journal homepage: www.intl.elsevierhealth.com/journals/ijmi

Determination of the effectiveness of two methods for usability evaluation using a CPOE medication ordering system

R. Khajouei[a,b,*], A. Hasman[a], M.W.M. Jaspers[a,c]

[a] Department of Medical Informatics, Academic Medical Center, University of Amsterdam, The Netherlands
[b] Kerman University of Medical Sciences, Kerman, Iran
[c] Department of Clinical Pharmacy, Academic Medical Center, University of Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 29 October 2010
Received in revised form
12 February 2011
Accepted 12 February 2011

Keywords:
User–computer interface
CPOE
Usability
User testing
Think aloud
Cognitive walkthrough
Usability evaluation methods

ABSTRACT

Objectives: To assess the effectiveness of two usability evaluation methods, cognitive walkthrough (CW) and think aloud (TA), for identifying usability problems and to compare the performance of CW and TA in identifying different types of usability problems.

Methods: A CW was performed by two usability evaluators and 10 physicians were recruited to perform a TA usability testing of a CPOE system (Medicator). The severity of identified usability problems was determined and the usability problems were categorized based on the User Action Framework (UAF). The potential of usability problems to cause medication errors was also determined. The thoroughness, validity and effectiveness of the two methods were compared.

Results: Fifty seven unique usability problems of different severity, spread over the four phases of interaction as defined by the UAF, were identified. The effectiveness of the TA method for identifying usability problems was 0.08 higher than that of the CW (0.70 vs. 0.62). The thoroughness (the extent to which a method can identify existing usability problems) of the TA was higher for the "Planning" and "Assessment" phases and lower for the "Translation" phase (as defined by UAF). The thoroughness of TA for identifying problems that may potentially result in medication errors was higher than that of CW (0.81 vs. 0.68). The number of usability problems identified by each of the methods was significantly less than the total number of detected real usability problems in Medicator ($p < 0.001$). The observed differences between the number of real usability problems identified by CW and TA (38 vs. 41), the difference between the average severity of the detected problems by CW and TA (2.37 vs. 2.41), and the difference for identifying problems potentially resulting in medication errors (15 vs. 18) were not statistically significant ($p > 0.4$).

Conclusions: This study shows that although TA showed a slightly better effectiveness, there is no significant difference between the performance of the CW and the TA methods in terms of number of usability problems identified and the mean severity of these problems. Since no single evaluation method will uncover all of the usability problems a combination

[*] *Corresponding author at:* J1B-115.2, Department of Medical Informatics, Academic Medical Center, PO Box 22700, 1105 AZ Amsterdam, The Netherlands, Tel.: +31 20 566 7874; fax: +31 20 691 9840.

E-mail addresses: rkhajouei@yahoo.com, r.khajouei@amc.uva.nl (R. Khajouei).

1386-5056/$ – see front matter

doi:10.1016/j.ijmedinf.2011.02.005

of methods is advised as the most appropriate approach, especially if usability problems can lead to potentially fatal outcomes.

1. Introduction

Computerized clinical applications have the potential to improve patient safety and to support clinicians in making clinical decisions [1] but the lack of adoption of these systems has become a key issue in healthcare [2,3]. Studies have shown that usability problems are among the factors negatively affecting the adoption [2,4] and effectiveness [5–7] of clinical applications. Usability problems of these systems can result in suboptimal user–system interactions, which in turn lead to user frustration, inefficiency and have the potential to negatively influence patient outcomes [8,9].

An important requirement in effective user interface design is to minimize the cognitive demands that a system imposes on the user when interacting with the system [10]. Thus, by revealing interface design characteristics that hinder users in different phases of interaction the system's usability can be improved. To identify these user-interface problems several usability evaluation methods (UEMs) from the field of Usability Engineering and Cognitive Psychology have been applied for evaluating interactive healthcare applications, individually or in combination with other UEMs. The selection of a specific method seems to depend on the difficulty to carry it out and the amount of resources available. However, when considering a specific UEM, one also should take into account its effectiveness to reveal user interface problems. The effectiveness of UEMs has been researched in other domains than healthcare. Studies that compared the application of several UEMs focused mostly on their cost-effectiveness [11]. Most studies counted the crude number of detected usability problems as a measure of UEMs' effectiveness without relating this number to the total number of usability problems as detected by all methods considered [12]. The number of identified problems for assessing the effectiveness of a method is only meaningful when it is compared to an estimate of the total number of existing problems (thoroughness).

The quality of UEM comparisons in terms of the number of revealed usability problems depends on the evaluator's behavior and precise analysis and classification of usability problems. Therefore a formal approach of analyzing and classifying usability problems seems essential when comparing the output of UEMs.

The User Action Framework (UAF) [13] is a structured knowledge base for the classification of usability concepts, issues and situations. UAF focuses on the different phases of user–system interaction and provides a consistent way to code usability problems. Classification of usability problems using UAF makes it possible to compare the effectiveness of the UEMs for identifying usability problems related to different phases of interaction.

The cognitive walkthrough (CW) (a usability inspection method) and think aloud (TA) (a user-testing method) methods are among the UEMs most often applied. Previous studies compared specific types of user-testing methods with CW [14,15] but those methods differed extensively from the TA method as defined by Ericssom and Simon [16]. From the eighteen studies that compared UEMs identified by Hartson [17] only one study [18] compared CW with TA. This study which evaluated the usability of a website merely focused on the severity of problems identified by each method. Also, studies focusing on the difference in effectiveness of these two methods in the domain of healthcare information systems are lacking. Usability problems of healthcare information systems may cause life threatening situations because they can potentially lead to medical errors. Existing UEMs do not consider these consequences. In this paper the results of two usability evaluation studies (a CW [8] and a TA [19]) of a computerized physician order entry (CPOE) system were used to assess the potential of each method for identifying certain types of usability problems.

