user-acceptable methods and method modifications to detect critical usability issues for
users. Importantly, improved mHealth designs for users could mean improvement in interactions, interaction performance, increased adoption, and long-term perhaps even increased adherence to interventions for chronic conditions.
Sammanfattning på svenska

Diabetes är en kronisk sjukdom som drabbar 422 miljoner patienter världen över enligt data från Världshälsorganisationen WHO med 30,3 miljoner i USA och 64 miljoner i Europa. Utbredningen visar på behovet av bättre sätt att stödja patienter i sin egenvårdsbehandling av sjukdomen. mHälsolösningar används alltmer för detta; men användbarhet är en utmaning som påverkar patienternas mHälsolösningar. Den senaste litteraturen inom området betonar behovet av ett ökat fokus på patientcentrerad diabetesvård, användarcentrerad hjälp för kroniska sjukdomar och standardiserades, systematiska tillvägagångssätt i utvärdering av användbarhet. Syftet med denna avhandling samt dess ingående, individuella studier var att inkludera dessa inriktningar i utvärderingen av två mobila system för diabetesegenvård.


Studie I betonade Viktigheten av att ta hänsyn till användarkaraktäristika i utförandet av utvärderingsuppgifter i mHälsa då dessa påverkade interaktionen och utkomsterna. Alla patienter hade svårigheter med de uppgifter som krävde flera steg. De patienter som var mer nyligen diagnostiserade kunde genomföra utvärderingsuppgifterna mer framgångsrikt, med färre fel och snabbare och hade högre tillfredsställelse; vilket var liknande resultat som för de mer erfarna användarna. Utbildningsnivå verklade, emellertid, inte ha någon påverkan på utförandet. I studie II bidrog användbarhetstestet med Think Aloud, djupintervjuer och enkäter till 19 konsoliderade problem, och triangulerade på fem kritiska problem för användarna. De kombinerade analysmetoderna resulterade i strukturerade, kategoriserade beskrivningar för att bidra till problemlösning. I studie III ledde den modifierade, strukturerade metoden till att expertutvärderarna fann och konvergerade på sjukdomsrelaterade, kritiska informationstillkortkommanden och poängterade potentiellt farliga användbarhetsproblem. Studie IV resulterade i att UC-CW metoden hittade fler kritiska problem för användarna jämfört med användbarhetstestet med TA även om båda metoder bidrog till liknande höga allvarlighetsbedömningar och brytande mot samma heuristiska kategorier. Den modifierade metoden visade sig mer effektiv per upptäckt
användbarhetsproblem och upplevdes som en lägre kognitiv belastning och som mer lättanvänd.

Dessa studier bidrar med olika ansatser som innefattar patientcentrerade, effektiva och också användarvänliga metoder och metodmodifieringar för att finna kritiska användbarhetsproblem för användarna. Långsiktigt kan förbättrade mHälsodesignlösningar för användarna betyda förbättrade interaktioner, ett förbättrat genomförande, ett ökat anammande av dem och kanske även ökad följsamhet vad gäller interventioner för kronisk sjukdom.
Included Studies


IV. Georgsson, M., Staggers, N., Årsand, E., Kushniruk, A. Using a User-Centered Cognitive Walkthrough to Evaluate a mHealth Diabetes Self-Management System Including a Case Study and External Validity Test. Manuscript

Additional related work not included in this thesis:


Summary of Studies

Overview

<table>
<thead>
<tr>
<th>Study design</th>
<th>Patient sample/ Evaluators</th>
<th>Data collection</th>
<th>Analysis methods</th>
<th>Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISO standard measures (efficiency, effectiveness, satisfaction) quantitatively compared to user characteristics</td>
<td>RCT sample: 2317, 18 clinics From control group: n=10 (female: 6, male: 4)</td>
<td>Questionnaires (demographics, ICT/Comp. knowledge), SUS, user tests</td>
<td>ISO standard (efficiency, effectiveness, satisfaction measures), grouped user characteristics</td>
<td>I (JAMIA)</td>
</tr>
<tr>
<td>Multi-method design (user test with TA, semi-structured interviews, questionnaires)</td>
<td>RCT sample: 2317, 18 clinics From control group: n=10 (female: 6, male: 4)</td>
<td>User test with TA, Semi-structured in-depth interviews, questionnaires</td>
<td>Coding and analysis with Framework Analysis (FA) and Usability Problem Taxonomy (UPT), Severity rating: Travis</td>
<td>II (JBI)</td>
</tr>
<tr>
<td>Heuristic evaluation (modified with in-depth severity factors) to guide analysis</td>
<td>3 dual domain evaluators (usability and health care)</td>
<td>Heuristic evaluation, modified</td>
<td>In-depth factor rating of usability problems (frequency, impact, persistence), HE violations, Severity rating: Nielsen</td>
<td>III (CIN)</td>
</tr>
<tr>
<td>UC-CW (modified cognitive walkthrough with users) validated with user test with TA</td>
<td>Purposive sampling n=12 Modified CW n=6, (female: 2, male: 4), Type 1, 2 diabetes User test with TA n=6 (female: 3, male: 3), Type 1, 2 diabetes</td>
<td>Cognitive walkthrough modified (UC-CW) and validated with TA</td>
<td>UC-CW: Usability problems (effectiveness), time (efficiency), (acceptance) Nasa RTLX, in-depth interviews Severity rating: Nielsen Validation: 6 TA sessions</td>
<td>IV (Manuscript)</td>
</tr>
</tbody>
</table>
Author Contributions

Study I Quantifying Usability: An Evaluation of a Diabetes mHealth System on Effectiveness, Efficiency, and Satisfaction Metrics with Associated User Characteristics. This is a summative usability evaluation study of a mobile health application for diabetes comparing patients with varying characteristics on a number of performance interaction outcomes. The results demonstrated that user characteristics need a more thoughtful consideration in the design of mHealth systems. This study was one of the featured articles in the JAMIA special issue Interactive Systems for Patient-Centered Care to Enhance Patient Engagement and one of the five most cited JAMIA articles of 2016. Georgsson was responsible for the study’s conception and design, the data collection and analysis of the data. He was also responsible for drafting the article. He was joined by his co-author in the interpretation of the data and critical revision, as well as final approval.

Study II An Evaluation of Patients’ Experienced Usability of a Diabetes mHealth System Using a Multi-Method Approach. This is an evaluation of patients’ experienced usability of a diabetes mHealth system focusing on a multi-method approach (user test, interviews and questionnaires) to evaluate a mHealth system. Novel analysis methods used in the study include the Framework Analysis (FA) method and Usability Problem Taxonomy (UPT) to structure, code and analyze the found usability problems. The result consisted of a list of in-depth usability problem descriptions that could aid in the communication and problem-solving and assist system development and design. It also provided an example of a standardized, structured and reproducible way of working in usability evaluation and analysis. Georgsson was responsible for the study’s conception and design, and the data collection. He was also responsible for drafting the article. He was joined by his co-author in the analysis of the data, interpretation of the data and critical revision, as well as final approval.

Study III A Modified, User-Oriented Heuristic Evaluation of a Mobile Health System for Diabetes Self-Management Support is a study that outlines an expert evaluation of a mHealth application for diabetes. It consists of a modified heuristic evaluation (HE) with a user-oriented approach using dual-domain experts (in usability and nursing) to evaluate the system, employing user-centered, validated tasks and realistic self-care scenarios, as well as in-depth severity factor ratings. The modified HE uncovered unique, critical issues as opposed to less severe issues of a non-critical nature, a common drawback of this method. This study was selected as one of the articles to be included in the highlights collection issue of 2016 for the journal CIN. Georgsson was responsible for the study’s conception and design, and data collection. He was also responsible for drafting the article. He was joined by his co-authors in data analysis, interpretation and critical revision, as well as final approval.

Study IV Using a User-Centered Cognitive Walkthrough to Evaluate a mHealth Diabetes Self-Management System Including a Case Study and External Validity Test (Manuscript). A proposal of the initial method modification was presented and
published as a peer-reviewed conference proceeding at pHealth 2016 and published in Studies in Health Technology and Informatics (see additional work not included in this thesis). In the described study the modified usability method was tested to determine its effectiveness and efficiency in finding usability problems as well as its acceptance to users. It was also validated against the golden-standard empirical usability test with users. Georgsson was responsible for the study’s conception and design, the data collection and analysis of the data. He was also responsible for drafting the article. He was joined by his co-authors in the analysis of data and interpretation, critical revision, as well as final approval.
# Table of Contents

Abstract ................................................................................................................................. i
Sammanfattning på svenska ................................................................................................. iii
Included Studies ....................................................................................................................... v
Summary of Studies ................................................................................................................ vi
Author Contributions .......................................................................................................... vii
Table of Contents .................................................................................................................... ix
List of Figures ........................................................................................................................ xi
List of Tables .......................................................................................................................... xii

## Background .......................................................................................................................... 1
  Diabetes - a Chronic Disease Condition of Large Proportions .............................................. 1
  Diabetes Etiology and Ways of Treatment .......................................................................... 1
  Information and Communication Technology, mHealth, Self-Management and Diabetes ..... 2
  mHealth Self-Management Systems and Usability ............................................................... 3
  User-Centered Design ......................................................................................................... 3
  Definition of Usability ......................................................................................................... 4
  Usability Evaluation Methods ............................................................................................. 5
  Common Advantages and Disadvantages with the Expert- and User-based Methods ......... 8
  Current Gaps in Diabetes mHealth Self-Management Research and Usability .................. 9
  Research Aims ..................................................................................................................... 10
  Overall Research Questions ............................................................................................... 11
  Theoretical Construct ......................................................................................................... 12
  System Descriptions .......................................................................................................... 12

## Materials and Methods ...................................................................................................... 17
  Study Samples and Settings ............................................................................................... 17
  Study Procedures ................................................................................................................ 18
  Data Collection Methods ................................................................................................... 23
  Data Analysis ...................................................................................................................... 26
  Ethical Considerations ....................................................................................................... 32

## Results .................................................................................................................................. 33
  Study I ................................................................................................................................ 33
  Study II .............................................................................................................................. 35
  Study III ............................................................................................................................ 41
  Study IV ............................................................................................................................ 45

## Discussion ........................................................................................................................... 53
  Study Approaches and Contributions ................................................................................. 53
  Lessons Learned and Recommendations .......................................................................... 61
  Limitations .......................................................................................................................... 63

## Conclusions .......................................................................................................................... 65
  Contributions and Implications of the Research ................................................................. 67
  Future Research .................................................................................................................. 69

## References ........................................................................................................................... 70
Study I.............................................................................................................................. 81
Study II ............................................................................................................................. 89
Study III .......................................................................................................................... 105
Study IV........................................................................................................................... 114
List of Figures

Figure 1 The usability components according to ISO 9241-11 .................................................. 4
Figure 2 System overview of the Care4Life system ................................................................. 13
Figure 3 Hierarchy of system views for the Care4Life system .................................................. 14
Figure 4 System overview of the Diabetes Diary application .................................................. 15
Figure 5 Hierarchy of system views for the Diabetes Diary application ................................. 16
Figure 6 Classification of categories according to the Usability Problem Taxonomy .......... 28
Figure 7 Application of the FA method and UPT ................................................................. 30
Figure 8 Modified heuristic evaluation process ................................................................. 31
Figure 9 Task success and error rate by task .................................................................. 34
Figure 10 Venn diagram with problem partitions for each data collection method .......... 37
Figure 11 Distribution of usability problems and severity ratings across system views ...... 38
Figure 12 Distribution of artifact and task component classifications on primary level ...... 41
Figure 13 Heuristic violations by heuristic category .......................................................... 42
Figure 14 Severity ratings across system views ................................................................. 43
Figure 15 Comparison of the number of usability problems, number of violated heuristics and average severity per system view for the UC-CW and TA ...................................................... 47
Figure 16 Heuristic violations across methods .................................................................. 48
Figure 17 Comparison of the number of detected usability problems, their classification on severity levels and average severity ratings across views for the UC-CW and TA .......... 49
List of Tables

Table 1 The included metrics for determining usability according to ISO 9241-11 ...............5
Table 2 Heuristic categories for usability evaluation ...........................................................................6
Table 3 Scenario and task in the evaluation for study I, II and III ......................................................21
Table 4 Scenario and task in the evaluation for the UC-CW part of study IV .........................................22
Table 5 Scenario and task in the evaluation for the TA part of study IV ..............................................23
Table 6 Scale for severity rating and severity factors ..............................................................................31
Table 7 Time on task per task ..................................................................................................................34
Table 8 Usability problems and classifications for the most severe issues .................................................39
Table 9 Effectiveness, efficiency and acceptance of the UC-CW and TA method ..................................45
Table 10 Problem descriptions, heuristic violations and severity rating per view for top five critical usability problems per method .................................................................50
Table 11 Time per method for whole evaluation procedure ......................................................................51
Table 12 Time per evaluation session part ...............................................................................................51
Background

Diabetes - a Chronic Disease Condition of Large Proportions

Chronic disease makes up a large disease burden and cause of death around the world [1]. Diabetes Mellitus (DM) is one of the chronic disease conditions affecting almost 422 million people globally according to the most recent figures from the World Health Organization (WHO) [2]. By the year 2030 it is expected that this disease will constitute one of the leading causes of death worldwide [3] with an increase of 205 million persons affected until the year 2035 [4]. Data from the National Diabetes Statistics Report of the Centers for Disease Control and Prevention (CDC) from 2017 show that currently in the United States the number of people who live with this disease is 30.3 million or 9.4% of the population and this figure is continuing to grow [5]. In Europe the corresponding numbers for this disease are 64 million who are affected by it [6] and for Sweden alone more than 400 000 have the disease according to the Swedish National Diabetes Register [7].

Diabetes Etiology and Ways of Treatment

Diabetes Mellitus is a chronic illness which occurs either when the pancreas is not producing enough insulin or when the body cannot effectively use the insulin produced. Since insulin is the hormone that regulates the blood sugar levels, an uncontrolled or raised blood sugar level (hyperglycemia) can lead to serious complications and damage especially when it comes to the body’s nervous system and blood vessels [3, 8].

The two most common types of diabetes are Type 1 and Type 2. Type 1 diabetes (or juvenile or insulin-dependent diabetes) is the form that requires daily insulin treatment through injections. The symptoms that occur during the onset of the disease include polyuria, and polydipsia, constant hunger, weight loss, decreased vision and extreme fatigue, which can all occur instantaneously [8]. Type 2 diabetes (also called adult-onset diabetes) is mainly the result of the body’s inability to effectively use insulin. This most common type of diabetes is largely due to reasons such as being overweight and having a sedentary lifestyle. The symptoms are similar to those of Type 1 diabetes but they are often less pronounced which can cause the disease to remain undiagnosed until the different complications occur [8].

The standard treatment for diabetes mellitus Type 1 and Type 2 normally consists of regulating the blood sugar levels with a combination of taking medication (insulin and/or oral medication), modifying nutritional intake, introducing some type of physical activity, and receive education about the disease [9, 10]. The continuous monitoring of the condition is considered very important to be able to avoid complications and/or secondary diseases [9]. Compliance is therefore essential for this complex disease and it requires a considerable amount of self-care by the patient [11, 12]. Since this can involve a number of different aspects that need to be controlled and/or changed this can be
considered a burden for the patient and contribute to difficulties in handling and coping with the disease [13]. Extensive support is therefore needed [14].

Type 2 diabetes, the most common form of the condition, affects almost 90% of those diagnosed [2]. As noted, this diabetes type is largely lifestyle-related and can also be self-managed by the patient to a certain degree along with the addition of the more conventional treatment procedures involving glucose-lowering treatments [15]. To avoid complications, continuous medical care and monitoring of the condition are vital. Poor glycemic control and insufficient self-management practices can lead to a worsening condition [12, 16]. Adherence is therefore of key importance but it has been found that adherence rates for needed life-style changes are often quite low [11].

Due to the complex nature of diabetes, with its significant changes in lifestyle, and extensive requirements with regards to self-management, it can be a burden and cause difficulties for patients trying to cope with their disease and its demands [13, 17, 18]. Patients currently perform almost 95% of their own diabetes care which emphasizes their vital role in disease management [13]. For this care to succeed and be effective, patients must have the ability to make well-informed decisions about how to live with their illness and also have the necessary tools available to them as well as strong health care provider support [14]. This puts heavy demands on patients as well as providers [14].

**Information and Communication Technology, mHealth, Self-Management and Diabetes**

Information and Communication Technology (ICT) is defined by UNESCO as “the combination of informatics technology with other related technologies, specifically communication technology” [19] and with ICT specifically referring to “forms of technology that are used to transmit, process, store, create, display, share or exchange information by electronic means” [20]. This broad definition of ICT includes those technologies that provide access to information through telecommunications such as the Internet, wireless networks, mobile phones, satellite systems and other communication mediums and technologies [20]. ICT is one of the most common ways to self-manage chronic disease conditions [21], and researchers highlight that its use in daily disease management is vital for improved patient outcomes [10, 22].

Current figures point to the fact that patients are becoming increasingly active and interested in monitoring their disease through different ICT technologies. A study by the PEW Research Center for the Internet and American Life Project showed that 40% of patients suffering from one chronic condition, track their disease regularly and for those patients with multiple chronic disease conditions the percentage was 62% [23]. Of these patients, 76% considered tracking to be positive for their health outcomes and intended to continue to do it on a regular basis [23].
As noted, ICT usage is rising, and different diabetes self-management support systems have also been developed for this purpose including mobile health or mHealth technologies [24-28]. The WHO defines mHealth as a part of eHealth or electronic health which consists of different information and communication solutions and services for health care. According to the definition, mHealth signifies the medical and health applications supported by mobile units and systems. These include, among others, mobile phones and similar forms of mobile technology, as well as other wireless communication devices [29]. According to the WHO Global Observatory for Health, mobile and wireless technologies are one important way to support health goals [29]. Another recent definition of mHealth focuses on its particular use in a health care context and on its benefits to the individual user. Kumar et al. describe it as “mobile health (mHealth) seeks to improve individuals’ health and wellbeing by continuously monitoring their status, rapidly diagnosing medical conditions, recognizing behaviors, and delivering just-in-time interventions, all in the user’s own environment” [30].

Different types of studies involving mHealth show that these solutions have proven successful and effective in improving glucose control for diabetes patients [27, 31-37] as well as aiding in achieving lifestyle changes [38]. Other interventions that have had beneficial effects on patients, for example, include algorithm-based systems for self-coaching with SMS messages that are individually tailored [39] and provide feedback related to patients' lifestyles [40].

mHealth Self-Management Systems and Usability

The worldwide penetration of mobile phones are at an estimated 7.5 billion subscriptions as of 2017 [41] with 396 million wireless subscriber connections solely in the United States [42]. This growing trend in mobile phone usage is expected to continue which contributes to the increasing importance of adequate usability for these devices [43].

Despite of the increasing number of available self-management tools in mHealth, many of the applications are deficient in how usable they are for the intended user [26], making usability of these solutions an important issue to address comprehensively [43, 44]. For the individual patient it is, for example, vital that the user interface is designed in a way that safeguards both clinical and user safety to avoid user errors which could discourage the user experience with the system [45-48]. Conducting usability evaluations can therefore be one effective way to determine how well the system or application meets the patients’ needs and expectations as well as the clinical requirements to assure both outcomes [49, 50].

User-Centered Design

Donald Norman and Stephen Draper [51] initiated the concept of user-centered design (UCD) and highlighted the necessity of having a good understanding of product users during the design and development processes of a product or service. In his definition from 1986 [52] Norman defined the central tenets of user-centered design as being: “the purpose of the system is to serve the user, not to use a specific technology, not to be an
elegant piece of programming. The needs of the users should dominate the design of the interface, and the needs of the interface should dominate the design of the rest of the system” [52]. Later Dennis Wixon, in an article by Karat et al in 1996 [53], added that the user-centered design process itself should be seen as: “... one that sets users or data generated by users as the criteria by which a design is evaluated or as the generative source of design ideas” [53].

**Definition of Usability**

Usability is an important part of user-centered design [54] and means that any part of a system should be easy to operate, easy to learn, easy to remember and also helpful to the user and this should also guide designers in the design process. It also means adhering to design principles such as an early focus on users and tasks, empirical assessments and iterative design cycles [55].

One way to define usability was outlined in the ISO 9241-11 standard [56]. The ISO 9241-11 is made up of specific components to determine how well a user fulfills certain specific goals, as well as the acceptability and satisfaction with the product that users feel when they interact with it in a specific context or more specifically “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.” [56] (see Figure 1).

![Figure 1 The usability components according to ISO 9241-11](image)

Essentially the ISO 9241-11 outlines what the user interaction should entail, with hands-on methods to indicate the overall usability of a product [57, 58]. The specific metrics in the standard can also be practically assessed which is commonly done through different representative tasks to determine how well the intervention supports the user in achieving specific task goals [59, 60]. Effectiveness is a measure commonly determined by measuring task completion success or counting the number of specific errors the user performs during the process of interaction. Efficiency signifies the level of effort and resources the user needs to achieve the goal(s); which is usually determined by timing
the tasks as well as averaging the times across users and/or tasks [56]. *Satisfaction* can be determined and measured in different ways (see Table 1) but one common method is using an instrument such as the System Usability Scale (SUS) by Brooke [61, 62]. This instrument assesses the overall satisfaction through a 10-item Likert scale and is commonly distributed at the end of the usability session to record users’ feelings and responses towards the system. By combining the instrument’s individual user scores, an estimate of the overall usability of the intervention is determined [61, 62]. These scores can be provided on a range from 0-100 where those of 70 and above are considered acceptable, those of 85 or higher signify a high usability level and those of 50 or lower imply a poor usability [63].

<table>
<thead>
<tr>
<th><strong>Table 1 The included metrics for determining usability according to ISO 9241-11 [56]</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Effectiveness:</strong></td>
</tr>
<tr>
<td>To what extent the user can achieve a goal with accuracy and completeness</td>
</tr>
<tr>
<td><strong>Efficiency:</strong></td>
</tr>
<tr>
<td>The level of effort and resource usage which is required by the user in order to achieve a goal in relation to accuracy and completeness</td>
</tr>
<tr>
<td><strong>Satisfaction:</strong></td>
</tr>
<tr>
<td>The positive associations and absence of discontent that the user experiences during the performance</td>
</tr>
</tbody>
</table>

**Usability Evaluation Methods**

Usability evaluations are commonly performed in the human-computer interaction (HCI) discipline and field of usability engineering where they can take on different forms; such as expert- and user-based methods. Heuristic evaluation and cognitive walkthrough are two common usability methods that involve expert assessments on the usability of a system [59, 64-66]. User-based evaluations typically utilize Think Aloud (TA) assessments with user tests [67, 68] or actual system users expressing their perceptions (e.g., interviews and questionnaires) when interacting with the system [59, 60, 69]. These methods are briefly explained next.

**Heuristic Evaluation**

Heuristic evaluation (HE) is one of the most frequently used usability inspection methods. In this method, that does not include users, experts apply their knowledge and understanding about design standards and principles to perform the evaluation [70]. Usability experts evaluate the system or application by systematically stepping through its different components using representative tasks to detect usability problems. Then, the experts assign a violated heuristic category or categories to the specific usability problem and ascribe a severity rating to a final master list of identified usability violations.
HE was initiated by Nielsen and Molich [47, 59, 64, 71] who defined a list of 10 different heuristic categories to choose from [64] to assign to each of the found usability problems/violations (see Table 2).

Table 2 Heuristic categories for usability evaluation [64]

<table>
<thead>
<tr>
<th>Visibility of system status</th>
</tr>
</thead>
<tbody>
<tr>
<td>The system should always keep users informed about what is going on, through appropriate feedback within reasonable time.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Match between system and the real world</th>
</tr>
</thead>
<tbody>
<tr>
<td>The system should speak the user’s language, with words, phrases and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>User control and freedom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Users often choose system functions by mistake and will need a clearly marked &quot;emergency exit&quot; to leave the unwanted state without having to go through an extended dialogue. Support undo and redo.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Consistency and standards</th>
</tr>
</thead>
<tbody>
<tr>
<td>Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform conventions.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Error prevention</th>
</tr>
</thead>
<tbody>
<tr>
<td>Even better than good error messages is a careful design which prevents a problem from occurring in the first place. Either eliminate error-prone conditions or check for them and present users with a confirmation option before they commit to the action.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Recognition rather than recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimize the user's memory load by making objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Flexibility and efficiency of use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accelerators -- unseen by the novice user -- may often speed up the interaction for the expert user such that the system can cater to both inexperienced and experienced users. Allow users to tailor frequent actions.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Aesthetic and minimalist design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dialogues should not contain information which is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Help users recognize, diagnose, and recover from errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Help and documentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the user's task, list concrete steps to be carried out, and not be too large.</td>
</tr>
</tbody>
</table>

After the usability problem has been determined and assigned a heuristic violation/s, the next step is to determine the severity of the problem by selecting a severity score between 0 (not a problem) and 4 (usability catastrophe) [64]. After the individual expert assessments have been completed, all the severity ratings are summed and then averaged.
through dividing the total number by the number of evaluators to receive a severity rating for each individual usability problem. The partitions for the averaged severity ratings consist of the value ranges: catastrophic (values \(\geq 3.5\)), major (2.5\(\leq\) value < 3.5), minor (1.5\(\leq\) value < 2.5) and cosmetic (value < 1.5).

**Cognitive Walkthrough**

Cognitive walkthrough (CW) is another expert-, and group-based, inspection method to evaluate usability. Similarly to the HE, experts evaluate the system by systematically stepping through its different components to detect usability problems. In this method, that similarly to the HE does not include users, experts apply their knowledge and understanding about design standards and principles to perform the evaluation. It was originally drafted and introduced by Polson and Lewis [72] and is based on theories of cognitive exploratory learning [66] aimed at supporting the user’s ability to learn through actions [72]. Within CW evaluation, understanding the cognitive aspects that are involved in the user’s learning when performing tasks, is essential. That these tasks can be completed with the system providing suitable and adequate feedback is of central importance. If this is not the case, it is necessary to find out where the system is deficient.

In CW evaluation, experts attempt to understand the intended users’ problem-solving processes and any usability problems that can be predicted as a result of the interaction (or how easy or difficult it is for the new user to accomplish tasks in the system). The process begins with a task analysis that specifies the sequence of steps or actions that a user needs to perform to accomplish a task as well as the system’s response to those actions. The designers, developers and evaluators of the software then walk through the steps together, asking themselves 4 questions at each step. These are: *Will the user try to achieve the effect that the subtask has? Will the user notice that the correct action is available? Will the user understand that the wanted subtask can be achieved by the action? Does the user get appropriate feedback?* [66]. The data are gathered throughout the walkthrough using specific designated forms, and afterwards a report of potential usability issues is compiled and shared [65].

**Usability Test with Think Aloud Protocol**

Usability test with Think Aloud protocol (TA) is a usability assessment method with a user-based focus, meaning that the user performs the evaluation. It is commonly applied to determine how the users think and reason as they perform the specific system tasks. Simon and Ericsson introduced the idea of verbal reports as data in 1984 with a revision occurring in 1993 [73]. The method has since then become well established within the field of human factors [59]. For the Think Aloud method, users are asked to talk aloud during their interactions with the system or application, expressing their reactions and what they think as they perform the specific tasks in it [74]. The usability session is often either recorded in some form and/or complemented with written notetaking by an observer during the evaluation process [60]. It is important that minimal intervention occurs from the usability evaluator to assure that the users’ thought processes are not interrupted, except to remind users to keep talking if they stop [60]. The main focus in
the method is on understanding the users’ decision making processes and how they experience the system or application expressed in their own words [59, 60].

**In-depth Interviews and Questionnaires**

In-depth interviews and questionnaires have also been used extensively as evaluation methods in the area of usability research for establishing the users’ ideas of any obstacles they might encounter in an evaluated system [60, 74]. These methods also include the users themselves in the evaluation process for their views either through interviewing them after they have used the system or having them fill out a questionnaire about their user experience [75]. Both of these ways of evaluating are used to obtain information and insight into how the system or application needs to be changed to accommodate users’ different skill levels, abilities and experiences [76, 77] and also to identify the usability problems that need to be corrected.

**Common Advantages and Disadvantages with the Expert- and User-based Methods**

Several advantages with expert-based methods exist. One is that expert-based methods can be applied early in design phases of a product. They also generate results quickly, efficiently and with low resource requirements and costs [47, 78] compared to user testing methods [79]. These methods also have the ability to detect many usability problems, which have added to their popularity and use in usability research [78].

Some of the common disadvantages mentioned with heuristic evaluation, for example, are that despite the fact that many problems can be found, they are typically low priority, minor interface design issues [78, 80, 81] or one-time problems [78, 82]. With the cognitive walk through a criticism is that the process often identifies many more problems than would have been found by the actual users in a test session. As well, the identified issues are not always relevant to the users [68]. In addition, the identified usability problems are commonly less general and less frequent than those found by other usability evaluation techniques [78].

A particular advantage to user testing is, as the name implies, the inclusion of the actual user in the evaluation. The method is also advantageous as it produces extensive, and detailed data, where only a small number of users (commonly between five to eight) is normally needed, to detect a large amount or 80 to 85% of the critical usability problems [83-86]. Usability testing with users is helpful at finding more severe and serious problems for users as well as more recurring and global problems [78, 80]. One of the main disadvantages of user tests is that they are more resource intensive, and also more costly and time-consuming to conduct [78, 79, 81].
Current Gaps in Diabetes mHealth Self-Management Research and Usability

*Increased Focus on Self-Management and Patient-Centeredness in Diabetes Care, User-Centeredness and Collaborative Approaches in Chronic Disease mHealth System Design*

Patient-centered goals and practices in diabetes self-management were recently highlighted in several guideline documents, among them the diabetes 2014 and 2017 National Standards for Diabetes Self-Management Education and Support [87, 88]. The American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD) also issued a joint statement recommending that diabetes treatment include patient preferences, needs and values as foci and that patient-centered approaches can result in more effective diabetes care delivery [89, 90]. Self-care is also brought up as optimizing the person-centered management of diabetes with the need for updating care approaches to promote self-management, increase person-centered diabetes knowledge and decision making, and deliver individualized care [91] also through the use of eHealth and mHealth technologies [92].

A concurrent call for more patient-centered care in chronic disease management is related to ICTs where recent reviews found a lack of ICT-enabled patient-centered care [93] asking for further research to be completed on this. Developers of self-management and self-monitoring systems are also encouraged to engage patients in their personal care [94] and include the chronic disease patient’s voice to a greater extent in the design of the devices [94, 95] as well as consider their needs more extensively in these kinds of self-monitoring solutions and applications [54, 96, 97].

Authors also specifically encourage researchers to adopt user-centered design foci both in the development of patient-centered technologies [98] and also in the implementation and evaluation of mobile health solutions for managing chronic conditions. In a recent review (2016), authors’ concluded that user-centered approaches, as well as collaborative and interdisciplinary team approaches, are imperative to obtain the required expertise and insight about users and to enhance feasibility, acceptability as well as usability of mHealth innovations [99]. Another conclusion was that developers should focus on involving users in diabetes application development as there is a lack of usability considerations especially for diverse patients with varied characteristics [100].

*Need for Rigorous, Standardized and Systematic Approaches in Usability Evaluation*

Concerns have also been raised about the performance of usability evaluation studies on systems in health care as a whole and more specifically in mHealth system design and diabetes mHealth research.

A recent systematic review by Ellsworth et al (2017) [101] which focused on published usability evaluations of electronic health records (EHRs), found that usability studies
generally lacked rigor. The authors found that questionnaires were the main method used to assess usability. When these were self-constructed, researchers failed to describe survey creation methods in such a way that it was possible for readers to assess survey psychometrics. When interviews, focus group or Think Aloud evaluations were conducted, almost all lacked sufficient details about the techniques used to moderate the sessions and the qualifications of the evaluators [101]. Also, the lack of consistent reporting when it comes to the background reporting frameworks lead to study limitations, according the authors, impeding the readers’ ability to make judgements about the reliability of the evaluation or being able to validate its outcomes. Ellsworth et al. concluded that the improvements needed in usability standardization as well as transparency have never been greater [101].

Similar sentiments have been raised in usability research on mHealth. Reviews of mHealth usability studies, in particular, emphasized the need for more rigorous, robust studies and methods in these evaluations. Zapata et al (2015) [102] pointed to the necessity of (1) utilizing two or more usability methods of different kinds, as well as (2) more systematic methods overall in mHealth evaluations, and also (3) using open-ended methods of inquiry with qualitative-oriented analyses to identify more comprehensive user concerns. They recommended that usability researchers also standardize processes [102]. Similar views have also been raised about diabetes mHealth systems. Lyles et al (2014) [103] concluded that user engagement, across all phases of design through evaluation, could benefit from a standardized approach and validated metrics as well as a greater focus on characteristics of the participant population, such as for example educational attainment, health literacy or self-reported technical knowledge [103].

International standards such as the ISO and other instruments for validation exist to support system design, but these have been used sparsely within health care. Several authors suggest that the ISO guidelines be used to a greater extent [104] to guide usability evaluations [105]. Standards such as the ISO 9241-11 can also be suitable for mHealth technologies. Bevan concluded that, overall, standards such as these should have a more frequent use in usability work due to that they are objective, define good practice, can ensure consistency as well as provide benchmarks for intervention design [104]. The adoption of more standardized, comprehensive approaches could also help with methodological consistency in mHealth evaluation making it possible to improve and also perhaps start to compare findings across different systems and applications [103] making usability testing more systematic and complete.

In sum, advantages and shortcomings have been raised about current usability evaluations, the need to improve the method results and outcomes in mHealth as well as adding more user-centered, patient-centered and systematic approaches for evaluation. Ways to address these specific areas are the foci of the included studies in this thesis.

**Research Aims**

The overall aim of this thesis was to determine different approaches of evaluating usability by including a more patient-centered focus to enhance patient engagement and
standardize, structure and streamline evaluation methods to allow for reproducible ways in usability evaluation. The specific aims of the individual studies were:

Study I: To examine the relationship between selected patient characteristics of a group of patients with diabetes on interaction outcomes (efficiency, effectiveness and satisfaction) through the use of standardized usability methods involving the ISO standard using a mHealth diabetes self-management system and show a way to quantitatively identify needed system modifications.

Study II: To determine the number of identified usability problems and how various usability methods contributed to the set of issues using a multi-method approach consisting of a usability test with Think Aloud, in-depth interview and questionnaire with diabetes patients about their usage of a mHealth diabetes self-management system. In addition, the combination of the two analysis methods, the Framework Analysis (FA) and Usability Problem Taxonomy (UPT) were used to code, structure and analyze the results to produce a list of in-depth categorized usability problem descriptions to be used as a communication tool in problem-solving. The intent was also to provide an example of a reproducible way to work in usability evaluation and analysis with a large amount of qualitative data.

Study III: To determine usability problems of a diabetes mHealth self-management system through the use of a modified heuristic expert evaluation method with a user-oriented approach in evaluation. This through dual-domain experts (usability and nursing) using validated patient-centered self-care tasks, realistic care process scenarios, and in-depth severity factor ratings to determine problem influence on the patient. The goal was to see whether this method modification could find the more critical problems from a patient point of view as opposed to less severe issues which is a common disadvantage of this method.

Study IV: To find a streamlined, patient-centered, and also evidenced-based process for cognitive walkthrough by developing and using a modified CW method (UC-CW) that incorporated patient users and then determine the method’s effectiveness and efficiency in finding relevant usability problems for diabetes patient users interacting with a diabetes mHealth self-management system. The number of problems were rated in terms of their heuristics and severity levels. In addition, the modified method was validated against the “golden-standard” user test with Think Aloud to determine how it compared in terms of the result and outcome and on user acceptance levels through NASA RTLX cognitive load scores and method opinions.

Overall Research Questions
1. How can more patient-centered approaches be included in usability evaluation methods of chronic disease and diabetes self-management systems to ensure patient engagement and in determining significant usability problems for users? (Study I, II, III and IV)
2. How can usability evaluations in mHealth become more structured, standardized and streamlined to accomplish reliability and validity throughout the data collection process and analysis and allow for a reproducibility of results? (Study I, II, III and IV).

**Theoretical Construct**

Activity theory (AT) was used as an overarching theoretical construct to draw inspiration from to guide the thoughts about the structure around the usability evaluation process and its inherent components. AT is commonly used for analytical approaches where a group of interconnected concepts can be used to identify and explore data. The theory states that all human effort has activity as a basic unit of analysis which is mediated through the use of tools. This theory is also applied in the human-computer interaction field and is there based on the components of subject, artifact or tool, object and outcome and seen as suitable to analyze actions and interactions with tools and artifacts in a contextual manner in a realistic setting [106-108]. A main reason for introducing the framework in the HCI field was that it was considered useful for thinking about the design of user interfaces and computer systems in the settings in which they were to be used [106]. It was also assumed that the theory could provide the contextual background that would allow for better designed and implemented technology.

AT originally came from Russian psychologists Vygotsky and Liontiev who used it to describe and explain school and work as activities in the developmental, cultural and historical context in which they occur [109, 110]. The framework was also later expanded by authors such as Engeström [111, 112] who provided an addition with concepts such as community, rules and division of labor to encompass additional aspects of the work context [112] and also aiming it towards primary care and medical practice [111]. Several authors within HCI have also used this framework. For example Kuutti [107] and Nardi [108] where Kuutti applied the framework to show how information technology can be used to support different kinds of activities at different levels [107] and Nardi who used the framework to show how it can be of value for examining data to come up with new sets of design concerns [108].

For the purposes of the performed studies, this analytical framework was used to think about the whole process of determining task difficulties and usability problems (the object) based on the diabetes patient, expert evaluator or similar (the subject), with the desired outcome or goal of improving the self-management process outcomes for the patients. Tasks were seen as a way to include the contextual factors and therefore it was important that these were realistic and based on actual needs of the diabetes patients and adherent to their context of use. Therefore, a lot of effort was put into designing and validating tasks that fit this purpose. All the included studies (I, II, III, IV) had these tasks as an essential evaluation component.

**System Descriptions**

System 1 (evaluated in study I, II, III) is an algorithm-based, mobile Short Messaging Service (SMS) combined with a web portal for diabetes patients called Care4Life
developed by Voxiva. Its main function is as an input-driven coaching support for patients, and as a self-management tool. It is also supposed to serve as an aid in the interaction between the patient and health care provider.

The parts of the interface have visual representations of measurements and goals which also consist of progress indicators in different colors such as red, yellow, and green. Patients can enter and track their results across these various parameters, such as, obtain their glucose readings or view their glucose trends in a diagram, see blood pressure values, determine their adherence when it comes to medication, exercise and weight goals as well as upcoming medical appointment visits. When patients send in their measurements they receive personalized responses from the system in terms of coaching support and system responses on their measurements and goals to be able to track their disease (see Figure 2).

![Figure 2 System overview of the Care4Life system](image)

The user interface consists of a Dashboard with sections containing access to different views such as 1) the Glucose Diary View; with sub-views: a) Submit Glucose Readings View and b) Glucose Readings View, 2) the Blood Pressure View; with sub-views: a) Submit Blood Pressure Readings View and b) Blood Pressure Readings View, 3) the Exercise Progress Goals View consisting of a) Submit Exercise Progress View and b) the Exercise Progress Readings View, 4) the Weight Tracker Goals View with a) Submit Weight Progress View and b) Weight Progress Readings View and a shared view: the Set Goals View, the 5) Medication Adherence Reminders View with sub-views a) Submit
Medication Adherence View, b) Medication Adherence Readings View and c) Set Medicine Reminders View, and finally a 6) Set Appointment Reminders View without any sub-views (see Figure 3).

