Author(s): Tyllinen, Mari & Kaipio, Johanna & Lääveri, Tinja & Nieminen, Marko H.T.

Title: We Need Numbers! - Heuristic Evaluation during Demonstrations (HED) for Measuring Usability in IT System Procurement

Year: 2016

Version: Post print

Please cite the original version:

Tyllinen, Mari & Kaipio, Johanna & Lääveri, Tinja & Nieminen, Marko H.T.. 2016. We Need Numbers! - Heuristic Evaluation during Demonstrations (HED) for Measuring Usability in IT System Procurement. CHI'16 Proceedings of the 34th Annual ACM Conference on Human Factors in Computing Systems. 13. DOI: 10.1145/2858036.2858570.

We Need Numbers! - Heuristic Evaluation during Demonstrations (HED) for Measuring Usability in IT System Procurement

Mari Tyllinen[a,c], Johanna Kaipio[a], Tinja Lääveri[b,c], Marko H.T. Nieminen[a]

[a] Aalto University
Espoo, Finland
{mari.tyllinen,
johanna.kaipio,
marko.nieminen}
@aalto.fi

[b] Helsinki University
Hospital and University
of Helsinki
Helsinki, Finland
tinja.laaveri@hus.fi

[c] Oy Apotti Ab
Helsinki, Finland
{mari.tyllinen,
tinja.laaveri}
@apotti.fi

ABSTRACT

We introduce a new usability inspection method called HED (heuristic evaluation during demonstrations) for measuring and comparing usability of competing complex IT systems in public procurement. The method presented enhances traditional heuristic evaluation to include the use context, comprehensive view of the system, and reveals missing functionality by using user scenarios and demonstrations. HED also quantifies the results in a comparable way. We present findings from a real-life validation of the method in a large-scale procurement project of a healthcare and social welfare information system. We analyze and compare the performance of HED to other usability evaluation methods used in procurement. Based on the analysis HED can be used to evaluate the level of usability of an IT system during procurement correctly, comprehensively and efficiently.

Author Keywords

Usability evaluation; public procurement; summative evaluation; measuring usability; healthcare and social welfare information system; electronic health record.

ACM Classification Keywords

H.5.2 Information Interfaces and Presentation (e.g. HCI): User Interfaces (Evaluation/Methodology); K.6.3 Software management (Software selection).

INTRODUCTION

Heuristic evaluation (HE) [39] is an established method in user-centered system design [21]. It has been widely used in software development and evaluation especially in formative [13,18] development contexts in which it is possible to adjust or change the functionality of the software through technical development. In this paper, we describe accommodating the method to summative use.

From the research perspective, HE may be considered an outdated, even obsolete, method that fosters little academic contribution. Over the years, HE has been extended to accommodate different purposes and situations (e.g. [48]). The increasing demands in public IT system procurement [29] of packaged/COTS (commercial-off-the-shelf) software, introduce new summative-type uses for this well-established, widely applied method. Moreover, COTS procurement is an understudied systems context [36]. Current and emerging procurement surroundings pose challenging new requirements for fitting methods to the context and applying them properly. Legislation on public procurement permits only decisions that are based on relevant objective criteria for choosing economically the most advantageous tender. This applies on both EU [12] as well as on the national level in Finland [17].

Public IT systems have a large-scale impact on a significant number of people. Therefore understandably, usability problems appear as a constant topic in national discussion, especially in the area of healthcare IT systems [50]. The role of users and user needs have been left to the background during the product development of packaged software [25]. Moreover, usability has not appeared as an explicit requirement, or target of evaluation in public procurement [29]. We expect this to change along with the changing regulation. Therefore, we propose new applicable practices for assessing usability during public IT system procurement of packaged software.

The method development and validation discussed in this paper has been devised during an ongoing large-scale public IT system procurement which we call ‘CAPIS’ (Client And Patient Information System). The object of procurement is a fully integrated information system for

CHI'16, May 07 - 12, 2016, San Jose, CA, USA

ACM 978-1-4503-3362-7/16/05…$15.00
DOI: http://dx.doi.org/10.1145/2858036.2858570

tertiary, secondary and primary healthcare as well as social welfare with approximately 40 000 professional end users.

HE has already been suggested as an evaluation method for IT system procurement [31]. However, we are still lacking proper ways to use it in a summative manner that would enable a score-based comparison between competing systems. Our paper addresses this uncovered research topic. The main contribution of our paper is the definition and real-life validation of the Heuristic Evaluation during Demonstrations (HED) -method for public IT procurement.