2. Background

A variety of analytic and empirical evaluation methods can be used to assess the usability of an interactive computer application, among which CW, an analytic method, and TA, an empirical method. Analytic methods are based on examination of the design characteristics of a prototype, or final system. Empirical methods are based on observation of the performance of the system in use [20]. Both CW and TA focus on users' tasks and have the goal of identifying usability flaws in an interactive computer system but differ in their approaches and resources required.

2.1. *Cognitive walkthrough*

CW is a usability inspection method which evaluates the ease with which a typical new system user can successfully perform a task using a given interface design [21]. It is especially appropriate for the development of applications where users must (or prefer to) master a new application or function by learning through exploration. The CW methodology is based on a theory of learning by exploration proposed by Polson and Lewis [22]. The input to a CW session includes a detailed design description of the user interface, a task scenario, explicit assumptions about the user population and the context of use, and a sequence of actions with which a user could successfully complete the task using the (prototype) system under evaluation. During the walkthrough process, the reviewers step through the actions, considering the behavior of the interface and its effect on the user, and attempt to identify those actions that would be difficult for the average member of the proposed user population to choose or to execute.

In a preparation phase basic system information is gathered and recorded. For example, the suite of tasks to be evaluated is identified and profile information concerning the typical user population is noted. Then an evaluator inspects the user interface by a stepwise approach, using knowledge about how a hypothetical user would process certain tasks while navigating through the system. Finally all information gathered and recorded from the CW process is interpreted.

2.2. *Think aloud*

This method is used in a number of social sciences such as cognitive psychology to gather data on the way humans solve problems but also in usability testing of interactive computer system designs. In the case of usability research, the primary concern is to support the development of usable systems by identifying design deficiencies [23]. The TA method was introduced in the usability field by Lewis [24] and was further refined by Ericssom and Simon [16]. TA protocols involve participants thinking aloud as they are performing a set of specified tasks. TA sessions can be held in the user's contexts of work or in simulation environments similar to the real working environment of the users. Equipment for video and audio recording of users, a prototype or final system, and a task scenario are required as input for a TA.

In the preparation phase participants are instructed how to think aloud. Then a representative sample of approximately 5–8 end users is asked to interact with the (prototype) system according to a predefined set of scenarios while verbalizing whatever thoughts come to their minds when performing the tasks in the system. During thinking aloud, interruption by evaluators is usually limited to reminding them to keep talking when a short period of silence (15–60 s depending on the goal) occurs. Video and audio data of the users and the screen shots of the system under use are recorded in this phase. Finally the data collected during the users' testing sessions are reviewed and analyzed by the usability evaluators to find the usability problems that participants experienced while interacting with the system.

We used these two methods to evaluate the usability of a CPOE (a medication ordering system) and studied the performance of each method in identifying different types of usability problems concerning each of the phases in the interaction.

2.3. *The User Action Framework*

The UAF is a standardized classification of usability problems. The UAF was built by adapting and extending Norman's theory of action model [25] into what is called the interaction cycle. The interaction cycle includes the concepts from all of Norman's stages of human–computer interaction, but organizes them pragmatically in a slightly different way. Like Norman's model, the interaction cycle depicts how the interaction between a user and a machine takes place in terms of cognitive and physical user actions. The UAF classification provides insight into what the users think, perceive and do throughout each cycle of interaction with a computer system. The purpose of the interaction cycle is to model the flow of user interaction in any interactive system. The UAF interaction cycle contains four phases; planning, translation, physical actions and assessment (Fig. 1). We use the example of 'retrieving the record of a patient' in a computerized patient record system (CPR) to explain the different phases of the interaction. In the planning phase, the user looks at the CPR interface and decides what to do to be able to start this task (e.g. searching for the patient's name or patient identification number). After planning what to do, in the translation phase, the user looks for the physical objects on the screen (e.g. entry fields and buttons) that can be used to carry out the task. When the user knows which objects to manipulate and which actions to do, in the physical action phase, (s)he executes the actions by manipulating the physical objects (e.g. entering the name of the patient in the entry field and pushing the search button). Finally, when the actions are completed, in the translation phase, the user evaluates the state of the system to evaluate if the action is performed by the system and if the appropriate results are achieved (displaying the patient's record or information).

3. Methods

3.1. *Study setting and the system*

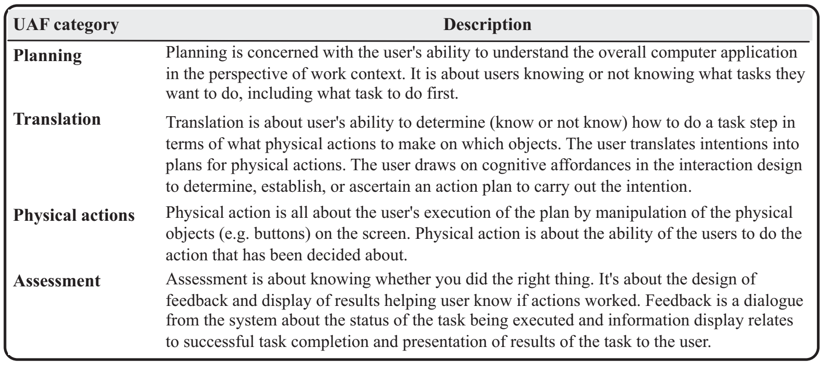
This study was conducted at the Academic Medical Center (AMC), a large university hospital in Amsterdam, the Netherlands. The AMC introduced in 2000 a commercially available CPOE system for medication ordering called Medicator®, which is used in 30 clinical departments of the AMC. Moreover, 15 other hospitals in the Netherlands also implemented the Medicator system. Medicator provides decision support in terms of alerts for medication interactions, overdose and double medication based on the pharmacy drug database and the national drug database (the Z-index of the Royal Dutch Association of Pharmacists). Next to the possibility to enter individual orders Medicator also provides the functionality to select predefined order sets based on clinical protocols. A more detailed description of Medicator is given by Kalmeijer et al. [27].

3.2. *Usability evaluation*

The results of two usability evaluation studies of a CPOE system (a cognitive walkthrough [8] and a think aloud [19] study) published elsewhere were used as input for this investigation.