**Figure 3 Hierarchy of system views for the Care4Life system**

System 2 (evaluated in study IV) is a diabetes self-management mHealth support system for Type 1 and 2 diabetes called the Diabetes Diary, developed by the Norwegian Centre for E-health Research.

With the Diabetes Diary, patients can enter and keep track of their blood glucose level, insulin, medication, dietary intake and activity level and can also analyze previous events and situations related to their food and medicine intake; all this to help them in their self-management decision-making for their diabetes. This self-management tool consists of five elements: a blood glucose data management system, a food habits data management system, a physical activity data management system, a personal goal-setting system, and a general diabetes information look-up system [27]. The blood glucose results can be transferred directly from the blood glucose monitoring system to the application via Bluetooth. The diet and physical activity systems enable manual data entry by the user. The glucose measurements can also be displayed visually in graphs, trend reports, and a feedback system through the color codings: red = below normal, green= normal, and yellow =above normal (see Figure 4).
In this application the user interface consists of a Main View with access to other views and sub-views consisting of: 1) the Glucose Entry View, 2) the Carbohydrate Entry View 3) the Insulin Entry View, 4) the Activity Entry View, 5) the Weight Entry View, 6) the Medication Entry View. All these views are connected to the sub-views: a) Date Setting View and b) Time Setting View. Also, the Carbohydrate Entry View and Insulin Entry View have access to the sub-view: a) Similar Situation List View and the 7) Glucose Graph View to the sub-views a) Plotted Graph View and b) Periodical Pattern Graph View. Additional views are the: 8) Preferences, 9) Tools and 10) List View (see Figure 5).
Figure 5 Hierarchy of system views for the Diabetes Diary application
Materials and Methods

Study Samples and Settings

Inclusion- and Exclusion Criteria
A sample of 10 patients (for each of the studies I and II) was randomly selected from a database convenience sample of 2317 patients representing 18 primary clinics in the Salt Lake City metropolitan area (Utah, USA). These patients were part of a large randomized controlled trial (RCT) on a mHealth intervention for diabetes [113] and were invited to participate in the usability evaluations for these studies. The inclusion criteria for each of the samples for the two studies (I, II) were that patients were diagnosed with Type 2 diabetes, that they had no cognitive impairment, had some knowledge of and had used computers, the Internet, and cell phone, and an ability to understand and speak the English language. A requirement was also that the patients had had no previous exposure to the mHealth system evaluated in the studies. The evaluation sessions were each conducted in a quiet room and or laboratory setting at a research institution called HealthInsight in Salt Lake City, Utah. Participants received a 20 dollar gift card for their participation.

Three expert evaluators performed the evaluation in study III. Due to the importance in this study of being familiar with the patient group and their needs when it comes to system support, each expert was selected based on their dual-domain competency which consisted of extensive usability experience in health informatics, being health care professionals (in this study Registered Nurses or RNs). They should also have knowledge and experience of diabetes patients’ requirements in terms of necessary tasks to support their self-management. The HE design of this study meant that only expert evaluators and no patients were involved, but the purpose of the method was to have a diabetes patient and these patients’ needs in mind when performing the evaluation. The expert evaluators each performed their evaluation part in their own offices at the University of Utah in Salt Lake City, Utah. They received no compensation.

A sample of 12 patients were selected by purposive sampling from a diabetes clinic, the Diabetes and Endocrinology Center at University of Utah Health in Salt Lake City, Utah for study IV. Six of the patients were randomized to a focus group setting performing a modified user-centered cognitive walkthrough method and 6 of the patients to individual usability sessions with Think Aloud. The inclusion criteria for the patient sample was that patients were diagnosed with Type 1 or Type 2 diabetes, that they had no cognitive impairment, that they had some knowledge of and had used computers, the Internet, and mobile phone, and were able to comprehend and speak English. Patients should also not have had any previous experience with the mHealth application for diabetes that they evaluated in the study. The usability evaluation sessions were either performed in a designated conference room with a projector (for the modified user-centered CW) or in a quiet, designated room at the clinic (the individual user sessions with TA). They received a 20 dollar gift card for taking part in the evaluation.
Study Procedures

After the patients had completed their written informed consent (study I, II, IV) they received an explanation of the evaluation process by researcher M.G. For study I patients filled out two validated pre-test questionnaires where one included specific questions about their background characteristics such as gender, age, length of their diabetes diagnosis, and their educational level and the other questions about their self-assessed IT/Computer, Internet knowledge, experience and use. For study II the evaluation also began with patients filling out a short pre-test questionnaire on similar demographic aspects as well as their diabetes information. They also filled out a validated post-test questionnaire on their IT/Computer, Internet, mobile phone knowledge and use as well as perceptions of these technologies. For study IV for each of the evaluation sessions, the modified, user-centered cognitive walkthrough (UC-CW) and the usability test with Think Aloud, patients also filled out validated questionnaires on their background demographic characteristics as well as their IT/Computer, Internet -and mobile phone experience and knowledge, frequency of use and perceptions.

For each of the studies I and II, standardized training was performed with the patients to simulate how a patient educational session on such a system would look in the actual health facility. This aspect was determined to be of essence to be able to decrease the individual variation in patients’ performance on evaluation tasks and to be certain of that all patients were exposed to the same system information [114]. When researcher M.G. completed the training, users could interact with the interface on their own to familiarize themselves with the system for about 10 minutes. For the UC-CW in study IV, M.G. instead demonstrated the mHealth application on a projector to the participant group following a standard procedure to decrease variability for a total of about 15 minutes where users also got to interact or ask for clarifications. For the TA, this demonstration process for the individuals on the computer (also according to a standard procedure) took 10-15 minutes per participant, or about 70 minutes in total, for the 6 patients. This in addition to M.G providing information at the beginning of the UC-CW session as well as the individual TA sessions about each of the specific methods participants were about to perform for 5 minutes. For study III, the dual domain experts also received identical instruction materials to learn about the system to ensure consistency among them. This consisted of a pre-recorded digital video on the different modules of the system, how the portal could be navigated, and an evaluation sheet with guidance for the evaluation.

For study I the patients interacted with the mHealth self-management system using eight specific designated scenarios and tasks outlined in a booklet where they had to enter and retrieve specific values and also read different graphs to be assessed on how effectively and efficiently they performed each assigned task. These interactions were also digitally recorded using Morae software. The researcher, M.G., also made observations and notations about patients’ individual performances on tasks directly into Morae. The session was completed by administering the System Usability Scale (SUS) instrument to assess patients’ level of satisfaction. The whole testing procedure, with all different parts, took about 1.5 hours per patient.
In study II the patients first performed a usability evaluation consisting of Think Aloud where they were also provided a booklet with eight specific scenarios and tasks to perform in the system. For each of the tasks, patients were asked to verbalize their thoughts out loud as they completed each task. If they fell silent, M.G., encouraged them to speak out loud about what they were thinking but any other interference with their train of thought was avoided. Similarly to study I interactions were digitally recorded using Morae where patients' navigation patterns on the different screens, facial expressions and voices were also captured. M.G. also here made observations and notations in the recording software about the individual performances on tasks. This usability test with Think Aloud was followed by an in-depth interview with an interview guide and interview topics where patients got to express their experiences with the system. For example, they could provide their opinion on what aspects they thought were easy and/or difficult with the system and any additional comments they had. This part was also digitally recorded and lasted for 15-20 minutes for each patient. The last part of study II consisted of a post-test questionnaire that also included patients' perceptions when it came to the system usability, where patients could also specify in writing what they found satisfying or dissatisfying with the system. The complete procedure for all the steps amounted to approximately two hours per patient.

In study III the evaluation process for the dual domain expert evaluators consisted of two parts. First, the evaluators became acquainted with the system and its use through the identical training materials they had received consisting of the system video and guidance for the evaluation. Then, they performed the modified heuristic evaluation. This entailed that each expert evaluator performed a set of eight scenarios and tasks provided in a booklet individually to detect usability problems. After this, they each assigned the detected problem/s one or several heuristics from Nielsen's 10 heuristic categories. These were then compiled into a master list where duplicate problems had been removed and the list was also verified among all evaluators for accuracy. As a final step, each evaluator assigned their severity rating to each problem in the master list with the rating factors: frequency, impact and persistence to determine the problem influence on the user. These severity rating factors were subsequently averaged into one severity rating for the individual usability problems.

In Study IV, after having received a short description of the modified CW (UC-CW) and demonstration of the application on a large screen, as well as having completed the pre-test questionnaires, each of the 6 patients was given a booklet of the 14 tasks with space to record their individual observations during the session and their answers to two simplified CW questions inspired by Spencer [115]; Would I (as a patient) know what to do at this step? If I do the right thing, do I (as a patient) know I have made progress toward my goal? These were also used for the subsequent group discussion. Each scenario and task was then read by one participant (all taking turns) and then the group guided the facilitator M.G. in how to perform the task on the screen with the projected mHealth application. At the end of each completed task, each patient first got to express their individual experiences about existing usability problem/s. Then to achieve a higher task level response, a group discussion was conducted to elicit additional thoughts about
possible deficiencies. After all tasks had been gone through, patients each filled out the post-test RTLX cognitive load instrument, and each individually answered four interview questions about how they experienced the method they had taken part in. The session was also digitally audio and video recorded through Morae software. The complete procedure for all the steps amounted to approximately 2 hours and 20 minutes for the whole group.

For the individual usability test with TA, in study IV, the participants also completed pre-test questionnaires after M.G. had explained the evaluation procedure and demonstrated the mHealth application on the computer screen to the individual participant. For the evaluation, they were each provided a booklet of the 9 tasks to work through. During the sessions, the individual patients were requested to think and verbalize aloud when performing each task in the diabetes application. If they fell silent M.G. encouraged them to keep talking, but avoided any other disturbance of the patient’s train of thought. Interactions in the system were digitally audio and video recorded using Morae software. M.G. also made notations and observations about the patients’ individual task performance directly into Morae. At the end of the evaluation each patient also filled out the post-test RTLX instrument, and answered the four interview questions about the method. The complete procedure with all 6 individual patients took 4 hours and 47 minutes in total.

**Usability Task Development and Validation**

All the scenarios and tasks used in the evaluations in study I, II, III, IV were developed to be as realistic as possible to simulate how the patients would use this kind of system in their self-management process.

Two of the studies (III and IV), in addition, included a content validity index rating for further validation. According to Lynn (1986) content validity is a crucial factor as it determines the extent to which items/elements included in an instrument (or in this case tasks) measure what they are intended to measure [116]. It determines the content representativeness or content relevance of the elements/items. The process consists of that items are developed and then rated in a judgement process by a number of experts to determine that the items are content valid. If there are five or fewer experts, all must agree on the content validity for their rating to be a reasonable representation [116]. This rating occurs using a 4 point ordinal rating scale where 1 signifies an irrelevant item and 4 an extremely relevant item. The actual content validity index is the proportion of items that receive a rating of 3 or 4 by the experts [116]. Authors recommend that to have excellent content validity the items together should reach a CVI of 1.00 and when including 3 to 5 experts have a CVI average of 0.90 or higher [117].

In study I the tasks were validated for face validity. This was done by three health care professionals as well as experts on usability, in a panel, to assure accuracy of task content and context. For study II and III a similar panel including a physician and a nurse with diabetes as a specialty, as well as a public health professional with experience of and expertise in chronic disease systems and a patient with diabetes also validated both
scenarios and tasks for face validity. In study III the scenarios and tasks were in addition also content validity indexed, achieving a content validity index average of 0.91 of 1.0. In study IV, the scenarios and tasks were first developed. Here they were based on the specifics of each method. This meant that for both the UC-CW and TA a review was first made of studies of applicable user tasks to include. The UC-CW evaluation then also included a validation in the form of a content validity rating.

Tasks of varying levels were developed by M.G. for the UC-CW, based on an extensive review of applicable guidelines for diabetes self-management [87, 92, 118] to assess the applicable aspects of the system and make sure that all important aspects for diabetes self-management were included. The tasks were then reviewed for accuracy by a panel consisting of three dual domain experts (usability and health care) and a diabetes patient and also content validity indexed by each. Two iterations were performed before reaching a total average score of 0.93/1.0 which was deemed acceptable [117]. For the TA, to adhere to the original method as close as possible for comparability and to prevent bias, a list of tasks were developed by M.G. based on similar studies using the user test with TA for diabetes mHealth self-management systems [119, 120].

Each set of tasks for study I, II, III and IV are provided below, along with an example scenario and task.

The eight tasks for study I, II, and III were disease-specific with varying levels of difficulty; where some included simple viewing while others consisted of several steps. These centered on registering and interpreting entries, managing entries, and self-management decision making. The specific tasks were to: (1) Upload glucose values into the system, (2) Interpret the glucose measurements shown in a graph, (3) Correct a recorded glucose measurement value, (4) Export the glucose value trends to a PDF for a provider visit, (5) Interpret a blood pressure measurement displayed in a graph view, (6) Set personal tracking goals for exercise and weight, (7) Set a medication reminder, and (8) Set a physician appointment reminder. An example of a scenario and task can be seen in Table 3.

**Table 3 Scenario and task in the evaluation for study I, II and III**

<table>
<thead>
<tr>
<th>Scenario:</th>
</tr>
</thead>
<tbody>
<tr>
<td>During your follow-up appointment with your provider, you agreed that a stronger commitment regarding weight loss and exercise would improve your diabetes condition. You now would like to activate the system’s support service for exercise tracking and weight tracking and put in your tracking goals regarding your exercise and weight.</td>
</tr>
</tbody>
</table>

Please complete the following tasks.  
1. Select and activate the service that you would like to use to set tracking goals for exercise and weight.  
2. Set your exercise goal to 3 times per week.  
3. Set your weight goal to 180 pounds. |
For study IV the validated tasks for the UC-CW were also disease-specific and of varying difficulty; where some included viewing and others consisted of several steps centering around the areas of registering entries, interpreting entries, managing entries and apply self-management decision making. These fourteen tasks consisted of: (1) Entering weight, (2) Entering physical activity, (3) Entering a glucose reading, (4) Entering carbohydrate intake, (5) Entering insulin, medication, (6) Interpreting the last entries for glucose, insulin, physical activity and carbohydrate intake on the main view, (7) Interpreting the last 24 hrs glucose entries on the Main View, (8) Interpreting entries in plotted graphs, (9) Interpreting entries in linear distribution graphs, (10) Interpreting entries in a list, (11) Searching and finding entries in a list, (12) Correcting and altering entries, (13) Performing a similar situations search, (14) Exporting, and sending entries to a health care provider. An example of a scenario and task can be seen in Table 4.

Table 4 Scenario and task in the evaluation for the UC-CW part of study IV

<table>
<thead>
<tr>
<th>Scenario:</th>
</tr>
</thead>
<tbody>
<tr>
<td>From the overview you notice that one glucose value was a little high during the last 24 hours. You would like to check this value more carefully and see a graphical overview of it and also to see if you have been within an acceptable range for most of the other values during the last week.</td>
</tr>
<tr>
<td>Please complete the following tasks.</td>
</tr>
<tr>
<td>1. Navigate to the view with the graphical overview of the plotted glucose readings for the last 24 hours in the Diabetes Diary with the high, low and acceptable ranges.</td>
</tr>
<tr>
<td>2. Locate the slightly high glucose value for the last 24 hours and state how high it was approximately.</td>
</tr>
<tr>
<td>3. Change views for the graph and look at all plotted values for the last week, and see if the values are mainly within an acceptable range and state this aloud.</td>
</tr>
<tr>
<td>4. When you consider yourself done with the task, finish and return to start mode or “home” in the Diabetes Diary app.</td>
</tr>
</tbody>
</table>

For the user test with TA the nine different tasks also involved simple viewing and different steps. These involved the areas of adjusting, personalizing the application, registering entries, interpreting entries, managing entries and decision making. These were: (1) Setting blood glucose measurement units, (2) Entering a glucose reading, (3) Entering carbohydrate intake, (4) Entering insulin, medication, (5) Interpreting entries in periodical pattern graphs, (6) Interpreting entries in plotted graphs, (7) Interpreting entries in a list, (8) Searching and finding entries in a list, (9) Exporting and sending entries to a health care provider. An example of a scenario and task can be seen in Table 5.
Table 5 Scenario and task in the evaluation for the TA part of study IV

Scenario:
After a brisk walk from the bus stop, you would like some lunch. You need to determine your glucose level to decide how much insulin you need for this meal. You measure your glucose value with the glucometer and get a value of 72 mg/dl. You also just remembered that you forgot to register your blood glucose value of last night in the Diabetes Diary, which you now also want to do.

Please do the following task at the same time as you talk out loud and describe what you are doing and what you are thinking.

1. Navigate to the view for entering you glucose in the Diabetes Diary.
2. Put in your new value of “72” mg/dl.
3. Enter a note "After a brisk walk” and save.
4. Enter the second glucose value of “85” mg/dl for yesterday night at “11” pm.
5. Enter a note "Forgot to put it in at night” and save.
6. When you consider yourself done with the task, finish and return to start mode or “home” in the Diabetes

Data Collection Methods
The data collection methods reflected each of the individual study purposes.

Usability Evaluation Using the ISO Standard, Background Demographics and Information Technology Experience Questionnaires
The data collection in study I consisted of two short pre-test questionnaires on demographics and IT/computers that were administered to patients to assess their background characteristics and IT experience. Patients answered specific questions about their background characteristics such as gender, age, length of their diabetes diagnosis, and their educational level and questions about their IT/Computer experience, self-assessed knowledge as well as their frequency of computer and Internet use.

The ISO standard usability measures of effectiveness, efficiency and satisfaction were then determined by users performing the usability test with the eight different tasks (described previously) and completing the accompanying System Usability Scale (SUS) instrument. Effectiveness was determined by the degree of task completion and total number of errors that the user performed and efficiency by timing the tasks. Morae software was used for audio-and video recording this interaction with the mHealth system. Satisfaction was determined by the SUS according to the guidelines developed by Brooke [61]. This instrument assesses the overall system satisfaction through a 10-item, 5 point Likert scale (from Strongly disagree to Strongly agree) and is commonly distributed at the end of the usability session to record users’ feelings and responses towards the system. The specific items to answer are: 1) I think that I would like to use
Think Aloud Usability Test, In-depth Interview, Pre-test Demographic and Post-test mHealth and Usability Experience Questionnaire

The data collection for study II also began with patients filling out a short demographic questionnaire on their gender, age, level of education, occupation, as well as length of diabetes diagnosis. The Think Aloud usability test then began consisting of patients performing the eight specified scenarios and tasks in the system. As described earlier, this consisted of that patients were asked to think aloud about what they were thinking and how they reasoned as they completed the prescribed tasks and researcher M.G also made recordings of individual performances on tasks in Morae. After the usability evaluation, patients were interviewed about their system experience through an open-ended interview with a topic guide and they also completed a post-experience questionnaire [121]. The interview was used to uncover comprehensive information from participants [122]. The three questions in the interview guide asked patients to comment on the parts of the system for which they experienced good usability and thought were well designed, and those those that they felt were inadequately designed as well as any other comments they wanted to express about the system usability.

For the post-interaction questionnaire on the mHealth system, patients first filled out a short-answer question section on their IT/Computer, Internet, mobile phone knowledge; then on their experience and use and on their thoughts about the usage of mobile and web service systems in health care; and what they thought about the particular mHealth system in terms of usability. They could also add any further comments they had about these topic areas. Patients were also asked about rating their preference for using the technology for work and/or leisure activities as well as in health care based on a 4 point Likert scale (Strongly agree (1) to Strongly disagree (4)). The last part of the post-interaction questionnaire consisted of open-ended questions about the particular system used in the study. Patients had to write down what usability/user experience they found satisfying and easy and which they experienced as dissatisfying or difficult to perform in the system. They also had the possibility to add further comments about the system on these aspects.

Modified Heuristic Evaluation

Study III consisted of the three dual domain experts who, based on validated scenarios and user tasks, performed a modified heuristic evaluation and in-depth severity factor rating. This to determine if more critical usability issues for users could be found. The
data collection and evaluation procedure began with the dual domain evaluators familiarizing themselves with the mHealth system and its usage. To ensure consistency among them as well as all experts having the same knowledge about functionality and user tasks, this also entailed being provided with instructional materials for the system in an identical format.

The information materials contained the different system modules shown by a digital video recording with instructions on how the portal was to be navigated, a manual for the study describing each specific scenario and the different tasks to be performed in the evaluation, an evaluation sheet as well as the user manual for the application.

Each dual domain expert evaluator then performed the eight different scenarios and tasks in the system independently.

**Modified User-Centered CW, and User Test with TA, Demographic and Computer/IT and Mobile Technology Experience and Perception Questionnaire, NASA RTLX and Method Acceptability Assessment**

The data collection for the UC-CW, and TA in study IV also consisted of two short pre-test questionnaires; the first on demographics such as gender, age, ethnicity, education, occupation, diabetes diagnosis and length of diagnosis. Participants also filled out a second on self-assessed computer/IT and mobile technology knowledge, experience and perceptions. Questions covered how often participants used a computer, mobile phone and the Internet, how often they used their mobile phone for calls, text messaging, using apps, and how they considered their level of computer-/IT-knowledge, and mobile phone knowledge. Participants also had to answer on a 5 point Likert scale (Strongly agree to Strongly disagree) if they enjoyed using the computer, mobile phone in their daily life, and if they considered eHealth and mobile health services useful in general and for themselves in their health care.

For each of the usability tests, participants performed the required tasks according to each specific method procedure (described previously) which was also digitally audio and video recorded through Morae and a video recording device in the case of the UC-CW session. Finally participants also had to fill in the NASA RTLX instrument consisting of 6 component scales and answer a short set of 4 questions individually to assess method acceptability for each performed method.

The NASA RTLX instrument developed by Byers et al (1989) [123] is a short and simplified form of the original NASA-Task Load Index created by Hart and Staveland (1988) [124]. It consists of six component scales or dimensions assessing participants on: Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration. Each of these reflect the contribution to the workload of a specific activity from the user’s perspective (with a total score from 0-100). For study IV the instrument was used to determine how easy or difficult the participants experienced the particular method they were a part of. In the original form these factors are also weighted by letting the individual subjects compare them pairwise based on their perceived importance.
This step is removed in the shorter NASA RTLX which is a version that has also received high experimental validity [125]. To get the overall workload of participants with the NASA RTLX, the six scale ratings are combined into an overall score. The higher the average, the higher the overall workload or cognitive load for the individual [124].

As a final part in data collection for study IV participants were asked how they individually experienced each method by replying to 4 questions in a short interview. These were: 1. What are your thoughts about the method you have just taken part in? 2. Was the method easy or hard to understand, in what way? 3. Was it easy or hard to perform the different parts of the method? 4. What is your overall experience of using/being a part of this method?

Data Analysis

Task Coding and Analysis, SUS Analysis and Comparison of Performance Measures to Grouped User Characteristics

After the usability evaluation in study I, the Morae recordings were carefully coded using specific metrics on the success rate for tasks, task errors, task times as well as satisfaction scores to adhere to the specifics of the ISO standard.

As stated previously, effectiveness was measured as to what extent the task had been completed and the amount of errors that the user performed per task. The task completion itself utilized three different coding determinations which consisted of (1) completed with ease - when the user performed the task without any assistance from the test administrator M.G., (2) completed with difficulty - when the task was achieved by the user but with minor difficulties and some hints were needed from the test administrator and (3) failed to complete - when the user was not able to complete the task, despite of having received some minor hints from the test administrator. An error was coded for the task if the user committed unsolvable errors or those types of errors that prevented further progression. Efficiency was measured by timing each individual task performed and then averaging the task time for the different patients. The timing process for the different tasks started when the user began performing it and was completed when the home button for the system was pressed. Deductions in time occurred if there was prolonged system loading and response time. For the SUS or satisfaction measurement, the data analysis was performed according to the instrument requirements [61, 62]. This consisted of that for all odd numbered items (1,3,5,7 and 9), a point was subtracted from the score the participant received on that item and for all even numbered ones (item 2, 4, 6, 8 and 10), five points were subtracted. To get the overall satisfaction value for the mHealth system as a whole the sum of all resulting item scores was then multiplied by 2.5. The SUS scores were then summarized and averaged across patients.

The effectiveness, efficiency and satisfaction results from the ISO standard were, as a final step, analyzed and compared against the grouped user characteristics from the
questionnaires. This was done to determine how the performance and user characteristics for different groups related to each other and to gain insight about the difficulties/ease users experienced compared to their characteristics. This part of the analysis highlighted aspects such as what tasks the users had difficulties with and also needed to be revised, as well as which characteristics could have to do with task performance.

**Multi-method Usability Problem Coding and Classification Utilizing Framework Analysis (FA) and the Usability Problem Taxonomy (UPT)**

For the coding of the multiple data collection sources and to assist in streamlining and structuring this process, two methods were combined to determine and define usability problems in study II; the Framework Analysis (FA) method as well as the Usability Problem Taxonomy (UPT) classification.

The FA method originates from Richie and Spencer in 1994 and is a standardized but also flexible framework for analyzing large amounts of qualitative data and also to aid the analysis of descriptive source data with the intent to produce qualitative results that are reliable as well as valid [126, 127]. It is a scientifically robust method [128] and has primarily been utilized in the social science research area but to a lesser extent in health care research [128, 129]. It consists of five different steps to analyze the data which are: to familiarize oneself with the data, identifying the thematic framework to be utilized in the analysis, indexing and applying the framework to the data, charting the data and as a final step mapping and interpreting the data [126]. Some of its advantages are its usefulness when it comes to data reduction, organization and interpretation. Another, that it also provides summarized data with a more structured output [129].

The UPT is a framework as well as a classification scheme, developed by Keenan et al [130] with the intent of being able to characterize usability problems on their different dimensional aspects allowing for clearly structured and defined usability problems. It was built on an empirical basis with over 400 described usability problems from real-world development projects [130] and has been used in the human-computer interaction field to structure and also classify the usability problems that can be encountered on graphical user interfaces including textual components [130]. With this method, problems can be classified on an individual problem level, also allowing them to be compared, analyzed and described to aid in problem distinguishability and correction [130, 131]. Classification occurs from two perspectives, the **task artifact** and the **task component** perspective. The artifact component defines the usability problems that arise upon user interaction with the interface and the task-component through the user movement through a task. The artifact component consists of three primary categories: **Visualness**, **Language** and **Manipulation** with associated subcategories and the task component of two primary categories: **Task-mapping** and **Task-facilitation** and associated subcategories [130] (see Figure 6).
Figure 6 Classification of categories according to the Usability Problem Taxonomy [130]

The data analysis process in study II included content analysis to code usability problems as well as applying both the FA and UPT. It started with that all the resulting data; the audio and video recordings and observations from the Think Aloud, as well as the post-test interview and questionnaire data on usability were transcribed, checked and imported into Nvivo 10 Qualitative Data Analysis Software.

The coding and analysis process itself consisted of applying the five FA method analysis steps [126] of data familiarization, identifying the themes and framework used and classifying usability problems using the UPT. Researcher M.G. performed the transcription, the importing of data and the initial coding of the data which was then also verified by his co-author N.S.. M.G. and N.S. then performed steps 2-5 in the analysis process together (see Figure 7).

These steps consisted of:

1) **Familiarization with the data.** This involved becoming accustomed with the transcribed data and immersing oneself in it which included reading it repeatedly.

2) **Identification of the themes and/or framework to be used.** This step consisted of identifying the themes and apply a suitable framework for the coding. In the original FA method this can either entail an inductive or deductive content analysis. Here the FA method was modified slightly and used to inductively determine usability problems and the UPT was used deductively to classify the problems across the different data collection methods. The usability problems that emerged could therefore also be designated as themes generated from the data. After the subsequent discussion and removal of duplicates in the data material, the usability issues were consolidated under their appropriate tasks.
3) **Indexation and application of the classification to the data.** The next step involved the indexation and application of the UPT classification framework and also to assign severity ratings to the final usability problem list. Each usability problem was, as described previously, classified according to the applicable component for the artifact (Visualness, Language or Manipulation) and/or suitable component for the task (Task-mapping and Task-facilitation). These categories as well as subcategories within both the artifact and task components are mutually exclusive which leads to one final categorization which can be a full classification (FC), a partial classification (PC) or no classification (NC) [130].

For the severity rating of the problems, Travis’ process was used consisting of three questions to ask about each usability problem. These are: (1) Does the problem occur on a red route - is it a frequent or critical task that the system needs to support? (2) Is the problem difficult for users to overcome? and 3) Is the problem persistent and does it keep on recurring [132]? The applicable severity ratings were determined through the scale (1) low, (2) medium, (3) serious or (4) critical and assigned to each individual problem [132]. The severity scores were then averaged per system view and also for the system as a whole.

4) **Charting of the data.** This step consisted of abstracting the resulting list of the usability problems that had been defined into their original context consistent with the FA method. This list was then charted based on the problem’s place of occurrence in the mHealth application, classification, as well as level of severity. To summarize the found issues for each method and patient, descriptive statistics were used.

5) **Mapping as well as interpretation of the data.** The last step in this data analysis process involved both mapping and also interpretation of the usability problems. Subsequently, after the final list of problems was charted, it was possible to determine the problems that were most prevalent as well as their different severity ratings for each of their respective system views. This facilitated the determination of the nature of the problems as well as what their classifications signified. The goal was for this list to serve as a guidance for designers when correcting the specific usability issues.
**Modified Heuristic Evaluation, In-depth Severity Rating of Usability Problems Based on the Factors Frequency, Impact and Persistence**

After each of the evaluators detected a usability problem in the mHealth system in study III, each of them assigned the problem with a heuristic violation or violations (if they felt that several were violated) from the list of Nielsen’s 10 heuristics. When all of them had completed this process, a master list was compiled, where the duplicate problems were also removed. This list was finally verified across all three evaluators for accuracy before being sent out for the specific in-depth severity rating process used.

For each individual problem in this master list, the evaluators determined the problem severity based on the frequency, impact and persistence severity rating factors [64]. The

---

**Figure 7 Application of the FA method and UPT**
The aim was to determine how each of these different factors for each problem would influence the user. This modified heuristic evaluation process is shown in Figure 8.

Separate averages were determined for each severity factor rating to allow for a greater problem specificity when it comes to its severity and the impact the problem could have on the diabetes patient user. The severity ratings consisted of ratings from 0 (not a usability problem) to 4 (usability catastrophe) [64]. These severity rating scales and severity factors are listed in Table 6.

After the evaluators had conducted their individual factor severity ratings according to the process described, all ratings were totaled and also divided by the number of evaluators. This was done to determine the average severity rating of each usability problem. This rating was then considered to be the overall severity rating for the issue.

**Table 6 Scale for severity rating and severity factors [64]**

<table>
<thead>
<tr>
<th>Severity rating</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 Not a usability problem at all</td>
<td></td>
</tr>
<tr>
<td>1 Cosmetic problem only</td>
<td>Need not to be fixed unless extra time is available</td>
</tr>
<tr>
<td>2 Minor usability problem</td>
<td>Fixing this should be given low priority</td>
</tr>
<tr>
<td>3 Major usability problem</td>
<td>Important to fix. Should be given high priority</td>
</tr>
<tr>
<td>4 Usability catastrophe</td>
<td>Imperative to fix this before product can be released</td>
</tr>
</tbody>
</table>

**Severity factors**

The *frequency* with which the problem occurs: Is it common or rare?

The *impact* of the problem if it occurs: Will it be easy or difficult for the users to overcome?

The *persistence* of the problem: Is it a one-time problem that users can overcome once they know about it, or will users repeatedly be bothered by the problem?
**Modified, User-Centered Cognitive Walkthrough and User Test with Think Aloud Usability Problem Coding and Analysis, Efficiency Determination and NASA RTLX and Interview Comparison**

The first part in the coding and analysis phase in study IV consisted of that two of the dual domain authors M.G. and N.S, for consistency as well as comparability between the methods, imported the recorded data from each method into a qualitative data analysis software (QDAS) program; in this case Nvivo 10. Then for both the UC-CW and TA the usability problems were coded through an inductive coding process [133]. The next step consisted of the analysis process where duplicates were removed and final lists of uniquely identified problems for each method were produced. At the end, each usability problem was assigned a heuristic violation/s based on Nielsen list of 10 heuristic categories [64] and severity ratings [64]. Author A.K. verified the coded and rated problems.

Just as with the previous methods, the demographics, Computer/IT and or mobile phone knowledge, experience and perceptions were summarized based on participants’ answers using descriptive statistics. For this study, the method effectiveness or the resulting usability problems, number and types of heuristics and severity ratings were determined (as described earlier). Also, the method efficiency (by determining the total method time consumption, time for the whole evaluation, as well as the evaluation process, and per identified problem). User acceptance was also determined through the resulting NASA RTLX scores, through descriptive statistics, and interviews coded into themes by M.G. through an inductive coding process by Hsieh and Shannon [133] with resulting themes verified by author N.S. All these measures were compared between the methods to determine the modified method’s feasibility and to validate it.

**Ethical Considerations**

Institutional Review Board (IRB) approval was received from the Western Institutional Review Board (WIRB), Olympia, Washington, USA for study I and II (HI-BEA-001#2011605) with individual written informed consent by each patient at Health Insight, Salt Lake City, Utah, USA. Study III did not involve any patients and therefore did not require ethical approval. IRB approval was also received for study IV (IRB_00101636) from the University of Utah Institutional Review Board, Salt Lake City, Utah, USA with written informed consent from each individual patient at the Diabetes and Endocrinology Center at University of Utah Health, Salt Lake City, Utah, USA.
Results

The summarized findings for each study are presented below. Study I includes patients’ characteristics and different usability evaluation tasks and metric results assessed through the ISO 9241-11 standard, as well as, the subsequent conclusions of comparisons between these. Study II and IV report on the included patient participants and the usability problems resulting from the structured multi-method approach and accompanying analyses as well as specific usability tests. The resulting usability problems are also reported for the modified heuristic expert evaluation in study III. As studies II, III and IV also focus on critical issues that have been coded and found based on severity ratings these are also specifically addressed.

Study I

The ISO 9241-11 was used as a main assessment method in this study to attempt to accomplish a structured and streamlined assessment (that could be easily reproducible) as well as providing an encompassing result by assessing task effectiveness, efficiency as well as satisfaction together. To incorporate a user-centered perspective and taking into account different patient users with different disease lengths, educational backgrounds and experience levels when it comes to IT/computers, the study also included participants of different backgrounds and characteristics to gain deeper insights into how these compared to task performance and satisfaction.

Participant Characteristics

Of the 10 included participants in this study six were female and four male. In this sample most, or 70%, were older adults between the ages of 50 to 69 years and the rest between 30-49 years old. Most were either college or university educated (80%). A majority of them, or sixty percent, had received a Type 2 diabetes diagnosis five years ago or more. In terms of their knowledge and use of IT and computers as well as preferences, half considered their IT knowledge level to be medium and 30% high. The greater majority of this sample, or 80% of the patients, stated that they used a computer and the Internet daily and this percentage of patients also either agreed or strongly agreed with enjoying to use computers in their work or during leisure.

Participant Results on Task Effectiveness, Efficiency and Satisfaction

When examining the task performance for these users, it was evident that of all the eight Tasks assessed (described previously) Task 3 and 4 were the most difficult for all participants to complete. These tasks consisted of correcting a glucose measurement value and of exporting a glucose measurement value. For both of these, task effectiveness measures indicated a 30% and 40% failure rate. This was a similar result with regards to the errors committed on these two tasks which reflected the task completion difficulty with 9 and 13 errors respectively. Upon examination of the kinds of errors committed it was evident that patients had particular difficulties with remembering the different steps, finding the correct options and how to perform the process steps. An in-depth
analysis further revealed that errors were particularly prevalent when it came to tasks involving: locating glucose values for correction, exporting data, remembering how values were deleted, selecting values and navigating to the right screens to accomplish this, and also remembering in what way and where to be able to add or export glucose values. Tasks 1, 2 and 5 which entailed sending in a measurement value, interpreting a glucose measurement value and interpreting a blood pressure measurement value were completed with ease by all participants and with no errors (see Figure 9).

**Figure 9 Task success and error rate by task**

In terms of *efficiency*, Tasks 3 and 4 also took the longest amount of time for participants to complete as might be expected due to the difficulties described above. In fact, it was evident that the tasks consisting of correcting and exporting values contained mean scores that were almost 2–3 times longer than those related to interpreting values such as in Tasks 2 and 5 for example (see Table 7).

**Table 7 Time on task per task**

<table>
<thead>
<tr>
<th></th>
<th>Task 1</th>
<th>Task 2</th>
<th>Task 3</th>
<th>Task 4</th>
<th>Task 5</th>
<th>Task 6</th>
<th>Task 7</th>
<th>Task 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>1.51</td>
<td>1.45</td>
<td>4.18</td>
<td>3.67</td>
<td>1.33</td>
<td>2.21</td>
<td>1.75</td>
<td>1.69</td>
</tr>
<tr>
<td>SD</td>
<td>0.48</td>
<td>0.45</td>
<td>1.55</td>
<td>1.71</td>
<td>0.37</td>
<td>0.72</td>
<td>0.56</td>
<td>0.57</td>
</tr>
<tr>
<td>Range</td>
<td>0.97–2.41</td>
<td>0.91–2.50</td>
<td>1.66–6.47</td>
<td>1.35–6.86</td>
<td>0.80–2.01</td>
<td>1.33–3.27</td>
<td>0.91–2.72</td>
<td>0.76–2.61</td>
</tr>
</tbody>
</table>
When it comes to the satisfaction scores, the average SUS score was 80.5 (SD 11.47) across users indicating good satisfaction. Despite this, there was a wide variation in scores with the lowest percentage scores at 62.5 and highest scores of 97.5 (a total range of 35 points). Thirty percent of the patient sample had a high score from 87.5 to 97.5 or excellent, and another 30% of the sample had the lowest scores from 62.5 to 72.5 considered "OK" to minimally acceptable [63].

**User Characteristics and Usability Metric Results**
The user characteristics in this study were also compared to the usability metric data for an insight into their effect on user performance. When examining the descriptive trends in the data, there were indications of differences among participants when it came to gender, age, as well as diabetes diagnosis, IT/Computer knowledge and experience. For example, it was observed that males in this study generally had higher average success rates, lower rates when it came to errors and higher mean scores for SUS than the female participants. Younger patients in this sample also had higher average completion rates when it came to tasks, only an average of one error, and also lower completion times for tasks and higher mean scores when it came to satisfaction.

Another observation was that patients with a diabetes diagnosis that was more recent (0 to 4 years) had a higher average success rate on tasks, fewer task errors, an average time that was shorter on tasks as well as a higher SUS score average. An educational level involving high school or college/university did not seem to have an influence on either performance or satisfaction scores. In addition, it was visible that the users that were more experienced with IT/computers were able to perform the tasks faster and without as many errors as those who were less experienced. Their average SUS scores were also higher compared to those users who were less experienced.

**Study II**
A structured and streamlined approach of performing a multi-method usability evaluation and find the usability problems in the system that were critical to patient users was also attempted in study II. Here the problems were also determined and analysed according to the steps of the Framework analysis (FA) process for added structure and classified according to the Usability Problem Taxonomy (UPT) for added problem definition.

**Participant Characteristics**
The 10 patients in study II also consisted of six women and four men. In this sample most of them were between 40 and 59 years old (with a range of 40-69). The majority had a university degree and 50% were employed. Sixty percent were diagnosed with Type 2 diabetes, and had had this diagnosis for five or more years in length. These participants were frequent users of technology. For instance, a large number used their computer daily (80%), the Internet daily (70%) as well as their mobile phone every day (90%) to either make or receive calls. All either strongly agreed (80%) or agreed (20%) with that
IT/Computers and mobile phone use in health care is a positive development and also thought this was advantageous for their own diabetes self-management needs.