RESEARCH BACKGROUND

In the context of this study three tracks of related research are of interest: usability evaluation of complex systems, usability in IT procurement of complex systems and measuring usability.

Healthcare has been identified as a domain for complex information systems [35]. Researchers have pointed out that there is a lack of appropriate usability evaluation methods [8,42]. In the context of clinical information systems, usability evaluation methods have been applied to guide further development (e.g. [27,32,40,41]). However, the evaluation has focused on a limited set of functionalities or parts of the system. The role of usability design and evaluation has not been established in social welfare domain [26].

Some researchers have argued that public organizations do not emphasize and include usability in IT system procurement because they are not prepared to take the responsibility for it [28]. In regard to electronic health record (EHR) systems, these organizations have assumed that the responsibility lies on the vendor [15]. However, the high configurability of these systems could be used to improve usability [37], hence the customer organization should also take responsibility for the usability in the EHR system implementation [15].

The importance of usability requirements [5] and understanding human factors [46] in IT system procurement were first highlighted two decades ago. Indeed, recent research proves that usability criteria can be used in public IT procurement [43,49] although this has not yet become a standard procedure [29]. There are examples of such procedures in the field of healthcare [31,33] and usability tests are recommended as the primary evaluation method when purchasing a healthcare IT system [31,45]. However, studies give only an overview of rankings, but the used evaluation criteria [31,33], formulation of numerical results or justification behind rankings have not been described in detail. Also, user testing requires significant resources: testing five tasks for two different user groups on two competing systems has taken one month of effort [43].

When purchasing large and complex information systems and evaluating several different systems from different vendors during the selection process, a more cost-effective approach is needed. Several researchers view HE as being a viable method for preliminary assessment of candidate systems in the healthcare field due to its low-cost and quickness [7,31,45]. CLIPS (clinical information processing scenarios) are suggested as a basis of heuristic evaluation when selecting information systems [31].

Product demonstrations have a major influence on the assessment of both the suitability of the products and the vendors [25] while the usability evidence provided by these traditional vendor demonstrations is seen as weak [31]. Authors [25] have argued that salesmanship related to demonstrations has been more important in decision-making than the actual IT system. However, recent legislation [12,17] requires objective criteria for choosing economically the most advantageous tender in public IT procurement. In addition to being useful demonstrations also provide deficiencies when used in procurement.

HE has been criticized for limited inclusion of use context, a very limited set of evaluated user interfaces that need to be selected before evaluation and not being able to reveal major missing functionalities in the evaluation [9,38]. While usability testing has been devised also for summative purposes, one could argue that inspection methods like HE have been designed for formative use because they provide mainly qualitative data on the usability problems [14]. The current practice on measuring usability indeed heavily relies on usability tests while inspection methods are not discussed [24]. Attempts to quantify HE have been presented [20] for evaluating the degree of usability of websites. In this model heuristics are divided into categories and problems are categorized accordingly resulting in a calculated usability score.

Based on the related research, we argue that there is a need for an inspection method for measuring usability during IT system procurement. We further argue that the value of a usability evaluation method for IT procurement is reflected in its ability to determine the level of usability of an IT system correctly, comprehensively and efficiently. *Correctness* includes *reliability* and *validity*, two common measures for examining an evaluation method [22]. *Comprehensiveness* is similar to *thoroughness* [2] relating to examining the system as broadly as possible. *Efficiency* [22] relates to the resources used to get as comprehensive an evaluation of the system as possible.

Our HED method responds to these challenges by adding the complex work context and process view to heuristic evaluation with the combination of demonstrations and user scenarios. HED quantifies the evaluation for comparison purposes in IT system procurement with a different calculation model than in [20].

OUR METHOD: HEURISTIC EVALUATION DURING DEMONSTRATIONS (HED)

Much like in the typical HE procedure [39], in HED the actual evaluation is performed in two steps. Before the evaluation steps, the process also includes preparing the user scenarios that the demonstrations, during which HED is performed, should be based on.

The user scenarios describe typical work processes that are supported by IT systems [6,16]. They are written by domain experts. The timespan of the described process can be days or even months. The user scenarios are provided to the competing vendors beforehand and the demonstrators are required to follow them. System functionalities outside the user scenario are not allowed to be demonstrated. The user scenarios are divided into shorter parts and the transitions between parts can be used for pre-defined short demonstration breaks.