3.2.1. *Cognitive walkthrough*

Two usability evaluators independently walked through different ordering tasks based on a realistic clinical scenario. The clinical scenario was designed by an expert in developing clinical protocols, validated by a clinical specialist and tested by two usability evaluators so that the scenario covers almost all Medicator functionalities for ordering different medications. This scenario required ordering two types of medication related to a clinical protocol for a leukemic patient. This scenario concerned the prescription of two medications, Idarubicine and ATRA (Vesanoid), for the first course of consol-



UAF category	Description
Planning	Planning is concerned with the user's ability to understand the overall computer application in the perspective of work context. It is about users knowing or not knowing what tasks they want to do, including what task to do first.
Translation	Translation is about user's ability to determine (know or not know) how to do a task step in terms of what physical actions to make on which objects. The user translates intentions into plans for physical actions. The user draws on cognitive affordances in the interaction design to determine, establish, or ascertain an action plan to carry out the intention.
Physical actions	Physical action is all about the user's execution of the plan by manipulation of the physical objects (e.g. buttons) on the screen. Physical action is about the ability of the users to do the action that has been decided about.
Assessment	Assessment is about knowing whether you did the right thing. It's about the design of feedback and display of results helping user know if actions worked. Feedback is a dialogue from the system about the status of the task being executed and information display relates to successful task completion and presentation of results of the task to the user.

Fig. 1 – User Action Framework phases defined by Andre [13,26].

idation chemotherapy. For this task all (correct) actions were stepped through and analyzed by the evaluators to determine potential usability problems a (novice) user with the expected clinical background may encounter in ordering medication in Medicator. The evaluators were taught how to use the system before the actual evaluation. CW was carried out in a laboratory situation with dummy patient data.

3.2.2. *Think aloud*

Ten physicians from the AMC's Hematology and Oncology department were asked to participate in the TA user testing sessions. Medicator experience of the physicians ranged from one week to more than three years. Four of them had less than one year experience with the system. TA sessions took place in the actual working place of the physicians. Before starting the TA sessions, participants were instructed to think aloud while performing the ordering tasks. Each participant was asked to order medications while verbalizing his/her thoughts. These tasks required the participant to order medications based on the same scenario used in CW. In performing the test in this way, the evaluation of all Medicator functionalities concerning both ways of medication ordering was made possible. Morae® version 2.0 (TechSmith Corporation, Okemos, Michigan), a usability evaluation tool, was used to record the sessions. Both video and audio tracks of the participants while performing the tasks as well as the corresponding screen sequences, changes and movements (e.g. mouse clicks and keystrokes) were recorded by Morae.

3.3. *Data collection*

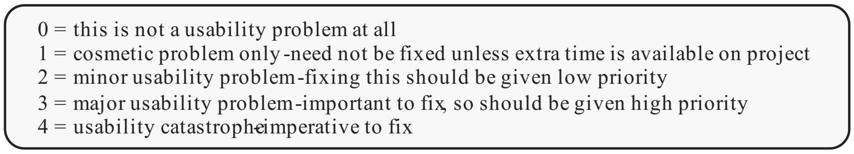
Each of the two CW evaluators independently evaluated the system by analyzing the execution of every action and resulting system state to identify usability problems and provided a list of usability problems and their descriptions. These two separate lists were compared by the two evaluators and merged into one master problem list. First, problems identified by both the evaluators were added to the list. Then, problems identified by either one of the evaluators were discussed and were added to this list if confirmed by the second evaluator.

Recordings of the TA sessions were reviewed and analyzed by two usability evaluators using the Morae® manager. Protocol analysis was performed on all verbal utterances of the participants [16]. The evaluators independently provided a list of usability problem descriptions encountered by the participants. Identified usability problem descriptions were merged into a unique master problem list and disagreements were resolved through review of video and audio data and subsequent discussion.

3.4. *Data analysis and measurements*

The analysis of the usability problem descriptions from the CW and the TA evaluations involved the following steps:

1. Usability problems were coded by two evaluators, having expertise in usability evaluation, using the User Action Framework (UAF) [13]. Two usability evaluators reviewed the usability problem descriptions to determine how the user was affected by the Medicator design when performing cognitive or physical actions and coded the usability problems for each phase of interaction. The two evaluators reviewed their codes in a meeting and discussed the discrepancies until they reached agreement. Any remaining disagreement was resolved through discussion with a third evaluator.
2. The severity of usability problems was determined according to Nielsen's classification [28], presented in Fig. 2. As the inter-rater reliability of an individual's rating is low compared to group rating [29], the severity rating of each problem type was assigned by consensus of three usability evaluators. For assigning a severity rating the frequency with which a problem (might) occur (red), the (potential) impact of the problem on the users and the (potential) persistence of the problem were taken into account [30]. Any interaction issue that was identified as usability problem by one evaluator but was not confirmed by the other evaluator(s) was not considered a real usability problem and removed from the analysis.
3. To determine the potential of each method for identifying usability problems that may result in medication errors, each problem description and the corresponding state of the system interface were reviewed. Usability problems that may result in the following medication errors were identified: wrong medication name, dosage, frequency and duration.
4. To compare the performance of CW and TA in identifying usability problems, the total set of usability problems



0 = this is not a usability problem at all
1 = cosmetic problem only -need not be fixed unless extra time is available on project
2 = minor usability problem-fixing this should be given low priority
3 = major usability problem-important to fix, so should be given high priority
4 = usability catastropheimperative to fix

Fig. 2 – Nielsen's severity rating for usability problems [28,30].

identified by both methods [17] was used as a standard-of-comparison usability problem set. The following measures were used for examining and comparing the two usability evaluation methods on their performance in identifying usability problems.

a. Thoroughness is the extent to which a usability evaluation method can identify real usability problems. Thoroughness is measured as the ratio of the number of real usability problems found using each usability evaluation method to the total number of real problems existing in the user interface of the system (given by the standard-of-comparison usability problem set) [31]:

$$\text{Thoroughness} = \frac{\text{number of real usability problems found by the evaluation method}}{\text{total number of real usability problems}}$$