**A Multi-Method Approach for Determining Usability Problems**

The three different evaluation methods; consisting of a usability test with Think Aloud, in-depth interview and questionnaire each contributed to 117 initial usability problems which in turn could be put into a list consisting of 19 consolidated usability issues. Overall, the usability test found the largest number (59) or 50% of the problems. Then came the in-depth interview which detected 34 of the usability problems. Finally, the questionnaire identified 24 usability problems.

In addition to being the method that was able to discover most of the usability problems, the usability test was also the method that detected most of the problems on its own as well as in combination with the other methods. For example, together the usability test and the in-depth interview detected 93 problems or 80%. The usability test and post-test questionnaire, on the other hand, found 83 or 71% of the problems combined. The in-depth interview and questionnaire detected 58 or 50% of the usability problems. Of all the included methods, however, the post-test questionnaire found the smallest amount of problems on its own as well as in combination with the other methods.

This result was also reflected in the 19 consolidated and unique usability problems. Here, the usability test found eight of these problems on its own and combined with the in-depth interview identified 18 and 9 of the resulting usability problems. All three methods in this study converged on five usability problems (see Figure 10).
Critical and Severe Usability Problems

In terms of problem location and problem severity, the highest number of usability problems (eight consolidated problems) and highest average severity ratings, with a severity rating of three, could be located in the Glucose Readings View. This view was followed by the Glucose Diary View and Dashboard with three problems each. The Glucose Diary View, Glucose Readings View as well as the Blood Pressure View also all had average severity ratings of three. Upon examining the problems, it was evident that those views with the highest problem numbers for participants were also those consisting of several steps in the task. In many cases, these were also vitally important diabetes management functions and received high severity ratings (see Figure 11).
Moreover, the found usability problems could be organized according to the most critical issues depending on the number of methods they were detected in. The two usability problems that were most severe of all (severity level 4) were present in both the Glucose Readings View as well as the Glucose Diary View. These problems required to first correct a glucose value through the removal of the erroneous value and then to add a new one. To perform these steps, patients needed to exit the Glucose Readings View, then navigate to the Glucose Diary View and after that add the correct value through the Add data button which was a confusing sequence to patients. They were also confused about the Delete data button’s dual function of both allowing the exporting and printing of data in the Glucose Diary View which was a name label that was rather unfitting in their opinion.

Other serious concerns (with severity level 3) were problems that were present in the Glucose Readings View when it came to exporting, which involved several hard, and difficult to understand sequences of steps. The exporting logic in this view was not clear to the users, and no support from the system was available for this process. If patients, for example, wanted to perform the exporting of a file, they were able to enter a file name to save, but it was not possible to specify where to save the file. Difficulties were also experienced for patients when it came to adjusting the value ranges in the Glucose Diary View and its accompanying graph. Another example involved difficulties with the navigation of the table located in the Glucose Readings View (that displayed the correct value ranges or values). Here, no navigation support was available to assist them. Another notable concern was the Blood Pressure View where it was only the systolic blood pressure that was distinguishable in the graph even if patients had also put in their diastolic blood pressure into the system. This display could be considered both incomplete as well as potentially harmful.

Figure 11 Distribution of usability problems and severity ratings across system views
Other difficulties with the system involved the various ways it was possible to access the reminders in the Dashboard View which included several different types of labels as well as multiple locations. The Medication Reminder, for instance, could be be reached through the Medication Adherence tab, the Message Settings and Medication Reminders pane as well as the Exercise, Weight and Medicine pane. The Appointment Reminder included no separate tab, as the others, but was instead listed under one of the menu items called Message Settings and Reminders. The exercise and weight “Tracking Goal” also had three different paths to access the item, such as for example, the Exercise and Weight Progress tab, the Exercise, Weight and Medicine pane and the Message Settings and Reminders pane (see Table 8).

Table 8 Usability problems and classifications for the most severe issues

<table>
<thead>
<tr>
<th>Usability Problem Description</th>
<th>Place of occurrence</th>
<th>% of pat. detecting per method</th>
<th>UPT Classification</th>
<th>Severity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficult knowing to exit the table view and click the “Add data” button to be able to add in a new value</td>
<td>Glucose Readings View</td>
<td>70 100 50</td>
<td>Manipulation- Cognitive aspects-Visual cues (FC)</td>
<td>4</td>
</tr>
<tr>
<td>Difficult to know to select “Delete data” for the table view to change, delete, export, or print the value list</td>
<td>Glucose Diary View</td>
<td>70 100 100</td>
<td>Language- Naming/labeling (FC)</td>
<td>4</td>
</tr>
<tr>
<td>Difficult to understand and perform the exporting action</td>
<td>Glucose Readings View</td>
<td>30 40 40</td>
<td>Manipulation- Cognitive aspects-Visual cues (FC)</td>
<td>3</td>
</tr>
<tr>
<td>The blood pressure graph shows only the systolic blood pressure value, and the diastolic blood pressure is missing</td>
<td>Blood Pressure View</td>
<td>10 20 -</td>
<td>Visualness- Presentation of Information/ results (FC)</td>
<td>3</td>
</tr>
<tr>
<td>Difficult to know how to adjust the range of values that</td>
<td>Glucose Diary View</td>
<td>90 - -</td>
<td>Manipulation- Cognitive aspects- Visual cues (FC)</td>
<td>3</td>
</tr>
</tbody>
</table>
are to be shown, retrieved

| No action support to specify where the exported file is to be saved | Glucose Readings View | 20 | - | Visualness-Non message feedback (FC) | Task mapping-Functionality (FC) | 3 |
| Difficult to know how to navigate within and adjust the table view to get to or show the right value range, value | Glucose Readings View | 10 | - | Visualness-Presentation of Information/results (FC) | Task mapping-Navigation (FC) | 3 |
| A confirmation is missing to indicate that the exported file was saved and its directory | Glucose Readings View | 10 | - | Visualness-Non-message feedback (FC) | Task facilitation-Keeping the user on track (FC) | 3 |

* TA = Think Aloud usability test, I = Interview, Q = Questionnaire

**Usability Problem Classifications**

In terms of the distribution for each UPT classification, all problems could be classified into either an artifact and/or task component. The usability problem classifications for the task component involved a number of different functional or navigational concerns as a part of Task-Mapping while the problems for the artifact component were mainly Language, Naming/labeling related or concerned Manipulation and Visualness. All of these types of issues signify that the majority of the usability problems were either connected to the design or structure of the system as a whole which was not particularly logical from the user’s point of view. Two problems categorized in the artifact component as well as five in the task component of the list of 19 problems couldn’t be classified. In the artifact component Visualness received the highest number of full classifications (7) while in the task component, Task-mapping received the highest amount (12) of fully classified problems (see Figure 12). Most of the subcategories for the system related to the presentation of information/result.
When it comes to the distribution of the more severe usability problems for each of the UPT classification levels, these usability problems could be found in the artifact components of Visualness, Manipulation, Cognitive aspects/Visual cues, Non-message feedback/Presentation of information/results. The two most critical problems of them all were in the artifact component Manipulation, Cognitive aspects/Visual cues and Naming/labeling deficiencies and task component Task-mapping/Navigation. As stated previously, these were also those issues that were brought up most frequently by patients across all the different data collection methods (also seen in Table 8).

**Study III**

The resulting usability problems, this time from the modified heuristic expert evaluation, were also specifically addressed in study III which also focused on critical issues that were coded and found based on severity ratings. To accomplish a streamlined approach and a user perspective, all expert evaluators followed specific steps and provided severity ratings based on how each specific problem would influence the user of the system based on the frequency, impact and persistence of each individual problem on users.

**Expert Characteristics**

The three experts performing the modified heuristic evaluation in study III were carefully chosen based on their dual-domain competency of extensive usability experience in health informatics, and being a health care professional (in this case registered nurses, RNs) and with experience with the patient group and their requirements, and diabetes self-management.
**Usability Problems**

The heuristic evaluation resulted in 129 usability problems as well as 274 heuristic violations. The number of usability problems for the different application views were between 12 and 34 problems. The *Dashboard View* generated the majority of usability problems (with 34). Next came the *Glucose Diary View* (with 21), the *Blood pressure View* (with 20), and the *Medication Adherence View* (with 15). The heuristic violations were from 25 to 69. The highest amount of heuristic violations were located in the *Dashboard View* (with 69), the *Glucose Diary View* (with 49), the *Blood Pressure View* (with 44), the *Medication Adherence View* (with 31), and the *Appointment Reminder View* (with 29). The average severity ratings for all views had a range between 2.7 to 3 on a scale of 0 to 4 where both the *Glucose Diary* and *Medication Adherence View* each had the highest severity ratings of 2.9 and 3.0. (see Figure 13).

![Figure 13 Usability problems, heuristic violations and averaged severity by system view](image)

**Heuristic Violations**

Of the 10 different types of violations, the heuristic categories Consistency and Standards (at 24.1%, n=66) and Match Between the System and the Real World (at 22.3 %, n=61) received the highest number of violations for all the views. These were followed by the categories Aesthetic and Minimalist Design (at 16.8 %, n= 46) as well as Recognition Rather than Recall (at 11.7 %, n=32). The categories Recover (at 1.4%, n=4) as well as Help (at 1.03% n=3) had the least violations for all views (see Figure 14).
Severity Ratings Across System Views

The problems that were most severely rated and of both major and catastrophic severity were in the Dashboard View (with n=16, n=4), the Glucose Diary View (with n=12, n=2), Blood Pressure View (with n=12, n=2) and Medication Adherence View (with n=6, n=4). A large number of the minor rated problems were also in the Dashboard View and Glucose Diary View (with n=14 and n=7). After these came the Appointment Reminder View with six problems. No cosmetic violations were present (see Figure 15).
Nature of Usability Problems and Prioritization

The modified HE evaluation uncovered that most of the catastrophic severity ratings for problems concerned different disease-related information deficiencies as well as specific system-related shortcomings regarding the display of necessary information for patients when it comes to their disease management. Some examples of the types of usability problems and the comments provided by the expert evaluators were:

Dashboard

“Each entry should, at minimum, have the time (not just the date) since many people with diabetes will do multiple glucose tests in one day.” (Total severity rating = 3.8, factor rating = 4 by all three evaluators for frequency as well as persistence).

Glucose Diary View

“It is very difficult to read the time line on the graph because the numbers are too crowded which makes it difficult to distinguish and read specific dates. This is especially cumbersome for diabetes patients who often have visual concerns.” (Total severity rating = 4.0, factor rating = 4 by all evaluators on all factors: frequency, impact and persistence).

Blood Pressure View

“It is disadvantageous for the patient to not see the diastolic blood pressure reading in the graph to compare against; only the systolic value is shown,” “when I hover I only can see the systolic value. If the whole BP is displayed and rated, what happens if only one value is abnormal?” (Total severity rating = 3.9, factor rating = 4 by all evaluators for frequency and persistence).

Medication Adherence View

“With the indications...in percentages, it is difficult for a patient or health care provider to determine what medication was taken or not, which day and what time,” “..Tallying up totals to say 100% of medications is an odd way to think about medication from a patient perspective. I wouldn’t say that I have taken 75% of my meds for the last month for instance. I need to know specifics and insulin or Metformin and that these are jointly displayed and tracked,” “..I found that this display did not match my mental model of medication compliance, need individual information regarding medication..” (Total severity rating = 3.7, factor rating = 4 by two evaluators on each factor for frequency, impact and persistence).
Study IV

Study IV consisted of testing a developed, user-centered CW (UC-CW) method also described elsewhere [134]. In short the modified method attempted to address method shortcomings of the original CW method. These modifications included: making the user the main evaluation contributor in finding the usability problems, a dual domain facilitator with the role of guiding the evaluation session and rating found usability problems, as well as a task development process consisting of validated tasks, a focus on higher level tasks in the evaluation and streamlining the evaluation process in terms of time and resources. CW was selected due to its group-based component and potential in including user groups.

The UC-CW was assessed on its effectiveness, efficiency and acceptance (see Table 9) on 6 patients compared to and validated against the “golden-standard” user test with Think Aloud on 6 patients. Apart from the method testing procedures, participants had to fill in a pre-test questionnaire with demographic questions, and their experience and perceptions of using IT, computers and mobile technology. The post-test questionnaire consisted in filling out a method experience instrument (NASA RTLX) on 6 component scales measuring cognitive work load and a brief set of 4 interview questions asking users how they each experienced the method they were a part of.

Table 9 Effectiveness, efficiency and acceptance of the UC-CW and TA method

<table>
<thead>
<tr>
<th>Effectiveness</th>
<th>Number and type of usability problems, heuristic determination, severity ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Efficiency</td>
<td>Total time for whole evaluation procedure, total time for evaluation process part, and time per detected usability problem</td>
</tr>
<tr>
<td>Acceptance</td>
<td>User acceptance through NASA RTLX (measuring method cognitive load), in-depth interviews</td>
</tr>
</tbody>
</table>

User Characteristics (UC-CW) and User Test with Think Aloud (TA)

Patient participants for the UC-CW evaluation session included two women and 4 men, between the ages of 25 and 75 years old (mean age of 48). One was African-American and five Caucasian. All had college or university as their highest educational level. Three were employed, two retired and one a student. Three participants had Type 1 and three Type 2 diabetes. Length of diabetes diagnosis varied from 3 up to 27 years (mean 14.8).

The individuals for the TA included 3 women and 3 men between the ages of 35 to 69 (mean age of 52). One was Latin-American, one African-American and 4 Caucasian. Four had college or university and two high school as their highest educational level. Two were retired and four employed. Three participants had Type 1 and three Type 2 diabetes. They had had their diagnosis between 1 and 33 years (mean 18.8).
IT, Computer, mHealth Experience and Use (UC-CW and TA)

When it comes to computer use for the UC-CW group, five participants used it every day and one 3-6 times a week. Two considered themselves to have high computer/IT knowledge and the rest medium knowledge. All participants used the Internet every day. Three considered their Internet knowledge to be high, two considered it to be medium and one small. Five participants used their mobile phone every day and one 3-6 times a week. Four participants used their mobile phone to receive and make phone calls on a daily basis, while two did this 3-6 days a week. Five used their mobile phone for text messaging every day, and one 3-6 times a week. When it came to using their mobile phone for e-mails, surfing and apps four expressed that they did this every day, one that they did this 1-2 times a week, and one never. When it came to mobile phone knowledge level two considered that they had high knowledge, three medium knowledge, and one small.

Computer use varied among the participants in the user test with TA group; three used a computer every day and three 3-6 times a week. One considered their computer and IT knowledge to be high, two felt it was medium, and three considered it small. Of the 6 participants three used the Internet every day, and three 3-6 times a week. One considered their Internet knowledge to be high, three considered it medium, and two small. Five used their mobile phone every day, and one 3-6 times a week. Five also used their mobile phone for making and receiving phone calls every day, while one did this 3-6 times a week. Five of the participants also used their mobile phone for text messaging every day, and one did this 3-6 times a week. When it comes to using their mobile phone for e-mails, surfing and using apps four stated that they did this every day, one that they did this 3-6 times a week, and one never. Regarding their mobile phone knowledge level, two considered it high, two medium, and two small.

Perceptions About eHealth and mHealth (UC-CW and TA)

Regarding questions about if they liked to use the computer in their daily life three participants strongly agreed, and the rest agreed in the UC-CW group. Three also expressed that they strongly agreed to liking to use their mobile phone in their daily life, while two agreed and one was neutral. Five participants strongly agreed that they thought eHealth and mobile health services within health care getting more common was a positive development, while one was neutral. On the statement if they could see advantages from themselves personally in using eHealth and mobile health services three strongly agreed, two agreed and one was neutral.

When it came to questions about if they liked to use the computer in their daily life in the TA group one strongly agreed, two agreed and three were neutral. Two strongly agreed to the statement if they liked to use their mobile phone in their daily life, while three agreed and one was neutral. Two participants strongly agreed that they felt eHealth services, mobile services and patient support services within health care being common was a positive development. Four agreed with this statement. Two participants strongly agreed to seeing advantages for themselves personally in using eHealth services and mobile health services, while one agreed and three were neutral.
Method Effectiveness

Usability Problem Numbers, Heuristics and Severity Ratings
When comparing the UC-CW against the TA on the number of problems, it was possible to see that it detected more or 26 compared to 20 problems. The UC-CW also had more problems and heuristic violations per view but both methods were similar on the views where this occurred. This included the Main View where the UC-CW had 11 problems and 29 heuristic violations and the TA 10 problems and 24 heuristic violations. For both methods, the Carbohydrate Entry View came next with the UC-CW with 4 problems and 14 heuristic violations and the TA with 3 problems and 9 heuristic violations. This view was followed by the List View with 3 problems and 8 heuristic violations for both methods. Both methods also had similar amounts of problems and heuristic violations for the Glucose Entry View with the UC-CW with 1 problem and three heuristic violations and the TA with 1 problem and 2 heuristic violations.

The methods also differed for some views. For example, in the Similar Situation View the UC-CW detected 2 problems and 6 heuristic violations while the TA none. The same for the Periodical Pattern View, where the UC-CW detected 1 problem and 3 heuristic violations while the TA none (see Figure 16).

Figure 16 Comparison of the number of usability problems, number of violated heuristics and average severity per system view for the UC-CW and TA

Another observation was that a total of 74 heuristic violations occurred in the UC-CW and 51 in the TA. Both the UC-CW and TA had the majority of their heuristic violations in the same heuristic categories. These included the category Match Between the System and the Real World at 25.7% (n=19) and 31.4% (n=16), followed by Consistency and
Standards at 22.9% (n=17) and 29.4% (n=15), and Recognition Rather than Recall at 17.6% (n=13) and 19.6% (n=10) (see Figure 17).

Figure 17 Heuristic violations across methods

When it comes to the application views on number of problems and severity rating levels, there were more catastrophic and major severity ratings for the UC-CW with 5 catastrophic and 10 major ratings versus 2 catastrophic and 8 major ratings for the TA. There were, however, more minor ratings in the TA with 10 versus 9 ratings. For the individual views, the UC-CW had the majority of the major problems in the Main View with 2 catastrophic and 2 major problems which was similar to the TA with one catastrophic and 3 major problems. For the List View the UC-CW had one catastrophic and 2 major problems and the TA had one catastrophic, one major and one minor problem. In addition, the UC-CW also had one catastrophic, one major and one minor problem in the Plotted Graph View where the TA only had one minor problem. The UC-CW also had one catastrophic and one major problem in the Similar Situation List View where the TA had none.

Both the UC-CW and TA were also very similar on their average severity ratings for the mHealth application as a whole with the UC-CW with a 2.7 and the TA with a 2.6 average severity rating. Both of these fall into the major severity span. Both methods were also similar on the severity ratings for the List View with a severity rating of 3.3 for the UC-CW and 3 for the TA, on the Time Setting View (3 for each) and Main View at 2.4 and 2.5 and somewhat similar for the Plotted Graph View at 2.3 and 2. They, however, varied in their severity ratings for the Glucose Entry View at 2 versus 3, Carbohydrate Entry View at 3 versus 2.3 and Periodical Pattern Graph View at 2 versus none. The UC-CW also had its highest average severity rating for the Similar Situations List View at 3.5 which was not present in the TA (see Figure 18).
Figure 18 Comparison of the number of detected usability problems, their classification on severity levels and average severity ratings across views for the UC-CW and TA

Usability Problem Characteristics - Most Severe and Critical Problems
With regards to the characteristics of the identified problems it was possible to see, especially when it came to the severe problems, that the UC-CW detected more severe problems than the TA with five problems versus two. The UC-CW also brought up two problems related to difficulties with patients not being able to personalize features according to their own particular disease needs which were not present in the TA. These were, in the Main View, that it on the top of the view was not possible to configure and personalize glucose ranges based on the patients’ own ones for the low, medium and high ranges visualized in the middle pane for the latest glucose measurements. In the Plotted Graph View it was also not possible to configure and personalize range levels (cut off levels) in the graphs showing the glucose values for the low, mid and high ranges visualized in the middle pane. Also, another problem only detected in the UC-CW was in the Similar Situation List View where there were unclarities for participants regarding how to interpret the relation between the number of IUs of insulin and mg/dl blood glucose units in the upper part of the image compared to the bottom part of the view for this function.

Both the UC-CW and the TA detected two critical problems involving navigational difficulties in the Main View and List View. In the Main View this involved that it was unclear for participants to know that it was necessary to swipe across the chosen graphical image on the bottom of the Main View to get to the Periodical Pattern Graph View where they were to check their glucose measurements for the past week, and in the List View participants felt it was unclear how to navigate in and understand where one was located when entering the list of all entries for registered food and glucose, medication and insulin. In addition, it was not possible to get a response on the sequence
of dates in this list which made patients confused (see Table 10 for a description of the most critical problems for both methods).

**Table 10 Problem descriptions, heuristic violations and severity rating per view for top five critical usability problems per method**

<table>
<thead>
<tr>
<th>Usability Problem Description</th>
<th>Place of occurrence</th>
<th>Detected by method</th>
<th>Heuristic Violations</th>
<th>Severity</th>
</tr>
</thead>
<tbody>
<tr>
<td>It is not possible to configure and personalize ranges based on the patients' own ones for the low, medium and high ranges that are visualized in the middle pane for the latest glucose measurements.</td>
<td>Main View</td>
<td>X</td>
<td>2, 7</td>
<td>4</td>
</tr>
<tr>
<td>Unclear to know that it is necessary to swipe across the chosen graphical image on the bottom of the Main View to get to the Periodical Pattern Graph View.</td>
<td>Main View</td>
<td>X</td>
<td>2, 4, 6</td>
<td>4</td>
</tr>
<tr>
<td>It is not possible to configure and personalize the range levels (cut off levels) in the graphs that show the glucose values for the low, mid and high ranges.</td>
<td>Plotted Graph View</td>
<td>X</td>
<td>2, 7</td>
<td>4</td>
</tr>
<tr>
<td>Unclear to know how to navigate in, understand where one is located and get a response on the sequence of dates in the list of all entries.</td>
<td>List View</td>
<td>X</td>
<td>1, 2, 6</td>
<td>4</td>
</tr>
<tr>
<td>Difficult to understand and interpret the relation between the number of IUs of insulin and mg/dl blood glucose units in the upper part of the image compared to the list of entries in the bottom part of the view for this function.</td>
<td>Similar Situation</td>
<td>X</td>
<td>1, 2, 4, 6</td>
<td>4</td>
</tr>
</tbody>
</table>

* UC-CW = User-centered Cognitive Walkthrough, TA = Think Aloud usability test

**Method Efficiency**
The whole evaluation procedure from preparation to final problem determination for the UC-CW included the literature review of diabetes self-management guidelines and construction of the evaluation tasks each taking 480 minutes for a total of 960 minutes. The task validation and content validity indexing process took 130 minutes, and the facilitation of the session 140 minutes. The coding of usability problems 480 minutes, and assignment of heuristic violations and severity ratings 120 minutes. This gave a total of 1830 minutes. Divided upon the 26 usability problems this resulted in 70.4 minutes per usability problem.
In contrast, for the TA task construction took 480 minutes in total, the session facilitator time for leading the sessions 281 minutes, as well as the usability problem coding 724 minutes for the 6 patient evaluation sessions. The heuristic violations and severity ratings took 120 minutes resulting in an overall time of 1605 minutes. When divided upon the 20 problems this gave 80.2 minutes per identified usability problem. This demonstrated that the UC-CW was almost 10 minutes faster per detected problem (see Table 11).

Table 11 Time per method for whole evaluation procedure

<table>
<thead>
<tr>
<th>Usability Evaluation Method</th>
<th>Task lit-review and evidence</th>
<th>Task construction</th>
<th>Task validation</th>
<th>Leading Evaluation Session</th>
<th>Data coding and UP determination</th>
<th>Assigning violations and severity ratings</th>
<th>Total time per method</th>
<th>Total time per identified usability problem</th>
</tr>
</thead>
<tbody>
<tr>
<td>UC-CW*</td>
<td>480</td>
<td>480</td>
<td>130</td>
<td>140</td>
<td>480</td>
<td>120</td>
<td>1830</td>
<td>70.4</td>
</tr>
<tr>
<td>TA*</td>
<td>N/A</td>
<td>480</td>
<td>N/A</td>
<td>281</td>
<td>724</td>
<td>120</td>
<td>1605</td>
<td>80.2</td>
</tr>
</tbody>
</table>

* UC-CW = User-centered Cognitive Walkthrough, TA = Think Aloud usability test

For the evaluation process it was possible to see that the UC-CW added up to a total of 5 minutes for informing about the method, 15 minutes for the demonstration of the application, and the evaluation itself took 120 minutes for the group of six participants resulting in a total of 140 minutes.

In comparison, the TA added up to a total of 281 minutes for all participants. This included the 5 minute information times 6 sessions which equaled 30 minutes total, and 10 to 15 minutes for the 6 sessions totaling 70 minutes (a mean of 11.7, and SD 2.6). The six evaluations sessions totaled 181 minutes; with the slowest participant taking 39 minutes and the fastest 22 minutes (a mean of 30.2, and SD 6.7). This showed that the UC-CW was faster for the evaluation process (see Table 12).

Table 12 Time per evaluation session part

<table>
<thead>
<tr>
<th>Usability Method</th>
<th>Evaluation</th>
<th>Method info (min)</th>
<th>App tutorial (min)</th>
<th>Evaluation Session (min)</th>
<th>Total time per method session</th>
</tr>
</thead>
<tbody>
<tr>
<td>UC-CW*</td>
<td></td>
<td>5</td>
<td>15</td>
<td>120</td>
<td>140</td>
</tr>
<tr>
<td>TA*</td>
<td></td>
<td>30</td>
<td>70</td>
<td>181</td>
<td>281</td>
</tr>
</tbody>
</table>

* UC-CW = User-centered Cognitive Walkthrough, TA = Think Aloud usability test
**Method Acceptance**
The NASA RTLX scores for each method showed a difference in participants’ experienced cognitive load. The overall result for the UC-CW showed a mean value for the participant group of 25.3 or more than half the cognitive load of the TA which had a mean value of 53.5. In addition the standard deviation scores (SD) also varied for both methods with the user-centered CW showing a minimal spread in scores of 3.2 compared to 14.2 for the user test with TA indicating a large spread among the participants.

The experienced cognitive difference was also partly expressed in the individual interviews where participants looked upon the UC-CW favorably in many regards such as that all felt it was easy to understand and perform (6) compared to the participants in the TA where some felt it easy to understand (4) and some felt the method difficult to perform (3), unusual (4) rather awkward (3), and cognitively demanding (4).

The comments expressed with the UC-CW were that patients felt that it was clear and seamless (6), a straightforward method (5) and a good experience (6). In addition they felt it positive to have an individual and collaborative part (5) and they felt it positive, and helpful with the mixed patient group of both Type 1 and 2 diabetes patients (6). Some of the patients also stated that they liked having the image of the mobile phone projected so that everyone could see the image at the same time (4). Some of the negative aspects expressed were that some felt that there was an overlap of questions (2) and some redundancy in participants’ answers and experiences (2).
Discussion

Study Approaches and Contributions

User-Centered Approaches in Usability Evaluation
The need to involve the user in the design process has been brought up extensively in the usability literature [95, 97] and also the user’s vital role in the process of evaluating self-management systems for their disease management so that their individual care needs are accounted for in these systems [94, 135]. One of the main goals of performing the methods and method modifications in the current usability evaluation studies was to find ways of incorporating the user.

Study I consisted of a usability evaluation with a user-centered and standardized approach as an overarching design influenced by the requirements of ISO standard 9241-11. To incorporate a patient and user-centered approach, the study compared the individual user performance measures to specific user characteristics to determine how each was affected by one another. In study II, an in-depth usability assessment was conducted with three data collection methods using two frameworks for data analyses and structuring of data. The multi-method approach had the intent to result in understanding patients’ interactions with the system in a more comprehensive manner and to provide triangulation on severe usability issues for users. In study III, a modified heuristic evaluation was performed with dual domain expert evaluators and user-centered scenarios and tasks to determine usability problems with an in-depth factor rating approach (based on problem frequency, impact, persistence) to determine the severity of each problem to users resulting in critical usability problems that needed correcting. In study IV, the step was taken to develop a modified, user-centered cognitive walkthrough (UC-CW) method incorporating patient-users as some of the common drawbacks with the original CW is that it relies only on experts’ understanding of user capabilities [136] and there is no direct involvement of the concerned user of the system during the evaluation of the system. Despite of the method being effective as a whole it can therefore be the case that usability problems that are found may not always be relevant to the users themselves [137] or can be of minor importance to users [79]. The modified method was also assessed and compared to determine the method’s effectiveness and efficiency in finding relevant usability problems for diabetes patient users as well as on its acceptability to these users.

Method Contributions on Standardized and Streamlined Approaches in Task Development, Performance, Data Collection and Analysis
A review conducted on usability studies on mHealth systems for diabetes showed a considerable variation in usability testing methods both when it came to the use of validated instruments and following standardized procedures [103]. Several recommendations were also provided on that researchers should utilize guidelines and
standards to obtain a more encompassing view of these kinds of applications’ usability [103, 138].

Study I attempted to address these areas in the literature through the performance of a usability study on a mHealth diabetes system assessing effectiveness, efficiency and satisfaction together. This by using the ISO standard 9241-11 and SUS instrument (evaluated for validity, reliability and sensitivity) [61, 62, 139-142] to determine usability metric performance compared to specific patient user characteristics. The results showed where patients experienced most difficulties in the system and where a re-design would be necessary in tasks in order to accomplish a wider usage among different users. The study was meant to serve as an example of how researchers or practitioners could generate benchmark data as well as a way to provide reproducible, and comparable results in their mHealth system evaluations.

Also in study II there was an aim to standardize and streamline approaches and on finding a way to structure data, this time through FA analysis. Particularly in data collection and data analysis involving a multi-method approach the FA method proved very beneficial in structuring and also coding this large amount of data [126]. Despite the need to perform a slight modification of the FA method to accommodate the UPT in usability problem classification, the FA was found very worthwhile and useful. The UPT in this study was also found to be of significant value in determining the classification for each problem and contributed in defining the usability problems on a hierarchical as well as detailed level. The resulting issues could therefore be more well-defined and it was possible to group the problems that were of a similar nature more easily [130]. It was also found that the addition of a problem severity rating could help aggregate similarly severe problems which was beneficial. All these could serve as aids in problem analysis and problem solving.

A standardized evaluation process was also used in study III and allowed for an efficient evaluation for the different experts that could also assist in reproducibility and generalizability. The process containing specific steps to follow in the evaluation process, also with a modified severity factor rating assisted in that all evaluators could think about each individual usability problem’s specific impact on the user and it also provided an in-depth analysis of each usability problem. In addition to the attempt to contribute with an objective method in determining the importance of the usability problems to the users in relation to other problems of a similar kind, this could also aid in problem prioritization.

One of the reasons to develop the user-centered modified CW in study IV was to find a streamlined way of working through the different evaluation phases (making the method efficient in terms of time). This was also the intent when it came to developing the tasks, which needed to be evidence-based but still effective, to serve as a solid foundation to base the evaluation on in detecting usability problems. This due to that the original CW, despite that it is an inspection method that can find numerous usability problems, [143] is cumbersome in the process of defining tasks and in the evaluation process as a whole.
The method modifications that concerned these particular aspects therefore involved streamlining the evaluation process in terms of time and resources, including a task development process consisting of validated tasks, and a focus on higher level tasks in the evaluation with usage of two evaluation questions instead of four. In terms of the UC-CW’s performance compared to the validation method, it turned out to be faster for most aspects except for the total time consumption for the method evaluation procedure as a whole where the user test with TA turned out to be faster. When determining the time taken per resulting usability problem, however, the UC-CW was almost 10 minutes faster than the user test with TA and also faster for the evaluation process itself.

**Methodological triangulation of usability problems, and method validations**

As stated in the background, in addition to emphasize more standardized procedures in evaluation, authors have also pointed to the need for employing different types of usability methods in mHealth evaluations and also using qualitative analysis and open-ended methods of inquiry to detect the more comprehensive concerns for users [102]. This in addition to more rigorous usability studies in mHealth and usage of several different methods to help validate findings [95, 103].

In Study II, in addition to applying the streamlined approaches in data collection and analysis described previously, multiple methods or methodological triangulation was used to detect the usability problems in the system to help with a more comprehensive understanding [144, 145] and to increase the depth of inquiry [146]. This use of multiple methods consisted of a combination of usability tests, in-depth interviews and questionnaires due to their suitability for research in health informatics [74]. Patton was the first who identified this type of triangulation where he advocated using observation, field notes and interviews in data collection [144, 147] to produce rich data [144, 148]. The multi-method approach was therefore seen as beneficial here especially in the identification of the initial usability problems, and also due to that the results from one method could be verified by another [149, 150], contributing to validating the problem [144, 151].

Also in health informatics where systems are becoming more complex, authors have advocated the importance of the use of multiple evaluation methods [60, 152]. In this study it was evident that user testing with TA identified the majority of problems due to that it found 18 of 19 consolidated usability problems. This finding is also consistent other literature on the topic which indicates that user testing not only identifies more usability problems, but it also finds more recurring problems, as well as identifies the underlying causes for the usability problems [78]. The next highest contributor, in terms of usability problems, was the in-depth interview with nine of the consolidated 19 problems. It contributed mainly to the triangulation of identified usability issues and not to any unique consolidated issues. Each usability testing method was, however, found to contribute to the overall number of usability issues in its own unique way [103] allowing for a comprehensive picture of usability challenges [152].
In study IV another important goal for developing the UC-CW, except to attempt to make the evaluation process more streamlined and efficient and user-centered, was adding important validation aspects within it and also validate the whole method. The whole evaluation process was therefore also compared to the TA in terms of its efficiency and also acceptance level, to also establish validity for the method as a whole. Several authors have particularly pointed to the need and importance of establishing reliability and validity in usability studies and to improve standardization and transparency [101] in addition to taking particular effort in validation against existing clinical assessments for these kinds of self-management applications [94]. For mHealth applications for diabetes, authors also emphasize more adherence to evidence-based guidelines [153, 154]. These various points were some of the aspects that the researcher also attempted to include in this method modification - both in terms of validating the method against a “golden-standard” method - as well as using particular validation elements, such as diabetes self-management guidelines in the task development and validation process in addition to context validity indexing.

**Task Performance Results, Satisfaction Measures and Usability Problems**

In study I, results indicated that patients had particular difficulties with Task 3 and 4 in the mHealth system. These tasks had to do with first updating and then exporting a glucose value. Why these tasks were difficult for users could have to do with the tasks themselves, and with user practice and experience. One aspect was that the two tasks were complex and required several steps compared to other tasks in the set such as task 1 and 5, for example, which only included one step. Another possible reason could be that this usability test was performed by users when they were practicing new and complex tasks. When examining the overall satisfaction with the mHealth system it was clear that almost one-third of them provided it with a rather poor usability rating (or "OK"). Users indicated confusion with the multiple and unclear steps present in Tasks 3 and 4 and even if the tasks ought to be clear and straight-forward to them, they were not since they required users to have to navigate to different parts of the system, or to delete values and then enter new ones. It was clear that all these multiple steps in the tasks needed to be re-designed and streamlined.

Upon examining the usability problems found in study II and the 19 final consolidated usability issues it was evident that the majority of these usability problems, also with high severity ratings (severity level 3), were located in the Glucose Readings View. Two other views, the Glucose Diary View and Blood Pressure View had fewer problems but also similarly high severity ratings. The problems entailed difficulties with the deletion and entering of glucose values, exporting as well as printing a glucose value (as well as salient blood pressure values) and also how a file was saved. The UPT classifications indicated that most problems were deficient in their visualization of Information/results. Other problem classifications that dominated were Manipulation, Cognitive aspects and Visual cues. This reinforced the draw-backs in the visualization and cognitive aspects for the system content. These results reinforced difficulties that patients had completing common tasks and actions that are important for patients to perform frequently for their diabetes self-management. With regards to the classification
of results for study II, several of the usability concerns could only receive a classification in the artifact component in the Naming/labeling category due to that they lacked sufficient descriptions for a task classification on a deeper level. This full or partial classification only meant that the problem description had more or less granularity, however, and no other impact [130].

In terms of the contributions of the multiple methods, user testing as well as the interview and questionnaire all triangulated on five problems and four shared (verified) problems. The interview did not contribute to any unique consolidated findings while the questionnaire did contribute to one. Despite of this each of the methods provided a lens for the different usability problems, especially the initial issues, where the user test contributed to 59, the interview to 34 and questionnaire to 24 problems and thus also contributed to the overall result [82, 155].

In study III 129 usability problems were found by the experts. Of the different application views the Dashboard View generated the majority of usability problems, closely followed by the Glucose Diary View, the Blood Pressure View, and the Medication Adherence View. The experts also detected 274 heuristic violations. The largest number of these were located in the same views. The majority of usability problems could also be categorized as major usability problems. The largest number of issues could be clustered in the heuristic categories Consistency and Standards and Match Between the System and the Real World that dominated for all the views. In terms of severity, the average severity ratings for all the views was from 2.7 to 3 on a 0 to 4 rating scale, with the Glucose Diary view as well as Medication adherence view receiving the highest severity ratings. Altogether, the difficulties with these categories and views point to that the system needs a better design to accommodate effective decision-making as well as action control for relevant patient tasks [156].

In study IV the UC-CW detected 26 usability problems, 5 catastrophic and 10 major severity ratings compared to 20 problems, 2 catastrophic and 8 major ratings for the user test with TA. More minor ratings occurred for the TA with 10 versus 9 ratings. Both the UC-CW and TA were very similar on their average severity ratings for the mHealth application as a whole with the UC-CW with a 2.7 and the TA with a 2.6 average severity rating. Both of these fall into the major severity span. Another observation was that for both the UC-CW and TA a majority of their heuristic violations could be found in the categories Match Between the System and the Real World, Consistency and Standards and Recognition Rather than Recall. When it comes to the application views on number of problems, heuristic violations and severity rating levels, the UC-CW had most in the Main View with 11 problems and 29 violated heuristics which was similar to the TA with 10 problems and 24 heuristic violations for this view. Then came the Carbohydrate Entry View and List View were both methods had a similar amount of problems and heuristic violations (UC-CW: 4 problems, 14 heuristics; TA 3 problems; 9 heuristics and 3 problems and 8 heuristics for both). The UC-CW also had some of the highest average severity ratings (major) for the Time Setting View which was similar to the user test with TA. Both methods were also similar on the high severity ratings for the List View. The
UC-CW had its highest average severity rating of all for the Similar Situations List View, however, which was not present in the user test with TA.

**Critical Usability Problem Results**

In Study II, III and IV critical usability problems could also be detected. These were the problems that several methods or experts triangulated on as the most severe.

In study II, the most critical issues, depending on the number of methods they were detected in, were two (with a severity level of 4) which occurred in the Glucose Readings View as well as the Glucose Diary View. These problems consisted of needing to correct a glucose value through the removal of the erroneous value and then adding a new one. To perform these steps, patients had to first exit the Glucose Readings View, then get to the Glucose Diary View. Here they had to find and click the Add data button to then add the new value. This was a confusing sequence to patients. The Delete data button was also confusing to patients due to its combined function of allowing data to both be exported and also printed in the Glucose Diary View. This label did also not make much sense to them. Other serious concerns (with a level 3 severity) were located in the Glucose Readings View, the Glucose Diary View, as well as the Blood Pressure View.