HED is based on documenting four different types of usability issues: heuristic violations, missing functionalities, omitted parts of the user scenario and positive findings; numeric scores are given for these during the demonstration. In contrast to typical HE procedure, missing functionalities in the system are revealed and documented. These aspects are an important addition to the procedure because they give a more accurate representation of the system's usability when real users use it for their tasks.

At the beginning the product to be demonstrated has 0 usability points which equals to the highest grade. Heuristic violations, missing functionalities and omitted parts of the user scenario give negative points and are called subtractions while positive findings give positive points. The total sum of points determines a usability grade for the demonstrated system. In theory, at the end of evaluation this sum could be positive, however, as HED is focused on revealing deficiencies a negative total sum is assumed. The process is depicted in more detail in the following.

Before the Evaluation

Selecting Heuristics

First, usability specialists, referred to as evaluators, select the list of heuristics to be used based on the evaluated system. When familiarizing with the chosen heuristics, the evaluators may decide to modify them, e.g. by emphasizing contextual aspects. The evaluators also define examples of violations that are special and typical for the type of systems that are to be evaluated in the contexts in focus.

Setting Scores for Usability Issues

The scoring comprises of the above-described types of *usability issues*: heuristic violations, missing functionalities, omitted parts of the user scenario and positive findings together with the *influence rating* of all four. *Influence rating*, a new concept, comprises *severity rating* of traditional heuristic evaluation and similarly *supportivity rating* for positive findings.

In the literature, there are two *severity rating* approaches for HE; a single scale and a two dimensional approach, with impact and frequency as the two dimensions of a table [38, p. 104]. Based on research, these dimensions are not correlated [44] and should be rated separately. In the context of risk management of ICT systems, similar measures exist [1]. The risk is calculated as the product of "loss due to an undesirable outcome" (L) and "probability of the undesired outcome" (P): R=R*P.

Similarly, we propose a scale of two dimensions for rating the heuristic violations in HED: the impact rating describing the impact of the violation on the users on a scale from -1 (minor usability violation) to -3 (major usability violation) and the frequency rating the rate with which the violation occurs in the demonstration on a scale from 1 (single) to 3 (prevailing/frequent). The severity rating is then calculated as the product of the impact and frequency ratings, and results in 6 different scores for heuristic violations: -1, -2, -3, -4, -6 and -9. Calculating the product gives a scale that emphasizes the more severe end. The documenting of cosmetic problems is neither possible nor necessary in a demonstration. Also, estimating the frequency of problems during the demonstrations can be viewed as representative of actual use.

Additionally, also an impact score of -10 is included, and can be given to a single *usability catastrophe*, in which case the frequency is evaluated as 1. The *supportivity rating* of positive findings, i.e. examples of good design solutions relating to the heuristics, includes an impact rating of +1 and evaluating the frequency as with heuristic violations.

Other subtractions than heuristic violations are missing functionalities and omitted parts of the user scenario. The evaluators should get familiar with the user scenarios. The most essential functionalities (e.g. summary view of patient's or customer's data) should be recognized during discussions with domain experts. The domain experts can explain which parts of the user scenarios are essential for completing their tasks. The existence of the required functionality is the prerequisite for evaluating their usability. When the demonstration does not include a main functionality, this results in a subtraction of either -5 (impact -5, frequency 1) or -10 (impact -10, frequency 1) depending on the functionality or level of deficiency.

Omitted parts of the user scenario are calculated to include the same average amount of heuristic violations as the demonstrated parts. For instance if only 75% of a user scenario were demonstrated and the heuristic violation points were -90, an additional subtraction of -30 would be added and the final heuristic violation points would be -120.

In our procedure, missing functionality and omitted parts of the user scenario reduce points instead of simply disqualifying a vendor for the following reasons: Firstly, in order to treat all vendors equally in relation to the user scenarios, the deficiencies need to affect the usability grade. Secondly, all functionalities in the user scenario are not necessarily included in the final system requirements due to the parallel ongoing negotiations that determine the content of the final request for proposal (RFP). Thirdly, the systems could be further developed after the vendor selection to eventually meet the requirements.

Determining Grade Boundaries

Grade boundaries are predetermined before the evaluation based on desired and/or acceptable result, by piloting the method and considering the length of the demonstration. For example, the grade *Fail* could be determined to be anything below -90, which corresponds to 10 major (impact -3) and prevailing (frequency 3) usability violations. It is more suitable to give grades instead of points as the result; the level of usability is more easily understood from the procurement viewpoint. This also simplifies determining the accepted minimum level as described above for the grade *Fail*. All grade boundaries should also include a so called *grey area* to mark those points that are so close to the boundary between two grades that the actual grade is not necessarily unambiguous and needs negotiation. For an example of defined grade boundaries see Table 3.