The thoroughness of each method was also determined for each UAF phase and for identifying problems that may lead to medication errors. In addition the thoroughness of the methods as a function of the severity level of the problems, the weighted thoroughness based on these severity levels (for both nominator and denominator each problem is multiplied by its severity and the products are summed) and the average severity of the problems were compared. The next formula is used for determining the weighted thoroughness where *s* is the severity of a usability problem (1, 2, 3, or 4):

$$\text{Weighted thoroughness} = \frac{\sum \text{`number of real usability problems found by the evaluation method with severity } s\text{'} \times s}{\sum \text{`total number of real usability problems at severity } s\text{'} \times s}$$

b. Validity is the extent to which a usability evaluation method accurately identifies usability problems. Validity is measured as the ratio of the number of the real usability problems found by a method to the number of issues (all system design aspects that may be considered as an obstacle to the effective and efficient accomplishment of a specified task by a specified user) the method (correctly or incorrectly) identified as usability problems:

$$\text{Validity} = \frac{\text{number of real usability problems found}}{\text{number of issues identified as usability problem}}$$

c. Effectiveness is the ability of a usability evaluation method to identify usability problems related to the user interface of a specific system:

$$\text{Effectiveness} = \text{Thoroughness} \times \text{Validity}$$

3.5. *Statistical analysis*

Data were analyzed descriptively using frequencies, means, standard deviations, and percentages. Differences in the total number of real usability problems in the Medicator system and those identified by the two usability evaluation methods were tested using the chi-square test (and Fisher's Exact Test when necessary). The Kruskal–Wallis test was used to compare the mean severity of the real usability problems identified by both methods and those identified by only one of the methods. A post hoc Bonferroni correction was carried out to adjust our alpha level. Differences between the methods in identifying problems that may lead to potential medication errors were tested by Fisher's Exact Test. A *p* value of <0.05 was considered significant. Al1 analyses were performed using SPSS 16.0 (SPSS Inc., Chicago, IL).

4. Results

The CW and the TA evaluation methods identified respectively 41 and 42 issues as usability problems in the Medicator system user interface. Three of the issues identified by CW and one of those identified by TA were rated severity "0" (not a usability problem) in the consensus meeting of evaluators and were removed from the lists of usability problems. In total 57 unique usability problems of different severity, which can be encountered by users in the four phases of interaction with the system, were identified in this study. Table 1 presents the number of real usability problems identified by the two methods, both as a function of the severity rating as well as of the UAF interaction cycle phases.

From the total number of identified usability problems ($n = 57$), 39% were found by both methods and 61% were identified by either one of the two methods. The difference between the total number of real usability problems in the Medicator system and those identified by each of the two methods separately was statistically significant at the $p < 0.001$ level according to the χ^2 test. Fisher's Exact Test showed that the observed difference between the number of real usability problems identified by CW and TA (38 vs. 41) was not significant ($p = 0.68$).

Nineteen of the usability problems were rated as highly severe (major and catastrophic) of which 32% ($n = 6$) were identified only by the TA and 16% ($n = 3$) only by the CW method.

Table 1 – Real usability problems identified by CW and TA in terms of number, severity and UAF phases.

	CW No (%)	TA No (%)	CW and TA No (%)	CW not TA No (%)	TA not CW No (%)
Usability problems	38	41	22	16	19
Severity					
1	2 (5)	3 (7)	0 (0)	2 (12)	3 (16)
2	23 (61)	22 (54)	12 (55)	11 (69)	10 (53)
3	10 (26)	12 (29)	8 (36)	2 (12)	4 (21)
4	3 (8)	4 (10)	2 (9)	1 (6)	2 (10)
Severity (average)	2.37 [SD = 0.71]	2.41 [SD = 0.77]	2.55 [SD = 0.67]	2.13 [SD = 0.72]	2.26 [SD = 0.87]
UAF					
Planning	3 (8)	4 (10)	1 (4)	2 (12)	3 (16)
Translation	13 (34)	12 (29)	9 (41)	4 (25)	3 (16)
Physical actions	7 (18)	7 (17)	5 (23)	2 (12)	2 (10)
Assessment	15 (40)	18 (44)	7 (32)	8 (50)	11 (58)

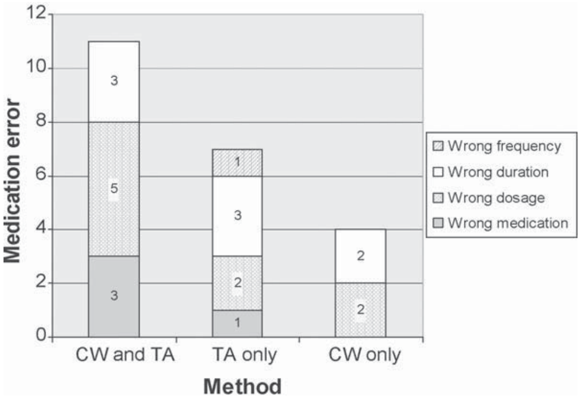


Fig. 3 – The number of identified problems by each method that potentially could lead to medication errors.

From the 38 low severity (cosmetic and minor) usability problems the number of problems that were identified only by one method was the same for the CW and the TA. Although the average severity of problems identified only by the TA method was higher than the average severity of those identified only by the CW (2.26 vs. 2.13), the Kruskal–Wallis test showed no significant difference between them and with the average severity of the total number of problems identified by the two methods ($p = 0.4$).

Fig. 3 presents the number of identified problems that potentially could lead to medication errors. Of the total 57 usability problems 22 could potentially lead to medication errors of which 11 problems were identified by both methods. From the remaining 11 problems 7 were identified only by the TA method and 4 were identified only by the CW method. Fisher's Exact Test showed that the observed difference between the CW and TA for detecting usability problems that could potentially result in medication errors (15 vs. 18) was not statistically significant ($p = 0.49$).