Problems were evident in the Glucose Readings View when it came to exporting with several hard-to-understand steps. The exporting logic was not apparent to the users, and no support was present in the system for this process. If patients found out how to export a file, they were able to enter a file name to save it, but it was not possible to specify where to save the file. Difficulties were also experienced when it came to adjusting the different ranges of values in the graph in the Glucose Diary View. Another difficulty for patients was navigating the table within the Glucose Readings View (which included the correct value ranges or values). Here, no navigation support was available to assist them. Another notable concern was the Blood Pressure View where it was only the systolic and not the diastolic blood pressure that was visible in the graph despite of patients entering both. This display could be considered to be both incomplete and potentially harmful.

In Study III the dual domain expert evaluators discovered specific disease-related deficiencies regarding the information in the system which also received the highest severity ratings. These included, for example, that the Dashboard View only allowed one daily entry and lacking a time for the glucose readings on the displayed meter. This was considered a fundamental flaw due to it being essential information related to patients’ self-management in the minds of the dual-domain evaluators since patients with diabetes often need to make several daily readings. In the Glucose Diary View, evaluators found the numbers on the timeline too small, crowded and difficult to read. This was considered as particularly problematic for individuals suffering from diabetes as they often might have visual acuity concerns. The likelihood of the users performing errors here was evident for both of these problems.

In study IV, the UC-CW also found deficiencies with important aspects and triangulated on severity (severity level 4) with the TA on two navigational concerns located in the
Main View and List View. These included that it was unclear for participants to know that it was necessary to swipe across the chosen graphical image on the bottom of the Main View to get to the Periodical Pattern Graph View where they were to check their glucose measurements for the past week. This problem received a high severity rating due to that it required a completely different action than all the previous steps that users had to perform in the application in addition to being an important aspect for a diabetes patient to keep track of; how their glucose values looked over time. Another problem of a similarly severe character was a problem in the List View where it was unclear to the participants how to navigate in and understand where one was located when entering the list of registered food and glucose, and medication and insulin entries, it was also not possible to get a response on the sequence of dates in the list of all entries which made patients confused. Due to that it is vital for the patients to be able to follow and be clear on their entered foods, their calculated insulin or medication and keep track of their glucose levels to be able to manage their disease effectively this problem also received this high severity rating. Apart from these problems the UC-CW found other critical issues not present in the TA. Patients highlighted that in the Similar Situation List View it was difficult to understand and interpret the relation between the number of IUs of insulin and mg/dl blood glucose units in the upper part of the image compared to the list of entries in the bottom part of the view for this function making it difficult to function as a disease support.

User Characteristics, Personalization and System Interaction

Authors recommend that more usability studies focus on patient engagement as well as product interaction [43, 44, 157, 158] and not just medical outcomes [26, 27, 159]. The literature also emphasizes more studies needing to be performed on the effects on users and also user demographics due to their predicted influence on requirements and design of different decision aids [160, 161]. It is thought that user characteristics can affect task performance and outcome but to what extent and how is not quite clear when it comes to mHealth systems for chronic disease [162]. Based on the current literature on the topic of usability and users there is a need to examine the influence of demographic characteristics as well as technology experience and use when it comes to usability assessments to aid in the design of future interventions that need to be targeted to certain populations [160–162].

Upon examining in study I which users with which characteristics in the sample who experienced ease versus difficulties with the system, trends in the data showed that males as well as those patients with more IT experience completed system tasks slightly better. They also had higher system usability satisfaction (SUS) scores or user system acceptance scores. The figures also showed that the younger patient users performed tasks faster with fewer errors. It was also visible that the results on performance measures and satisfaction measures were congruent which meant that those participants who were more successful in completing tasks were also faster and had fewer errors; and those who performed better on tasks also had higher satisfaction scores. Conversely, the trends in the data pointed to those areas that were problematic for certain users, especially those with characteristics such as being older, and less experienced.
Conclusions that can be drawn from this are that designers may need to tailor the system interactions for these users. This study aspect also echoes other literature in the area which states that there is a need for mHealth interventions to be adapted and suited to a wider variety of users in order to encourage usage for a higher number [163].

Individualized usage and personalization of mHealth functions and features to tailor these more to the individual user and their skill level to accommodate individual adaptations has been brought up and demonstrated in several studies [97, 164]. Recent systematic reviews on mobile applications for control and self-management of diabetes have also indicated the need for a focus on personalized mobile health uses [165] and that mobile applications should also be customized to the specific type of diabetes [166]. This along with the importance of providing patient-centered approaches to assist patients’ individual adherence strategies to lead to better outcomes [167]. These were aspects that became particularly visible in study IV and in the UC-CW evaluation. There were two critical usability problems that surfaced from the group that had to do with important deficiencies about personalization aspects that patients thought were needed for their individual disease management. For example, in the Main View patients identified that it was not possible to configure and personalize ranges based on their own ones for the low, medium and high ranges in the middle pane of the application for their latest glucose measurements. Also in the Plotted Graph View the glucose range delimiters were also not able to personalize and configure based on the patient’s own ones (for low, medium and high glucose value ranges) which would be necessary as these can differ for individual patients. These two problems about the importance of personalization of these types of disease intervention may have been a direct result of the UC-CW and its group based dimension of several different users collaborating on finding issues, as these severe problems were not found in the individual TA sessions.

**Method Acceptability to Users**

As it was important to also determine how the users themselves experienced the methods they were part of in study IV, the UC-CW as well as the other method it was compared to (the TA) were assessed using the NASA RTLX instrument to measure the cognitive load on participants. In addition, individual interviews with participants were also performed. The NASA RTLX scores showed that the UC-CW had a mean value for the participant group of 25.3 (SD 3.2) or more than half the cognitive load of the TA with a mean value of 53.5 (SD 14.2) indicating a preferable outcome for the UC-CW. The experienced difference also seemed partly expressed in the individual interviews where participants looked upon the UC-CW favorably in many regards such as that it was easy to understand and perform compared to the participants in TA, for example, who felt that method was difficult to perform, rather awkward, and cognitively demanding for them. The UC-CW group also stated that it was a straightforward and seamless method, felt it positive to have an individual and collaborative part in it and that it was helpful with the mixed patient group of diabetes patients of different types, genders and ages when performing the method. Some less liked aspects were that some felt that there was an overlap of questions and some redundancy in participants’ answers and experiences.
Lessons Learned and Recommendations

Study I demonstrated how the application of systematic measures and usability methods and relevant patient characteristics could be taken into account for a particular mHealth system. The results showed how the ISO standard for usability could be used to assess the effectiveness, efficiency and satisfaction measures together for an increased understanding of the experienced system usability for users, their interaction needs and task performances. Moreover, this study showed how patients’ characteristics had an influence on this performance. These aspects could provide insights into how more usable eHealth and mHealth systems could be developed, also for a larger number of users. These considerations are also important due to the anticipated growth in mHealth solution usage numbers [43].

In Study II the findings verified the usefulness of utilizing several data collection methods to generate more comprehensive lists of usability issues, that at the same time also contributed to triangulate severe usability problems in the system. Each of the individual methods contributed to the initial problems. For the consolidated issues it was possible to determine the percentage of users who brought up a specific concern in the different methods which provided an important indication of its severity to the users. This type of triangulation as well as severity indication to users can provide a clear benefit in assisting developers to locate the usability problems that should be prioritized. One of the most significant contributions of study II, however, lies in the methods used for data analysis. The FA method was very useful for structuring the qualitative data. It contributed both in standardizing problem descriptions as well as promoting a consistent, and streamlined analysis among the researchers. In essence it was valuable because it provided a systematic way in both the managing and mapping of data [126]. It also created reproducible steps [129], and supported both the inductive (usability problem coding) and deductive (UPT) parts [148]. When combining both the FA and UPT framework it was also possible to create lists of in-depth, two-dimensional classifications for the found usability problems which in itself could aid in solution generation and contribute to problem-solving for designers [130].

This study provides several additions to the literature. It shows a multi-method approach for data collection on usability problems, it aims to standardize data analyses, and to contribute to the reproducibility of results when it comes to qualitative usability studies [102]. The FA method can also be recommended for future studies utilizing several researchers to code data because of its structured process [126]. Here it was also found effective with just two researchers due to being able to follow linear steps in the coding process. It allowed the researchers to work together in a seamless way in the analysis of the large amounts of data and was also found advantageous in its clear structure to be able to achieve detailed usability problems that could then be classified using UPT.

The in-depth severity factor ratings, the dual domain evaluators and the validated user tasks for the modified, user-oriented heuristic evaluation in study III enabled detecting a large number of severe system deficiencies. This detection of major usability problems through the HE method is in contrast to the work of other authors both within and
outside the domain of the health care who have mainly found minor interface issues with this method [78, 81, 82]. The identified severe usability problems found here would require re-design to be able to achieve the self-management needs for patients. The information provided in the modified HE can be used during formative evaluation, or summative testing and/or for comparing different versions and kinds of applications.

It is probable, that the methodological changes to the HE in this study could have aided in enabling the evaluators to find important usability concerns. As shown, conducting usability tests with users is very helpful in system design, but as was seen in this study, the dual domain experts could also add important, additional dimensions to the evaluation. In Nielsen’s view, three to five single domain expert evaluators of usability are expected to find between 74% and 87% of the usability problems [168] present in a system. Dual domain experts usually find a higher amount or 81% to 90% of the problems with two to three evaluators [168]. These dual domain experts can therefore be seen as especially suitable when evaluating complex systems as the ones found in the health care area, due to their specific domain and usability knowledge [169, 170]. Here they were especially useful in understanding the user context in terms of needed tasks for the chronic disease self-management system, as well as how the specific problem would influence the diabetes patient in terms of its frequency, impact and persistence. They were also able to detect specific system design concerns regarding particular needed information related to these kinds of self-management tasks. Dual domain experts can therefore be recommended and can add an important value in future heuristic chronic disease management systems evaluations to identify unique and critical usability problems, cognitive support and disease-related concerns [34, 170, 171].

In study IV the importance of performing validity measures, and comparisons when developing a method or performing a method modification was apparent, as it gave important insights. It was evident that the addition of several task validation steps in the UC-CW to add to its validity clearly impacted the overall time it took to perform the method compared to the TA, making it take longer to perform as a whole. However, when dividing the total amount of time taken for the evaluation procedure as a whole compared to the number of problems found in each method the UC-CW performed faster. Both the UC-CW and TA performed largely similar results on the number of usability problems found and the severity of problems, with one important exception – the UC-CW was able to identify critical personalized disease support needs in the application, which might be a possible result of the patient group dimension of that method where users were able to identify their specific needs in disease support. This also provides evidence that this method modification can contribute to finding the more severe and important problems for users which authors have emphasized has been one of the drawbacks with the original CW method [28, 54] as well as the ability to encourage the identification of more recurring problems as opposed to one-time, low priority problems [28]. This implies that the UC-CW identified as severe and relevant problems as did the TA. In addition, the user acceptance measures also proved to be an important addition. The NASA RTLX instrument, for example, showed that there was a large difference in how users perceived the UC-CW compared to the TA method, perceiving it
as considerably less cognitively demanding, which was also expressed in the interviews. Its less perceived cognitive demands may mean the UC-CW may be an applicable method to scale up both on the number of participants and user groups performing the evaluation sessions, adding to the evidence of results. Overall this implies that the UC-CW has a beginning validity, but these results would also need to be verified with other mHealth systems and applications and other patient groups.

**Limitations**

**User Samples**

All studies involved small user samples for patients (I, II, IV) and small numbers of dual domain experts (III). For study I and II the small user samples were randomly selected from a database of a convenience sampled diabetes population of a larger mHealth diabetes intervention study [113]. Even if these two study sample sizes of 10 diabetes patients each are fitting for usability testing studies and also a higher number than the sample sizes that are recommended by researchers such as Nielsen, Landauer and Virzi and Monk [84-86] formative usability evaluations, such as these two studies, state that five to eight users can usually find 80-85% of the usability issues [84-86]. This meant that for study II consisting of a multi-method approach large amounts of data were generated despite of the small sample. It may be possible, however, that these users do not represent all the kinds of users of mHealth diabetes systems or the general diabetes population. One could, for example, note that the samples in each of study I and II were more highly educated and might use technology more than diabetes patients from a lower income or less technology savvy population. This can make it difficult to generalize these findings to the general population. For Study III, although the modified heuristic evaluation process detected many major and severe usability problems, it may be the case that other problems could also have surfaced if the dual domain evaluator numbers had been increased [168].

In study IV even if having a purposive sampling technique, there were variations that may have impacted the overall results for the different methods, such as for example more education, more experience with the technology and a slightly more positive view of its possibilities for the individual in the UC-CW group. In addition, the case study design with independent groups to compare the methods could also have created a variability. The variability was largely unavoidable as the same patients could not be exposed to both the UC-CW and TA without creating a learning bias. An attempt was therefore made to create groups as representative and comparable as possible. The variability between groups could also have been so significant that no differences between the methods would be seen. Despite this, it was evident that both the UC-CW and TA largely produced similar results in terms of the number and severity of usability problems, and what views in the application that had most issues. The UC-CW group had slightly more severe problems and issues around personalization features that would be needed for diabetes patients, but both methods were otherwise very congruent on the characteristics of problems and severity level of these.
**Efficiency and Effectiveness of Methods**

In terms of method performance, particularly in study IV which tested and compared a new method modification (the UC-CW) the addition of several task validation steps clearly impacted the overall time it took to perform the method compared to the TA; making it take longer to perform as a whole. However, this may need to be compared with the total amount of time taken per found problem in each method, as well as for the evaluation process itself where with such a mindset that method actually performed faster. It is also evident that there will be variations when performing a group-based method versus individual-based methods where it might have been the case that for the modified CW, despite writing their individual thoughts down, participants might have streamlined their responses more with the group which might not have been the case to the same extent in the individual TA test for example, as it lacked the group dimension. Still, the group dimension with its mix and input was one of the aspects that participants in the UC-CW stated that they really appreciated based on their comments about how they experienced the method. When looking at the NASA RTLX scores on the whole there was a clear indication that participants felt that the UC-CW was less cognitively demanding than the TA. Future researchers intending to use the UC-CW may therefore have to weigh the aspects they gain and potentially lose before deciding upon if it is the most applicable method for their evaluation.
Conclusions

The purpose of study I was to exemplify the use of systematic usability methods by utilizing the measures in the ISO standard guidelines for usability for the assessment of effectiveness, efficiency as well as satisfaction with the validated System Usability Scale (SUS). It was perceived that these, combined with assessments of user characteristics, could provide an increased understanding when it comes to system usability for users. In this quantitative approach there was a notable variation across users and tasks and while the results indicated good, acceptable (but not excellent) satisfaction with the mHealth system it was evident that almost one-third of the tested users considered the system to have rather poor usability. Study I also attempted to address a gap in the literature through the examination of patients with various characteristics and skill levels when it came to technology on a number of performance measures. The trends in the participant socio-demographic, IT knowledge and experience data showed that age, gender, duration of disease and IT experience level of the participants may also have an influence on the interaction outcomes. These results intended to display objective data and point to those areas needing to be corrected by the developers which were especially those tasks that participants considered confusing. Objective data is important since both researcher and designer perceptions about specific tasks and user performance on these may not always be congruent with the actual performance data. It is also important that system change lists are data-based to determine the priority corrections. The trends such as those found in these results can help in the interpretations of user needs and in the design of more usable mHealth systems. To increase mHealth usage on a larger scale and promote both more extensive use and acceptance of this technology, the characteristics of users will most likely need a greater consideration in future mHealth system design.

In addition to applying a systematic and streamlined approach in data collection and analysis, study II took into consideration recent systematic reviews performed on mHealth self-management tools for diabetes that emphasize the need for more studies both on patient interaction as well as system usability. The structured data analyses performed in this study allowed for reproducible steps as well as validation of data (through triangulation) to determine the problems that were most severe for users. For data analyses the FA method, which can also guide analyses across multiple researchers, provided a more standardized and structured way to determine usability problems especially from a large amount of qualitative data which was the case here. The UPT was also advantageous here both as an in-depth classification scheme for problems as well as in the determination of the severity of usability problems. In addition, it was especially helpful in the categorization of the types of problems where similar kinds of problems were detectable which could be useful for designers in problem solving. Both of these together could be considered a new combination of methods for health IT. In addition, study II also provided a multi-method design example in the detection of various problem types and levels of severity. The usability testing with Think Aloud was the most important of the methods in surfacing usability issues while the in-depth interview, as well as the questionnaire, both allowed for data triangulation on severe usability
problems. It is, however, clear that all the data collection methods contributed to the usability problems in a more comprehensive way.

In study III specific modifications were made to the HE consisting of dual domain experts, the use of validated, and user-centered self-care tasks along with separate in-depth severity factor ratings. This was done to attempt to accomplish a fast and resource-efficient, streamlined heuristic evaluation process related to diabetes patients’ self-management. The results demonstrate that a modified heuristic evaluation such as this can detect critical, medically related issues in this context. The provided modifications, demonstrate that Nielsen’s original methods can be improved upon to attempt to achieve improved results. These can be considered important when determining how the chronic disease system fits patient self-management needs. The dual domain experts can also provide important information related to potential patient safety concerns as well as determine how the application adheres to usability guidelines.

In study IV the different measures to compare the UC-CW against proved useful. This is important, not only because it gives a perception of how the method performs when tested in practice, but also provides an objective measure when being able to compare its results against the other methods it is attempting to distinguish itself from; adding to its validity. Even if the modification did produce almost the same results as the TA in terms of the total number of usability problems, severity ratings and significance, it was possible to see that more critical disease personalization aspects were mentioned. In addition, the actual time taken to perform the evaluation both in terms of the evaluation process itself but also when compared to the amount of problems found, turned out to be less even if taking longer as a whole. The follow-up interview indicated that participants were also very comfortable with the UC-CW indicating it was easy to perform and enjoyable, despite some minor comments about certain redundancy. This could be compared to their comments about the TA that many found rather awkward and cognitive demanding. The results for the NASA RTLX also showed that participants experienced the UC-CW almost half as cognitive demanding compared to the TA which was an interesting result, speaking favorably towards its ease of use.

The study results not only, demonstrate important outcomes in terms of actual performance numbers and the importance of validity to determine actual method results but also highlight the importance of including validity measures in the whole evaluation method process. Performing validity measures, determining actual method results, is one part but of equal importance is the users. Including users in the method improves the relevance of determined usability problems contributing to the validity of measures, but also including a dimension of what the users think of the method, as here through the NASA RTLX and short interview, is vital when it comes to implementation and especially when deciding and selecting between suitable methods. This specific component may also be a new addition to this kind of method evaluation process and comparison in the health informatics field. All in all, the results and external validation thus provide beginning evidence of the UC-CW’s utility as an effective, efficient and acceptable evaluation method.
Contributions and Implications of the Research

The main contributions of this thesis and the included studies relate to:

- The empirical results on the identified usability problems for the evaluated diabetes mHealth systems and applications (Study II, III, IV).

- The empirical results on the critical usability problems to users in the evaluated diabetes mHealth systems and applications (Study II, III, IV).

- Incorporating, and considering user characteristics in connection with task performance results (Study I)

- Methodological modifications to usability assessment methods to include a user-centered emphasis, and that at the same time are fast and efficient in finding usability problems (Study III, IV)

- Evaluation method applications to conduct standardized (Study I), streamlined and structured data collection and analysis processes (Study II, III, IV)

- Novel application of two methods: Framework Analysis (FA) and Usability Problem Taxonomy (UPT) for usability problem determination, classification and analysis (Study II)

- Comparison and validation of a modified usability evaluation method (Study IV)

- Determining users’ experienced cognitive load and acceptance of different usability methods through a novel application of the NASA RTLX instrument and post-evaluation interviews (Study IV)

The included research demonstrates the necessity to take a critical approach in assessing both the strengths and weaknesses of different methods to determine those that are most appropriate and practical for carrying out usability evaluations in the complex context of health care [172]. When there are only limited resources available it can be a challenge to select evaluation methods. Evaluators must then balance resources with the need to be comprehensive and adhere to user, clinical and usability perspectives to provide the most value in terms of consumed time, effects and outcomes [80].

Adopting a user-centered approach to technology design, as also evidenced in this research, is clearly a crucial element in increasing the likelihood of obtaining the needed fit between an individual user and the technology. Human factors techniques such as the ones used in this research have an important place and have shown to be useful when applied to actual users in their context of use. They assist in understanding users’ specific
needs and needed features for self-care of their health conditions [173]. However, as shown here, existing techniques can also be improved upon in several ways to attempt a better fit.

It is also clear, as researchers point out that gaps still exist and improved strategies on how health care providers should evaluate and recommend chronic disease mHealth applications to their patients are still needed, as well as both patient and clinician input on the usability and clinical utility of these applications [174]. This research can be seen as a first step to try to find ways that accommodate users and clinicians to incorporate these aspects into usability evaluation methods and processes to attempt to produce better outcomes.

This research makes evident the importance of approaching this area from a transdisciplinary mindset to achieve new perspectives and in trying to advance the field. It has a firm basis in methods from human factors research and is applied in a clinical context with the intent to make it more pertinent [175]. As other researchers point out about the growing area of mHealth systems, as these become more prevalent and necessary in the delivery of health care, scientific, engineering and regulatory challenges will continue to emerge. The future success of mHealth will, therefore, to a large degree depend on these transdisciplinary approaches, such as the ones here, across researchers in the fields of computing, engineering, and the health sciences [30].
Future Research

Based on the included studies of this thesis it is clear that future research could advantageously continue to explore mHealth technologies for chronic diseases such as diabetes, usability testing techniques and applicable method modifications to incorporate the user.

All studies included in this research could be replicated on a greater scale with larger user samples with more varied characteristics. In study I this would consist of randomly selected user samples to find out the performance and satisfaction with the system and other user characteristics compared to the stated ones or similar performance measures. With regards to study II, researchers could use purposive sampling to select users with various characteristics to achieve maximized variability. Future research foci might also include usability testing with larger user samples, also testing in earlier stages of the development life-cycle, as well as testing other classification schemes and frameworks such as the ones in study II and also investigate how reproducible results are. Future researchers may also consider using more data collection methods to determine if these produce more usability problems and more unique issues. In study III the goal was to provide a cost-effective, efficient method of modifying the HE that would also find the more serious usability concerns that could have an influence on the disease management of patients. In future research both HE and user tests may also be combined to determine the optimal number of usability issues even if this might signify higher costs. In study IV, in addition to adding more participants with more varied characteristics in the UC-CW groups and usability testing groups due to its perceived lower cognitive demand, additional validation tests could be performed to compare methods against as well as more user acceptance measurements.

It is probable that continued methodological improvements in usability studies for mHealth could lead to more comprehensive identification of usability concerns, in addition to more specific redesign recommendations that are based upon user-centered data. More reliable data analyses could also improve upon the reproducibility of determined results. Importantly, improved mHealth designs for the intended users of these systems could mean improved interactions, interaction performance, increased adoption of mHealth applications and long-term perhaps even increased adherence to interventions for chronic conditions.
References


115. Spencer, R., The streamlined cognitive walkthrough method, working around social constraints encountered in a software development company, in


Study I

Quantifying usability: an evaluation of a diabetes mHealth system on effectiveness, efficiency, and satisfaction metrics with associated user characteristics

Mattias Georgsson1,2, Nancy Staggers1,3

ABSTRACT

Objective Mobile health (mHealth) systems are becoming more common for chronic disease management, but usability studies are still needed on patients’ perspectives and mHealth interaction performance. This deficiency is addressed by our quantitative usability study of a mHealth diabetes system evaluating patients’ task performance, satisfaction, and the relationship of these measures to user characteristics.

Materials and Methods We used metrics in the International Organization for Standardization (ISO) 9241-11 standard. After standardized training, 10 patients performed representative tasks and were assessed on individual task success, errors, efficiency (time on task), satisfaction (System Usability Scale [SUS]) and user characteristics.

Results Tasks of exporting and correcting values proved the most difficult, had the most errors, the lowest task success rates, and consumed the longest times on task. The average SUS satisfaction score was 80.5, indicating good but not excellent system usability. Data trends showed males were more successful in task completion, and younger participants had higher performance scores. Educational level did not influence performance, but a more recent diabetes diagnosis did. Patients with more experience in information technology (IT) also had higher performance rates.

Discussion Difficult task performance indicated areas for redesign. Our methods can assist others in identifying areas in need of improvement. Data about user background and IT skills also showed how user characteristics influence performance and can provide future considerations for targeted mHealth designs.

Conclusion Using the ISO 9241-11 usability standard, the SUS instrument for satisfaction and measuring user characteristics provided objective measures of patients’ experienced usability. These could serve as an exemplar for standardized, quantitative methods for usability studies on mHealth systems.

Keywords: diabetes, mobile health, self-management, usability evaluation, user-centered design

BACKGROUND AND SIGNIFICANCE

To assist patients with diabetes, different types of support systems for self-management have been developed using Information and Communication Technology (ICT) including mobile health (mHealth) technologies.1,2 Studies show that mHealth in particular has helped improve chronic disease3 and glucose management for these patients,4,5 but the usability of mHealth interventions still needs to be addressed more fully.6,8

Ensuring adequate usability is of the essence for the individual patient and because of the worldwide penetration of mobile phones.8 Mobile phone availability is approaching the 7 billion mark globally, with a recent figure showing an estimated 6.9 billion subscriptions.9 An estimated 335 million wireless subscriber connections (mobile phones, tablets, etc.) are in the United States alone.10 Given the volume of mHealth applications and insufficient usability assessments, the potential impacts on user interactions are not clear but are likely substantial. Attention is needed to performing standardized, systematic mHealth usability assessments.11

International standards and validated instruments are available to aid in the design of mHealth systems, although these are not yet widely used in health care. For example, an October 2014 review of usability testing studies on mHealth technology for diabetes showed great variations in usability testing methods. Of the list of 23 applicable studies, only four used a validated instrument.11 Only one study followed a completely standardized procedure including the use of validated instruments.11

Authors recommend a more widespread use of International Organization for Standardization (ISO) guidelines and techniques12,13 to guide usability evaluations.14 Expanding usability testing to be more systematic and complete can help build a science of mHealth usability. More standardized, comprehensive approaches would improve methodological consistency, making it possible to begin comparing findings across mHealth application evaluations.11,12

Researchers need to assess the full set of recommended measures—effectiveness, efficiency, and satisfaction—to obtain a more thorough picture of the usability of any application.11,14 In that vein, this paper employs ISO 9241-11 and techniques to address this gap for mHealth systems.

Studies also need to be conducted on the effects of relevant user demographics because they can affect requirements including the design of future decision aids for chronic disease.15,16 Specific user characteristics can produce differences in task performance, but how and in what way is not completely understood for mHealth systems.17

Correspondence to Mattias Georgsson, MSc, RN, Faculty of Computing, Blekinge Institute of Technology, 371 79 Karlskrona, Sweden; mattias.georgsson@bth.se; Tel: +46 455385000. For numbered affiliations see end of article.

© The Author 2015. Published by Oxford University Press on behalf of the American Medical Informatics Association. This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs licence (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial reproduction and distribution of the work, in any medium, provided the original work is not altered or transformed in any way, and that the work property cited. For commercial re-use, please contact journals.permissions@oup.com.
In this study, our goal was to complete an evidenced-based approach to assessing a mHealth system. To address noted gaps in current mHealth usability testing, we employed a standardized method assessing the full set of major aspects of usability recommended by ISO 9241-11. We used instruments with established psychometrics such as the System Usability Scale (SUS) instrument, and we measured relevant user characteristics associated with mHealth task performance to obtain a broader understanding of users, their tasks, and performance outcomes in mHealth interactions.

Overview of Usability Testing
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.” Usage testing is commonly used in the usability engineering and human-computer interaction disciplines, and more recently, in the health care arena. The purpose of usability testing is to determine obstacles to effective and efficient product use as well as product acceptability and satisfaction for representative users as they interact with the intervention in a specific context.

Usability can be measured by several different methods. One of the most common is heuristic evaluation, an expert-based approach originally introduced by Nielsen. In contrast, user testing typically employs actual users such as health consumers. A particular advantage to usability testing is that resource requirements can be fairly modest. The number of users needed to perform user tests varies, but a small sample of five to eight users can commonly identify 80–85% of usability problems. User test sessions include representative tasks and scenario-based evaluation in an environment suitable for conducting the testing. This can be either a laboratory environment under controlled conditions or in the field, either of which allows researchers to obtain an understanding about how the intervention is used in a specific context. The observation and task/scenario-based process involves observing and assessing how users interact with the intervention and how well the interaction supports users in achieving their specific task goal(s).

Measuring Usability
ISO 9241-11
One of the quantitative methods to determine usability is outlined in ISO 9241-11, a standard consisting of specific metrics about how well a user fulfills specific goals (see Figure 1). This standard includes the main concepts of user-centered design (UCD), which entails incorporating users early and often throughout all steps of the design and development process.

ISO 9241-11 describes in depth how users should interact with a product, employing hands-on methods to indicate its overall usability. One common technique is to record users as they perform representative tasks during the interactions. If the indicated measures of effectiveness, efficiency, and satisfaction are fulfilled adequately, the product can be considered to have attained an acceptable level of usability. Appropriately terms are defined in Table 1.

Standards such as the ISO 9241-11 are especially suitable to apply to new technologies such as mHealth systems and applications. In support of this notion, Bevan concludes that these standards should be used more frequently in usability work as they define good practice, are objective, can ensure consistency in the work, and can provide benchmarks for intervention by designers.

Measuring Effectiveness and Efficiency
Effectiveness is typically measured by task completion success and by counting the number of errors performed during the interaction. Efficiency includes the level of effort and resource use by the user to achieve usability goal(s) and is typically measured by timing each task and averaging times across users and/or tasks.

Measuring Satisfaction
Satisfaction can be objectively measured with available instruments, including the SUS instrument used extensively outside health care. Developed and designed by Brooke, the SUS measures overall usability, allowing comparisons across a range of contexts and systems.

This 10-item Likert scale instrument is typically administered immediately after interaction, allowing users to record their initial feelings and responses. Instrument items have a range of 0–4. The SUS has been evaluated for validity, reliability, and sensitivity. SUS scores range from 0 to 100 providing an estimate of overall usability of the intervention in the minds of users. Scores of above 70 are considered to be acceptable or good while scores of 85 or above indicate a high level of usability or excellent score. Scores of 50 or below indicate poor or unacceptable usability.

MATERIALS AND METHODS
The measures for this work were determined by employing a standardized usability assessment along with patient characteristics. The evaluation process is described below.

Study Sample and Setting
The study received approval (HI-BEA-001#20111605) from the Western Institutional Review Board, Olympia, Washington, USA. Study participants were selected from a database of 2317 patients involved in a larger randomized controlled trial on a diabetes mHealth intervention. The larger study used a convenience sampling technique and

Figure 1: Usability framework according to ISO 9241–11.

Table 1: Included attributes for usability according to ISO 9241–11.

| Effectiveness | To what extent the user can achieve a goal with accuracy and completeness |
| Efficiency | The level of effort and resource usage which is required by the user in order to achieve a goal in relation to accuracy and completeness |
| Satisfaction | The positive associations and absence of discontent that the user experiences during the performance |
involved patients from 18 primary care clinics in the Salt Lake City metropolitan area, Utah. Inclusion criteria for the larger study were (1) adults of 18 years old or older with (2) a type 2 diabetes diagnosis and glycated hemoglobin (HbA1c) ≥ 8%, (3) not pregnant, (4) with cell phone access, and (5) able to speak and understand the English language. Ten patients were randomly selected from the larger study database for inclusion in our usability study. These patients had no prior experience with the mHealth web system evaluated in our current study. Inclusion criteria for our usability study sample were (1) patients diagnosed with type 2 diabetes; (2) no cognitive impairment; (3) some knowledge and use of computers, the Internet, and cell phones; and (4) the ability to speak and understand the English language. After obtaining informed consent, patients were scheduled for individual usability sessions in a quiet room at HealthInsight, one of the US Beacon Communities in Salt Lake City, Utah.

System Description
Care4Life® is an interactive algorithm-based mobile short messaging service management system with an accompanying web portal for patients with diabetes (see Figure 2). The system was designed as a personalized, self-care management tool meant to function as input-driven coaching support for patients’ self-care management and to aid in the collaboration between patients and their health care providers. Patients interact with the system by being prompted to send in their self-management measures via text messaging. Typical transmitted measures are blood glucose, blood pressure, weight, exercise, and medication adherence.

User Assessment Tasks
Representative user tasks consisted of (1) sending in measurement values to the system, (2) interpreting glucose measurements displayed in a graph, (3) recording glucose measurement values, (4) exporting glucose measurement value trends to a portable document format, (5) interpreting blood pressure measurement in a graph view, (6) setting and tracking goals regarding exercise and weight, (7) setting medication reminders, and (8) setting a physician appointment reminder. It was vital to include a variety of relevant tasks of different lengths and levels of difficulty to obtain representative and accurate performance. Tasks were based on real case scenarios to simulate how patients would interact with the system in a real-life situation according to the care and self-management process and were validated by a panel of three health care professionals and usability experts for content and context accuracy.

Instruments and Outcome Measures
Demographics and IT/Computer Knowledge and Experience Questionnaires
Two short pretest questionnaires were administered to assess patients’ background characteristics and information technology (IT) experience. These were divided to make them easy for patients to complete. Validation was performed by three health care professionals and usability experts. Patients were asked about their age, gender, educational background, how long they had had diabetes, their experiences and knowledge about computers/IT including self-assessed IT knowledge, and if and how often they used a computer as well as the Internet.

ISO Standard Usability Measures
ISO measures of effectiveness and efficiency were measured as follows: Effectiveness was determined by degree of task completion and the total number of errors per task. Task completion was coded using three categories: (1) completed with ease when the user was able to perform the task without any help from the test leader, (2) completed with difficulty when the subject achieved the task with minor difficulties and or with minor hints from the test leader, and (3) failed to complete when the subject was unable to complete the task, even with some minor hints. An error was coded when the subject performed errors they could not solve or committed errors preventing further progress.

Efficiency was determined by timing each individual task and computing the average time for each task across patients. Efficiency or time on task began when the user started performing the task and ended when they pressed the participant/home button. Time was deducted for prolonged loading and response time for the system.

Satisfaction was measured by administering the SUS. Scores were calculated according to Brooke’s guidelines. This consisted of summing the scores on each of the 10 individual items. 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 satisfaction value.

Study Procedure and Data Collection
After completing the informed consent, patients filled out the pretest surveys on demographics. Standardized training is recommended to decrease individual variability in task performance due to knowledge about a system. Thus, standardized training was done to simulate an actual patient educational care process in the health clinic, and users interacted with the system on their own. Researchers explained the proposed evaluation process. Patients then interacted with the system using the described tasks to enter and retrieve values as well as read different graphs. Morae® software was used to video- and audio-record patients interactions in the system. Subsequently, the recordings were carefully coded using the specific metrics as defined above on task success rate, errors, and time on task. The session ended with the administration of the SUS. The testing procedure took ~1.5 h per patient, and patients received a gift card for $20 after finishing the session.

Data Analysis
Results were calculated using a Microsoft Excel spreadsheet. Descriptive statistics such as means and standard deviations were calculated in SPSS® version 22™ on effectiveness, efficiency, and satisfaction results. Aggregated data were compiled for the specific tasks across patients. Descriptive statistics results were compared to the
sociodemographic data to distinguish differences and similarities between users and their performance data. To make comparisons between user characteristics and user performance and satisfaction, we compared gender, age, education, diabetes diagnosis, and IT/computer knowledge and experience (based on rated experience vs rated inexperience) against effectiveness, efficiency, and SUS mean scores. Statistical correlations were not performed due to the small sample size.

RESULTS
Patient Demographics, IT/Computer Knowledge, and Use (Pretest Questionnaire)
The 10 patients had a variety of different backgrounds and characteristics. Six participants were female and four male. Most (70%) were older adults (50–69 years), while 30% were between the ages of 30–49. Most patients had college or university education (80%). Sixty percent were diagnosed with type 2 diabetes ≥5 years ago while 40% were diagnosed ≤4 years ago.

IT and computer knowledge use and preferences differed among patients. Half rated their IT knowledge level at medium and 30% at high while the remainder indicated little knowledge. A majority (80%) of patients reported using a computer and the Internet every day. Eighty percent also agreed or strongly agreed that they enjoyed using the computer in their work or leisure time.

Effectiveness
The task effectiveness results are presented in Figure 3. Tasks 3 (correcting a glucose measurement value) and 4 (exporting a glucose value) were the most difficult to complete with 30% and 40% failure rates, respectively. Tasks 1, 2, and 5 (send in measurement value, interpret glucose measurement value, and interpret blood pressure measurement value) were completed with ease by all.

The error rates mirrored task completion difficulty with Tasks 3 and 4 having the largest number of errors while Tasks 1, 2, and 5 were completed without any errors (see Figure 3). The largest number of errors was quite high at 9 and 13. The kinds of errors committed included patients having difficulties in remembering steps, finding correct options, and performing the different steps in the process. Errors were especially prevalent with tasks related to (1) locating glucose values for correction, (2) exporting these data, (3) remembering how to delete values or select values, (4) navigating to the correct screens to accomplish these tasks, and finally (5) remembering how and where to add or export their glucose values.

Efficiency
As may be seen in Table 2, Tasks 3 and 4 consumed the longest amount of time, as might be expected given the difficulties with task success and errors mentioned above. On the other hand, Tasks 2 and 5 took the shortest times. Tasks of correcting and exporting values (3 and 4) had mean scores 2–3 times as long as those related to interpreting values.

Satisfaction
The average SUS score for the whole group was 80.5 (SD 11.47) indicating good satisfaction across these mHealth system users as seen in Figure 4. However, wide variations in scores existed with a low value of 62.5 and high score of 97.5 (a 35-point range). The highest ratings ranged from 87.5 to 97.5 or excellent (30% of the patient sample) to the lowest from 62.5 to 72.5 or “OK” to minimally acceptable (30% of the patient sample).

User Characteristics, IT Knowledge, and Internet Skills and Usability Metrics
User characteristics and objective data were assessed for additional insights (Table 3). Descriptive trends indicate a difference across genders, ages, diabetes diagnosis, and IT/computer knowledge and experience. Males in this sample had higher average success, lower error rates, and higher mean SUS scores than females. Younger patients also had higher average task completion rates on tasks, only one error average on tasks, lower task completion times, and higher mean satisfaction scores.

Patients with a more recent diabetes diagnosis had a higher average task success rate, fewer errors on tasks, shorter average time on tasks, and higher average SUS scores. High school or college/
Our results indicated that males and those with more IT experience performed slightly better and had higher SUS scores. Younger patients also performed faster and had fewer errors. These trends point to problematic areas, especially for users with specific characteristics. Therefore, designers may also need to tailor interactions for targeted users; for example, older, less experienced users. This aspect of the study is congruent with other literature indicating mHealth interventions need to be better adapted to a wide variety of users to facilitate wider usage for a larger number of users.

Results across the examined performance and satisfaction variables were congruent, although this is not always the case in usability studies. That is, participants who completed tasks more successfully were also faster and committed fewer errors. Those who performed better had higher satisfaction scores.

Contributions to the Literature
To our knowledge, this is the first usability study on mHealth diabetes interventions assessing usability effectiveness, efficiency, and satisfaction together by using validated measures. In addition, we used the ISO usability standard and SUS instrument to compare usability metric performance outcomes to pertinent patient user characteristics.