Scoring Alignment across Evaluators

Before the evaluation, the evaluators should have an alignment discussion on the scoring, for a common understanding on the meaning, use and application of the impact and frequency ratings. Also, before HED is used for real, the evaluators should practice using the method together for a few times.

First Step: Analysis during the Demonstration

Evaluators individually document the observed positive findings, heuristic violations, missing functionality and omitted parts of the user scenario as they occur during the demonstration. The pre-indicated breaks during the demonstration can be used to assess and document the impact and frequency of observed usability issues. Because the analysis is conducted during a fast-paced demonstration, the detailed descriptions of the documented heuristic violations and their relation to a specific heuristic cannot be documented in contrast to traditional HE [38] and the method presented by González et al. [20].

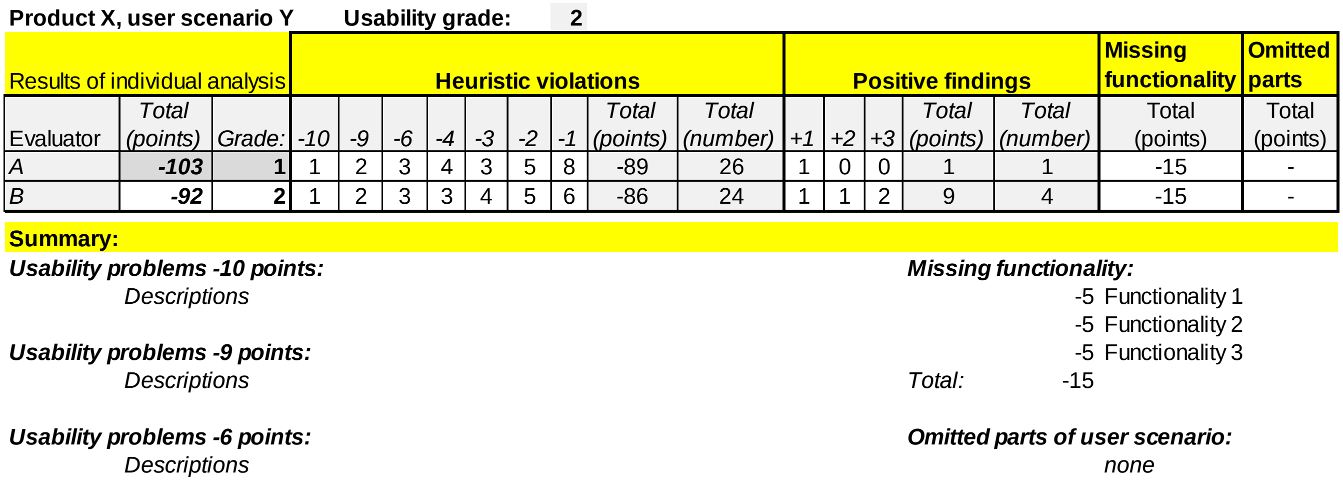
Immediately after the demonstration the evaluators should determine with one to three domain experts, who also followed the demonstration, subtractions for the missing functionalities and omitted parts of the user scenario. This is done to avoid unwarranted subtractions i.e. necessary functionalities were demonstrated in a way that the domain experts were not able to describe to the evaluators beforehand. Accordingly, these subtractions are the same for both evaluators. The resulting score Us(v_s) is calculated from the analysis for each evaluator:

Us(v_s) = $\sum(I_n \times F_n)$, where, I_n is the impact and F_n the frequency of each usability issue *n* for vendor *v* for user scenario *s*. Usability issues include the heuristic violations, other subtractions and positive findings.

Second Step: Aggregation of Results and Defining the Usability Grade

After each demonstration, the evaluators meet to negotiate the final grade for the system based on their individual Us(v_s) scores. They present individual results to each other and compare the scores and resulting grades. They collect a comparative table that includes the number of heuristic violations and violations for each severity rating for all evaluators. The table similarly includes positive findings and their supportivity ratings as well as notes on the other subtractions (Figure 1).