Fig. 4 presents the thoroughness of the two methods as a function of the severity of problems, the UAF phases, and their potential for causing medication errors. The thoroughness of the TA method in identifying problems of severity level 1, 3 and 4 was noticeably higher than that of the CW, but there was not much difference between the methods in weighted thoroughness (Fig. 4(a)).

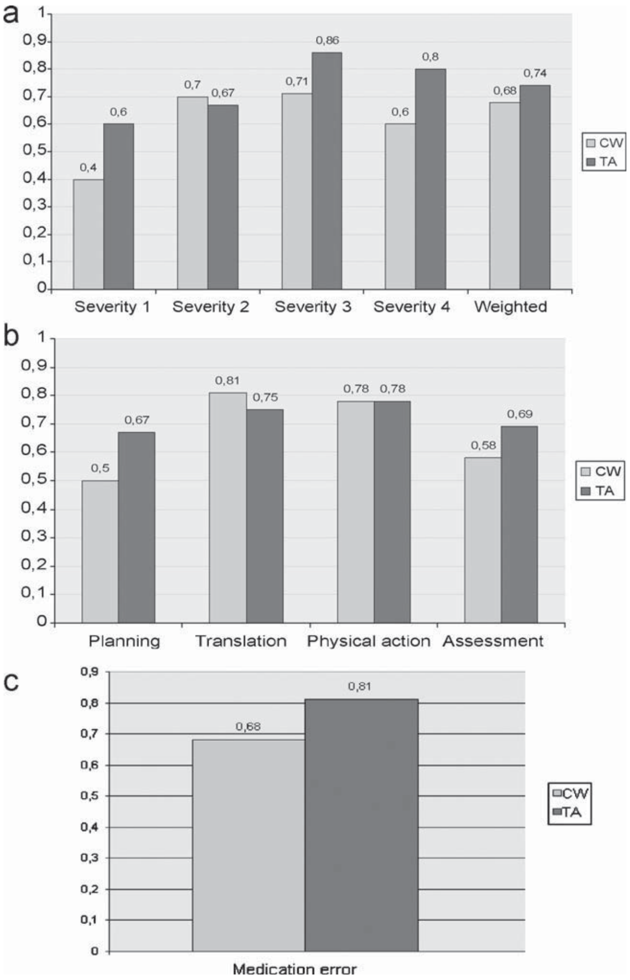


Fig. 4 – The thoroughness of two usability methods (a) as a function of the severity of usability problems and the weighted thoroughness, (b) as a function of the UAF phases, and (c) as a function of the potential for causing medical errors.

Inspection of the categorization of usability problems as a function of the UAF phases showed that among the problems identified by only one method the number of problems identified by TA concerning the phase "Assessment" was higher than that identified by the CW. Looking at Fig. 4(b) the thoroughness of the TA was higher for the phases "Planning" and

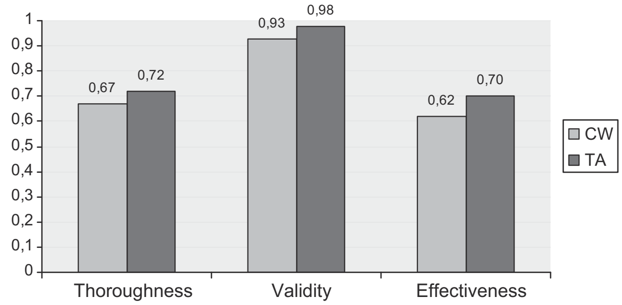


Fig. 5 – Thoroughness, validity and effectiveness of two usability methods.

"Assessment", and lower for the phase "Translation". The thoroughness of TA for identifying problems that may potentially result in medication errors was higher than that of CW (0.81 vs. 0.68). On a scale from zero to one, overall thoroughness and validity of the TA method were 0.05 and its effectiveness was 0.08 higher than those of the CW method (Fig. 5).

On average problems that were identified by both methods were encountered twice as much (8.5 vs. 4.2 per problem) by the physicians during TA than problems that were identified with the TA approach only. Qualitative analysis of the usability problems identified specific themes in those problems that were difficult to detect by one of the evaluation methods. The following themes were identified in the problems revealed by the CW but not by the TA method:

Previous experience with the same or a similar system: CW identified some issues as usability problems that were no problem for users who used the system previously and thus TA did not detect them. For example, a problem with the location of the functional buttons for initiating, changing and canceling orders, which was the opposite of the usual place of buttons in a user interface (for Left-to-Right languages) was identified by CW but it did not present any problem to users in the TA sessions. Participants in TA sessions had previous experience with the system and were familiar with the location of these buttons whereas this is not the case for an evaluator simulating a novice user.

Different possible ways of carrying out a task: If a task can be carried out in several ways there may be usability problems associated with some of the ways to carry out this task. This problem can be identified by CW because the evaluators explore all possible ways the task can be carried out. Users in TA will not experience this problem if they carry out the task in a way not associated with problems. For example, to initiate an order a patient can be searched via the patient's name, the date of birth or the patient number. CW identified a usability problem in the search procedure using the date of birth. This problem was not identified by TA because all users searched patients by their patient numbers.

Situation dependent problems: Some problems were identified by CW but none of the TA users experienced those problems. For example, CW revealed that the system allows entering an administration date in the past. This problem was not revealed by TA because users did never enter a date in the past.

The following themes were identified in the problems revealed by the TA but not by the CW method:

Problems concerning user preferences: Users in TA mentioned that they prefer a different way of information presentation than the current one. This sort of problems was not revealed by CW. For example, TA users preferred to see an overview of the total number of medication administration days over the start and end date of administration. Presentation of the administration start and stop dates was not identified as a usability problem by CW.

Situation dependent problems: A CW evaluator might not anticipate all situations a user might experience. For example, in selecting an item from the list of clinical protocols one of the TA users inadvertently clicked on an adjacent item. CW evaluators did not expect this problem to occur.