This study helps fill several gaps in the literature. Previous authors recommended more usability studies focused on patient engagement and product interaction as opposed to exclusively medical outcomes. Other authors have reported the need to explore the influence of demographic characteristics and technology on usability assessment scores to assist in designing future interventions targeted to specific populations. Finally, previous authors argued about the importance of assessing effectiveness, efficiency, and satisfaction measures together to obtain a fuller picture of the usability as was done in the current study. This study fulfilled these recommendations.

This study adds to the literature on mHealth interventions as an example of an evidence-based approach resulting in a more inclusive quantitative usability assessment. We employed actual users during testing, adhering to recommendations from the ISO usability standard. The results provide developers with specific data about where patients experience the most difficulty in the product and where tasks should be redesigned to accommodate a wider variety of users. This study could serve as an exemplar to others evaluating mHealth systems who want to generate benchmark data and provide reproducible comparable results.

Usability testing results do not have to show overall poor usability to be helpful. Moreover, we need positive examples of usability testing with mHealth systems. In the past, authors focused mainly on negative examples. This study uses a standard methodology to explicate the core usability issues that need to be addressed and provides specific techniques others may employ to evaluate and improve the usability of mHealth systems.

### Table 2: Time on task per task

<table>
<thead>
<tr>
<th>Task</th>
<th>Time per task (min) Mean (SD)</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1</td>
<td>1.51 (0.48)</td>
<td>0.97–2.41</td>
</tr>
<tr>
<td>Task 2</td>
<td>1.45 (0.45)</td>
<td>0.91–2.50</td>
</tr>
<tr>
<td>Task 3</td>
<td>4.18 (1.55)</td>
<td>1.66–6.47</td>
</tr>
<tr>
<td>Task 4</td>
<td>3.67 (1.71)</td>
<td>1.35–6.86</td>
</tr>
<tr>
<td>Task 5</td>
<td>1.33 (0.37)</td>
<td>0.80–2.01</td>
</tr>
<tr>
<td>Task 6</td>
<td>2.21 (0.72)</td>
<td>1.33–3.27</td>
</tr>
<tr>
<td>Task 7</td>
<td>1.75 (0.56)</td>
<td>0.91–2.72</td>
</tr>
<tr>
<td>Task 8</td>
<td>1.69 (0.57)</td>
<td>0.76–2.61</td>
</tr>
</tbody>
</table>

### Figure 4: Overview of modified SUS rating table with inserted value ranges.

<table>
<thead>
<tr>
<th>ADJETIVE</th>
<th>Grade</th>
<th>Scale</th>
<th>Range</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Worst</td>
<td>WORST</td>
<td>0</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>Imaginable</td>
<td>IMAGINABLE</td>
<td>10</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>Poor</td>
<td>POOR</td>
<td>20</td>
<td>30</td>
<td>20</td>
</tr>
<tr>
<td>OK</td>
<td>OK</td>
<td>30</td>
<td>50</td>
<td>30</td>
</tr>
<tr>
<td>GOOD</td>
<td>GOOD</td>
<td>50</td>
<td>60</td>
<td>50</td>
</tr>
<tr>
<td>Excellent</td>
<td>EXCELLENT</td>
<td>60</td>
<td>70</td>
<td>60</td>
</tr>
<tr>
<td>IMAGINABLE</td>
<td>IMAGINABLE</td>
<td>70</td>
<td>80</td>
<td>70</td>
</tr>
<tr>
<td>BEST</td>
<td>BEST</td>
<td>80</td>
<td>90</td>
<td>80</td>
</tr>
<tr>
<td>Acceptable</td>
<td>ACCEPTABLE</td>
<td>90</td>
<td>100</td>
<td>90</td>
</tr>
<tr>
<td>Minimal</td>
<td>MINIMAL</td>
<td>100</td>
<td>110</td>
<td>100</td>
</tr>
<tr>
<td>High</td>
<td>HIGH</td>
<td>110</td>
<td>120</td>
<td>110</td>
</tr>
<tr>
<td>Low</td>
<td>LOW</td>
<td>120</td>
<td>130</td>
<td>120</td>
</tr>
<tr>
<td>Acceptable</td>
<td>ACCEPTABLE</td>
<td>130</td>
<td>140</td>
<td>130</td>
</tr>
<tr>
<td>Not Acceptable</td>
<td>NOT ACCEPTABLE</td>
<td>140</td>
<td>150</td>
<td>140</td>
</tr>
</tbody>
</table>

Interpretation of Task Performance Results, Satisfaction, and Demographic Trends
Possible reasons that Tasks 3 and 4 (updating and exporting a glucose value) proved difficult might be related to the tasks, to user practice, and also to needed redesign. These two tasks were more complex, requiring several steps, while tasks 1 and 5 were more straightforward, with only one step each. Another possible reason is that this usability test assessed users while they were still practicing these new, more complex tasks.

The overall satisfaction results indicated good, although not excellent, usability. In fact, nearly one-third of the participants gave the system a rather poor usability rating (“OK”). The results indicate areas for system improvements. For example, users were confused about the multiple steps and current, nonintuitive steps in Tasks 3 and 4. Although this task ought to be straightforward, the current design is not, requiring users to navigate to different areas of the system to delete and enter new values. Clearly, the multiple steps can be streamlined, so this usability test presented objective data about these current, cumbersome tasks.
The study used methods recommended in the ISO standard to assess effectiveness, efficiency and satisfaction with validated tools such as the SUS instrument. Together, these methods provide an increased understanding of system usability and could serve as an exemplar for methodological approaches by designers and researchers.

The present study addressed a gap in the literature by examining patients with varied characteristics and skill levels on a suite of performance outcomes. Trends such as those seen herein can help developers interpret user needs in designing more usable mHealth systems. Trends in participant socio-demographic and IT knowledge data also indicated that gender, age, disease duration, and IT experience level may influence interaction outcomes. To increase the scale of mHealth use to promote wider use and wider acceptance, these kinds of user characteristics will likely need to be considered more thoughtfully in system design in the future.

FUNDING
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

COMPETING INTERESTS
The authors have no competing interests to declare.

CONTRIBUTORS
Conception and design: M.G.; data collection: M.G.; analysis of the data: M.G., N.S.; interpretation of the data: M.G., N.S.; drafting of the article: M.G.; and critical revision of the article for important intellectual content and final approval of the article: M.G., N.S.

ACKNOWLEDGEMENTS
The authors thank Korey Capozza, MPH, Consumer Engagement Director at HealthInsight, for access to the system for evaluation and for logistical support prior to and during the study.

LIMITATIONS
Our study involved a smaller user sample from a specific diabetes population, which may make it difficult to generalize findings to the general population. Although these users were randomly selected from a larger database of a mHealth diabetes intervention study, the larger study used a convenience sampling technique. The sample size of 10 subjects is appropriate for usability testing and is greater than sample sizes recommended by Nielsen and Landauer, and Virzi. However, these users may not be representative of all mHealth users of diabetes systems. In particular, the participants in this randomized sample were more highly educated than most samples of patients with diabetes, especially those in urban areas. Future research in the area could explore other mHealth technologies as well as repeating this study on a greater scale with larger, randomly selected user samples to determine performance and satisfaction outcomes. Other user characteristics could be compared to these performance measures. In summary, this system focused on diabetes specifically, but the findings may also have broader applications for product designs of other chronic disease applications and concomitant usability testing.

CONCLUSION
The purpose of this study is to act as an exemplar for systematic usability methods. While results showed good (not excellent) perceived usability satisfaction, nearly one-third of users rated the system as having rather poor usability. The results do show objective data for developers and point to needed corrections, especially for confusing tasks. The variations across users and tasks are notable. Objective data are important because designer and researcher perceptions about tasks and user performance may not be congruent with performance data. In addition, system change lists can then be data-based to determine priority corrections.

This study demonstrates the use of a thorough quantitative approach by taking into account varied needs of users who interact with mHealth systems for disease management. It also shows the usefulness of performing several different kinds of assessment measures.

### Table 3: Comparison between user characteristics and usability metrics

<table>
<thead>
<tr>
<th>User characteristics</th>
<th>Success rate, mean (SD)</th>
<th>Error rate, mean (SD)</th>
<th>Time on task, mean (SD)</th>
<th>Satisfaction, mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>93.75 (12.50)</td>
<td>2.25 (2.63)</td>
<td>1.68 (0.39)</td>
<td>83.13 (2.39)</td>
</tr>
<tr>
<td>Female</td>
<td>89.58 (12.29)</td>
<td>4.17 (2.71)</td>
<td>2.58 (0.41)</td>
<td>79.69 (14.98)</td>
</tr>
<tr>
<td>Age (years)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30–49</td>
<td>100 (0.00)</td>
<td>1.00 (1.00)</td>
<td>1.95 (0.27)</td>
<td>88.33 (8.04)</td>
</tr>
<tr>
<td>50–69</td>
<td>87.50 (12.50)</td>
<td>4.43 (2.57)</td>
<td>2.34 (0.68)</td>
<td>77.14 (11.50)</td>
</tr>
<tr>
<td>Diabetes (years)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0–4</td>
<td>100 (0.00)</td>
<td>1.50 (1.14)</td>
<td>2.05 (0.30)</td>
<td>88.13 (6.57)</td>
</tr>
<tr>
<td>&gt;5</td>
<td>85.42 (12.29)</td>
<td>4.67 (2.73)</td>
<td>2.34 (0.74)</td>
<td>75.42 (11.56)</td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High school</td>
<td>93.75 (8.84)</td>
<td>3.50 (0.71)</td>
<td>2.39 (0.23)</td>
<td>81.25 (12.37)</td>
</tr>
<tr>
<td>College/university</td>
<td>90.63 (12.94)</td>
<td>3.38 (3.07)</td>
<td>2.18 (0.67)</td>
<td>80.31 (12.13)</td>
</tr>
<tr>
<td>IT/Comp. Experience</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Less experienced</td>
<td>87.50 (17.68)</td>
<td>6.50 (3.54)</td>
<td>2.71 (0.22)</td>
<td>76.25 (19.44)</td>
</tr>
<tr>
<td>More experienced</td>
<td>92.19 (11.45)</td>
<td>2.63 (2.07)</td>
<td>2.10 (0.61)</td>
<td>81.56 (10.43)</td>
</tr>
<tr>
<td>Sample (n = 10)</td>
<td>91.25 (11.86)</td>
<td>3.40 (2.72)</td>
<td>2.22 (0.60)</td>
<td>80.50 (11.47)</td>
</tr>
</tbody>
</table>
REFERENCES


AUTHOR AFFILIATIONS

1Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah, USA
2Faculty of Computing, Blekinge Institute of Technology, Karlskrona, Sweden
3School of Nursing, University of Maryland, Baltimore, Maryland, USA

RESEARCH AND APPLICATIONS
Study II

An evaluation of patients’ experienced usability of a diabetes mHealth system using a multi-method approach

Mattias Georgsson, Nancy Staggers

Article history:
Received 8 June 2015
Revised 14 October 2015
Accepted 19 November 2015
Available online 27 November 2015

Keywords:
Usability
mHealth
Diabetes
Multi-method evaluation
Framework analysis
Usability Problem Taxonomy

Abstract

Objective: mHealth systems are becoming more common to aid patients in their diabetes self-management, but recent studies indicate a need for thorough evaluation of patients’ experienced usability. Current evaluations lack a multi-method design for data collection and structured methods for data analyses. The purpose of this study was to provide a feasibility test of a multi-method approach for both data collection and data analyses for patients’ experienced usability of a mHealth system for diabetes type 2 self-management.

Materials and methods: A random sample of 10 users was selected from a larger clinical trial. Data collection methods included user testing with eight representative tasks and Think Aloud protocol, a semi-structured interview and a questionnaire on patients’ experiences using the system. The Framework Analysis (FA) method and Usability Problem Taxonomy (UPT) were used to structure, code and analyze the results. A usability severity rating was assigned after classification.

Results: The combined methods resulted in a total of 117 problems condensed into 19 usability issues with an average severity rating of 2.47 or serious. The usability test detected 50% of the initial usability problems, followed by the post-interview at 29%. The usability test found 18 of 19 consolidated usability problems while the questionnaire uncovered one unique issue. Patients experienced most usability problems (8) in the Glucose Readings View when performing complex tasks such as adding, deleting, and exporting glucose measurements. The severity ratings were the highest for the Glucose Diary View, Glucose Readings View, and Blood Pressure View with an average severity rating of 3 (serious). Most of the issues were classified under the artifact component of the UPT and primary categories of Visualness (7) and Manipulation (6). In the UPT task component, most issues were in the primary category Task-mapping (12).

Conclusions: Multiple data collection methods yielded a more comprehensive set of usability issues. Usability testing uncovered the largest volume of usability issues, followed by interviewing and then the questionnaire. The interview did not surface any unique consolidated usability issues while the questionnaire surfaced one. The FA and UPT were valuable in structuring and classifying problems. The resulting descriptions serve as a communication tool in problem solving and programming. We recommend the usage of multiple methods in data collection and employing the FA and UPT in data analyses for future usability testing.

1. Introduction

Diabetes is a condition affecting 29.1 million people in the United States [1] with concomitant health care expenditures of an estimated $198 billion [2]. The most common form of diabetes is Type 2 (T2DM) affecting a majority, 90%, of those with the condition [3]. A healthy diet, regular physical exercise, and maintaining a normal body weight in addition to medication treatment are seen as important [3] whereas poorly regulated glycemic measurements and poor self-management practices are contributors to a worsening condition [4,5]. Due to its complexity, Type 2 diabetes puts heavy demands on both patients and providers [6]. Different types of support systems for diabetes self-management have been developed recently using Information
and Communication Technology (ICT) [7,8]. Studies show that mobile health (mHealth) technology has been particularly successful in improving glucose management for these patients [9–11] and has aided in regulating lifestyle changes [11]. Recent interventions having beneficial effects are, for example, those focusing on self-coaching by allowing users to interact with algorithm-based systems that provide individually tailored messages [12] and feedback messages with automatically updated text messages based on clinical guidelines and patients’ lifestyles [13]. Although mHealth applications have been beneficial in diabetes self-management, researchers mention the need for improved usability to allow patients to feel confident in their interactions with mHealth applications [14]. Unfortunately, relatively few studies exist on mHealth usability [15]. Of the available studies, most lack methodological rigor [16,17] and use a single method for data collection [17,18]. Also, qualitative data analyses in mHealth studies typically lack a structured approach, making study reproducibility difficult, especially when using multi-method approaches for large data sets with multiple researchers. No usability studies are yet available using the Framework Analysis (FA) method or the Usability Problem Taxonomy (UPT) to address this gap. Our approach acknowledges these current limitations and tests the feasibility of a multi-method approach for data collection and novel, structured techniques for data analyses.

Methodological improvements in mHealth usability studies could result in more comprehensive identification of usability issues, more specific redesign recommendations based upon user-centered data, more reliable data analyses and they could potentially improve the reproducibility of results across studies. For users, improved mHealth designs could result in improved interactions, user performance, adoption of mHealth applications and perhaps even increased adherence to suggested interventions for chronic care conditions.

1.1. Review of the literature

Recent reviews of mHealth usability studies point to the need for more rigorous studies and the use of multiple methods to help validate findings [16,17]. Authors commented that each usability testing method makes its own unique contribution to the overall identification of usability issues [17]. Thus, a robust usability evaluation should employ a combination of methods [16,17]. Specifically, Zapata et al. emphasized the need for (1) employing two or more different types of usability methods, (2) using more standardized methods and tools in mHealth evaluations and (3) using open-ended inquiry methods and qualitative analysis to identify more comprehensive user concerns and recommended that future researchers conducting usability evaluations need to standardize processes and use a combination of more than one method [17,18].

1.1.1. Multi-method approach to data collection

Multiple methods in data collection can increase the depth of inquiry while improving the reliability and validity of findings [19]. The use of multiple methods also assists with methodological triangulation because it can allow for a more comprehensive understanding of a phenomenon [20,21]. This type of triangulation was first identified by Patton who advocated data collection using observation, field notes and interviews [20,22]. These kinds of qualitative methods produce rich data [20,23] and are especially suitable in health informatics research [24]. Pertinent to mHealth, multiple methods in data collection could allow for a more comprehensive understanding of the mHealth user experience and pin-point specific usability problems for redesign.

In the current study, we test the feasibility of overcoming common methodological issues in data collection during a mHealth usability study. We completed a multi-method approach to data collection with user-based testing, Think Aloud protocol, open-ended interviews and a short answer questionnaire.

1.1.2. Structured approaches for data analyses

The main shortcomings of qualitative research methods are a lack of standardization in data analytic techniques, meaning that replicating the results of analyses can be difficult [25]. Two methods can assist in overcoming this issue, the Framework Analysis (FA) method and the Usability Problem Taxonomy (UPT) classification scheme. Neither has yet been employed in mHealth or other health informatics usability evaluations to our knowledge.

The FA, originally created by Richie and Spencer in 1994, is a standardized yet somewhat flexible framework for analyzing qualitative data. Its purpose is to assist with the analysis of descriptive, textual source data to produce reliable and valid qualitative results [26,27]. Importantly, the FA is seen as scientifically robust [28]; it has been used in social sciences research and to a lesser extent in health care research [28,29]. Recent applications of the FA have been mainly in health care (nursing research) and in multidisciplinary studies to manage large sets of qualitative data [28,29].

The FA consists of five stages (1) familiarizing oneself with the data, (2) identifying a thematic framework to be used, (3) indexing and applying the framework to the data, (4) charting the data and (5) mapping and interpretation [26]. This analysis method is especially useful in organizing, reducing, and interpreting data because it provides clear steps to follow and it produces more structured output with summarized data [29]. Moreover, the process can be used by several researchers simultaneously [28] such as multi-disciplinary team members [29]. The FA promotes, in particular, data/decision transparency [28]. It is useful for both experienced and less experienced coders due to its available audit trail [28]. The framework can accommodate different analytic tools such as paper, post-it notes, Microsoft Word, or NVIVO [28]. Primary difficulties with this method are those usually inherent to qualitative analysis e.g., it is time consuming and requires a committed manner of analysis [29].

Finding ways to structure and classify usability problems has been explored within the Human–Computer Interaction (HCI) field. Keenan et al. [30] found no framework for classifying usability problems on an individual problem level, particularly using various perspectives that would allow problems to be compared, analyzed and described. Such an analysis would provide better information for problem correction [30]. These authors thought a new approach would either have a trained individual examine the whole set of usability problems to look for trade-offs, contradictions, and consistency issues [31], or an expert might think about the problems from a more global perspective to look for problem clusters [32]. Keenan et al. considered using heuristic analysis [33,34], but found that technique lacked sufficient problem distinguishability, mutual exclusiveness and specificity [35]. Thus, the UPT was built empirically using over 400 usability problem descriptions collected on real-world development projects [30].

UPT is a classification scheme and framework for characterizing usability problems according to their dimensions, providing a clear structure in usability problem definition. It was initially used to classify usability problems found on graphical user interfaces with textual components where usability problems were easily detected, classified, and analyzed [30]. The UPT is based on the notion that usability problems should be examined from two perspectives: the task-artifact approach proposed by Carroll et al. [36] and the Object-Action Interface Model by Shneiderman [37] to enhance problem definition. The artifact component defines usability problems arising when the user interacts with the interface while categories under task component focus on usability problems that surface when a user moves through a task [30]. Problem
classification occurs from both an artifact and a task component perspective and is then divided into five primary categories: Visualness, Language and Manipulation, Task-mapping and Task-facilitation (see Fig. 1).

The UPT classification could also aid developers in several ways (1) the categories are based on problem characteristics versus only problem type, meaning classifications occur on two levels (artifact and task); (2) UPT offers a method of identifying problem clusters using different ways and levels of categorizing problems, (3) it is beneficial for examining problem sets at varying levels of abstraction, and (4) problems distributed across UPT categories make visible the kinds of issues encountered most often. It can thus be used by developers to assess problem scope and frequency as well as defining the types of problems that both global and local solutions may be considered. The UPT has been used only sparingly since its creation, but its classification methods could provide an excellent description of the dimensions in detected usability problems. We concluded the UPT is a lesser known but robust technique.

1.2. Study aim

Our purpose was to test the feasibility of (1) using a multi-method approach for data collection during a mHealth usability evaluation and (2) applying structured approaches to data analyses by using the FA method and UPT. These comprise a novel, multi-method approach to data analyses and, to our knowledge, are the first applications of both FA and UPT in health informatics research.

2. Materials and methods

2.1. The evaluated system

The system evaluated in this study is an interactive, SMS-based mobile intervention for patients with diabetes, designed as a personalized self-care management tool. It is a commercially available tool in current use in several organizations. The system consists of a combined mobile phone solution and web service. Patients interact by either sending in or being prompted to send in their Type 2 diabetes, self-management values, e.g., morning blood glucose via text message. Using the web service, they can enter and/or review their results across various parameters. Patients can, for example, obtain their glucose readings or blood pressure values visualized in a meter format, see medication adherence levels, exercise and weight progress and view scheduled medical appointments. When sending in their measurements patients receive personalized coaching and tailored responses from the system to track the progress of their disease. The web portal contains different features, main views and sub-views for patients to perform different actions and task results (see Fig. 2). In this study we concentrated our evaluation mainly on the web service part of the solution because of its inherent complexity and because it had no previous usability evaluation.

2.2. Study sample and setting

Institutional Review Board approval was obtained (WIRB, Olympia, Washington, USA). A larger randomized controlled trial for a diabetes mHealth intervention study involving 18 primary care clinics in the Salt Lake City metropolitan area served as the study population for our usability study. A set of 2317 patients met the inclusion criteria for the larger study [38]. Ten patients were randomly selected from this larger randomized controlled trial on the mHealth intervention by computer randomization and invited to take part in our usability evaluation. Inclusion criteria for our usability study sample included (1) patients diagnosed with Type 2 diabetes (2) no cognitive impairment; (3) familiarity and some knowledge and use of computers, the Internet, and cell phone; and (4) the ability to speak and understand the English language. Patients had no previous exposure to the mHealth web system evaluated in our study. The evaluation sessions were conducted in a quiet laboratory setting at HealthInsight, a U.S. Beacon Community in Salt Lake City, Utah.

2.3. Data collection methods

Usability can be evaluated by several different methods. Inspection methods such as heuristic evaluation [39,40], and cognitive-walkthrough [41,42] are expert methods meaning that experts go through the system to identify usability issues. Think Aloud protocol [43,44] is a user-related method for evaluating usability where users express their perceptions out loud as they interact with the system. Other user-related methods include administering

Fig. 1. Classification of categories according to the Usability Problem Taxonomy [30].
Appendix A

The main goal of user-related methods is to involve actual users in the evaluation and obtain their perspectives [45], to gain insight into how the intervention needs to be adapted for different users’ abilities and experience levels [46,47] and to identify usability problems for correction.

2.3.1. Usability test with think aloud protocol

Think Aloud is a usability assessment method commonly employed to determine users’ thoughts and opinions while they perform a list of specified tasks with a system. The method originated in 1984 in psychology when Simon and Ericsson thought of verbal reports as data. Revised in 1993 [48], the technique is well established within the Human Factors field [39]. Think Aloud asks users to talk aloud during their interactions, to express their reactions and thinking and to explain what they are doing as they perform specific, representative tasks [24]. The resulting data are normally audio- and/or video-recorded and/or an observer takes thorough, written notes [49]. Minimal intervention from the usability tester assures users’ thought processes are not interrupted except to remind them to keep talking [49]. The focus is on understanding users’ decision making processes and on how users experience the system in their own words [39,49]. Because the method provides extensive, detailed data, only a small sample of five to eight users is needed in usability testing to detect 80–85% of usability problems [50–52] to gain a thorough understanding of task behavior [48] and to identify the main usability problems [53]. Representative tasks for the specific domain are also essential, and they should be as realistic as possible [54]. Also, testing should be conducted in the actual user’s context or one as close to the natural environment as possible [49,55]. Authors indicate that Think Aloud provides complete and detailed descriptions of patients’ thought processes during system interactions and the technique generates many usability problems [56].

2.3.2. Representative tasks

Tasks for this study were based on common patient user interactions with the system; they were disease-specific and had varying levels of difficulty to simulate patient usage in a clinic or at home. Tasks were validated by a panel consisting of a physician and a nurse whose specialties were diabetes, a public health professional with chronic patient intervention systems expertise, and a diabetes patient. The specific tasks patients had to perform consisted of (1) uploading glucose values into the system, (2) interpreting a glucose measurement in a graph view, (3) correcting a recorded glucose measurement value, (4) exporting glucose measurement value trends to a PDF to simulate material to take to a provider visit, (5) interpreting a blood pressure measurement in a graph view, (6) setting personal tracking goals for exercise and weight, (7) setting medication reminders, and (8) setting a physician appointment reminder.

2.3.3. In-depth interview and usability questionnaire

Both interviews and questionnaires have been used extensively in usability research to determine users’ opinions about the difficulties they experience in an evaluated system [49]. In our study patients completed both an open-ended interview and a post-experience questionnaire [57]. The open-ended interview guide asked users to talk about aspects of the system with good or poor usability. This kind of interview is especially useful in uncovering comprehensive information from participants [58]. The three questions in the interview asked patients to comment on sections of the system they thought were well designed, sections that were inadequately designed and any further comments they might have about system usability (see Appendix A).

Patients also completed a post-interaction questionnaire on the mHealth system (see Appendix B). The first section consisted of short answer questions about: patients’ IT/computer, mobile phone, and internet experience and use; their experience and perceptions about web and mobile service systems in health care; what they thought about the specific mHealth system evaluated in the session in terms of usability and any further comments they might have about these topics. Patients were also asked to rate their preferences in technology usage for work and leisure time using a Likert scale with 4 points ranging from strongly agree (1) to strongly disagree (4).

The second part of the post-questionnaire included open-ended questions about the specific system patients used in this study. Patients specified in writing their thoughts about what they found difficult and easy about the system and its navigation, they listed usability/user experiences they found satisfying or dissatisfying and then a final question asked for any further comments about the system.

Both the interview guide and post-experience questionnaire were assessed for face validity by a panel of three health care
professionals, three usability experts and one patient. The resulting format was finalized via discussion and consensus.

2.4. Study procedure

First, patients were asked for informed consent. Then, patients were walked through the different steps of the evaluation procedure and asked if they had any questions. The evaluation started with patients filling out a brief demographic questionnaire. Topics included age, gender, educational level, occupation, and how long they had been diagnosed with diabetes.

Next, standardized training was performed to simulate an actual patient educational process in a health clinic. This was important to decrease individual variability and to ensure that patients all had the same information about the system [59] because none had used the system before. After the training was completed by the first author, users interacted with the system on their own to get familiar with it for an average of about 10 min.

The Think Aloud and usability evaluation session was conducted by the first author. Patients were given a booklet outlining the specific tasks to perform. During the session patients were asked to think aloud as they completed the prescribed tasks in the system. If they became silent, the researcher encouraged them to think aloud as they completed the prescribed tasks in the specific tasks to perform. During the session patients were walked through the different steps of the evaluation procedure and asked if they had any questions. The evaluation started and the Usability Problem Taxonomy (UPT) was avoided. Interactions were digitally audio-and video recorded using Morae software [60]. The recording showed patients’ navigation on screens, their facial expressions and captured their voices. The researcher made observations and notations about the individual task performances directly into Morae.

Afterwards, patients were interviewed about their experiences and perceptions about the system, by the first author, using the in-depth interview guide with interview topics. Patients were provided the opportunity to express freely what they thought was easy and difficult with the system and add any further comments. This session was also digitally audio-recorded and lasted for 15–20 min. Finally, patients completed the post-test questionnaire on their perceptions about the usability of the system.

The complete testing procedure for all the steps averaged approximately two hours with a range of 1.5–2 h. Patients received a gift card for $20 after completing the session.

2.5. Data analyses steps utilizing the Framework Analysis (FA) method and the Usability Problem Taxonomy (UPT)

The audio-and video recordings from the sessions, Think Aloud comments and observations, post-test interview and post-test questionnaire usability data were transcribed, checked for accuracy and imported into Nvivo 10 Qualitative Data Analysis Software [61]. Data analyses included content analysis in addition to applying the FA and UPT.

Coding and analysis included the five steps in the FA method [26] (1) data familiarization, (2) identifying themes and framework used and (3) classifying usability problems using the UPT, (4) the results were organized or charted into the place of occurrence within the mHealth application and then (5) mapped and interpreted. Descriptive statistics were used to summarize data.

The first author transcribed, imported the data and completed the initial coding which were verified by the second author. Both authors conducted step 2–5 in the analysis together.

Framework Analysis and Usability Problem Taxonomy process

![Diagram of the Framework Analysis and Usability Problem Taxonomy process](image)

Fig. 3. Application of the FA method and UPT.
See Fig. 3 and the accompanying description of the analysis process below.

1. Familiarization with the data

The first step was to become familiar with the transcribed textual data through immersion. This occurred by reading the uploaded textual documents several times.

2. Identifying the themes and/or framework to be used

The second step was to identify themes and apply a framework to code the key issues. In the original FA method, researchers can select either an inductive or deductive content analysis approach to perform the coding. We modified the FA method slightly. We used FA to generate usability problems inductively across the various data collection methods, but we used the UPT deductively as well to classify the problems (see Fig. 3). The usability problems can be considered themes derived from the data. After a discussion and removal of duplicates, we consolidated the usability issues under their appropriate tasks. The application of the UPT is described in more detail below, and we provide examples of two coded usability problems in Table 1.

3. Indexing and applying the classification to the data

The third step consisted in indexing and applying the UPT classification framework as well as assigning severity ratings to our final usability problem list. We classified each usability problem into an artifact component (Visualness, Language and Manipulation) and/or a task component (Task-mapping and Task-facilitation) preceding in the classification scheme as far as possible. The categories and subcategories within the artifact and task components at any level are mutually exclusive resulting in one final categorization [30]. An example of a classified usability problem from our data is shown in Table 2.

For the severity rating, we used a process defined by Travis that asks three questions about each usability problem (1) Does the problem occur on a red route; i.e., is it a frequent or critical task the system needs to support? (2) Is the problem difficult for users to overcome? and (3) Is the problem persistent and does it keep recurring [62]? Ratings were assigned using the scale (1) Low, (2)

<table>
<thead>
<tr>
<th>Usability problem description</th>
<th>Sample comments and notes</th>
<th>Questionnaire</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Difficult knowing to exit the table view and click the &quot;Add data&quot; button to be able to add in a new value</td>
<td>&quot;first you have to delete it and then you have to exit out of that back to the dashboard and then click on add if I remember right...but that seemed a little awkward, like excess steps, too many steps&quot;</td>
<td>Going from screen to do one procedure was a bit frustrating</td>
</tr>
<tr>
<td></td>
<td>&quot;There should be a button somewhere here that would tell you how to add...but there is no...I can’t find a button here that says to add...so it’s a little bit difficult...&quot;</td>
<td>When changing/updating information I need to go into a few other screens to do so. It was not simple to do</td>
</tr>
<tr>
<td>(9) Difficult to know to select &quot;Delete data&quot; for the table view to change, delete, export, or print the value list</td>
<td>&quot;I was going to export data but it said to delete when I see delete it’s like a big red flag you know you always hope that you...that you get that question are you sure you want to do this (?) because once it’s gone it’s pretty much gone.&quot;</td>
<td>The wording was difficult like having to go to &quot;delete&quot; when I was to export</td>
</tr>
<tr>
<td></td>
<td>&quot;I don’t want to delete it, I want to print it...so choosing the wrong terminology&quot;</td>
<td>When exporting data I needed to click the &quot;delete&quot; button. It should be a simple Export button to select</td>
</tr>
<tr>
<td></td>
<td>(Observer comment/s): The subject comments this before clicking the delete data button before exporting the values on the Glucose Reading View)</td>
<td></td>
</tr>
</tbody>
</table>
medium, (3) serious or (4) critical for each problem [62]. The resulting severity scores were averaged per system view and for the whole system.

(4) Charting the data

This step consisted of abstracting our final list of usability problems back into their original context as is consistent with the FA method. We arranged the order according to the problem’s place of occurrence, classification and severity level. Descriptive statistics were used to summarize issues within each method and per patient.

(5) Mapping and interpretation of the data

The final step involved mapping and interpreting the usability problems. After the resulting list of problems were charted, we were able to identify the most prevalent problems and their severity ratings in their respective views. By doing this, we could determine the nature of the problems and what their classifications implied. The latter served as guidance for designers in correcting specific usability issues.

3. Results

The sample of 10 patients included six women and four men (see Table 3). Most were between the ages of 40–59 with a range of 40–69. The majority were university educated and half were employed. Sixty percent had a Type 2 diabetes diagnosis of five or more years.

The sample used technology regularly. For example, 80% used a computer daily, 70% used the internet daily, and 90% indicated they used their mobile phone to make and receive calls every day. A majority agreed or strongly agreed that using IT/Computers and mobile phones in health care is a positive development and thought it was positive for their own diabetes self-management.

3.1. Usability problems

A total of 117 initial usability problems were detected by the ten patients across the different data collection methods where 59 problems were detected during the usability test, 34 in the post-interview and 24 using the questionnaire. The problems were consolidated into a list of 19 unique usability issues (see Appendix C). The average severity rating for the whole site was 2.47 or serious.

Sample problem descriptions are presented in Table 4. They are organized according to the most critical issues and listed in their place of occurrence.

The usability problems and severity ratings were the highest for the Glucose Readings View which had eight (consolidated) problems and a severity rating of three. The Glucose Diary View and Dashboard had three problems each. The Glucose Readings View, the Glucose Diary View and the Blood Pressure View all had severity ratings of three. The views with the highest number of problems were those with several steps in a task. These were vitally important diabetes management functions; they received high severity ratings (Fig. 4).

3.1.1. Critical and severe usability problems

The two most critical usability problems (severity level 4) were located in the Glucose Readings View and the Glucose Diary View. These required correcting a glucose value by removing the erroneous value and adding a new one. To complete the task, patients had to exit the Glucose Readings View, navigate to the Glucose Diary View and add the new value by clicking the Add data button. This sequence was confusing. Patients were also confused about the
Table 4
Example of usability problems and classifications for the most severe issues.

<table>
<thead>
<tr>
<th>Usability problem description</th>
<th>Place of occurrence</th>
<th>% of pat. detecting per method</th>
<th>UPT classification</th>
<th>Severity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficult knowing to exit the table view and click the “Add data” button to be able to add in a new value</td>
<td>Glucose readings view</td>
<td>70 100 50</td>
<td>Manipulation-Cognitive aspects-Visual cues (FC)</td>
<td>4</td>
</tr>
<tr>
<td>Difficult to know how to select “delete data” for the table view to change, delete, export, or print the value list</td>
<td>Glucose Diary View</td>
<td>70 100 100</td>
<td>Language-Naming/labeling (FC)</td>
<td>4</td>
</tr>
<tr>
<td>Difficult to understand and perform the exporting action</td>
<td>Glucose Readings View</td>
<td>30 40 40</td>
<td>Manipulation-Cognitive aspects-Visual cues (FC)</td>
<td>3</td>
</tr>
<tr>
<td>The blood pressure graph shows only the systolic blood pressure value, and the diastolic blood pressure is missing</td>
<td>Blood Pressure View</td>
<td>10 20 –</td>
<td>Task mapping-Interaction (FC)</td>
<td>3</td>
</tr>
<tr>
<td>Difficult to know how to adjust the range of values that are to be shown, retrieved</td>
<td>Glucose Diary View</td>
<td>90 – –</td>
<td>Task mapping-Functionality (FC)</td>
<td>3</td>
</tr>
<tr>
<td>No action support to specify where the exported file is to be saved</td>
<td>Glucose Readings View</td>
<td>20 – –</td>
<td>Visualisation-Non message feedback (FC)</td>
<td>3</td>
</tr>
<tr>
<td>Difficult to know how to navigate within and adjust the table view to get to or show the right value range, value</td>
<td>Glucose Readings View</td>
<td>10 – –</td>
<td>Task mapping-Functionality (FC)</td>
<td>3</td>
</tr>
<tr>
<td>A confirmation is missing to indicate that the exported file was saved and its directory</td>
<td>Glucose Readings View</td>
<td>10 – –</td>
<td>Task facilitation-Keeping the user on track (FC)</td>
<td>3</td>
</tr>
</tbody>
</table>

* TA = Think Aloud usability test, I = Interview, Q = Questionnaire.

Fig. 4. Distribution of usability problems and assigned severity ratings across places of occurrence.

Delete data button as it had a dual function of allowing data to be exported and printed in the Glucose Diary View. The latter was an example of a label that did not make sense to them. These problems were classified into both (artifact) Manipulation, Cognitive aspects/navigation, and Naming/labeling deficiencies. These were issues brought up most frequently by patients across the different data collection methods.

Other serious issues (level 3 severity ratings) were in the Glucose Readings View, the Glucose Diary View and the Blood Pressure View. Several examples illustrate these issues. Problems occurred in the Glucose Readings View for the exporting task which included many difficult, hard-to-understand steps. The logic of exporting was not apparent, and system support was not available for this process. Then, if patients discovered how to export a file, they could enter a file name to save the exported file, but they were unable to specify where the file would be saved. Patients also had difficulty adjusting the range of values in the Glucose Diary View and its graph. In another example, patients experienced difficulty navigating within the table in the Glucose Readings View (which allowed the display of correct value ranges or values). No navigation support was available. The Blood Pressure View was notable because only the systolic blood pressure was visible in the graph although patients also entered their diastolic blood pressure. This was an incomplete and potentially harmful display.

Other difficulties included the various ways of accessing reminders in the Dashboard View due to labels and multiple locations. The Medication Reminder could be accessed through the Medication Adherence tab, the Exercise, Weight and Medicine pane, the Message Settings and Medication Reminders pane. The Appointment Reminder had no separate tab but was listed instead under the menu item titled Message Settings and Reminders. The "Tracking Goal" (exercise and weight) item had three different paths: the Exercise and Weight Progress tab, the Exercise, Weight and Medicine pane and the Message Settings and Reminders pane.

3.2 Results for methodological approaches

The different methods contributed to the 117 different initial usability problems. The usability test detected the majority (59) or 50% of the problems followed by the in-depth interview which identified 34 or 29% of the problems. The questionnaire identified substantially fewer usability problems at 24. Further, when looking at how the methods performed in combination with one another, the usability test plus the in-depth interview detected the highest number of problems (93) at 80%, while the usability test and post-questionnaire uncovered 83 or 71% of the usability problems. The combined questionnaire and the in-depth interview identified only 58 or 50% of the problems. These data indicate that the usability test detected most of the usability problems both on its own and in combination with the other methods while the post-test questionnaire detected the least volume of problems on its own and in combination.

Fig. 5 shows the unique and shared consolidated 19 usability problems between the different methods as depicted in the Venn diagram. As can be seen, the usability test and in-depth interview allowed us to identify 18 and 9 of the resulting usability problems respectively. The usability test alone detected eight consolidated usability problems. This was more than the others at zero or one. All three methods converged on five usability problems.
3.3. Problem classifications

Fig. 6 indicates the distribution of problems for each UPT classification level and the depth of classification. The most critical usability problems were classified into the artifact components Manipulation, Cognitive aspects/Visual cues, and Visualness, Non-message feedback/Presentation of information/results. Task classifications involved a variety of functionality/navigational concerns as well as one non-classification. The additional problems were classified under the artifact components Language and Naming/labeling and task components Task-Mapping and Navigation. In terms of the distribution of problems, we were able to classify all problems into the artifact or task component or both. However, two problems in the artifact component and five problems in the task component could not be classified. Visualness received the highest volume of full classifications (7) in the artifact component while in the task component, 12 problems received a full classification in Task-mapping.