If the individual scores are within the same grade and relatively close to each other, the resulting grade can be accepted as is. By contrast, if the grades differ from each other, closer comparison of the individual analysis should be done to reach a common understanding, which will then be transformed into a final grade. The *grey areas* defined for the grade boundaries can assist in the discussion. During the negotiation evaluators should go through the documented most severe heuristic violations (-10, -9 and -6 points), the numbers of less severe heuristic violations (-4, -3, -2 and -1 points), the numbers of positive findings (+1, +2 and +3 points) and the total subtraction points and



Product X, user scenario Y **Usability grade:** **2**

Results of individual analysis			Heuristic violations									Positive findings					Missing functionality	Omitted parts
Evaluator	*Total (points)*	*Grade:*	*-10*	*-9*	*-6*	*-4*	*-3*	*-2*	*-1*	*Total (points)*	*Total (number)*	*+1*	*+2*	*+3*	*Total (points)*	*Total (number)*	Total (points)	Total (points)
A	***-103***	**1**	1	2	3	4	3	5	8	-89	26	1	0	0	1	1	-15	-
B	***-92***	**2**	1	2	3	3	4	5	6	-86	24	1	1	2	9	4	-15	-

Summary:

Usability problems -10 points:

Descriptions

Usability problems -9 points:

Descriptions

Usability problems -6 points:

Descriptions

Missing functionality:

-5 Functionality 1
-5 Functionality 2
-5 Functionality 3

Total: -15

Omitted parts of user scenario:

none

Figure 1. Example of a comparative table, that summarizes the results of HED. All numbers are examples and not related to actual evaluations in CAPIS.

positive points. Based on this the evaluators should jointly decide which grade is most justified. Finally, descriptions of the most severe heuristic violations (Figure 1), is included based on the individual evaluations.

APPLYING THE HED METHOD: CASE CAPIS

The real-life validation of the HED method took place in spring 2014 during the dialogue stage of the CAPIS procurement project.

During summer 2013 we piloted the first version of HED as part of a pilot on the whole evaluation procedure using one of the currently used EHR systems as the evaluated system. During the pilot, three usability specialists applied the first version of the method. The main findings from the pilot study that influenced the HED method were: Documenting on a paper chart was too slow. Training and practicing are important; the evaluators should thoroughly know the heuristics and the user scenario, as well as be familiar with the application domain; in the pilot case clinical work and EHR systems and at later stages of CAPIS also social welfare. The findings start to repeat which should be taken into consideration when determining the grade boundaries for demonstrations with different lengths. Conversely, the same issues are witnessed several times and the findings can be specified over time which aids in documenting during the fast-paced demonstration.

In spring 2014 the aim was to select 2-4 candidates that meet the minimum requirements, including usability, to continue in the procurement process. The HED method was one part of the usability evaluation. Additionally, perceived usability questionnaires for future end-users were used during demonstrations. Also, a more traditional HE was used for a limited user group and context of use.

The evaluation was based on comprehensive user scenarios. The four competing IT vendors were required to strictly follow the user scenarios during product demonstrations. Of the nine user scenarios, six were assessed with the HED method. Three of these user scenarios covered healthcare and three of them social welfare. The user scenario based demonstrations were also used in the procurement for the evaluation of the coverage and quality of the system functionalities by subject domain experts, i.e. future end-users such as physicians, nurses and social welfare professionals.

The final result for the usability of a vendor's solution in CAPIS was determined by a weighted combination of scores from all the different methods and user scenarios. The details of this final scoring are not in the scope of this paper.

User Scenarios and Demonstrations

The six user scenarios chosen for usability evaluation with HED covered the central areas of healthcare and social welfare information system use from the professionals' perspective. They depicted typical workflows and IT system use between seamlessly cooperating university hospital level specialized medical care, primary health care and social welfare. The user scenarios were similar to clinical information processing scenarios (CLIPS) [16,23,34]; the method was adapted also for social welfare. They were written by healthcare and social welfare professionals with the guidance of domain experts working at the procurement office.

User Scenario	Length (approx.)
1. Emergency department (ED) – Intensive care unit (ICU) – Operating room (OR)	6 hours
2. Maternity clinic – Labor and delivery – Child health clinic	3 hours
3. Inpatient ward	2 ½ hours
4. Social assistance	2 ½ hours
5. Child welfare	2 hours
6. Services for people with disabilities	2 hours

Table 1. The user scenarios used in CAPIS and their lengths.

The user scenario contexts and reserved times for the demonstrations are depicted in Table 1. The user scenarios had different lengths, and in total of about 18 hours of usability evaluation during demonstrations were done for each vendor's solution.