Domain knowledge dependent problems: In TA, a user may encounter problems because of insufficient domain knowledge. Because CW evaluators suppose that the potential users of the system know their own domain, the CW might miss this type of usability problems. For example, when a user searched for ATRA the system retrieved other names (generic and trade name) of the medication (Tretinoine and Vesanoid). This caused confusion for a TA user because he did not know that these were other names of the same medication.

5. Discussion

5.1. *Principal findings*

The results of this study showed that the number of usability problems identified by each of the methods was significantly less than the total number of usability problems of the system determined by both methods. The TA method showed 0.08 better effectiveness in identifying usability problems than the CW method but these two methods do not differ significantly in terms of number of identified usability problems and the average severity of these problems. Dependent on the users' characteristics and context of use some of the problem may not be detectable by either CW or TA.

Although there was a slight difference in the effectiveness of the CW and the TA methods, there was no significant difference in the number and mean severity of problems identified by these methods in our study. Karat [15] and Jeffries [14] compared the effectiveness of empirical usability testing with some analytic methods including CW for office system user interfaces. The results of Karat [15] showed that usability testing detected a higher number of usability problems and a significantly higher number of relatively severe problems than CW. The results of Jeffries [14] were consistent with our study as Jeffries identified almost the same number of usability problems using either CW or usability testing. Jeffries pointed out that the higher mean severity of problems for usability testing may reflect a bias on the part of the raters. Our study differs from Jeffries' [14] and Karat's study [15] in the sense that in the TA method used in our study evaluators extracted users' data from the audio and video recordings of TA sessions while in the usability testing method used by Jeffries and Karat users merely described the usability problems they encountered to the usability evaluators during the sessions. Distraction of users in our TA was limited to reminding them to think aloud if a user went silent, while in the usability testing by Jeffries and Karat users had to stop the interaction and explain each problem they encountered to the evaluators. In this type of usability testing the user's role is more like that of an evaluator.

The ability of the CW and the TA method to identify low-priority problems was the same but TA identified 16% ($n = 3$) more serious usability problems than CW. Of the problems identified by one method only, TA identified twice as much high severity problems (major and catastrophic) than CW (6 vs. 3) and the same number of low severity problems (cosmetic and minor) (13 vs. 13). This is consistent with Jeffries [14] findings where usability testing identified twice as much severe problems than CW and substantially less least severe problems. Karat [15] also found that empirical testing was better than CW in identifying relatively severe problems. We considered cosmetic and minor usability problems as having low-priority, and major and catastrophic problems as serious.

Many of the studies comparing usability evaluation methods used the number and severity of identified usability problems to compare their effectiveness. We broke down the identified problems using the UAF classification [13] as well to determine whether different UEMs have a different effectiveness in detecting problems in the various phases of user interaction. In our study the capability of both methods to identify problems in the planning, translation and physical action phases was the same. Only for the assessment phase the TA method had a better performance.

Usability problems of healthcare information systems may affect patient outcomes. Therefore, we assessed the potential of each method for identifying problems that may lead to medication errors. Although the difference between CW and TA for identifying this type of problems was not significant, the TA method showed a relatively higher thoroughness. This can be because TA method can find more severe problems than CW as it was shown by the results of this study and the study by Jeffries [14]. However, when assigning severity to usability problems, which is usually according to Nielsen's classification [28], their potential effect on task outcome is not taken into account. We therefore suggest making an investigation of the impact of usability problems on patient outcome a standard component of the usability evaluation of health information systems.

The small difference in the validity of the CW and TA results can be explained as follows. Thinking aloud takes place automatically while a user solves a problem [32]. The evaluators of the TA sessions analyze what usability problems a real user experienced during system interaction while CW evaluators are hypothesizing about what problems a particular novice user with a certain background could encounter during system interaction. In CW the evaluator determines whether the system supports a potential user in selecting and executing the correct actions and if the user will notice and understand the results of the action. Therefore, CW may also identify some issues as problems that would not actually hinder users [31].

Many of the low and high severity problems found in the study were identified by only one of the two methods, but not both. For example in the TA method users with different profiles (e.g. regarding domain and system knowledge and experience) may encounter different types of usability problems. Since in a CW the evaluators only use the users' basic computer knowledge as input to the test, they can miss some of the problems that might be experienced by users having less domain knowledge or experience with a different interface of a similar application. The walkthrough process focuses on learnability by exploration and the evaluators examine each individual step in the correct action sequence and try to find a plausible reason why the prospective user would choose each action. This approach does not predict what activities the user might engage in when carrying out a step [33] but focuses on the usual sequence of actions in the interaction process where users are likely to encounter problems [31]. As users in TA may not follow the correct action sequence and sometimes perform a wrong series of actions, the TA analysis may identify problems that are not detected by CW. For example later in the ordering process a user may recognize that he has carried out a wrong action in a previous step and try to undo that action while the system does not provide an undo button at the current step. It is unlikely that evaluators in the CW will identify this sort of problems because they follow the correct sequence of actions. Likewise some problems were not detected by the TA, for example when there are different ways to perform some (sub) tasks (e.g. initiation of a new order in Medicator). In CW when the system offers alternative correct action sequences for carrying out a task, all sequences are analyzed [34]. Users usually take the most straightforward or the most familiar sequence to accomplish a task. Therefore CW identified a number of usability problems, related to alternative correct action sequences that were not encountered by users, who followed one specific sequence for their actions. Moreover TA protocols are not necessarily complete because a subject may verbalize only part of his thoughts [32] increasing the risk of overlooking specific usability problems. Usability problems with low persistence (problems that burden the user only in the first encounter) may not hinder users in TA if they have become familiar with the system and had previous encounters with similar situations. Due to the focus on the interaction behavior of a beginning system user, these types of problems are usually detected by the CW. This suggests that when selecting a usability evaluation method the users' characteristics, such as previous computer experience, earlier experience with similar system(s) and domain knowledge levels should be seriously taken into account. The same problem can be encountered during TA several times. Problems that were only identified by TA were encountered less often than the problems that were identified by both methods.