As can be seen in Fig. 7, most subcategories for the whole system related to the presentation and visualization of information. For the six total Manipulation problems, five were categorized in the subcategory of Cognitive aspect and further broken down into Visual cues and Direct manipulation. The Language problems all related to the Naming/labeling subcategory. Most of the Task-mapping problems were associated with Navigation and Functionality. This indicates that most usability problems were related to system design and structure as a whole as it was not logically organized from a user perspective.

4. Discussion

In this study we conducted an in-depth usability assessment of a mHealth system using a combination of data collection methods and by employing two new frameworks for data analyses to structure and classify data. The multi-method approach in data collection resulted in a more comprehensive understanding of patients’ interactions with the system and provided triangulation on severe usability issues. The structured data analyses provided an audit...
trail for analytic decisions and resulted in clear descriptions of the usability problems.

4.1. Usability problem types and classifications

A total of 117 usability problems were identified across the different methods and were consolidated into 19 final classified usability issues. The majority of problems were located in the Glucose Readings View and had the highest severity ratings (serious or 3). Two other views, the Glucose Diary View and Blood Pressure View, had fewer usability problems but also high severity ratings. Usability problems included issues with deleting and entering glucose values, exporting and printing a glucose value (including salient blood pressure values) and information about how to save a file.

Based on the UPT classifications, the majority of problems were deficient visualization of Information/results. Manipulation, Cognitive aspects and Visual cues also dominated the usability issue classifications, reinforcing the visualization issues and cognitive aspects for system content. The results pointed to difficulties patients had completing common tasks. These were actions patients need to perform on a frequent basis for diabetes self-management, so these concerns need to be resolved quickly.

4.2. Method contributions

4.2.1. Data collection methods

Using a multi-method approach is beneficial for identifying usability problems, especially initial usability problems. Results from one method can be verified by another [63,64], contributing to the role of triangulation [26]. In our methods feasibility study, user testing identified the vast majority of problems followed by the interview and then the questionnaire. User testing plus the interview and questionnaire triangulated on five and four shared (verified) problems, respectively. The interview did not contribute unique consolidated usability problems in this study while the questionnaire produced one unique consolidated finding. However, each method provided a lens to view different usability issues, especially initial usability problems where usability testing contributed 59, interviews 34 and questionnaires 24 problems and thus contributed to the overall results [66,67].

Think Aloud was essential in usability testing in our study as it detected 18 of 19 consolidated usability issues. This is consistent with past literature indicating that user testing identifies more problems, identifies more recurring usability problems, and defines the underlying causes for usability problems [68]. The in-depth interview contributed the next highest volume of issues and nine of the 19 consolidated problems. Its contribution in particular was for triangulation on identified usability problems although the method did not contribute any unique consolidated issues. The reason for the latter is not completely clear but it may be that patients echoed only the issues they voiced during testing. Perhaps these particular issues resonated with them, and they did not have insights into other, new problems during the interview. Open-ended interview questions are beneficial in that they provide an opportunity for the user to express thoughts freely and provide perspectives about problems that may not surface during user testing. Of course, open-ended questions can vary from study to study but the primary value of interviewing in the evaluation process lies in users expressing perspectives in their own words. Also, multiple lenses might generate more unique issues in a more complex application than the one studied here. In the future, researchers may want to follow-up the high-level questions we used with more specific probes to uncover unique issues during interviewing. Future researchers may also categorize issues in a more granular fashion than in this study.

The post-questionnaire generated fewer issues but verified significant usability problems. A reason for the lower volume of issues with the questionnaire might be due to study sequencing. We began with the usability test and ended with the questionnaire which meant that some problems reverberated during each step of the process. Also, questions could have seemed similar to users or perhaps they became fatigued at the end of the two-hour testing session. Perhaps users mentioned the same issues each method because they were seen as especially severe to them. This may indicate the value of using all three methods. The significance of the post-questionnaire may lie specifically in eliciting users’ views in writing as they had time to ponder and write about issues they thought were difficult.

Five shared usability problems were detected by all three methods. Four were critical issues while one was less so. Thus, different methods may surface different issues having varying levels of severity. The multiple methods produced a more comprehensive set of usability problems than did one individual method, although more so at the initial problem identification step. This notion is also highlighted by other authors in the health informatics technology literature. In particular, authors indicated that multiple evaluation methods are important because systems are becoming more complex [49]. Based on our results, we also recommend the use of multiple methods but advocate a minimum of three methods to capture a variety of issues of varying kinds and at differing severity levels. However, the levels of questions in the interview and questionnaire should be different. This recommendation will need to be tempered by the available time participants have for testing, e.g., few physicians will be available for two-hour usability evaluation sessions.

4.2.2. Structured data analyses

The FA method provided a beneficial way of structuring data from our multi-method approach and was particularly useful for coding large amounts of data as noted by previous authors [26]. Despite our need to modify FA to accommodate the UPT for usability problem classification, we found the FA very worthwhile overall. Researchers may find a similar modification for the FA suitable during future usability evaluations.

Sometimes vague problem descriptions or the problem nature meant a problem could not be fully classified in the UPT. As Keenan et al. noted, this can result in a null or partial (versus full) classification [30]. In our study several usability issues could only be classified in the artifact component level in the Naming/labeling category as they did not have sufficient descriptions that might lead to a deeper task level classification. Whether issues are fully or partially classified does not mean that issues are categorized as “better” or worse but merely that the description is more or less granular.

We found the UPT of significant value. Determining the UPT classification for each problem assisted in defining the usability problems on a more detailed as well as a hierarchical level. Similar to Keenan et al. [30], we found the resulting issues were well described overall, and we could group problems of a similar nature easily. We also found that adding a severity rating to the problems and aggregating the frequency of the problem would likely aid in problem analysis and future approaches for solutions.

4.3. Contributions to the literature

Our findings verify the usefulness of multiple data collection methods in generating more comprehensive lists of issues and serving as a means of triangulation for severe usability issues. Each method generated unique initial usability problems. For consolidated usability problems we were able to calculate the percentage of patients who raised a specific issue across the different methods,
providing an indication of its importance to users. This is a clear benefit as triangulation can assist developers in locating priority user problems. The multiple methods approach was somewhat less beneficial in identifying unique consolidated usability problems. Thus, the most significant contributions of our study lie in the data analysis methods. The FA method was useful for structuring qualitative data as it assisted in standardizing problem descriptions and promoted consistent analyses across researchers. Likewise, the FA was valuable because it (1) provided a systematic way of managing and mapping the data [26], (2) created reproducible steps [29], and (3) supported both our inductive (usability issue coding) and deductive (UPT) approaches [23]. By combining the FA method with the UPT framework we were also able to create an in-depth, easy-to-follow classifications for each identified usability problem. We think this will aid in solution generation and could contribute to problem-solving for designers [30].

Overall, this study adds to the body of literature because it provides an example of a multi-method approach for collecting data on usability problems [18], for standardizing data analyses, and contributing to the reproducibility of results in qualitative usability studies. This more standardized and structured approach could assist in building a science around mHealth usability.

4.4. Lessons learned and recommendations

We found the FA analysis method valuable because the multi-method design generated a large amount of data even with a sample of only 10 users. However, coding using both FA and UPT was still a laborious process. Its utility is in generating a clear audit trail and decision path for data categories (usability problem classifications in our case).

The FA method is recommended for studies employing several researchers because of its more structured process [26]. We found it effective with just two researchers because of the linear steps in the coding process and the resulting descriptions, allowing researchers to work together in data analysis. We also found it advantageous to have detailed usability problems that we could then classify using UPT.

The UPT was also beneficial for classifying usability by the artifact and task components. However, it was helpful to have two researchers to discuss classifications before reaching consensus. Keenan et al. [30] indicate the UPT is easy to use for individuals familiar with the terminology in HCI and those with some background in usability. Classification is also easier when the researcher is familiar with the particular context of use of the system. The two researchers in this study fit those criteria. Future researchers will want to take these criteria into consideration before analyses.

Once the analysis process was completed, we had a list of defined and classified usability problems with severity ratings that could be aggregated into similar UPT categories. We recommend using both frameworks for data analyses in future usability studies, especially those with multi-method approaches.

4.5. Limitations and future research

The sample size of 10 patients was small but fitting for usability and methods feasibility testing. The smaller sample is consistent with sample sizes recommended by Nielsen and Landauer [50], Virzi [51], Monk et al. [52]. For formative usability tests, such as this one, five to eight users are able to detect 80–85% of usability problems [50–52]. Moreover, an extensive amount of data were generated from the multi-method approach. Thus, 10 users was sufficient for methods feasibility testing in this study.

Our study involved a randomized sample from a larger, convenience sample of patients with diabetes. The convenience sampling frame for the larger study may make it difficult to generalize findings to the diabetes population as a whole. For instance, the study included patients who are more highly educated and who use technology more than an urban, lower income or rural sample might. For future studies, researchers may want to use purposeful sampling to select users with different characteristics to maximize variability.

Patients were new to using the system. It is possible that more practiced users may adapt to the observed issues after they interact more with the system. However, the user interactions did surface critical usability issues and pointed to particular issues that made the current version non-intuitive. In future mHealth research, it is imperative that users be involved early and often in the development process of mHealth applications. Future research might include testing with larger samples, testing earlier in the development life-cycle, testing of other available classification schemes and frameworks and/or repeating this study to investigate how reproducible results are with the tools employed here. Future researchers may also want to test data collection methods using a larger sample and more methods to determine whether they yield more unique usability issues.

5. Conclusions

Recent systematic reviews of mHealth self-management tools in general and for diabetes in particular speak to the need for more studies on patient interaction and system usability. Finding standardized, structured and reproducible ways to work in usability evaluation is important for providing evidence. This study provides an example of a multi-method design for both data collection and data analyses. Multiple data collection methods resulted in a more comprehensive set of usability problems and helped trianguulate data. The structured data analyses allowed reproducible steps and data validation (triangulation), a method of determining the most severe problems for users.

Usability testing with Think Aloud was essential for surfacing usability issues. The in-depth interview and questionnaire allowed data triangulation for severe usability issues, but both uncovered a smaller volume of consolidated usability issues. For data analyses, the more structured method, using the FA, provided a more standardized and feasible way to derive usability problems from a large volume of qualitative data. The FA method can guide analyses across multiple researchers. The UPT was advantageous as an in-depth classification scheme and for determining severity ratings for usability problems. It also assisted in categorizing specific types of problems which could be useful for designers. We recommend the use of multiple data collection methods to uncover a variety of problem types and severity levels. We also recommend the use of the FA and UPT methods during data analyses.

Conflict of interest

The authors declare that there are no conflicts of interest.

Appendix A

Start-up topics/questions for in-depths interviews

- What parts of the system did you think were well designed (?)
- Which parts of the system did you think were inadequately designed (?)
- Do you have any other comments about the system functions and regarding its usability (?)
Appendix B

1. How do you consider your level of computer-/IT-knowledge?
   - High
   - Medium
   - Small
   - None

2. How often do you use a computer (approximately)?
   - Every day
   - Several times a week
   - Once in a while
   - Never

3. How often do you use the Internet (approximately)?
   - Every day
   - Several times a week
   - Once in a while
   - Never

4. How do you consider your mobile phone knowledge level?
   - High
   - Medium
   - Small
   - None

5. How often do you use your mobile phone to make, receive phone calls?
   - Every day
   - Several times a week
   - Once in a while
   - Never

6. How often do you use your mobile phone for text messaging?
   - Every day
   - Several times a week
   - Once in a while
   - Never

7. How often do you use your mobile phone for e-mails, surfing?
   - Every day
   - Several times a week
   - Once in a while
   - Never

To what level do you agree or disagree with the following statements?

8. I like to use the computer in my work/leisure time.
   - Strongly agree
   - Agree
   - Disagree
   - Strongly disagree

9. I like to use my mobile phone in my work/leisure time.
   - Strongly agree
   - Agree
   - Disagree
   - Strongly disagree

10. Web services, mobile services such as different patient support and information services within health care are getting more common. I believe that this is a positive development.
    - Strongly agree
    - Agree
    - Disagree
    - Strongly disagree

11. I can see advantages for me personally in using web services, mobile services within health care.
    - Strongly agree
    - Agree
    - Disagree
    - Strongly disagree

12. What usability/user experiences were you satisfied with regarding the mHealth system?
    ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

13. What usability/user experiences were you dissatisfied with regarding the mHealth system
    ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

    ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
**Appendix C**

Usability problem descriptions and classifications (complete list)

<table>
<thead>
<tr>
<th>Usability problem description</th>
<th>Place of occurrence</th>
<th>% of pat. detecting per method</th>
<th>UPT Classification</th>
<th>Severity</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Difficulty knowing to exit the table view and click the “Add data” button to be able to add in a new value</td>
<td>Glucose Readings View</td>
<td>TA: 70, I: 100, Q: 50</td>
<td>Manipulation-Cognitive aspects-Visual cues (FC)</td>
<td>4</td>
</tr>
<tr>
<td>(2) Difficult to understand and perform the exporting action</td>
<td>Glucose Readings View</td>
<td>TA: 30, I: 40, Q: 40</td>
<td>Manipulation-Cognitive aspects-Visual cues (FC)</td>
<td>3</td>
</tr>
<tr>
<td>(3) No support to specify where the exported file is to be saved</td>
<td>Glucose Readings View</td>
<td>TA: 20, I: –, Q: –</td>
<td>Visualness-Non message feedback (FC)</td>
<td>3</td>
</tr>
<tr>
<td>(4) Difficult to know how to navigate within and adjust the table view to get to or show the right value or value range</td>
<td>Glucose Readings View</td>
<td>TA: 10, I: –, Q: –</td>
<td>Visualness-Presentation of Information/results (FC)</td>
<td>3</td>
</tr>
<tr>
<td>(5) No message about the lack of capability to save an exported file</td>
<td>Glucose Readings View</td>
<td>TA: 10, I: –, Q: –</td>
<td>Visualness-Non message feedback (FC)</td>
<td>3</td>
</tr>
<tr>
<td>(6) Difficult to know to go to the “Export” tab to view, export, and/or print the data table or its values</td>
<td>Glucose Readings View</td>
<td>TA: 50, I: 30, Q: 30</td>
<td>Manipulation-Cognitive aspects-Direct manipulation (FC)</td>
<td>2</td>
</tr>
<tr>
<td>(7) Difficult to know to choose the Action-tab for deleting a value</td>
<td>Glucose Readings View</td>
<td>TA: 50, I: –, Q: –</td>
<td>Manipulation-Cognitive aspects-Visual cues (FC)</td>
<td>2</td>
</tr>
<tr>
<td>(8) Difficult to find the Export button for the specific export command</td>
<td>Glucose Readings View</td>
<td>TA: 20, I: –, Q: –</td>
<td>Visualness-Object appearance (FC)</td>
<td>2</td>
</tr>
<tr>
<td>(9) Difficult to know to select “Delete data” for the table view to change, delete, export, or print the value list</td>
<td>Glucose Diary View</td>
<td>TA: 70, I: 100, Q: 100</td>
<td>Language-Naming/labeling (FC)</td>
<td>4</td>
</tr>
<tr>
<td>(10) Difficult to know how to adjust the range of values that are to be shown, retrieved</td>
<td>Glucose Diary View</td>
<td>TA: 90, I: –, Q: –</td>
<td>Manipulation-Cognitive aspects-Visual cues (FC)</td>
<td>3</td>
</tr>
<tr>
<td>(11) Incorrect rendering of graph values for the last 30 or 90 days as it lacks the correct delimiters</td>
<td>Glucose Diary View</td>
<td>TA: 20, I: –, Q: –</td>
<td>Visualness-Presentation of Information/results (FC)</td>
<td>2</td>
</tr>
<tr>
<td>(12) Difficult to find, access “Medication Reminder” as four different paths existed with similar but different names</td>
<td>Dashboard</td>
<td>TA: 40, I: 20, Q: –</td>
<td>Language-Naming/labeling (FC)</td>
<td>2</td>
</tr>
<tr>
<td>(13) Difficult to find, access “Appointment reminder” as it was listed under a pane with a different name</td>
<td>Dashboard</td>
<td>TA: 30, I: –, Q: 10</td>
<td>Language-Naming/labeling (FC)</td>
<td>2</td>
</tr>
<tr>
<td>(14) Difficult to find the “Tracking goal” (exercise, weight) item as three different paths existed with similar but different names</td>
<td>Dashboard</td>
<td>TA: 40, I: –, Q: –</td>
<td>Language-Naming/labeling (FC)</td>
<td>2</td>
</tr>
<tr>
<td>(15) The blood pressure graph shows only the systolic blood pressure value; the diastolic blood pressure is missing</td>
<td>Blood Pressure View</td>
<td>TA: 10, I: 20, Q: –</td>
<td>Visualness-Presentation of Information/results (FC)</td>
<td>3</td>
</tr>
<tr>
<td>(16) Weights lack conversion to the metric system</td>
<td>Submit Weight Progress View</td>
<td>TA: 10, I: 10, Q: –</td>
<td>Visualness-Presentation of Information/results (FC)</td>
<td>(NC)</td>
</tr>
<tr>
<td>(17) Difficult to detect and distinguish the Update and Submit buttons</td>
<td>Set Goals View</td>
<td>TA: –, I: –, Q: 10</td>
<td>Visualness-Object appearance (FC)</td>
<td>2</td>
</tr>
</tbody>
</table>

(continued on next page)
Appendix C (continued)

<table>
<thead>
<tr>
<th>Usability problem description</th>
<th>Place of occurrence</th>
<th>% of pat. detecting per method</th>
<th>UPT Classification</th>
<th>Severity</th>
</tr>
</thead>
<tbody>
<tr>
<td>(18) Times are only available for whole hours, not minutes or half hours</td>
<td>Set Medication Reminders View</td>
<td>10 10 –</td>
<td>Manipulation (PC)</td>
<td>Task-mapping- Functionality (FC)</td>
</tr>
<tr>
<td>(19) Only dates and not times are available for appointment reminders</td>
<td>Set Appointment Reminders View</td>
<td>20 10 10 (NC)</td>
<td>Task-mapping- Functionality (FC)</td>
<td></td>
</tr>
</tbody>
</table>

* TA = Think Aloud usability test, I = Interview, Q = Questionnaire.

### References


Study III

A Modified User-Oriented Heuristic Evaluation of a Mobile Health System for Diabetes Self-management Support

Mattias Georgsson, MSc, RN, Nancy Staggers, PhD, RN, Charlene Weir, PhD, RN

Mobile health platforms offer significant opportunities for improving diabetic self-care, but only if adequate usability exists. Expert evaluations such as heuristic evaluation can provide distinct usability information about systems. The purpose of this study was to complete a usability evaluation of a mobile health system for diabetes patients using a modified heuristic evaluation technique (1) dual-domain experts (healthcare professionals, usability experts), (2) validated scenarios and user tasks related to patients’ self-care, and (3) in-depth severity factor ratings. Experts identified 129 usability problems with 274 heuristic violations for the system. The categories Consistency and Standards dominated at 24.1% (n = 66), followed by Match Between System and Real World at 22.3% (n = 61). Average severity ratings across system views were 2.8 (of 4), with 9.3% (n = 12) rated as catastrophic and 53.5% (n = 69) as major. The large volume of violations with severe ratings indicated clear priorities for redesign. The modified heuristic approach allowed evaluators to identify unique and important issues, including ones related to self-management and patient safety. This article provides a template for one type of expert evaluation adding to the informaticists’ toolbox when needing to conduct a fast, resource-efficient and user-oriented heuristic evaluation.

KEY WORDS: Diabetes, Heuristic evaluation, Mobile health, Patient self-management, Usability evaluation

Recent figures from the World Health Organization show that 347 million people in the world are affected by diabetes; this chronic disease is predicted to be the seventh leading cause of death in the world by the year 2030. Data from the Department of Health and Human Services and the Centers for Disease Control and Prevention show that in the United States alone, the number of people living with the disease is 29.1 million. This number continues to grow rapidly. Of the people diagnosed with this disease, 90% have type 2 diabetes. Factors such as poorly regulated glycemic levels have a large influence on patients’ conditions and are therefore vital to monitor for controlling the disease. Type 2 diabetes is also lifestyle related to a large degree and can be self-managed to a certain extent in addition to using more conventional treatment procedures.

Self-management is becoming increasingly important in diabetes care; researchers found that self-management support should be integrated into patients’ everyday lives to achieve desired, improved patient outcomes. As an adjunct to diabetes management, researchers highlighted the use of information and communication technology (ICT) and the development of applications for day-to-day self-care and disease management. In support of that goal, the mobile health (mHealth) system for this study was developed as an individually based mobile and web support system for type 2 diabetes patients’ self-management.

Usability of eHealth systems
To facilitate their adoption, mobile healthcare applications and systems for chronic disease management must be usable. Despite the increasing availability of self-management tools, many of the patient-operated ICT applications are still deficient in terms of usability. Completing an evaluation process in a self-management context for these kinds of systems requires an understanding of users and their needs when performing tasks in such a system. Having this focus can help ensure that mobile applications are safe for clinical and patient use and possibly prevent user errors. Authors thus argue that it is vital for user interfaces to be designed in a way that does not contribute to errors as this can also be a factor negatively affecting users’ experiences with the system. Usability evaluations can therefore help to appropriately determine how well the application or system meets the clinical need and patients’ expectations and in safeguarding both the quality of care and patient outcomes. These evaluations can be expert based, such as heuristic evaluation (HE) or cognitive

Author Affiliations: Department of Biomedical Informatics, University of Utah, Salt Lake City (Mr Georgsson and Dr Staggers and Weir); Faculty of Computing, Blekinge Institute of Technology, Karlskrona, Sweden (Mr Georgsson); School of Nursing, University of Maryland, Baltimore (Dr Staggers); VA Salt Lake City Health Care, Utah (Dr Weir).

The authors have disclosed that they have no significant relationship with, or financial interest in any commercial company pertaining to this article.

Corresponding author: Mattias Georgsson, MSc, RN, Faculty of Computing, Blekinge Institute of Technology, 371 79 Karlskrona, Sweden (mattias.georgsson@bth.se).

This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 3.0 License, where it is permissible to download and share the work provided it is properly cited. The work cannot be changed in any way or used commercially.
walk-through, or empirical and user based evaluations, such as think-aloud methods involving user tests with actual system users.

**Heuristic Evaluation**

Heuristic evaluation is one of the most common usability inspection methods completed by usability experts. Users are explicitly not part of this kind of method. Instead, experts apply the knowledge they have about usability principles, processes, and standards to evaluate systems. Heuristic evaluation was first defined by Nielsen and Molich. In this technique, usability experts evaluate an application to find usability problems, assign them to a specific category of heuristic and ascribe a severity rating. Nielsen originally defined 10 heuristic categories and recommended assigning severity scores to a master list of usability violations.

Author have attempted to modify and extend Nielsen’s techniques in different ways to achieve better results in various contexts. These include Zhang et al, who came up with 14 heuristics by combining Nielsen’s 10 heuristics with Shneiderman’s eight “golden rules” to evaluate infusion pumps. Allen and colleagues employed a more simplified use of the HE inspection method by having evaluators select only those heuristics they deemed appropriate for their assessment and assigned severity ratings for the usability problems on the fly instead of first creating a master list of all usability problems. Chatratichart and Brodie extended the method to a technique they called HE-Plus, a directed approach using usability problem profiles to help evaluators focus their evaluations on specific types of problem areas to provide more consistent and reliable evaluation results.

Heuristic evaluation or expert usability evaluation can be useful because it provides a unique perspective and distinct information and because it is a discount usability technique, meaning it is relatively quick, cost effective, and resource efficient. However, as other authors have shown, the original method by Nielsen can be improved upon for better results. Critics of the technique, for example, indicate that many problems found with HE can be minor interface design problems or of a more general nature. User tests, in comparison, involve actual users and identify problems of a critical, qualitative nature. On the other hand, these are also more costly and time-consuming.

In sum, current expert techniques require improvements to be able to find more severe usability problems of a critical nature for users. In this article, we addressed this gap. Our approach to accomplish this is by using a modified HE technique using its beneficial aspects and also focusing on the patient user and their needs in disease management and system information and interaction requirements to provide enhanced evaluation results. Our modifications involve (1) employing dual-domain experts (healthcare professionals and usability experts combined) as evaluators, (2) using realistic, validated user tasks with appropriate scenarios related to patients’ diabetes self-care, and (3) making severity ratings specific and in-depth across three severity rating factors by predicting each problem’s influence on patients with factors of impact, persistence, and frequency. Our intent was to explore whether the technique would be able to detect both crucial and context-related problems in patient self-management in addition to the more common, minor usability issues.

**MATERIAL AND METHODS**

**System Description**

The mobile system we evaluated was designed as a low-cost, convenient, personalized self-care management and tracking tool for use by a large number of patients with diabetes. It was also meant to function as support for conversations between patients and their healthcare providers. The system combined a web service and mobile phone solution for patients to send in self-management values, that is, glucose and blood pressure, via text messaging. The web user interface is divided into sections consisting of a Dashboard, Glucose Diary, Blood Pressure, Medication Adherence, an Exercise and Weight progress, and Appointment reminder view. Each has graphical representations of the different measurements and goals with progress indicators in red, yellow, and green. For example, the sections include a meter to visualize glucose readings, a blood pressure bar with systolic and diastolic values, a medication adherence section indicating how much of the prescribed medication was taken, and an exercise and weight progress section to show exercise and weight measures. Using their cell phones, patients can retrieve, enter, and edit their values and goals. Scenarios and tasks were developed based on these kinds of patient interactions and uses.

**Expert Evaluators**

The HE was performed by three expert evaluators who identified heuristic violations listed in Nielsen’s taxonomy. According to Nielsen, three to five single-domain usability expert evaluators find, on average, between 74% and 87% of usability problems. The number of usability problems found by dual-domain experts is even higher at 81% to 90%. Only two to three dual-domain evaluators are then deemed necessary. These types of experts are seen as especially suitable in evaluating complex systems, such as those in the healthcare area, because they have usability expertise and extensive knowledge in the specific domain of application. Each expert for this study was thus carefully selected based on dual-domain competency consisting of (1) extensive usability experience in health informatics, (2) being healthcare professionals (registered nurses [RNs]), and experience with (3) the patient group and their task.
requirements, and (4) diabetes self-management. As this was a HE evaluation, it involved only expert evaluators and no patients. Therefore, institutional review board approval was not required for this study.

**Use Scenarios and Tasks**

Scenarios and tasks outlined specific steps that evaluators used to interact with the diabetes self-management system in the HE. Tasks were based on real case scenarios to simulate how patients would use the system in a self-management process in a clinic or at home. To ensure that these were as realistic as possible, a panel also evaluated both scenarios and tasks. The panel included a physician with a diabetes specialty, a diabetes RN, a public health professional with chronic patient intervention systems expertise, and a diabetes patient. The panel verified and validated tasks for content validity and accuracy (content validity index of 0.91 of 1.0). The eight tasks and scenarios were disease specific and had varying levels of difficulty. For example, tasks consisted of viewing and locating glucose values on graphs, identifying and correcting collected glucose values, setting weight and exercise goals and medicine and appointment reminders, and viewing summary statements about medical measurements. Table 1 includes an example of a scenario and task.

**Nielsen’s Heuristics**

Similar to other method modifications, we selected Nielsen’s 10 heuristics for this study because they have been thoroughly tested, are widely accepted by user experience experts, and are fast and easy to apply. To attend to some of its shortcomings, our approach was to apply our specific modifications of (1) dual-domain experts (healthcare professionals, usability experts), (2) validated scenarios and user tasks related to patients’ self-care, and (3) in-depth severity factor ratings to determine if more critical issues could be found. The HE categories are listed in Table 2. Part of the original work with Molich, the categories are Nielsen’s published work from 1994. The evaluators used the 10 heuristics to categorize usability problems by employing the specific modifications.

**Severity Rating Scale and Factors**

Typically, HE techniques include assigning a single severity score. Instead, we divided severity ratings into factors of frequency, impact, and persistence. The focus was on how each of the three different factors for each usability issue would influence the user in different ways, and separate averages were calculated for each. Subsequently, separate averages were calculated for each severity factor. This allowed for greater specificity about the severity of the problem and its impact on the specific diabetes patient users. Severity ratings ranged from 0 (not a problem) to 4 (usability catastrophe). Specific descriptors for the scale for the severity rating and the severity factors are listed in Table 3.

After evaluators conduct their individual factor severity ratings, all ratings are summed and divided by the number of evaluators to arrive at an average severity rating for each usability problem. This rating is considered the overall severity rating, as shown in Table 4.

**Evaluation Procedure**

Evaluators had identical instructional materials to learn the system and to ensure consistency across evaluators. Information materials consisted of a digital video on system modules, how to navigate the portal, a study design manual detailing each specific scenario and tasks to be performed, an application user manual, and an evaluation guide sheet. The study design manual also included materials on how to conduct the evaluation, the scenarios, and a usability task manual outlining how to navigate tasks. Providing specific scenarios and tasks to simulate the diabetes patient care process ensured that all experts had the same knowledge level about the functionality and user tasks.

The procedure itself was a two-part process. The evaluators first familiarized themselves with the system and its usage using the materials and training described above. Then, they performed the modified HE as visualized in Figure 1.

Each dual-domain expert evaluator performed the eight scenarios and tasks independently. After the evaluators detected a usability problem, they assigned each problem to a heuristic violation/s from the categories in Table 2. A master list was compiled, duplicate problems were removed, and

---

**Table 1. Example of a Scenario and Task Used in the Evaluation**

<table>
<thead>
<tr>
<th>Scenario:</th>
</tr>
</thead>
<tbody>
<tr>
<td>During your follow-up appointment with your provider, you agreed that a stronger commitment regarding weight loss and exercise would improve your diabetes condition. You now would like to activate the system’s support service for exercise tracking and weight tracking and put in your tracking goals regarding your exercise and weight.</td>
</tr>
</tbody>
</table>

Please complete the following tasks.

1. Select and activate the service that you would like to use to set tracking goals for exercise and weight.
2. Set your exercise goal to 3 times per week.
3. Set your weight goal to 180 pounds.
4. When you consider yourself done with the task, finish and return to “Participant Home.”
the list was verified across the evaluators for accuracy. Then, each evaluator individually assigned severity scores to each problem by using the severity rating factors of frequency, impact, and persistence. These were also averaged by factor and combined into one severity rating for each usability problem as described above. Descriptive statistics were used to summarize heuristic violations and associated severity scores.

**RESULTS**

The HE resulted in a total of 129 usability problems and 274 heuristic violations. The usability problems by place of occurrence (view), number of heuristic violations, and mean severity ratings are summarized in Figure 2. The number of usability problems ranged from a low of 12 to a high of 34 across application views. The Dashboard view generated the most usability problems (34), followed by the Glucose Diary view (21), the Blood pressure view (20), and the Medication adherence view (15). Heuristic evaluation violations ranged from 25 to 69. The largest number of heuristic violations was on the Dashboard view (69), the Glucose Diary view (49), the Blood pressure view (44), the Medication adherence view (31), and the Appointment reminder view (29).

The average severity ratings ranged from 2.7 to 3 on a scale of 0 to 4, with the Glucose Diary view and

---

**Table 2. Heuristics for Usability Evaluation According to Nielsen**

<table>
<thead>
<tr>
<th>Heuristics for Usability Evaluation According to Nielsen</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Visibility of system status:</strong></td>
</tr>
<tr>
<td>The system should always keep users informed about what is going on, through appropriate feedback within reasonable time.</td>
</tr>
<tr>
<td><strong>Match between system and the real world:</strong></td>
</tr>
<tr>
<td>The system should speak the user’s language, with words, phrases, and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order.</td>
</tr>
<tr>
<td><strong>User control and freedom:</strong></td>
</tr>
<tr>
<td>Users often choose system functions by mistake and will need a clearly marked “emergency exit” to leave the unwanted state without having to go through an extended dialogue. Support undo and redo.</td>
</tr>
<tr>
<td><strong>Consistency and standards:</strong></td>
</tr>
<tr>
<td>Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform conventions.</td>
</tr>
<tr>
<td><strong>Error prevention:</strong></td>
</tr>
<tr>
<td>Even better than good error messages is a careful design which prevents a problem from occurring in the first place. Either eliminate error-prone conditions or check for them and present users with a confirmation option before they commit to the action.</td>
</tr>
<tr>
<td><strong>Recognition rather than recall:</strong></td>
</tr>
<tr>
<td>Minimize the user’s memory load by making objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.</td>
</tr>
<tr>
<td><strong>Flexibility and efficiency of use:</strong></td>
</tr>
<tr>
<td>Accelerators—unseen by the novice user—may often speed up the interaction for the expert user such that the system can cater to both inexperienced and experienced users. Allow users to tailor frequent actions.</td>
</tr>
<tr>
<td><strong>Aesthetic and minimalist design:</strong></td>
</tr>
<tr>
<td>Dialogues should not contain information which is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility.</td>
</tr>
<tr>
<td><strong>Help users recognize, diagnose, and recover from errors:</strong></td>
</tr>
<tr>
<td>Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution.</td>
</tr>
<tr>
<td><strong>Help and documentation:</strong></td>
</tr>
<tr>
<td>Even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the user’s task, list concrete steps to be carried out, and not be too large.</td>
</tr>
</tbody>
</table>

---

**Table 3. Scale for Severity Rating and Severity Factors**

<table>
<thead>
<tr>
<th>Severity rating</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Not a usability problem at all</td>
</tr>
<tr>
<td>1</td>
<td>Cosmetic problem only—Need not be fixed unless extra time is available</td>
</tr>
<tr>
<td>2</td>
<td>Minor usability problem—Fixing this should be given low priority</td>
</tr>
<tr>
<td>3</td>
<td>Major usability problem—Important to fix. Should be given high priority</td>
</tr>
<tr>
<td>4</td>
<td>Usability catastrophe—Imperative to fix this before product can be released</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Severity factors</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>The frequency with which the problem occurs</td>
<td>Is it common or rare?</td>
</tr>
<tr>
<td>The impact of the problem if it occurs</td>
<td>Will it be easy or difficult for the users to overcome?</td>
</tr>
<tr>
<td>The persistence of the problem</td>
<td>Is it a one-time problem that users can overcome once they know about it, or will users repeatedly be bothered by the problem?</td>
</tr>
</tbody>
</table>
Medication adherence view having the highest at 2.9 and 3.0 respectively.

**Heuristic Violations Across System Views**

Of the 10 types of HE violations depicted in Figure 3, the categories of Consistency and Standards and Match Between the System and the Real World dominated at 24.1% (n = 66) and 22.3% (n = 61) respectively, followed by Aesthetic and Minimalist Design at 16.8% (n = 46) and Recognition Rather Than Recall at 11.7% (n = 32). The heuristic categories Recover 1.4% (n = 4) and Help 1.03% (n = 3) had the fewest violations across all views.

**Severity Ratings Across System Views**

Most severity ratings across views (Figure 4) consisted of major and catastrophic severity ratings. The most severely rated problems were located in the Dashboard view (n = 16, n = 4), Glucose Diary view (n = 12, n = 2), and the Blood pressure view (n = 12, n = 2). Most of the minor usability problems were similarly located in the Dashboard view (n = 14) and the Glucose Diary view (n = 7). The Appointment reminder view, however, came next with six issues. There were no cosmetic violations.

**Nature of Usability Problems and Prioritization**

The modified HE evaluation revealed that most catastrophic ratings concerned disease-related task deficiencies and specific system-related shortcomings in displaying necessary information for patients. Some examples of these types of usability problems and comments provided by the evaluators are as follows:

- **Dashboard:**
  
  "It is superfluous with the years showing on the time line, it makes the reading crowded and difficult to read."

- **Glucose Diary view:**
  
  "It is very difficult to read the time line on the graph because the numbers are too crowded which makes it difficult to distinguish and read specific dates. This is especially cumbersome for diabetes patients who often have visual concerns."

- **Blood Pressure view:**
  
  "It is disadvantageous for the patient to not see the diastolic blood pressure reading in the graph to compare against; only the systolic value is shown."

**Table 4. Example of Severity Rating Scoring Table Including Nielsen’s Three Factors**

<table>
<thead>
<tr>
<th>Place of Occurrence</th>
<th>Usability Problem Description</th>
<th>Heuristics Violated</th>
<th>Factors</th>
<th>Severity Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Glucose Diary View</td>
<td>“It is superfluous with the years showing on the time line, it makes the reading crowded and difficult to read.”</td>
<td>H7, H8</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

**FIGURE 1. Modified HE process.**
With the indications… in percentages, it is difficult for a patient or health care provider to determine what medication was taken or not, which day and what time.

… Tallying up totals to say 100% of medications is an odd way to think about medication from a patient perspective. I wouldn’t say that I have taken 75% of my meds for the last month for instance. I need to know specifics and insulin or Metformin and that these are jointly displayed and tracked.

…I found that this display did not match my mental model of medication compliance, need individual information regarding medication. (Total severity rating of 3.7 factor rating of 4 for two evaluators on each factor of across frequency, impact, and persistence.)

**DISCUSSION**

In this study, most usability problems were categorized as major issues. The largest volume of problems clustered in the categories of Consistency and Standards and Match Between the System and the Real World. Both of these categories indicate that the system requires better design to support effective decision making and action control for relevant patient user tasks. 31

This evaluation uncovered specific concerns related to the disease-related information deficiencies of the system. These included, for example, that the Dashboard view allows only one daily entry and lacks a time for glucose readings on its current meter. These are fundamental issues for a diabetes application. Since patients with diabetes often have multiple readings in one day, this is a major usability problem related to essential information needed or patients’ self-management tasks. In the Glucose diary, the numbers on the timeline were too small, crowded together, and difficult to read. This is especially problematic for individuals with diabetes as they often have visual acuity issues. The likelihood of users performing errors is increased for both of these issues.
Using the in-depth severity factor ratings, the use of dual-domain evaluators and validated user tasks related to the care process enabled the evaluators to find a large volume of severe deficiencies in the system. Finding major issues with HE is in contrast to other authors’ work both within and outside the health domain who mainly found minor interface issues.\textsuperscript{23–25} The identified major and severe usability problems found here require immediate attention for redesign to fulfill patients’ self-management needs. The methodological changes to HE may have enhanced evaluators’ abilities in finding important usability problems.

Usability testing with users is very helpful, but as was seen here, dual-domain experts add important, additional dimensions to any usability evaluation. In particular, they were able to understand tasks as part of the context of use for a chronic disease self-management system and were able to identify specific system design concerns and uncover distinct information needs related to these self-management tasks. Dual-domain experts may also be able to assist in identifying possible patient safety issues that patients may not identify. A case in point was that the medicine reminder was too nonspecific and provided insufficient information, making it possible for patients to commit errors in insulin dosages.

We recommend that dual-domain experts be employed in future HEs for chronic disease management systems whenever possible. These kinds of experts can identify unique and critical usability problems as well as deeper cognitive support issues and specific disease-related concerns. Dual-domain experts provide added value in uncovering pertinent issues.\textsuperscript{26,29,32} This information can be used both during the iterative design process, during formative evaluation, summative testing, and for comparing different versions and applications.

Other methodological modifications can assist future usability evaluations. A standardized evaluation process and use of specific scenarios and tasks allowed for efficient evaluations across the experts and can also aid in reproducibility and generalizability. The modified severity factor rating method in this study also proved important as it allowed the evaluators to think about the specific impact of each individual usability problem and provided a more in-depth analysis of the specific usability problems. This modification provided an objective method to determine the importance of the usability problem in relation to others of a similar nature and could aid in problem prioritization.