Before the Evaluation

The HED procedure was conducted by two usability specialists (the first two authors) who have extensive knowledge and experience with usability evaluation methods in general and in usability of healthcare and social welfare information systems in particular. We decided to use Nielsen's 10 heuristics [38], because they are well aligned with heuristics presented for the field of health informatics (e.g. [23,51]) and both usability evaluators had applied them in several evaluations previously. Additionally, we combined our own experiences with cases from literature to compile a set of examples of heuristic violations specific to health information systems. An illustration on the nature of these examples is presented in Table 2. We found the list applicable also for the evaluation of social welfare user scenarios.

The usability of the system for each user scenario $Us(v_s)$ was converted to grades on a scale of 3 (good) – 2 (fairly good) – 1 (acceptable) – 0 (fail). The grade boundaries were determined for each user scenario based on the length of the user scenario and piloting the method: for a two hour user scenario the absolute minimum points were -170; for user scenarios up to five hours in length every hour increased the minimum points by -60 points and for the fifth hour and the following hours every hour increased the minimum points by -40 points; over 90 % of the minus points resulted in a failed grade; and the grade boundary for grade 2 was at 60% and for grade 3 at 30 %. This resulted in the grade

Heuristic by Nielsen [38]	Example of violation specific to health information systems
Simple and Natural Dialogue	Optimal amount of information in one screen regarding the task at hand, possibility to easily drill down to details for example from a summary view of patient's situation.
Minimize User Memory Load	Patient's / customer's identification information (e.g. name, unique ID) is prominently displayed on screens in order to minimize the memory load and avoid documentation to the wrong patient / customer file.
Prevent Errors	Consistent use of colors and other methods to highlight the abnormal values / information throughout the system. If used inconsistently, the user may not notice the visually differentiated important information or it can be misunderstood.

Table 2. Examples of heuristic violations.

boundaries presented in Table 3. The *grey areas* around the grade boundaries were determined to be +- 5 %.

The usability evaluators jointly wrote more detailed definitions for the impact and frequency ratings and the *usability catastrophe* so that the scoring during the demonstration would be equal. A couple of days before evaluation, domain experts guided the usability evaluators through the user scenario from the IT system point of view. They were also usually able to tell what types of screens were to be expected at different parts of the user scenarios. The user scenario was printed on paper and the parts of the story with the essential functionality that would result in subtractions if missing from the demonstration were highlighted.

The usability evaluators practiced HED a week before the actual evaluations for an hour during a training session with the EHR currently in use at the university hospital (not one of the contending systems).

First Step: Analysis during the Demonstrations

The two evaluators participated in 24 user scenario sessions (demonstrations): three healthcare (1, 2, 3) and three social welfare (4, 5, 6) user scenarios were demonstrated by all four vendors (I, II, III, IV) during a period of five weeks according to the schedule in Table 4. There was a two-week break between the healthcare and social welfare demonstrations.

Both evaluators had own laptops to document the observed usability issues according to the HED process. After the demonstration of the user scenario, if time remained, the vendors were asked to re-demonstrate those parts of the story where the domain experts felt something was unclear. The two usability evaluators could not affect this process, but were able to make complementary documentation. Finally, immediately after the demonstration was over, the other subtractions were decided according to the process.

User Scenario	Grade 3	Grade 2	Grade 1	Grade 0
1	0 … -111p.	-112 … -222p.	-223 … -333p.	-334 … -370p.
2	0 … -69p.	-70 … -138p.	-139 … -207p.	-208 … -230p.
3, 4	0 … -60p.	-61 … -120p.	-121 … -180p.	-181 … -200p.
5, 6	0 … -51p.	-52 … -102p.	-103 … -153p.	-154 … -170p.

Table 3. Grade boundaries and related points.

Second Step: Aggregation of Results and Defining the Usability Grade

After each demonstration, both evaluators reviewed their own documentation and made final necessary adjustments such as combined multiple documentations of the same problem into one with an appropriate frequency rating. After this, the evaluators compiled a summary of the number of violations with different severity ratings and positive findings. The most severe violations (with rating of -10, -9 and -6 points) were reviewed in more detail and it was marked whether both evaluators had documented the same problem.

The grade negotiations were usually quite straightforward following the process described in the previous section. In case of a difference in the resulting grade the first step was to pinpoint the main reason for the difference, such as the other evaluator had documented substantially more small violations than the other or the other had documented more positive issues. This analysis required in some cases a more detailed review of the documented issues, for example if the positive findings did not overlap between the evaluators this could be considered as a raising factor from a lower grade. The common analysis resulted in a final grade that both evaluators agreed on. This step took on average one hour.