This study has two limitations. First, since we are not sure that we have discovered all usability problems present in the evaluated CPOE system, according to a suggestion by Hartson et al. [17] we used the set of usability problems as identified by either method as a standard-of-comparison usability problem set. Therefore, the reported thoroughness of the methods may not be totally correct. Second, like Jeffries [15] we compared the usability evaluation methods by evaluating a single system. The inclusion of more systems in the usability assessment would have resulted in more generalizable results.

To our knowledge this is the first study that compared CW and TA in the domain of healthcare. Our results did not differ much from results obtained in other domains. However, as stated earlier, in the healthcare domain attention should also be paid to medical errors resulting from the usability problems.

5.2. *Meaning of the study and directions for future research*

About 40% of the usability problems were found by both methods and the users in the TA method encountered these problems twice as much as the problems that were identified by the TA method only. It is promising that either of the methods can reveal the problems that users encounter often, but selecting either one of the two usability evaluation methods may result in missing quite a number of low to high severity problems. The results of this study suggest that a combination of these methods gives the best results. However, decisions to employ only one method should be made with caution as the evaluation outcome is highly dependent on factors such as the interest for revealing specific types of problems, context of use and the resources available for evaluation. Some authors introduced "usability testing" as the gold standard for making a standard usability problem set that can be used for evaluating other methods [17]. Our study and studies by others [14,15] do not support this conclusion as the results of these studies showed that TA failed to reveal many low-severity and serious problems, which were identified by the CW method.

An issue that should be taken into account in any usability evaluation is that the way in which problem descriptions are written down affects the judgment of the evaluators. In our study first evaluators independently provided two separate lists of identified problems. When trying to merge the lists and resolve discrepancies between these evaluators we realized that part of the disagreement was caused by a different description and interpretation of the same problem. Karat [15] likewise observed that the language that was used by walkthrough evaluators to write down the problem description was difficult to understand and interpreted differently by other evaluators. In this study the evaluators had expertise in both usability evaluation methods and the users in the TA sessions had previous experience with the Medicator system. Although some studies have been done concerning the effect of usability expertise [35,36], further study is necessary to determine to what extent the users' system experience influences the results of evaluation studies.

6. Conclusions

This study shows that although TA showed a slightly better effectiveness, there is no significant difference between the performance of the CW and the TA methods in terms of number of usability problems identified and the mean severity of these problems. None of the methods can detect all usability problems that exist in a system. Since no single evaluation method will uncover all of the usability problems a combination of methods is advised as the most appropriate approach, especially if usability problems can lead to potentially fatal outcomes.

Summary points
What was already known:

- While the effectiveness of some usability evaluation methods (UEMs) has been studied in certain domains, studies on the effectiveness of these methods in the health care domain are lacking.
- Previous studies mostly compared usability evaluation methods on their cost effectiveness and resources needed. Although these studies compared the effectiveness of the methods based on the crude number of usability problems encountered, they did not take into account the thoroughness of the methods and they also did not study the nature of the usability problems that were commonly identified by the evaluated methods.
- It has been shown that some usability evaluation methods are better able to detect problems of a specific severity but the effectiveness of the methods for detecting usability problems related to the different phases of user interaction was not studied.

What the study has added to our knowledge:

- UEMs are not uniformly effective during all phases of user interaction. For example, think aloud (TA) has a better performance for identifying usability problems in the "Assessment" phase of the interaction.
- Because of the user characteristics and context of use some problems may not be detectable by certain UEMs. For example, TA may not identify certain problems if users have previous experience with the same or a similar system.
- Problems that are commonly found by the CW and the TA methods are likely to be encountered by users more often than the problems that are detected by only one of the methods.
- In the domain of health care the total effectiveness of the TA method is slightly better than that of the CW method but the methods do not differ significantly in terms of the number and severity of the detected problems. However, in evaluation studies of health information systems the effectiveness of methods for identifying usability problems that affect patient outcome should also be investigated.

Authors' contributions

R. Khajouei and M.W.M. Jaspers contributed to the conception and design of the study, acquisition of data, analysis and interpretation of data, and preparation of the first draft of the manuscript. Arie Hasman critically revised the content of the manuscript. All authors participated actively in the editing of the manuscript and approved its final version.

Conflict of interests

The authors have no conflicts of interest to declare.

Acknowledgement

The authors would like to acknowledge the cooperation of the participants in think aloud sessions. We would also like to thank Martijn Harpe for his cooperation in designing the clinical scenario. We are grateful for the assistance of Dennis de Jongh in conducting a part of the think aloud sessions.

REFERENCES

[1] R. Kaushal, K.G. Shojania, D.W. Bates, Effects of computerized physician order entry and clinical decision support systems on medication safety: a systematic review, Arch. Intern. Med. 163 (12) (2003) 1409–1416.

[2] E.M. Campbell, K.P. Guappone, D.F. Sittig, R.H. Dykstra, J.S. Ash, Computerized provider order entry adoption: implications for clinical workflow, J. Gen. Intern. Med. 24 (1) (2009) 21–26.

[3] E.W. Ford, A.S. McAlearney, M.T. Phillips, N. Menachemi, B. Rudolph, Predicting computerized physician order entry system adoption in US hospitals: can the federal mandate be met? Int. J. Med. Inform. 77 (8) (2008) 539–545.

[4] J.S. Ash, D.F. Sittig, R. Dykstra, E. Campbell, K. Guappone, The unintended consequences of computerized provider order entry: findings from a mixed methods exploration, Int. J. Med. Inform. 78 (Suppl. 1) (2009) S69–S76.

[5] J.S. Ash, M. Berg, E. Coiera, Some unintended consequences of information technology in health care: the nature of patient care information system-related errors, J. Am. Med. Inform. Assoc. 11 (2) (2004) 104–112.

[6] R. Koppel, J.P. Metlay, A. Cohen, B. Abaluck, A.R. Localio, S.E. Kimmel, et al., Role of computerized physician order entry systems in facilitating medication errors, JAMA 293 (10) (2005) 1197–1203.