**Limitations**

Although the modified HE process uncovered many major usability problems, other problems might have been detected had the number of dual-domain evaluators been expanded over the recommended two to three evaluators.\textsuperscript{27} Both HE and user tests could also be combined to detect an optimal number of usability problems,\textsuperscript{33} although it would entail higher costs. We wanted to provide an efficient, cost-effective method of modifying HE to also be able to identify more serious usability issues that could have an impact on patients in their disease management.

**CONCLUSION**

Consumer health systems and applications in mHealth should be evaluated for usability as well as medical adequacy. This article describes useful modifications to HE by modifying and deepening Nielsen’s techniques. Specifically, modifications were using dual-domain experts; employing validated, patient-centered self-care tasks and realistic care process scenarios; and using separate in-depth severity factor ratings. In particular, dual-domain experts can provide unique information related to the salient tasks for patient self-care and identify potential patient safety issues as well as determine how an application adheres to known usability guidelines.
This modified heuristic method can be used by other informaticists in healthcare who need to conduct a fast and resource-efficient heuristic process related to patient self-management. The results of the study show that a modified HE can uncover unique, critical issues in this context. This kind of evaluation may be done at any point in the system lifecycle. As an expert evaluation technique, HE should be included in any usability evaluation and is quite suitable for nHealth applications designed for chronic disease patients. Thus, it has an important place in usability evaluations. With the modifications provided here, Nielsen’s original techniques can be improved to achieve improved results. Techniques like the ones described here can be an important addition to any informaticists’ toolbox when determining chronic disease systems’ adequacy for patient self-management needs.

Acknowledgments
The authors acknowledge Korey Capozza, MPH, Consumer Engagement Director at HealthInsight, Salt Lake City, UT, for access to the system for evaluation.

References
Study IV

IV. Georgsson, M., Staggers, N., Årsand, E., Kushniruk, A. Using a User-Centered Cognitive Walkthrough to Evaluate a mHealth Diabetes Self-Management System Including a Case Study and External Validity Test. *Manuscript*
Using a User-Centered Cognitive Walkthrough to Evaluate a mHealth Diabetes Self-Management System Including a Case Study and External Validity Test

Mattias Georgsson, a, b, 1 Nancy Staggers, a, c Eirik Årsand, d, e, Andre Kushniruk f
a Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah, USA
b Faculty of Computing, Blekinge Institute of Technology, Karlskrona, Sweden
c Summit Health Informatics, Salt Lake City, UT, USA
d The Norwegian Centre for E-health Research, University Hospital of North Norway, Tromsø, Norway
e UiT The Arctic University of Norway, Department of Clinical Medicine, Tromsø, Norway
f School of Health Information Science, University of Victoria, Victoria, Canada

Abstract

Background: Self-management of chronic diseases with mHealth systems is becoming common. With the move toward patient-centered care, it is vital to incorporate patients in the development and evaluation of these systems. Current methods for usability evaluation such as formal usability testing can be very costly and time-consuming. Other methods may be more efficient but lack a user focus (e.g., heuristic evaluation [HE] and cognitive walkthrough [CW]). We propose a modified method to address identified deficiencies in the original CW technique and then assess its effectiveness and user acceptance. In this case study we use the modified, user-centered CW (UC-CW) method with diabetes patient users of a mHealth self-management system and validate it against the “golden-standard” user test with Think Aloud (TA) on the number, types and severity of usability problems, as well as the consumed time, and user experience of the method’s cognitive load.

Materials and methods: A total of 12 diabetes patients were included: 6 in the UC-CW group evaluation session, and 6 in the individual user tests with TA. The setting for the assessments was the Diabetes and Endocrinology Center at University of Utah Health in Salt Lake City, Utah, USA. The modified UC-CW method consisted of: making the user the main evaluation contributor in detecting usability problems, a dual domain facilitator with the role of guiding the evaluation session and in rating found usability problems, a task development process resulting in validated tasks, a focus on higher level tasks in evaluation and in streamlining the evaluation in terms of time and resources. Users interacted with the mHealth application for both the modified method and user test with TA. Participants also filled in a pre-test questionnaire of demographic questions and one on their experience, knowledge and perception about information technology. Post-evaluation assessments included the NASA RTLX instrument and a set of brief interview questions about the different methods.

Results: Participants in both methods were a similar mix of males and females with similar types of diabetes. They were equally knowledgeable and experienced in using mobile phones. A total of 26 usability problems were found with the UC-CW and 20 problems with the TA. Participants in both methods gave the application similar severity ratings for all views (UC-CW = 2.7 and TA= 2.6) and a similarly high number of problems in similar views (Main view [UC-CW=11, TA=10], Carbohydrate Entry view [UC-CW=4, TA= 3] and List view [UC-CW=3, TA=3] and with similar heuristic violations (Match Between the System and Real World [UC-CW=19, TA=16], Consistency and Standards [UC-CW=17, TA=16]).

1 Faculty of Computing, Blekinge Institute of Technology, 371 79 Karlskrona, Sweden.
Both methods converged on 8 usability problems. When comparing the severity of the problems, the UC-CW detected 5 critical issues (severity level 4) while the TA detected two. The distinct issues in the UC-CW compared to the TA were personalized features required for patients’ individual disease needs. In terms of efficiency, the whole evaluation procedure took less time to perform for the TA than the UC-CW. However, when the number of usability problems is considered, the UC-CW was faster. The UC-CW was also faster for the evaluation process. The NASA RTLX scores indicated that participants experienced the UC-CW as half as demanding in terms of cognitive load. Common themes were that the UC-CW was perceived as easy to perform and enjoyable while the TA was considered somewhat awkward and more cognitively demanding.

Conclusions: The modified method proved useful for finding severe and recurring usability issues for users and highlighted the need for personalized user features. The UC-CW also proved efficient with a high user acceptance. These results indicate that the UC-CW can be a useful method to evaluate a mHealth self-management system for diabetes. The external validation that compared the method outcomes provided beginning evidence of its utility as an effective, as well as efficient and acceptable method.

Keywords: Chronic Disease, Cognitive Walkthrough, Diabetes, Mobile Health, Self-Management, Think Aloud, Usability, User Test, User-Centered Design,

Introduction

Chronic Disease Self-management Systems

Levels of chronic diseases are at an all-time high globally. The World Health Organization (WHO) estimates that deaths from non-communicable chronic diseases such as cardiovascular, cancer, chronic respiratory disease and diabetes were 40 million annually in 2017 [1]. One of the ways to address the rising chronic disease levels is by implementing information and communication technology (ICT) self-management solutions for patients. In recent years the prevalence of these solutions is rising, but there is still a low adoption level of ICT-enabled, patient-centered solutions in chronic disease management [2, 3]. Self-management and self-monitoring systems have the potential to improve disease self-management, but this will not be realized until patients are included in the design and development of these solutions [4, 5]. It is critical that developers of self-monitoring solutions such as eHealth and mHealth systems engage with patients, ensuring that the technology meets their needs [6] and that user-centered design approaches are employed throughout the development and evaluation of these technologies [7].

Methods are available for user-centered approaches; however, current methods for usability evaluation, such as formal usability testing with Think Aloud (TA), can be very costly and time-consuming to conduct. Other methods may be more efficient but they lack a user focus (e.g., heuristic evaluation and cognitive walkthrough that utilize experts). In this study we propose improvements to the original CW and provide a case study to establish the utility and feasibility of the modified method. Specifically, we use the modified method during patient interactions with a mHealth diabetes self-management application. We then compare and validate the modified method with a user test with Think Aloud using the same mHealth application to determine the number and types of identified usability problems, and method efficiency. In addition we assess how the users themselves experience each performed method.
Background

User-Centered Design

User-centered design (UCD) is an important concept in health informatics due to its focus on incorporating the user in system development and evaluation and it is deemed critical for the ICT tools patients’ use [8]. The term UCD was first created by Norman and Draper [9] and also defined by Norman [9] describing it as: “the purpose of the system is to serve the user, not to use a specific technology, not to be an elegant piece of programming. The needs of the users should dominate the design of the interface, and the needs of the interface should dominate the design of the rest of the system.” [9]. Other authors such as Dennis Wixon, added to this, in an article by Karat et al [10], that user generated data should be the prerequisite by which a design is evaluated.

Usability Evaluation Methods

Usability, central to UCD, is similarly 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” [11]. To design for usability also means that the system should also be easy to use, learn, and remember as well as helpful to users [12].

Usability can be assessed through user- and expert-based methods. The user-based ones involve user testing where the user is key in performing the evaluation. One of the most commonly employed user test methods is Think Aloud or TA. Here, users are provided a set of tasks to complete while verbalizing what they find deficient or confusing in the system, resulting in a list of usability problems. In expert-based methods, experts perform the tasks in the evaluation and determine a list of usability problems in the system. Two of the most common expert methods include heuristic evaluation and cognitive walkthrough [13]. Each of the different methods are briefly described below:

User Test with Think Aloud Protocol

The user test with Think Aloud protocol (TA) is a technique commonly employed in the human computer engineering and human factors fields. Introduced by Lewis [14] and further refined and clarified by Lewis and Rieman [15], it was based on work by Ericsson and Simon [16-18]. The goal is to gain a thorough understanding of task behavior, [18] users’ thought processes during system interaction and to identify the main usability problems [19].

In this method users are asked to verbalize their thoughts during their interactions, express their reactions and explain their actions as they perform representative tasks. A TA session is normally audio- or video-recorded. An observer can also take thorough, written notes throughout. This results in instantly generated data for analysis. Throughout the session there is minimal intervention from the observer to assure users’ thought processes are not interrupted except to remind them to keep talking if they fall silent. The focus in this method is on understanding users’ decision making processes and on how users experience the system as expressed in their own words. The method provides extensive, detailed data, and therefore only a small sample of five to eight users is normally needed to detect 80–85% of usability problems [20-22].

Heuristic Evaluation

Heuristic evaluation (HE) is also a common usability method; this one performed by experts. It was first defined by Nielsen and Molich [23]. In this method, experts systematically step through the system or application to find usability problems. After duplicates are removed, the resulting master list of problems is sent out anew to each evaluator to assign each problem with a specific heuristic violation/s and also provide a severity rating. Nielsen originally defined 10 heuristic categories [24].
Cognitive Walkthrough
The cognitive walkthrough (CW) was originally introduced by Polson and Lewis [25, 26]. This group-based method has its foundation in theories of cognitive exploratory learning, involving a user’s ability to learn through his or her actions. Using CW, experts attempt to grasp users’ problem-solving and correlated system deficiencies. This is particularly important for those systems about which users have limited primary knowledge, i.e., handling the system for the first time or where they have to be assured of immediate and adequate cognitive support and feedback from the system in their interaction.

The CW analysis process consists of an initial task analysis where the experts specify the sequence of actions needed to be performed by a user to accomplish a specific goal, as well as the system response to these. Then the designers and/or developers of the software, as a group, walk through the different steps. Here they use four questions (described in Table 1) [25-27]:

<table>
<thead>
<tr>
<th>Analysis questions for the Cognitive Walkthrough [27]</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Original CW-questions:</strong></td>
</tr>
<tr>
<td>- 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?)</td>
</tr>
<tr>
<td>- Will the user notice that the correct action is available? E.g. is the button visible?</td>
</tr>
<tr>
<td>- 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.</td>
</tr>
<tr>
<td>- Does the user get feedback? Will the user know that they have done the right thing after performing the action?</td>
</tr>
</tbody>
</table>

By answering these specific questions for each of the subtasks, usability problems are identified. This process also consists of collecting the resulting data in specific, designated forms. At the end of the session a list of the potential issues is compiled and reported [26].

The CW, similarly to the HE, is considered valuable for its ability to generate results quickly, and economically compared to other usability methods such as formal usability testing consisting of the user test with TA. Another advantage is the ability to apply this method early on in the product design phase [28]. Some of its major drawbacks is that it does not employ real users. It also often identifies many more problems than would be found in a session with a single user. These identified issues might also be minor, low-priority interface design concerns or one-time issues [29-31] and not always relevant to the user [15].

A Review of the CW Literature and Method Deficiencies
After a detailed review of the scientific literature on the CW method, we were able to determine the following major deficiencies related to the user involvement, the task development processes (and focus on lower level tasks) as well as the evaluation process with its time and resource demands.

User Involvement
As mentioned earlier, one of the critical drawbacks of the CW is that the method does not involve testing with the real system users. Three sets of previous authors have therefore offered modifications to overcome this deficiency. Bias [32] described a method called pluralistic walkthrough with 6-10 system developers, 2-3 usability engineers and 6-10 intended system users who all interact in a group session. Just as in the original CW process, the group goes through a series of tasks within the system and are encouraged to take on the role of intended users. They write down their thoughts based on questions arising from the interaction. A discussion then follows with the stipulation that the users express their views before the other groups determine the usability
problems in the system. Gabrielli, Mirabella and Kimani et al. [33] developed a technique that included performing the conventional expert CW sessions but added video recordings showing the users’ interactions with the system. This modification was aimed at discovering contextual factors to aid the experts in understanding the users’ needs and situation better when determining usability problems [33]. Last, Granollers and Lorés [34] proposed adding a separate Think Aloud session with users to the experts’ conventional walkthrough and analyses in the determination of usability problems. All these method modifications intend to include users but none of them incorporate them fully and on equal terms but rather as an “add on” to the original method.

Task Development Process
Tasks are of central importance for any evaluation to be successful in finding issues or system limitations, and it is important that these encompass the right areas of the system and are sufficient in their number and specification. The CW has, typically, a very limited and strict task analysis (as seen earlier) and tasks are broken down into minute actions [35]. Also, there are no real guidelines for the task development process, task selection, or how to prioritize tasks or system aspects in the evaluation [36].

Currently, the method trusts the expert’s ability to understand the users’ characteristics as well as predict the users’ needs in the development of evaluation tasks. This can result in variability due to that it may be difficult for the experts to fully understand the experiences and needs of the end users [37]. A variability can also occur due to individual differences among experts and users and even among the experts themselves [38]. A mismatch can therefore arise between what the experts’ perceptions of and the users’ real task needs are [34]. This mismatch can lead to deficiencies in results that are potentially important; such as that adverse end user problems can be overlooked and missed.

Last, a risk of bias can exists in the task development process as it may often be the case that the experts who develop the tasks are also part of the system’s development team [39-41]. If the person chosen to lead the evaluation session is also part of the design and development team there may be a risk that he or she consciously or subconsciously favors certain aspects in the system and their integration. Moreover, the quality of the entire evaluation rests on the experience and knowledge of the evaluators. This can also create a variability in results across evaluators, which can also affect the validity of findings [38].

Focus on Lower Level Tasks
As stated earlier, the CW is limited to lower level, detailed task examinations and those aspects observed during the evaluation [36]. This is a critical limitation as higher level deficiencies can be missed especially regarding the general context of use and understanding the system as a whole from the user’s perspective.

The Evaluation Process
The evaluation process itself includes several additional time and resource demands.

Time and Resource Demands
Despite being efficient as an expert usability method, the CW evaluation process itself is time-consuming and resource intensive in its different evaluation parts [27, 36]. In each part, evaluators (i.e., experts) need to document their answers to the four previously described questions in specific forms for each task step. These can mount to a very large number of forms that in the end have to be categorized and summarized into different lists.
There is also a large number of professionals that are involved (i.e., session leader, note taker, logger, session evaluators) and considerable time needed for the whole team to complete the structured walk through and analyses. The minute steps in the sequence of machine operations required to complete the tasks also makes the data collection process inflexible [35].

Suggested Improvements and CW Method Modifications
We offer detailed suggestions below about how the CW method can be improved. The overall goal is to develop a modified, user-centered CW by including actual system users, employing dual domain facilitators, improve the task development process and streamline the different parts of the evaluation process.

User Involvement
Similar to the modifications proposed by authors Bias [32], Gabrielli, Mirabella and Kimani et al [33] and Granollers and Lorés [34] we propose including the user in the modified CW method. Different from these approaches, however, we suggest that the users not be an “add-on” but the main actor/s in the CW session and in determining usability problems. In our proposed method modification these actual users are patients or the end users of the system so that they can contribute their own experiences and context of use. Preferably they are real chronic disease patient users of the system or application with various severity levels of the disease, with different genders, ethnic backgrounds, educational backgrounds, ICT skill levels and experiences to achieve a representative user group. Our intent is that this results in more realistic notions and discussions during the evaluation sessions and potentially leads to improved data collection and findings. A major advantage in using actual users is also their important contribution to the validity of findings. They can also contribute to decreased variation, decreased potential bias from expert evaluators and increased validity during data collection. This can improve the quality and reliability of the raw data, analyses and ultimately CW findings.

Dual Domain Facilitator
We recommend that dual-domain experts be used for facilitating and moderating the evaluation session with users. A dual domain expert is one with knowledge and experience from human computer interaction (HCI), health care and the disease in question, and preferably also with some system knowledge. This expert should not be part of the design or development team to decrease the risk of bias.

The Task Development Process
To provide for better support in the task development process, we recommend adding contextual factors for task creation. Mainly, this would involve needs analyses based on the specific user group with information and guidelines about the specific disease, and central parts in the daily patient self-management routine. For the system side, this could also consist of deeper knowledge about the system’s inherent parts, capacity and system functionality as well as obtained information from manuals, earlier performed demand or technical specifications.

To decrease the risk of insufficient user representation in task definition, our suggestion is to validate the developed, representative tasks used in the evaluation by a panel consisting of dual domain experts and real chronic disease patient users. This could for example be done by utilizing a content validity index rating.

Include a Focus on Higher Level Tasks
To improve the focus on higher level activities, we suggest that the modified method include a whole system contextual perspective to avoid being limited to lower level deficiencies, granular activities
and only those parts that are observed in the evaluation session. This could be accomplished through
the inclusion of more holistic questions about the system during the group discussions, and make
questions more general but still inclusive. A suggestion is that after the users have shared their
individual thoughts from their own notes about problems, two general questions are used to initiate
the group discussion between the real users/patients which could address the specific problems
related to the whole system. For this Spencer’s two simplified questions could be adapted and used
(see Table 2).

**Table 2 Simplified questions for the Cognitive Walkthrough adapted from Spencer [42]**

<table>
<thead>
<tr>
<th>Simplified CW questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Would I (as a patient) know what to do at this step?</td>
</tr>
<tr>
<td>- If I do the right thing, do I (as a patient) know I have made progress toward my goal?</td>
</tr>
</tbody>
</table>

The Evaluation Process

Time and Resource Demands

To increase throughput, and minimize the burden of detailed low level sub-tasks, the repetitive work
of filling in system form answers with the 4 typical CW questions the use of Spencer’s two simplified
questions could also result in time efficiencies [42].

The volume of resources can also be simplified through the dual domain session facilitator who can
demonstrate aspects and tasks in the system, understand the user group needs/questions and
translate these findings to the system developers. The facilitator can also provide economies in terms
of demonstrating the system before the evaluation and also be part of the post-evaluation coding
and analysis.

Another way to decrease the time and resource requirements is to use streamlined and transparent
digital handling in the different parts of the evaluation process with digital video-recording
technology and software. This can aid in capturing the interaction in the application and the walk-
through of the application’s different parts by the session facilitator. This also reduces the need to fill
out an extensive number of forms during the session, as everything can be digitally captured. The
problem coding phase can then also be performed after the evaluation session which can make it
more efficient.

As all data will be digital, uniform and synchronized, it can be more easily imported into a Qualitative
Data Analysis System (QDAS) for the subsequent analyses. In addition, data can be shared more
easily and quickly so that more expert evaluators can be added in the analysis if needed. If the
material is analyzed in group sessions a projector, computer or digital systems for eMeetings can
contribute to decreasing the required time and resources. A last important point is that as all experts
in the analysis will be presented with identical captured digital material this can help assure
transparency, reliability of the material, contextual validity and validity in the analyses among experts
(see Table 3 for a summary of the method).
Table 3 Summary of Method Modifications

<table>
<thead>
<tr>
<th>Component</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>User involvement</strong></td>
<td>Use a purposive sample of chronic disease patient users of the self-management system/application who are diagnosed with the specific chronic disease condition with various severity levels of the disease, different genders, ethnic backgrounds, ICT skill levels and experiences to have a representative group of patient users (ideally 5 to 8 users per session)</td>
</tr>
<tr>
<td><strong>Dual domain facilitator</strong></td>
<td>Have a dual-domain expert facilitator who moderates and provides support to the patient users throughout the whole evaluation process and throughout all different steps of the self-management system evaluation.</td>
</tr>
<tr>
<td><strong>Task development process</strong></td>
<td>Use tasks with acceptable validity (measured using a content validity index). Have representative and patient-centered tasks developed based on disease-specific self-management guidelines and validated by a panel consisting of dual domain experts with expertise within HCI/Usability and the applicable health care domain as well as real chronic disease patient users.</td>
</tr>
<tr>
<td><strong>Include focus on higher level tasks</strong></td>
<td>Use both an individual- and group-mediated collaborative evaluation approach with the patient users first individually noting responses to tasks followed by a group discussion to achieve a higher level task focus. The evaluation is moderated by the dual-domain expert facilitator.</td>
</tr>
<tr>
<td><strong>Evaluation questions</strong></td>
<td>Have two short, open, simplified evaluation questions adapted from Spencer answered by the patient user evaluators individually for every task. Use these also as “triggers” for the subsequent group discussion, and collective focus for task performance.</td>
</tr>
<tr>
<td><strong>Digitally captured data</strong></td>
<td>Use digital recordings for data collection and all other parts of the CW evaluation session.</td>
</tr>
<tr>
<td><strong>Uniform coding, classification and problem rating</strong></td>
<td>Perform content analysis, coding, classification and rating of the usability problems in a purely digital environment for digital data collection/materials from recordings and answering sheets from the modified CW evaluation process using a Qualitative Data Analysis Software (QDAS).</td>
</tr>
</tbody>
</table>

Application of Improvements and Testing the Modified CW in a Case Study on a Diabetes mHealth Application

**Study Objectives**

We wanted to determine if we could improve upon the original CW method to address some of its shortcomings such as the user involvement and cumbersome process in some of its procedures but take advantage of its beneficial aspects, such as being group-based and effective as a whole compared to other methods. Our method modification was designed to address the lack of the inclusion of the user, and also determine if it was possible to streamline the evaluation and task development process further, with the inclusion of important validity aspects, to detect critical usability problems for users.

**Materials and Methods**

**Study Design**

In the following case study we tested the modified CW (UC-CW) on a diabetes mHealth application by applying the improvements discussed in Table 2. The intent was to determine the method’s effectiveness, efficiency and acceptance, and compare it against the “golden-standard” usability test with Think Aloud with two groups of 6 participants as a form of beginning validity.

To measure **effectiveness** the number of problems and problem types for each method were determined, as well as their heuristic violations and severity ratings. Each method was also compared on **efficiency** in terms of the time consumed for the whole evaluation procedure, the evaluation
process part itself and per detected usability problem. User NASA RTLX scores and participant opinions about the methods were compared to determine user acceptance levels. The comparison measures are provided in Table 4.

Table 4 Effectiveness, efficiency and acceptance of the modified method

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Effectiveness</td>
<td>Number and type of usability problems, heuristic determination, severity ratings</td>
</tr>
<tr>
<td>Efficiency</td>
<td>Total time for whole evaluation procedure, total time for evaluation process part, and time per detected usability problem</td>
</tr>
<tr>
<td>Acceptance</td>
<td>User acceptance through NASA RTLX (measuring method cognitive load), in-depth interviews</td>
</tr>
</tbody>
</table>

Study Setting and Sample

Participants

The study was conducted at the University of Utah, Salt Lake City, USA. After Institutional Review Board approval was received for the study (IRB_00101636) 12 patients were recruited. The setting for the assessments was the Diabetes and Endocrinology Center at University of Utah Health in Salt Lake City, Utah where the UC-CW session was conducted in a conference room fitted with a projector and individual TA tests in a designated, quiet room at the clinic. All participants filled out an individual written informed consent at the beginning of their respective evaluation session.

Inclusion- and Exclusion Criteria

The 12 diabetes patients were recruited by the first author M.G. at the Diabetes and Endocrinology Center using purposive sampling to ensure that users with different characteristics were represented in the method evaluations. The inclusion criteria were: (1) patients diagnosed with Type 1 or Type 2 diabetes (2) no cognitive impairment; (3) familiarity and some knowledge and use of computers, the Internet, and mobile phone; and (4) the ability to speak and understand the English language. Patients should also have no previous exposure to the mHealth application being evaluated (i.e. the Diabetes Diary). Patients were then randomized into the two test groups of 6 participants each to either perform the modified CW evaluation or individual usability tests. All participants received $20 gift cards for their participation.

The Diabetes mHealth Application

The evaluated mHealth application, called the Diabetes Diary, has been developed by the Norwegian Centre for E-health Research as a self-management support for Type 1 and Type 2 diabetes. This application allows patients to monitor their blood glucose, medication/insulin, food, and activity. They can also analyze previous situations involving their food and medicine to assist their decision making.

The patient’s blood glucose measurements can either be directly transferred to the application from a glucometer via Bluetooth or entered manually. The food and physical activity related data is entered manually. The glucose measurements can be displayed in graphs, and as trends reports with feedback provided through different color designations with red indicating levels below normal, green as normal, and yellow levels above normal (see Figure 1).
The application consists of a combination of views and sub-views. The Main View is connected to all views: 1) the Glucose Entry View, 2) the Carbohydrate Entry View, 3) the Insulin Entry View 4) the Activity Entry View 5) the Weight Entry View 6) the Medication Entry View, 7) the Glucose Graph View 8) the Preferences, 9) Tools and 10) List View. All the first 6 views are connected to the sub-views: a) Date Setting View and b) Time Setting View. Views 2 and 3 (the Carbohydrate Entry View and the Insulin Entry View) are in addition connected to the sub-view a) Similar Situation List View and view 7 (the Glucose Graph View) to the two sub-views: a) Plotted Graph View and b) Periodical Pattern Graph View (see Figure 2).

Figure 2 System view hierarchy of the Diabetes Diary
Study Instruments

Demographic Form and ICT and mHealth Experience and Perception Questionnaire
Participants completed a pre-test demographic form and a questionnaire about their experience and perceptions of IT, computers and mobile technology. Questions in the demographic part included background details on age, gender, education, ethnicity, occupation and disease information such as type of diabetes and length of diabetes diagnosis.

Participants were also asked about their frequency of computer use and knowledge, and frequency of use and knowledge of the Internet. Questions about their mobile phone usage involved frequency of use, how often they used their phone to make and receive calls, how often they used it for text messaging and how often they used it for e-mails, surfing and using apps. They were also asked how they considered their mobile phone knowledge level.

In addition, participants were assessed on a Likert scale of 5 points (from Strongly agree to Strongly disagree) if they agreed to liking to use their computer and mobile phone in their life on a daily basis, as well as if they felt eHealth services and mobile health services getting more common within health care was positive and if they could see personal advantages in the usage of eHealth and mobile health services.

The questionnaire was validated for face validity by a panel of three health care professionals, three usability experts and one patient.

NASA RTLX Instrument and In-depth Interview
Post-evaluation, participants filled in an instrument of their experience with the method (NASA RTLX) based on 6 component scales and answered a set of 4 questions in a short in-depth interview about how they each experienced the particular method.

The NASA RTLX instrument developed by Byers et al [43] was used to determine method acceptability to users or how easy or difficult they felt it was. This short and simplified version was based on the original NASA-Task Load Index by Hart and Staveland [44]. The six components or dimensions that are assessed are: 1) Mental Demand, 2) Physical Demand, 3) Temporal Demand, along with 4) Performance, 5) Effort, and 6) Frustration. Each scale signifies how a specific activity contributes to the workload based on the user’s perspective (with the combined score of the scales from 0-100). In the original NASA-Task Load Index these factors are also weighted through a pairwise comparison by the user on their perceived importance. This addition has been removed in the shorter NASA RTLX instrument, which researchers point to has also made this particular version achieve a high experimental validity [45]. The participant’s overall workload is determined by adding the total scores and divide it on the six scales. The higher the resulting average, the higher the experienced workload or cognitive load [44].

A set of four questions were used to determine how each individual subjectively experienced the method they were a part of. These consisted of what their thoughts were about the method, if they felt it easy or hard to understand (and if so in what way), if they felt it was easy or hard to perform the different method parts and what their overall experience was of the method (?).

Procedures
The different evaluations started with all 6 patients filling out their individual written consent forms, and their pre-test questionnaires. After a demonstration of the system to users, they interacted with the mHealth application using the different methods. After the evaluation sessions were completed,
Each participant filled in the NASA RTLX and took part in the short in-depth interview. Each method evaluation procedure is described below:

a) The UC-CW Evaluation

The UC-CW evaluation began with that suitable scenarios and tasks were developed for the evaluation, and then validated in a judgment process consisting of two steps. Author M.G. created the tasks based on an extensive review of diabetes self-management guidelines [46-48] to ensure that all important self-management aspects were included. These tasks were then reviewed by a panel of three dual domain experts (in usability and healthcare) as well as a diabetes patient for accuracy and also content validity indexed. Two iterations were conducted before reaching an acceptable total score of 0.93/1.0 [49]. The resulting 14 validated tasks are listed in Table 5.

Table 5 Categories and task descriptions for the UC-CW evaluation

<table>
<thead>
<tr>
<th>Category, Scenario</th>
<th>Task</th>
<th>Task name/Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Register entries</td>
<td>1</td>
<td>Enter weight</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Enter physical activity</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Enter glucose reading</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Enter carbohydrate intake</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Enter insulin, medication</td>
</tr>
<tr>
<td>Interpret entries</td>
<td>6</td>
<td>Interpret last entries for glucose, insulin, physical activity and carbohydrates intake on the main view</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>Interpret the last 24 hrs glucose entries on the main view</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>Interpret entries in plotted graphs</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>Interpret entries in linear distribution graphs</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>Interpret entries in list</td>
</tr>
<tr>
<td>Manage entries (part 1)</td>
<td>11</td>
<td>Search and find entries in list</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>Correct, alter entries</td>
</tr>
<tr>
<td>Apply self-management decision making</td>
<td>13</td>
<td>Perform a similar situations search</td>
</tr>
<tr>
<td>Manage entries (part 2)</td>
<td>14</td>
<td>Export, send entries</td>
</tr>
</tbody>
</table>

Then, the actual evaluation occurred with the dual domain session facilitator M.G. and a mix of Type 1 and Type 2 diabetes patients of varying ages, genders, educational backgrounds and ethnicities. M.G. followed a standard procedure to decrease variability which entailed informing about the method for about 5 minutes and then demonstrating the mHealth application on a large screen for 15 minutes. Each patient was then provided a booklet of the 14 tasks. Each scenario and task consisted of a space to record their individual observations, and answers to the two previously stated simplified CW questions for the subsequent group discussion. Participants took turns in reading out each scenario and task and the group then guided the facilitator in performing the task on the screen showing the projected application. After each task, patients first expressed their individual thoughts about usability problem/s in the application. Then, for a higher task level response, the group added additional thoughts about possible flaws. The whole session was digitally audio and video recorded through Morae software and took 2 hours and 20 minutes.

b) The User Test with TA Evaluation

The TA evaluation session was performed as described by Lewis [14, 15] based on Ericsson and Simon [16-18]. Due to the importance and necessity of adhering closely to the original method for comparability and also to prevent bias, a task list was similarly created by M.G. based on mHealth system studies that utilized user tests with TA for diabetes self-management [50, 51]. This resulted in a list 9 tasks for patients to perform (see list of tasks in Table 6). Then the actual evaluation occurred. This consisted of 6 individual sessions with diabetes patients of varying genders, ethnicities, ages and
Table 6 Categories and task descriptions for the TA evaluation

<table>
<thead>
<tr>
<th>Category</th>
<th>Task</th>
<th>Task name/Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjust, personalize application</td>
<td>1</td>
<td>Set blood glucose measurement unit</td>
</tr>
<tr>
<td>Register entries</td>
<td>2</td>
<td>Enter glucose reading</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Enter carbohydrate intake</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Enter insulin, medication</td>
</tr>
<tr>
<td>Interpret entries</td>
<td>5</td>
<td>Interpret entries in periodical pattern graphs</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>Interpret entries in plotted graphs</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>Interpret entries in list</td>
</tr>
<tr>
<td>Manage entries</td>
<td>8</td>
<td>Search and find entries in list</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>Export, send entries</td>
</tr>
</tbody>
</table>

After M.G. had performed a similar standardized procedure of informing about the method for about 5 minutes and demonstrating the mHealth application to each individual participant on the computer screen for about 15 minutes the individual patient received a booklet of the 9 tasks to perform. During the individual sessions, patients were asked to think aloud as they performed each task in the application and if they fell silent they were encouraged by M.G. to keep talking. Any other interference with their thought processes was avoided. The system interactions were also digitally audio and video recorded with Morae software. M.G. also recorded any observations and notations about the patient task performances directly into Morae. The individual TA sessions took a total of 4 hours and 47 minutes for all sessions.

Data Analyses
Demographic Questionnaires and NASA RTLX
The demographic characteristics, technology experience and perceptions as well as the NASA RTLX scores were analyzed using descriptive statistics.

Usability Problem Determination and Rating per Method
A problem analysis and coding phase occurred for the UC-CW and then for the TA. This process was completed by authors M.G. and N.S. to ensure consistency and comparability between the methods.

The first part of this coding and analysis phase consisted of importing the recorded data into a QDAS program, Nvivo 10. The coding process for the UC-CW and TA consisted of inductively arriving at usability problems [52]. Then, duplicates were removed and final master lists of identified problems were produced for each method. Data analysis then occurred where problems were assigned an appropriate heuristic category and severity rated. For this Nielsen’s list of 10 heuristic categories and severity rating scale of 0 to 4 [24] were used. The coded and rated problems were verified by Author A.K. (see Appendix A for the list of heuristic categories and severity rating levels).

Interviews
The individual interviews were transcribed and entered into Nvivo 10. These were then thematically coded by author M.G. using an inductive coding process from Hsieh and Shannon [52] to derive at common themes. The coded themes were verified by author N.S.

Results
Below follow the participant characteristics, the IT/computer, mHealth experience and use and perceptions about eHealth and mHealth for the UC-CW group along with the effectiveness and
efficiency and acceptance results for this method. This section is followed by a comparison of the UC-CW against the validation method, the TA.

The UC-CW

User Characteristics
The patient participants included two women and four men, all between 25 and 75 years old. One participant was African-American and the rest Caucasian, all with college or university education. Three of them were employees, two retired and one a student. Three had Type 1 and three Type 2 diabetes. The length of their diabetes diagnosis varied from 3 up to 27 years.

Computer/IT, Mobile Phone Experience and Use
Five of the participants used computers daily, and one 3-6 times a week. Two of them considered their computer/IT knowledge to be high and the rest medium. All participants used the Internet daily. Three thought that they had high Internet knowledge, while two thought it to be medium, and one small. Five of them used their mobile phone daily and one 3-6 times a week. Two considered their mobile phone knowledge level to be high, three medium and one small. Four used their mobile phone to make and receive phone calls every day, and two 3-6 times a week. Five used their mobile phone to text message daily while one 3-6 times a week. In addition, four used their mobile phone for e-mails, apps and surfing on a daily basis, one 1-2 times a week, and one never.

Perceptions about eHealth and mHealth
Of the participants, three strongly agreed and rest agreed to liking to use the computer in their daily life. Three also strongly agreed, two agreed and one was neutral to enjoying to use their mobile phone in their daily life. In addition, five strongly agreed and one was neutral to that it was a positive development that eHealth and mobile health services are getting more common. Three also strongly agreed, two agreed and one was neutral to seeing advantages personally in the usage of eHealth and mobile health services.

Identified Usability Problems and Heuristic Violations across Application Views
Twenty-six unique usability problems were identified in the UC-CW. These usability problems are shown by view, number of heuristic violations and mean severity ratings in Figure 3 below. The usability problems ranged from 1 to a high of 11 problems across the different views. Problems included a lack of individual personalization possibilities, navigational difficulties and unclarities when it came to the display of features. The Main View generated most usability problems (11) and violated 29 heuristics, followed by the Carbohydrate Entry View with 4 problems and 14 heuristics, and the Plotted Graph View and List View with 3 problems each and 8 heuristics each. The lowest number of usability problems were in the Time Setting View, the Glucose Entry View and the Periodical Pattern Graph View 1 problem each and 3 violated heuristics each.

The average severity ratings ranged from 2 to 3.5 on the scale of 0 to 4 with the Similar Situations List View the highest and catastrophic rating followed by the List View, the Time Setting View, and Carbohydrate Entry View with major severity ratings.
Figure 3 Usability problems by view, number of heuristic violations and average severity ratings for the UC-CW

Heuristic Violation Categories

Seventyfour heuristic violations were present for all views. Match Between the System and the Real World dominated at 25.7% (n=19), followed by Consistency and Standards at 22.9 % (n=17) and Recognition Rather than Recall at 17.6 % (n=13). User control and freedom at 1.4% (n=1) had the fewest violations across all views (see Figure 4).

Figure 4 Heuristic violation categories for the UC-CW
Severity Ratings Across System Views

The mean severity rating for all views in the application was 2.7 which is a major severity. As may be
seen in Figure 5, the heuristic violations with the highest severity ratings were in the *Main View* with
2 major and 2 catastrophic problems. *The Carbohydrate Entry View* had 4 major problems while the
*List View* had 2 major problems and one catastrophic problem and *the Similar Situations List View*
one major and one catastrophic problem. In addition, the *Plotted Graph View* had one catastrophic
but also one minor and cosmetic problem and the *Time Setting View* had one major problem.

![Figure 5 Severity ratings per view UC-CW](image)

**Figure 5 Severity ratings per view UC-CW**

Usability Problem Characteristics

As seen, the five problems that were most severe (with a severity rating of 4) were located in the
*Main View, Plotted Graph View, List View, and Similar Situation List View*.

The two problems in the *Main View* entailed one problem about that it was not possible to configure,
nor personalize ranges based on patients’ own ones for low, medium and high values visualized in
the middle pane for the latest glucose measurements, an important aspect for their individual
disease needs. The other problem was a navigational concern that it was unclear to participants that
they needed to swipe across the bottom graphical image in this view to navigate to the *Periodical
Pattern Graph View* to check last week’s glucose measurements. Here a completely different action
was needed than the previously performed user steps.

Another personalization problem was in the *Plotted Graph View* which consisted of that range or cut-
off levels in the graphs showing the low, mid, and high glucose values could also not be adjusted for
the individual patient which was a problem of similar importance as the previously stated one. The
problem in the *List View* was that participants felt it unclear how to navigate in and understand their
location when they entered the list of registered food, glucose, medication and insulin. It was also
not possible to obtain a response on the sequence of dates in the all entries list, confusing patients.
Due to these difficulties these problems also received a high severity rating.

The last of the most severe problems was in the *Similar Situation List View* where participants
experienced it difficult to comprehend and also interpret how the number of IUs of insulin as well as
mg/dl blood glucose units in the view’s upper part related to the list of entries in the sections on the bottom. This is a vital aspect to clarify as this function allows patients to view previous insulin dosages and their number of carbohydrates to then determine their insulin requirements.

Time Estimates for Entire Evaluation Procedure, per Detected Usability Problem, and Evaluation Session
The literature review of the diabetes self-management guidelines for the UC-CW took 480 minutes, and the same amount of time for the construction of the evaluation tasks. The task validation and content validity indexing process consumed a total of 130 minutes (divided upon the dual domain researchers and the diabetes patient). The time taken for facilitating the session was 140 minutes in total. In addition, the coding of usability problems took 480 minutes and assigning heuristic violations and severity ratings 120 minutes. This gave an overall time of 1830 minutes. Divided on the 26 usability problems this entailed a total of 70.4 minutes per usability problem.