Week	Mon	Tue	Wed	Thu	Fri
1	1 / I			3 / I	2 / I
2	1 / II			3 / II	2 / II
3	1 / III	1 / IV		3 / III, 2 / IV	2 / III, 3 / IV
4	4-6 / I		4-6 / II		
5		4-6 / III	4-6 / IV		

Table 4. Evaluation schedule for the user scenarios (1-6). The number of competing vendors was four (I, II, III, IV). However, the process was planned for six candidates.

ANALYSIS AND EVALUATION OF RESULTS

We analyze the performance of the HED method by assessing the grades documented by both evaluators and by relating it to other usability evaluation methods applied in the CAPIS procurement process.

Evaluator-specific grades are further analyzed with inter-rater agreement and inter-rater reliability. Finally, the correlation of grades for usability resulting from different methods (HED and user questionnaires) is presented as an initial analysis of the correctness of the method. The criteria of comprehensiveness and efficiency are discussed briefly.

Other Usability Evaluation Methods during Procurement

User Questionnaires

Three different questionnaires were applied for evaluating perceived usability: two short questionnaires during the breaks (six statements each, total 12 statements) and a summative questionnaire at the end of each user scenario (10 statements). The questionnaires' design utilized the established usability questionnaires, such as SUMI (Software Usability Measurement Inventory) [30], SUS (System Usability Scale) [4], and QUIS (Questionnaire for User Interaction Satisfaction) [9], as well as a tailored usability questionnaire for EHR systems [50]. None of the established questionnaires were suitable as is, because the questions were not to be answered based on experience of using the system but based on seeing the system being demonstrated.

The statements were answered using four-point Likert scale: Fully agree (3) – Fully disagree (0). The statements were either on traditional usability issues (consistency, logic, status, complexity and visual appearance of the system) or EHR-related (compatibility of system and clinical tasks, support for collaboration in clinical work). Examples of these statements: "*I perceive the arrangement of the fields and functionalities on-screen logical*", "*The system supports collaboration and information exchange between involved parties*" and "*This system is highly suitable for my daily work tasks*". In the short questionnaires 6 statements were on usability issues and 6 statements EHR-related. The summative questionnaire was similar to SUS [4], but replaced questions 4, 9 and 10 with EHR-related questions.

Table 5 includes a summary of the numbers for questions and subject domain experts that responded to the questionnaires in each user scenario.

Traditional Heuristic Evaluation

We conducted also a more traditional HE procedure [39] with a similar scoring process as with HED for one small area of the CAPIS procurement: the client and patient portal from the citizen's perspective. The evaluation covered 10 tasks. The two usability specialists first performed individual evaluations with access to the system and finally aggregated the results similarly as in HED. The time spent conducting the evaluation was not restricted but in practice took approximately 2 ½ hours per vendor. The aggregation of the results took approximately an additional 45 minutes.

Scenario	Questionnaires	Respondents
1	(2 x 6 questions) x 2 + 10 questions	22-23 (physicians and nurses)
2	2 x 6 questions + 10 questions	9-10 (physicians and nurses)
3	2 x 6 questions + 10 questions	17 (physicians and nurses)
4	2 x 6 questions + 10 questions	29-31 (social workers)
5	2 x 6 questions + 10 questions	29-31 (social workers)
6	2 x 6 questions + 10 questions	29-31 (social workers)

Table 5. Usability questionnaires and respondents.

Statistical Analysis

For the initial analysis of the correctness of HED, reliability and validity measures were calculated from the grade data. For the analysis of reliability of the assessments performed by two different evaluators, two types of calculations were used: the inter-rater agreement (IRA) relates "to the extent to which different raters assign the same precise value for each item being rated" [19]. The simplest index of IRA is percent agreement. Inter-rater reliability (IRR) relates "to the extent to which raters can consistently distinguish between different items on a measurement scale" [19]. With two evaluators and ordinal data weighted kappa and intra-class correlation coefficient (ICC) are suggested. For the calculation of weighted kappa the R statistics software irr-package was used and for the calculation of ICC the R statistics software psych-package was used. For analyzing validity, correlation was calculated. The Pearson correlation to be used with interval data was calculated with the R statistics software.

Inter-rater Agreement (IRA) and Inter-rater Reliability (IRR)

The individual grades given by the two evaluators proved the same or adjacent in all scenario sessions as can be seen from Table 6. The grades given ranged from 0 to 3, thus the whole range of grades was in use. In all sessions with adjacent grades one or both of the evaluators were on the *grey area* near the grade boundaries. This was also the case in some of the sessions where the grade was the same.