[7] A.W. Kushniruk, M.M. Triola, E.M. Borycki, B. Stein, J.L. Kannry, Technology induced error and usability: the relationship between usability problems and prescription errors when using a handheld application, Int. J. Med. Inform. 74 (7–8) (2005) 519–526.

[8] R. Khajouei, J.D. de, M.W. Jaspers, Usability evaluation of a computerized physician order entry for medication ordering, Stud. Health Technol. Inform. 150 (2009) 532–536.

[9] N. Staggers, B.M. Jennings, C.E. Lasome, A usability assessment of AHLTA in ambulatory clinics at a military medical center, Mil. Med. 175 (7) (2010) 518–524.

[10] D.L. Cuomo, C.D. Bowen, Understanding usability issues addressed by three user–system interface evaluation techniques, Interact. Comput. 6 (1) (1994) 86–108.

[11] M.W. Jaspers, A comparison of usability methods for testing interactive health technologies: methodological aspects and empirical evidence, Int. J. Med. Inform. 78 (5) (2009) 340–353.

[12] W.D. Gray, M.C. Salsman, Damaged merchandise? A review of experiments that compare usability evaluation methods, Hum. Comput. Interact. 13 (3) (1998) 203–261.

[13] T.S. Andre, H.R. Hartson, M.S. Belz, A.F. McCreary, The user action framework: a reliable foundation for usability engineering support tools, Int. J. Hum. Comput. Stud. 54 (2001) 107–136.

[14] R. Jeffries, J.R. Miller, K.M. Uyeda, User interface evaluation in the real world: a comparison of four techniques, in: CHI Conf. Proc., 1991, pp. 119–124.

[15] C.-M. Karat, R. Campbell, T. Fiegel, Comparison of empirical testing and walkthrough methods in user interface evaluation, in: CHI Conf. Proc., 1992, pp. 397–404.

[16] K.A. Ericssom, A.H. Simon, Protocol Analysis: Verbal Reports as Data, revised ed., MIT Press, Cambridge, MA, 1993.

[17] H.R. Hartson, T.S. Ander, C.R. Wilinges, Criteria for evaluating usability evaluation methods, Int. J. Hum. Comput. Interact. 15 (1) (2003) 145–181.

[18] T. Beer, T. Anodenko, A. Sears, Apair of techniques for effective interface evaluation: cognitive walkthroughs and think-aloud evaluations, in: Proc. Hum. Factors Ergon Soc. Annu. Meet., 1997, pp. 380–384.

[19] R. Khajouei, N. Peek, P.C. Wierenga, M.J. Kersten, M.W. Jaspers, Effect of predefined order sets and usability problems on efficiency of computerized medication ordering, Int. J. Med. Inform. 79 (10) (2010) 690–698.

[20] D. Hix, H.R. Hartson, Formative evaluation: ensuring usability in user interfaces, in: L.J. Baas, P. Dewan (Eds.), User Interface Software, Willey, New York, 1993, pp. 1–30.

[21] K. Wharton, J. Bradford, R. Jeffries, M. Franzke, Applying cognitive walkthroughs to more complex user interfaces: experiences, issues, and recommendations, in: CHI Conf. Proc., 1992, pp. 381–388.

[22] J.P. Polson, H.C. Lewis, Theory-based design for easily learned interfaces, Hum. Comput. Interact. 5 (2) (1990) 191–220.

[23] M.T. Boren, J. Ramey, Thinking aloud: reconciling theory and practice, IEEE Trans. Prof. Commun. 43 (3) (2000) 261–278.

[24] H.C. Lewis, Using the “Thinking Aloud” method in cognitive interface design, IBM RC-9265, 1982.

[25] D.A. Norman, Cognitive engineering, in: D.A. Norman, S.W. Draper (Eds.), User Centered System Design: New Perspectives on Human–Computer Interaction, Lawrence Erlbaum Associates, Hillsdale, NJ, 1986, pp. 31–61.

[26] VirginiaTech., The User Action Framework (UAF) and its Tools [Internet], Virginia Tech., Blacksburg, 2003 March [cited 2010 December 18], Available from: http://research.cs.vt.edu/usability/projects/uaf%20and%20tools/index.htm.

[27] M.D. Kalmeijer, W. Holtzer, D.R. van, H.J. Guchelaar, Implementation of a computerized physician medication order entry system at the Academic Medical Centre in Amsterdam, Pharm. World Sci. 25 (3) (2003) 88–93.

[28] J. Nielsen, Usability Engineering, Academic Press, Boston, 1994.

[29] J. Nielsen, Heuristic evaluation, in: J. Nielsen, R.L. Mack (Eds.), Usability Inspection Methods, John Wiley and Sons, Inc., New York, 1994, pp. 25–62.

[30] J. Nielsen, Severity Ratings for Usability Problems [Internet], Jacob Nielsen, Fremont, 2005 April [cited 2010 December 18], Available from: http://www.useit.com/papers/heuristic/severityrating.html.

[31] A. Sears, Heuristic walkthroughs: finding the problems without the noise, Int. J. Hum. Comput. Interact. 9 (3) (1997) 213–234.

[32] W.M. van Someren, F.Y. Barnard, A.C.J. Sandberg, The Think Aloud Method: A Practical Guide to Modelling Cognitive Processes, Academic Press, London, 1994.

[33] J. Rieman, M. Franzke, D. Redmiles, Usability evaluation with the cognitive walkthrough, in: CHI’95 Mosaic of Creativity, 1995, pp. 387–388.

[34] J.P. Polson, H.C. Lewis, J. Rieman, K. Wharton, Cognitive walkthroughs: a method for theory-based evaluation of user interfaces, Int. J. Man-Mach. Stud. 36 (5) (1992) 741–773.

[35] S. Kirmani, Heuristic Evaluation Quality Score (HEQS): defining heuristic expertise, J. Usability Stud. 4 (1) (2008) 49–59.

[36] J. Nielsen, Finding usability problems through heuristic evaluation, in: CHI Conf. Proc., 1992, pp. 373–380.