The evaluation session itself took 5 minutes to inform about the method, 15 minutes to demonstrate the application and 120 minutes to perform the evaluation. Together this totaled 140 minutes for all 6 participants.

NASA RTLX Scores and Interviews
The NASA RTLX scores showed participants’ experienced cognitive load. The mean value for the UC-CW participant group was 25.3 with a standard deviation score (SD) of 3.2.

Participant Perceptions about the Method
The post-test questions on method experiences gave subjective data about how participants experienced the method. All six patients found the UC-CW easy to understand and easy to perform. They thought the method was clear and seamless (6), provided the ability for the participants to express themselves, stimulated discussion (6), was a straightforward method (5), and a good experience (6). They liked having individual and collaborative parts where they could write down their own thoughts and then discuss with the other patients (5). They felt that the discussion component triggered additional thoughts and helped them find more usability problems (6) and also liked the ability to work together to solve the tasks (6). In addition, the patients emphasized that it was positive to them to have mixed patient groups with different diabetes types, ages, and years with the disease (6). Four of the patients also liked having the image of the mobile phone projected so that everyone could see the image at the same time. Negative comments included an overlap in some questions (2) and also a redundancy in answers regarding participants’ task experiences (1).

Comparisons between Evaluation Methods
A series of analyses were completed to compare the results of the UC-CW against the TA method.

The methods were first compared on their user characteristics, computer/IT and mobile phone experience and use and perceptions about eHealth and mHealth. Then they were compared on the effectiveness measures: number of detected usability problems, violated heuristics, their severity levels as well as the types of problems. They were also compared on the efficiency measures of the total time for the whole evaluation procedure, and time per detected problem as well as the time for the evaluation part. Finally the user acceptance was compared through the NASA RTLX and in-depth interview results.
User Characteristics, Computer/IT, Mobile Phone Experience and Use, Perceptions about eHealth and mHealth Compared

When comparing the UC-CW results with the TA, the UC-CW had 4 men and two women while the TA sample had an equal number of men and women (3 each). The UC-CW also had the youngest and oldest participants of 25 and 75 years old (mean age 48) compared to the ages between 35 and 69 (mean age of 52) in the TA sample. In terms of ethnicity, the UC-CW had a less varied group of five Caucasians and one African-American compared to the TA with four Caucasians, one African-American and one Latin-American participant. In terms of education, the UC-CW group had more – all college or university compared to four with college or university, and two with high school as their highest education in the TA group. In terms of occupation three were employed, two retired and one a student in the UC-CW group compared to the TA group where four were employed and two retired. In both methods half of the participants had Type 1 and half Type 2 diabetes. The diabetes diagnosis length varied slightly; the UC-CW group from 3 to 27 years (mean 14.8) and the TA group 1 to 33 years (mean 18.8) (see Table 7).

Table 7 User characteristics UC-CW and TA

<table>
<thead>
<tr>
<th>User characteristics</th>
<th>UC-CW*</th>
<th>TA*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Male</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>2</td>
</tr>
<tr>
<td>Gender</td>
<td>Female</td>
<td>3</td>
</tr>
<tr>
<td>Age</td>
<td>Years, Mean (SD)</td>
<td>48(19.6)</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>White/Caucasian</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Black/African American</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Hispanic/Latino</td>
<td>-</td>
</tr>
<tr>
<td>Education</td>
<td>High School</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>College/University</td>
<td>6</td>
</tr>
<tr>
<td>Occupation</td>
<td>Retired</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Employed</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Student</td>
<td>1</td>
</tr>
<tr>
<td>Years with diabetes</td>
<td>Years, Mean (SD)</td>
<td>14.8(8.5)</td>
</tr>
<tr>
<td>Type of diabetes</td>
<td>Type 1</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Type 2</td>
<td>3</td>
</tr>
</tbody>
</table>

* UC-CW = User-centered Cognitive Walkthrough, TA = Think Aloud usability test

Computer/IT, mobile experience and use

Regarding computer/IT experience and use, as described earlier, five participants in the UC-CW group used the computer daily with one 3-6 times a week. This compared to the TA group where computer use varied among the participants with half using a computer daily and half 3-6 times a week. In addition, two in the UC-CW group considered themselves to have high computer/IT knowledge and the rest medium while in the TA group one experienced their knowledge as high, two as medium and three as small. When it came to the Internet, all participants in the UC-CW group used the Internet every day. In the TA group, however, half used the Internet on a daily basis, and half 3-6 times a week. Three in the UC-CW group also experienced their Internet knowledge to be high while two thought it medium and one small. Of the participants in the TA group one considered their knowledge high, three medium, and two small.

For mobile phone use in the UC-CW group, five participants used theirs every day and one 3-6 times a week. These were the same figures in the TA group. When it came to mobile phone knowledge,
two in the UC-CW group considered their knowledge high, three medium, and one small. These were similar figures to the TA group, with two perceiving it high, two medium, and two small. Four in the UC-CW group used their mobile phone to make and receive phone calls every day, and the rest 3-6 days a week. Five in the TA group also used their mobile phone to make and receive phone calls daily, and one 3-6 times a week. Similarly, five in the UC-CW group as well as the TA group used their mobile phone daily for text messaging, and one did this 3-6 times a week. When it came to using it for apps, e-mail and surfing, four in the UC-CW group expressed doing this every day, one 1-2 times a week, and one never. For the TA group four similarly stated doing this daily, one 3-6 times a week, and one never (see Table 8).

<table>
<thead>
<tr>
<th>Computer/IT, mobile phone experience and use</th>
<th>UC-CW*</th>
<th>TA*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency of computer use</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Every day</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>3-6 times/week</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Computer/IT-knowledge</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Medium</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Small</td>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td>Frequency of Internet use</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Every day</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>3-6 times/week</td>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td>Internet-knowledge</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Medium</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Small</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Frequency of mobile phone use</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Every day</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>3-6 times/week</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Mobile phone-knowledge</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Medium</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Small</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Mobile phone use for phone calls</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Every day</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>3-6 times/week</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Mobile phone use for text messaging</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Every day</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>3-6 times/week</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Mobile phone use for e-mails, surfing, apps</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Every day</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>3-6 times/week</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>1-2 times/week</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>Never</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

* UC-CW = User-centered Cognitive Walkthrough, TA = Think Aloud usability test

Perceptions about eHealth and mHealth

Regarding questions in the UC-CW group about if they liked using the computer in their daily life half of the participants strongly agreed and half agreed while in the TA group one strongly agreed, while two agreed and three were neutral to this statement. In the UC-CW group, three also strongly agreed, two agreed and one was neutral to like using their mobile phone daily. In the TA group, similarly, two strongly agreed three agreed and one was neutral to the statement. In the UC-CW group five participants strongly agreed while one was neutral to eHealth and mobile health services within health care getting more common was positive. Two participants in the TA group also strongly agreed and four agreed with this statement. Regarding whether they could see advantages for themselves personally in using eHealth and mobile health services three strongly agreed, two agreed
and one was neutral in the UC-CW group. In the TA group, however, two strongly agreed, while one participant agreed and three were neutral (See Table 9).

### Table 9 Computer/IT, mobile phone, eHealth and mHealth perceptions

<table>
<thead>
<tr>
<th>Computer/IT, mobile phone, ehealth and mHealth perceptions</th>
<th>UC-CW*</th>
<th>TA*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enjoy using computer in daily life</td>
<td>Strongly agree</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Agree</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Neutral</td>
<td>-</td>
</tr>
<tr>
<td>Enjoy using mobile phone in daily life</td>
<td>Strongly agree</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Agree</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Neutral</td>
<td>1</td>
</tr>
<tr>
<td>E-Health services, mobile health services for support,</td>
<td>Strongly agree</td>
<td>5</td>
</tr>
<tr>
<td>information within health care is a positive development</td>
<td>Agree</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Neutral</td>
<td>1</td>
</tr>
<tr>
<td>Can see advantages personally using E-health services,</td>
<td>Strongly agree</td>
<td>3</td>
</tr>
<tr>
<td>mobile health services within health care</td>
<td>Agree</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Neutral</td>
<td>1</td>
</tr>
</tbody>
</table>

* UC-CW = User-centered Cognitive Walkthrough, TA = Think Aloud usability test

Comparison of the Number of Detected Usability Problems, Number of Heuristics Violated and Average Severity for the UC-CW and TA

As may be seen in Figure 6, the highest number of usability problems and heuristic violations for the UC-CW as well as for the TA were in the Main View containing 11 problems and 29 heuristic violations versus 10 problems and 24 heuristic violations. In both methods, the Carbohydrate Entry View followed with 4 problems and 14 heuristic violations and 3 problems and 9 heuristic violations, and the List View at 3 problems and 8 heuristic violations for each. Both methods also had a similar number of problems as well as heuristic violations for the Glucose Entry View with 1 problem and three heuristic violations and 1 problem and 2 heuristic violations. For the Main View and the Time Setting View the severity ratings were also similar at 2.45 and 2.5 and 3 (for each method) as well as somewhat similar for the Plotted Graph View and List View at 2.3 and 2 and 3.3 and 3 respectively.

The methods also differed for some of the views. In the Similar Situation View, for example, 2 problems and 6 heuristic violations were detected by the UC-CW while none with the TA. Also in the Periodical Pattern Graph View the UC-CW found 1 problem and 3 heuristic violations and the TA none. The UC-CW also had the highest severity rating for the Similar Situation View at 3.5.
When it comes to the heuristic violations, a total of 74 occurred for all views in the UC-CW while there were 51 in the TA. As can be seen, both the UC-CW and TA had a majority of their heuristic violations in the category Match Between the System and the Real World at 25.7% (n=19) and 31.4% (n=16). This was followed by the category Consistency and Standards with the UC-CW at 22.9% (n=17) and the TA at 29.4% (n=15) and Recognition Rather than Recall with the UC-CW at 17.6% (n=13) and the TA at 19.6% (n=10). User control and freedom had the fewest violations for both methods across views at 1.4% (n=1) and 3.9% (n=2) except for the categories Recover and Help that had none for either method (see Figure 7).
Comparisons on the Usability Problems, Heuristics and Severity Ratings

The UC-CW detected a total of 26 unique usability problems compared to 20 problems for the TA. As shown (see Figure 8), the UC-CW also had more catastrophic and major severity ratings overall compared to the TA with 5 catastrophic and 10 major ratings versus 2 catastrophic and 8 major ones. More minor ratings occurred in the TA with 10 compared to 9 ratings, but also two cosmetic ratings in the UC-CW.

![Figure 8 Comparison on usability problems, their classifications on heuristics and severity levels for the UC-CW and TA](image)

When it comes to individual views, the UC-CW had a majority of the most severe problems in the Main View with 2 catastrophic and 2 major problems. This was similar to the TA with one catastrophic and 3 major problems here. The List View contained one catastrophic and 2 major problems for the UC-CW while the TA had one catastrophic, one major and one minor problem in this view. In addition, one catastrophic, one major and one minor problem were also in the Plotted Graph View for the UC-CW where the TA had one minor problem. The UC-CW also had one catastrophic as well as one major problem in the Similar Situation List View. Here the TA had none.

The majority of the minor usability problems in the UC-CW were also located in the Main View (n=6), followed by the Glucose Entry View, Plotted Graph View, and Periodical Pattern Graph View with one minor rating each. Two views also had cosmetic ratings which were the Main View and Plotted Graph View with one such violation each. Similarly to the UC-CW, the majority of the minor usability problems were also located in the Main View (n=6) for the TA, followed by the Carbohydrate Entry View with 2 minor ratings and the Plotted Graph View and List View with one minor rating each. There were no cosmetic problems in the TA.

As stated earlier, the UC-CW had the highest average severity ratings for the Similar Situation List View at 3.5 not present in the TA. The methods also varied in their severity ratings for the Glucose Entry View (at 2 versus 3), Carbohydrate Entry View (at 3 versus 2.3) and Periodical Pattern Graph View (at 2 versus none). They, however, had similar severity ratings for the List View with a severity rating of 3.3 and 3. Both methods also had similar severity ratings on the Time Setting View (3 each)
and Main View at 2.4 and 2.5. In addition to this, both methods also had similar average severity ratings for the whole mHealth application with 2.7 for the UC-CW and 2.6 for the TA; both indicating a major severity, and converged on a total of 8 usability problems (see the list of all problems and descriptions in Appendix B).

Comparison between Methods on Usability Problem Characteristics

As stated previously, the most severe and critical problems for the UC-CW were five and for the TA two. The problems that were unique for the UC-CW entailed the need to have the ability to personalize features for patients own particular disease needs. This involved that in the current version of the application the low, medium and high ranges that were visualized in different ways for the latest glucose measurements could not be configured and personalized based on the patients’ own. (Main View and Plotted Graph View). In addition, there were also unclarities for participants regarding the relation between the insulin and blood glucose units displayed in the upper part compared to the lower part for the Similar Situation List View.

The navigational difficulties involved that it was unclear with the action of needing to swipe across the graphical image on the lower part of the view to get to where participants could check their glucose measurements for the previous week (Main View), and difficulties on how to navigate in and comprehend where one was located when entering the list of registered food, glucose, and medication. Here participants could also not get a response on the sequence of dates in the list of all entries making them confused (List View). These two problems were present in both the UC-CW and the TA (see Table 10 for the most severe and critical problems).

<table>
<thead>
<tr>
<th>Table 10 Usability problem descriptions, detection by method, heuristic violations and assigned severity rating for top five detected usability problems</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Usability Problem Description</strong></td>
</tr>
<tr>
<td>---------------------------------------------------------------</td>
</tr>
<tr>
<td>It is not possible to configure and personalize ranges based on the patients' own ones for the low, medium and high ranges that are visualized in the middle pane for the latest glucose measurements.</td>
</tr>
<tr>
<td>Unclear to know that it is necessary to swipe across the chosen graphical image on the bottom of the Main View to get to the Periodical Pattern Graph View.</td>
</tr>
<tr>
<td>It is not possible to configure and personalize the range levels (cut off levels) in the graphs that show the glucose values for the low, mid and high ranges.</td>
</tr>
<tr>
<td>Unclear to know how to navigate in, understand where one is located and get a response on the sequence of dates in the list of all entries.</td>
</tr>
<tr>
<td>Difficult to understand and interpret the relation between the number of IUs of insulin and mg/dl blood glucose units in the upper part of the image compared to the list of entries in the bottom part of the view for this function.</td>
</tr>
</tbody>
</table>

* UC-CW = User-centered Cognitive Walkthrough, TA = Think Aloud usability test
Comparison between Method Time Estimates for Whole Evaluation Procedure

As described previously for the UC-CW, the literature review of the diabetes self-management guidelines as well as the construction of evaluation tasks took 480 minutes each, the task validation as well as content validity indexing a total of 130 minutes, and the session facilitation 140 minutes while the coding of usability problems took 480 minutes and the assigning of heuristic violations and severity ratings 120 minutes in total. This gave an overall time of 1830 minutes for the whole evaluation. When dividing the time on the 26 problems this was 70.4 minutes per resulting usability problem. In comparison, the TA took 480 minutes in total for the task construction, and 281 minutes for the facilitation of the sessions. Here, the coding of the usability problems took 724 minutes for the 6 individual patient evaluation sessions and assigning heuristic violations and severity ratings 120 minutes. This resulted in a total time of 1605 minutes. Divided upon the 20 resulting usability problems this turned out to be 80.2 minutes per usability problem. This showed the UC-CW as almost 10 minutes faster per found problem for the whole evaluation procedure (see Table 11).

Table 11 Time in minutes for whole evaluation per method and per identified usability problem

<table>
<thead>
<tr>
<th>Usability Evaluation Method</th>
<th>Task literature review and evidence</th>
<th>Task construction</th>
<th>Task validation</th>
<th>Leading Evaluation Session</th>
<th>Data coding and UP determination</th>
<th>Assigning violations and severity ratings</th>
<th>Total time per method</th>
<th>Total time per identified usability problem</th>
</tr>
</thead>
<tbody>
<tr>
<td>UC-CW*</td>
<td>480</td>
<td>480</td>
<td>130</td>
<td>140</td>
<td>480</td>
<td>120</td>
<td>1830</td>
<td>70.4</td>
</tr>
<tr>
<td>TA*</td>
<td>N/A</td>
<td>480</td>
<td>N/A</td>
<td>281</td>
<td>724</td>
<td>120</td>
<td>1605</td>
<td>80.2</td>
</tr>
</tbody>
</table>

* UC-CW = User-centered Cognitive Walkthrough, TA = Think Aloud usability test

Comparison between Method Time Estimates for Evaluation Process

As stated previously, the UC-CW consumed 5 minutes for the method information, 15 minutes for the application demonstration, and 120 minutes for the evaluation session with the six participants. This gave a total of 140 minutes. In comparison, the TA consumed 30 minutes for method information with 5 minutes for each of the 6 participants. The application demonstration took 70 minutes; including 10 to 15 minutes for each of the 6 participants (mean 11.7, SD 2.6). The individual six evaluations took 181 minutes in total, with the slowest participant taking 39 minutes and the fastest 22 minutes for this part (mean 30.2, SD 6.7). This amounted to a total of 281 minutes. This showed the UC-CW to faster for the evaluation process (See Table 12).
Comparison between method acceptability to users

NASA RTLX scores
The NASA RTLX scores showed a mean of 25.3 for the UC-CW group with a standard deviation score (SD) of 3.2 while the mean for the TA participants was 53.5 and with a standard deviation score of 14.2. This showed almost half the cognitive load for the UC-CW in comparison to the TA.

Participant Perceptions
As indicated previously, all six patients found the user-centered CW easy to understand and easy to perform including aspects such as that it was clear and seamless, provided the ability for the participants to express themselves, and a good experience. In addition they all felt that the method stimulated discussion and that the discussion component triggered additional thoughts and helped them find more usability problems. They also liked the ability to work together to solve the tasks and felt it positive having mixed patient groups with different diabetes types, ages, and years with the disease. Five of the participants expressed that they also liked having individual and collaborative parts where they could write down their own thoughts and then discuss with the other patients, and five also felt that the method was straightforward. Four enjoyed having the image of the mobile phone projected so that it was visible for all at the same time. Two negative comments included that patients felt there was an overlap in some questions and one patient also expressed that there was some redundancy in the answers and experiences regarding the tasks.

In comparison, four of the six patients thought that the TA was easy to understand and one patient that it was not difficult to perform. Four also stated that they felt the method was unusual, especially the think aloud component and that it constituted a heavy cognitive load while three felt it was difficult to perform and three that it was rather awkward.

Discussion
We developed a modified user-centered CW to attempt to address method shortcomings as identified during a detailed literature review. Our method modifications consisted of: making the user the main contributor in the evaluation in detecting usability problems, a dual domain facilitator with the role of guiding the evaluation session and rating found usability issues, as well as a task development process with validated tasks, a focus on higher level tasks in the evaluation and on streamlining the evaluation process with regards to used time and resources.

As seen, the patient groups for the modified UC-CW and TA were similar on characteristics (type of diabetes, length of diagnosis, gender, age, occupation and fairly similar on ethnicity). They were also similar on mHealth use (mobile phone use for calls, and text messaging, usage of apps, e-mails and

---

### Table 12 Time per evaluation session part

<table>
<thead>
<tr>
<th>Usability Evaluation Method</th>
<th>Method info (min)</th>
<th>App tutorial (min)</th>
<th>Evaluation Session (min)</th>
<th>Total time per method session</th>
</tr>
</thead>
<tbody>
<tr>
<td>UC-CW*</td>
<td>5</td>
<td>15</td>
<td>120</td>
<td>140</td>
</tr>
<tr>
<td>TA*</td>
<td>30</td>
<td>70</td>
<td>181</td>
<td>281</td>
</tr>
</tbody>
</table>

* UC-CW = User-centered Cognitive Walkthrough, TA = Think Aloud usability test
They, however, differed slightly on some personal characteristics (education, computer- and IT experience) where the UC-CW group had slightly more education and computer experience. On knowledge and experience of mobile phones, the groups were similar on levels of knowledge but differed on types of knowledge (computer, IT) where the TA group had less knowledge. The TA group also had a slightly more critical view towards the technology compared to the UC-CW group and slightly less Internet knowledge. The groups were similar on liking to use the mobile phone in their daily life and about thinking that eHealth and mobile phone health services getting more common within healthcare. They both saw advantages for themselves in using these services (where more were neutral in the TA group). The groups again differed on computer use in their daily life where more in the UC-CW were positive about this statement.

Even if our purpose was not to compare user groups on their performance, but to compare our method modification to the established usability methods, the user groups’ characteristics may have influenced their overall perception of the methods they were involved in. For example, the slightly higher education and comfort with computers as well as the slightly more positive view towards the use of mobile health for their condition in the UC-CW group compared to the TA group might have made the UC-CW group more comfortable with their evaluation method and being more positive towards the use of these technologies overall. This was similar to the findings of Sarkar et al [52] who found that while their diverse group of patients of lower socioeconomic status were interested in using mHealth technology they had difficulties with its usage due to that they were not accustomed to its different features, and struggled with it.

Determined Usability Problems and Usability Problem Characteristics

The number and severity of identified usability problems with the UC-CW were higher compared to the TA method (5 of a critical nature compared to two). The UC-CW identified important deficiencies about personalization features that patients needed for their individual disease management such as that ranges could not be configured and personalized for low, medium and high glucose values in the middle pane for patients latest measurements in the Main View and it was also not possible to configure the range delimiters in the graph based on the patient’s own cut off levels (for the low, medium and high glucose value ranges) in the Plotted Graph View. The importance of individualization of mHealth interventions has been brought up in the literature where Scheibe et al [53] found similar needs for their evaluated diabetes patient groups. There participants also stated that the applications needed to be individually adaptable to suit their specific needs. Due to that these personalization aspects were only found in the UC-CW and not in the TA they might have been an immediate result of the user group component of the UC-CW as it allowed both Type 1 and 2 patients to discuss the problems on a higher level and how the mHealth application applied to each of them. This also provides evidence that this modified method can be suitable and contribute in finding those important, severe usability problems for users which authors have pointed to has historically been one of the major drawbacks with the original CW method [29, 54]. This along with the ability contribute to identifying the more recurring usability issues as opposed to one-time, low priority ones [29].

Except for the above findings, the UC-CW and TA were comparable on the severity of the identified problems and their classification into heuristic categories: Match between the System and the Real World, as well as Consistency and Standards and Recognition Rather than Recall. Most of the usability problems were also in the Main View, the Carbohydrate Entry View as well as List View for both methods, and the two methods also had 8 problems in common. Importantly, they also had similar major severity ratings for the whole mHealth application. This signifies that the UC-CW was able to detect as severe and as relevant problems as did the TA. Overall, it is therefore possible to
conclude that the UC-CW has a beginning validity. These results would, however, also need to be verified with other mHealth systems, applications and also other patient groups.

Another goal for developing the user-centered CW was to attempt to make the process of evaluation more efficient, streamlined as well as perform important validation measures during the process. The importance of including validations against clinical guidelines for self-management devices has been brought up in the literature [5] as well performing validation measures in usability evaluation [55]. Here the results indicated that adding validation aspects into the evaluation procedure for the UC-CW made it take longer as a whole when compared to the TA. However, the UC-CW identified more usability problems and also more critical problems and once the time was divided per usability problem, the UC-CW turned out as almost 10 minutes faster per detected problem. Also for the evaluation process itself the UC-CW was faster at 140 minutes compared to 281 minutes for the TA. All in all this indicates that our method modification, despite the addition of some important and time consuming validation aspects such as guideline review and content validity index rating turned out to be faster.

Method Acceptability to Users
When examining the results, the user-centered CW was considered half as cognitive demanding as the user test with TA. The subjective comments indicated that the UC-CW was also easier to understand and perform for participants when compared to the statements about the TA which they experienced as rather awkward as well as cognitive demanding. The user acceptance when it comes to the method itself has currently not been explored to a great extent in the literature on usability methods but appears to be an important dimension to add when deciding between different methods. If the method is experienced as less cognitive demanding this can also mean that it is easy to scale up and validate with more user groups.

Limitations
This study had several limitations. First, the case study design implies that participants’ individual characteristics may have created variability between the groups. While the comparisons between methods did offer some clear differences, it might be possible that those differences were due to individual variations in group composition. This was something that was unavoidable because the same patients could not be exposed to both these methods without creating a learning bias. Therefore, an attempt was made to create groups as similar as possible by using a purposive sampling technique. In fact, the variability could also just as easily have been so pronounced that no differences between methods would have been seen.

The addition of several validation steps for tasks in the UC-CW clearly had an impact on the time it took overall to conduct the method, and contributed to make it take longer as a whole. However, once the total time was divided by the number of identified problems, the UC-CW was actually faster. Both of these are important aspects to consider when choosing among available methods. Last, it is evident that variations exist between group- and individual-based methods where participants might be influenced by the group and modified their responses based on group comments. Still, the group aspect with its mix of patients with different input was one of the areas that the UC-CW participants stated that they appreciated and might have contributed to making the UC-CW considered less cognitive demanding. Future researchers who intend to use the modified CW may therefore have to weigh those aspects they gain and or potentially loose before making a decision on the most appropriate method for their evaluation.
Conclusion

The overall purpose of conducting this case study was to determine how our modified UC-CW performed on effectiveness, efficiency and user acceptance levels compared to an established usability method. Our study results demonstrate improved method performance and suggest improvements in the validity of findings. It indicates the importance of considering the validity of measures in the entire method evaluation process. The inclusion of users also improves the relevance of usability problem identification. Lastly, it is visible that the validity of measures can be improved by adding a component about users’ perceptions about method experience; the assessment of cognitive demands which was done here to increase the depth and breadth of findings. To our knowledge these modifications indicate important aspects to include in a method evaluation process of this kind in the health informatics field. The results reiterate the importance of including and comparing usability measures on all aspects of the method, effectiveness, efficiency as well as acceptance, not just one to get an encompassing view.

Acknowledgement

The authors would like to thank the Diabetes and Endocrinology Center at University of Utah Health in Salt Lake City, Utah for their support, and also all the diabetes patients who contributed to this study.

References


Appendix A

Heuristic categories for usability evaluation according to Nielsen [24]

<table>
<thead>
<tr>
<th>Visibility of system status</th>
<th>The system should always keep users informed about what is going on, through appropriate feedback within reasonable time.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Match between system and the real world</td>
<td>The system should speak the user’s language, with words, phrases and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order.</td>
</tr>
<tr>
<td>User control and freedom</td>
<td>Users often choose system functions by mistake and will need a clearly marked &quot;emergency exit&quot; to leave the unwanted state without having to go through an extended dialogue. Support undo and redo.</td>
</tr>
<tr>
<td>Consistency and standards</td>
<td>Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform conventions.</td>
</tr>
<tr>
<td>Error prevention</td>
<td>Even better than good error messages is a careful design which prevents a problem from occurring in the first place. Either eliminate error-prone conditions or check for them and present users with a confirmation option before they commit to the action.</td>
</tr>
<tr>
<td>Recognition rather than recall</td>
<td>Minimize the user's memory load by making objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.</td>
</tr>
<tr>
<td>Flexibility and efficiency of use</td>
<td>Accelerators -- unseen by the novice user -- may often speed up the interaction for the expert user such that the system can cater to both inexperienced and experienced users. Allow users to tailor frequent actions.</td>
</tr>
<tr>
<td>Aesthetic and minimalist design</td>
<td>Dialogues should not contain information which is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility.</td>
</tr>
<tr>
<td>Help users recognize, diagnose, and recover from errors</td>
<td>Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution.</td>
</tr>
<tr>
<td>Help and documentation</td>
<td>Even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the user's task, list concrete steps to be carried out, and not be too large.</td>
</tr>
</tbody>
</table>

Scale for severity rating according to Nielsen [24]

<table>
<thead>
<tr>
<th>Severity rating</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0 Not a usability problem at all</td>
<td></td>
</tr>
<tr>
<td>1 Cosmetic problem only</td>
<td>Need not be fixed unless extra time is available</td>
</tr>
<tr>
<td>2 Major usability problem</td>
<td>Fixing this should be given low priority</td>
</tr>
<tr>
<td>3 Major usability problem</td>
<td>Important to fix. Should be given high priority</td>
</tr>
<tr>
<td>4 Usability catastrophe</td>
<td>Imperative to fix this before product can be released</td>
</tr>
</tbody>
</table>
## Usability problem descriptions, detected by method, heuristic violations and assigned severity rating for all usability problems

<table>
<thead>
<tr>
<th>Place of occurrence</th>
<th>Usability Problem Description</th>
<th>Detected by method</th>
<th>Heuristic Violations</th>
<th>Severity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Main View</td>
<td>The bottom right side of the Main view has a hidden icon for entering weight which is not shown beside the other icons for entries on the view.</td>
<td>X</td>
<td>2, 4, 6</td>
<td>3</td>
</tr>
<tr>
<td>Main View</td>
<td>The visible icons for different entries on the top of the Main View cannot be configured based on the patient’s needs (is instead reached via a separate plus icon on the bottom).</td>
<td>X</td>
<td>2, 7, 8</td>
<td>3</td>
</tr>
<tr>
<td>Main View</td>
<td>The unit of measurement for the latest added values and average values are not displayed.</td>
<td>X</td>
<td>1, 5, 6</td>
<td>2</td>
</tr>
<tr>
<td>Main View</td>
<td>Unclear in the display of figures for the latest performed activity or latest meal if the figure signifies the beginning or end of the occurrence.</td>
<td>X</td>
<td>2, 4, 6</td>
<td>2</td>
</tr>
<tr>
<td>Main View</td>
<td>It is not possible to configure and personalize ranges based on the patients’ own ones for the low, medium and high ranges that are visualized in the middle pane for the latest glucose measurements.</td>
<td>X</td>
<td>2, 7</td>
<td>4</td>
</tr>
<tr>
<td>Main View</td>
<td>It is not possible to change the color on the blood drop icons that symbolize the summary of values for the latest added glucose measurements.</td>
<td>X</td>
<td>2, 6</td>
<td>1</td>
</tr>
<tr>
<td>Main View</td>
<td>Unclear to know that it is one of the three identical blue arrows that needs to be clicked to get to the Graph view.</td>
<td>X</td>
<td>2, 4, 5</td>
<td>2</td>
</tr>
<tr>
<td>Main View</td>
<td>Unclear to know that it is necessary to swipe across the chosen graphical image on the bottom of the Main View to get to the Periodical Pattern Graph View.</td>
<td>X, X</td>
<td>2, 4, 6</td>
<td>4</td>
</tr>
<tr>
<td>Main View</td>
<td>Unclear to know that when the right graphical view has been swiped forward on the bottom of the Main View, the blue arrow also has to be clicked to get to the main graph view (the large one)</td>
<td>X</td>
<td>1, 4, 6</td>
<td>2</td>
</tr>
<tr>
<td>Main View</td>
<td>Unclear to know and distinguish between the menu choices in the dropdown menu to find the function for sending the list of entered values.</td>
<td>X, X</td>
<td>2, 4</td>
<td>2</td>
</tr>
<tr>
<td>Main View</td>
<td>Unclear to know and to distinguish between the choices in Tools in the drop-down menu to send the list of entered values.</td>
<td>X, X</td>
<td>2, 4</td>
<td>2</td>
</tr>
<tr>
<td>View</td>
<td>Description</td>
<td>X Coordinates</td>
<td>Severity</td>
<td></td>
</tr>
<tr>
<td>---------------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>---------------</td>
<td>----------</td>
<td></td>
</tr>
<tr>
<td>Main View</td>
<td>Unclear to know how to find and use the drop-down menu to reach the Glucose Unit Setting View.</td>
<td>2, 4, 6</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Main View</td>
<td>Unclear how to distinguish between the choices in the drop-down menu to reach and perform changes in the Glucose Unit Setting View.</td>
<td>2, 4, 6</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Main View</td>
<td>Unclear how to find and choose the right path and icon to reach the Glucose Entry View.</td>
<td>2, 6</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Main View</td>
<td>Difficult to perform the swiping action across the chosen graphical image on the bottom of the Main View to get to the Periodical Pattern Graph View.</td>
<td>2, 4</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Main View</td>
<td>Unclear to know how to navigate, interact via one of the three identical blue arrows in the Main View to get to the List View.</td>
<td>2, 4, 6</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Main View</td>
<td>Unclear to distinguish that it is the top one of the three identical blue arrows in the Main View that needs to be clicked to get to the List View.</td>
<td>4, 5</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Main View</td>
<td>Unclear to know, find and use the drop-down menu to reach the Tools-menu and send the entered values.</td>
<td>2, 4</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Time Setting View</td>
<td>Difficult to understand and interact with the view to set hours, minutes, am/pm.</td>
<td>1, 2, 4</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Time Setting View</td>
<td>Unclear to know that it is necessary to press the ok button to save the selected date, that it is not enough to just exit the view.</td>
<td>3, 5</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Glucose Entry View</td>
<td>Unclear to know that one has to click on the visible value in blue for the date and time to get to the view to change these.</td>
<td>4, 5, 6</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Glucose Entry View</td>
<td>Unclear to know to use the check-mark icon to save selected and entered measurement values.</td>
<td>3, 5</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Carbohydrate Entry View</td>
<td>Unclear to know that after having added the selected meal from the meal database, it is also necessary to click the plus-icon to add the meal.</td>
<td>1, 2, 4, 6</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Carbohydrate Entry View</td>
<td>Unclear to know how grams of carbohydrates for the selected and added meal from the database is presented in relation to the division mark and the whole meal.</td>
<td>2, 4, 6</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Carbohydrate Entry View</td>
<td>Unclear to know to use the check-mark icon to save selected and entered measurement values.</td>
<td>3, 4, 5</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Carbohydrate Entry View</td>
<td>Unclear and many steps with deficient, confusing visualizations to get to the Similar Situations function.</td>
<td>1, 2, 6, 8</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Carbohydrate Entry View</td>
<td>Unclear to know and select the Food Composition Search Entry field, easy to confuse it with the Note Entry field.</td>
<td>4, 6</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Carbohydrate Entry View</td>
<td>Unclear to know that one should not click the box labelled grams or IUs to enter these, but instead enter them into an empty box next to it.</td>
<td>X</td>
<td>2, 4, 5</td>
<td>2</td>
</tr>
<tr>
<td>-------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>---</td>
<td>-----</td>
<td>---</td>
</tr>
<tr>
<td>Plotted Graph View</td>
<td>Unclear to know that one should not click the box labelled grams or IUs to enter these, but instead enter them into an empty box next to it.</td>
<td>X</td>
<td>1, 2, 6</td>
<td>2</td>
</tr>
<tr>
<td>Plotted Graph View</td>
<td>Unclear to know to press the clock with arrow icon to change from/between the 24 hr, weekly and monthly view for the graph.</td>
<td>X</td>
<td>2, 7</td>
<td>4</td>
</tr>
<tr>
<td>Plotted Graph View</td>
<td>It is not possible to configure and personalize the range levels (cut off levels) in the graphs that show the glucose values for the low, mid and high ranges.</td>
<td>X</td>
<td>4, 5, 8</td>
<td>1</td>
</tr>
<tr>
<td>Plotted Graph View</td>
<td>Risk for misunderstanding when the clock with the arrow icon on the top is not possible to click but the identical clock on the bottom is to change between the 24 hr view, weekly and monthly view.</td>
<td>X</td>
<td>1, 4, 5</td>
<td>2</td>
</tr>
<tr>
<td>Periodical Pattern Graph View</td>
<td>Unclear visualization in the Periodical Pattern graph if the view is showing the daily or weekly summary.</td>
<td>X</td>
<td>1, 2, 6</td>
<td>4</td>
</tr>
<tr>
<td>List View</td>
<td>Unclear to know how to navigate in, understand where one is located and get a response on the sequence of dates in the list of all entries.</td>
<td>X</td>
<td>2, 4</td>
<td>3</td>
</tr>
<tr>
<td>List View</td>
<td>Unclear to know that it is necessary to click the blue arrow to execute the search instead of the check-mark.</td>
<td>X</td>
<td>2, 4</td>
<td>3</td>
</tr>
<tr>
<td>List View</td>
<td>Unclear and difficult to know that one should use the blue arrows in the search pane to step through the search hits and how these arrows work.</td>
<td>X</td>
<td>2, 4, 6</td>
<td>2</td>
</tr>
<tr>
<td>List View</td>
<td>Unclear to know and difficult to locate the search field on the top of the list of all entries in the List View.</td>
<td>X</td>
<td>1, 2, 4, 6</td>
<td>4</td>
</tr>
<tr>
<td>Similar Situation List View</td>
<td>Difficult to understand and interpret the relation between the number of IUs of insulin and mg/dl blood glucose units in the upper part of the image compared to the list of entries in the bottom part of the view for this function.</td>
<td>X</td>
<td>1, 2, 4, 6</td>
<td>4</td>
</tr>
<tr>
<td>Similar Situation List View</td>
<td>Unclear to know that after having activated the suitable Similar Situation and its number of IUs of insulin, this has not been selected but one has to add the number of IUs of insulin manually in a previous view and save.</td>
<td>X</td>
<td>1, 5</td>
<td>3</td>
</tr>
</tbody>
</table>

* UC-CW = User-centered Cognitive Walkthrough, TA = Think Aloud usability test
ABSTRACT

Diabetes is a chronic disease affecting 422 million patients worldwide according to World Health Organization data with 30.3 million in the United States and 64 million in Europe. The prevalence speaks to the need for improved ways to support patients in disease self-management. mHealth solutions are increasingly used for this; however, usability is a current challenge affecting patients’ mHealth use. Recent literature emphasizes an increased focus on patient-centeredness in chronic disease mHealth system design and standardized, systematic approaches for usability evaluation. The aim of this thesis and its individual studies was to incorporate these foci into the evaluation of two mobile health self-management systems for diabetes.

Study I used ISO standard 9241-11 to examine the relationship between selected group characteristics of diabetes patients on specific interaction outcomes to quantitatively identify needed system modifications. Study II utilized a multi-method design to assess diabetes patients’ mHealth usage and combined two novel analytic methods to structure and analyze results. Study III used a modified, user-oriented heuristic evaluation (HE) method, validated tasks and in-depth severity factor ratings to identify critical problems from patients’ point of view. By developing and employing a modified, user-centered cognitive walkthrough method (UC-CW), study IV assessed its effectiveness and efficiency in finding relevant usability problems for users as well as patients’ acceptance. The modified CW was validated against the golden-standard user test with Think Aloud.

Study I emphasized the importance of considering user characteristics in mHealth performance as these influenced interaction outcomes. All patients had difficulties with multiple-step tasks. Patients more recently diagnosed were able to perform tasks more successfully, with fewer errors and at faster times and had higher satisfaction scores; similar outcomes to the more experienced users. Educational level did not, however, seem to influence performance. In study II, the usability test with Think Aloud (TA), in-depth interviews and questionnaires contributed to 19 consolidated issues, and triangulated on 5 critical usability problems for users. The combined analysis methods resulted in structured, categorized descriptions to aid in problem-solving. In Study III, the disease-related, critical information deficiencies found by expert evaluators using the modified, structured method also converged on and highlighted potentially adverse user concerns. Study IV demonstrated that the UC-CW found more critical user problems compared to the user test with TA despite both methods producing similar major average severity ratings and violations of heuristic categories. The modified method was more efficient per detected problem and experienced as less cognitively demanding and with a higher ease of use.

These studies offer different approaches that include patient-centered, efficient and user-acceptable methods and method modifications to detect critical usability issues for users. Importantly, improved mHealth designs for users could mean improvement in interactions, interaction performance, increased adoption, and long-term perhaps even increased adherence to interventions for chronic conditions.