The IRA for our grade data was 70.8 % indicating that the agreement between evaluators was substantial and high confidence in grades given to vendors in demonstrations were correct. For the IRR Kappa was 0.696 ($p < 0.001$ and z= .000337 and z-value = 3.59). The result for ICC was 0.73 ($p<0.001$) lower-bound = 0.69 and upper-bound = 0.91. The IRR calculations indicate that the reliability of the

User Scenario	Vendors and the grades			
	I	II	III	IV
1	same	same	same	adjacent
2	adjacent	same	same	adjacent
3	same	same	same	adjacent
4	adjacent	same	same	adjacent
5	same	same	same	same
6	same	same	same	adjacent

Table 6. HED grades given were close to each other in all sessions. Sessions in which one or both evaluator's points were in the *grey area* are marked.

grades was substantial as the values are between the suggested 0.61-0.80 [19].

Correlation between HED and user questionnaires

For the comparison purposes of this paper all the statements in the questionnaires were treated as equal and the points given by subject domain experts to a user scenario were calculated as an arithmetic mean of all answers. Although the respondents were same in all sessions across vendors, some did not provide all answers. All given responses are included in the calculations in this paper. This differs from the calculation used in CAPIS, where respondents who did not give scores to all vendors were excluded because of equal treatment required in procurement. For the differential between the arithmetic mean points given by the subject domain experts in each user scenario and the final grade given by the usability specialists, see Table 7. Overall, the arithmetic mean points are higher than the grades, evidenced by only a few negative values. When interpreting the table it should be noted that HED produced integers while the arithmetic mean points of the subject domain experts include decimals.

Based on correlation cor=0.79 ($p<0.001$, df = 22 and t = 6.097), the results of the perceived usability questionnaires and HED seem to be well aligned.

Comprehensiveness and Efficiency of the Method

Comprehensiveness and efficiency are interrelated measures. The execution of HED took 18 hours per vendor for evaluation, with an additional 6 hours for the aggregation of results. Thus the required total resources per vendor for the execution with two evaluators were 48 hours or 6 days of effort. This comprised 10 different use contexts (two of the user scenarios had three different use contexts). This results in ~5 hours per use context. As compared with traditional heuristic evaluation, which took a total of 6.5 hours per vendor (3.25 hours per evaluator) and evaluated one use context, HED was more comprehensive (more contexts) and more efficient by 26 %. Also, when considering the fact that markedly more tasks (more comprehensive) were evaluated with HED than with traditional HE per use context, the efficiency is further increased.

User Scenario	Differential of Results			
	I	II	III	IV
1	-0.01	0.23	0.58	0.83
2	0.04	0.24	0.18	1.09
3	1.28	0.12	-0.42	0.05
4	0.8	0.73	0.65	1.38
5	0.68	0.47	0.61	1.15
6	0.78	0.34	0.64	1.38

Table 7. Differential between usability evaluation methods: user questionnaires and HED. A negative value indicates that HED evaluated the usability higher and a positive value indicates that user questionnaires evaluated the usability higher in that user scenario.

The effort used for getting the results from the user questionnaires was approximately 396 ½ hours per vendor or about 40 hours per use context (based on Tables 1 and 5). In our case the number of subject domain experts varied between 9 and 31. With the minimum number of subject domain experts (9) attending each user scenario the effort per use context would be 16 hours (2 days), thus making HED also more efficient than user questionnaires by 68%. This calculation does not take into account the fact that domain experts did not solely use their effort during the demonstration for perceived usability evaluation but also assessed the coverage and quality of the IT system. The comprehensiveness of questionnaires is the same as with HED, as the same use contexts and tasks are evaluated.

DISCUSSION

In public IT procurement any usability evaluation method is useful only if it produces consistently comparable numeric results that give an accurate view of the usability of the whole system. This should also be done as efficiently as possible. Our findings suggest that HED responds to these requirements well.

HED differs from traditional heuristic evaluation of usability by including the context of use into the evaluation with user scenario based demonstrations and by producing a quantitative usability score for the demonstrated system. The method was applied and validated in the context of a major public IT procurement of a client and patient information system. Our results show that HED gives correct usability grades to the evaluated systems, is aligned with the results of perceived usability and makes it possible to evaluate the IT system more comprehensively and efficiently than with other methods.

Importance of Findings

As the definition in ISO9241-11 indicates, usability is context-dependent. Traditional heuristic evaluation, however, has been criticized for not addressing context [10]. For complex IT systems one challenge is the complexity of user tasks, workflows and the user environment (e.g. [23]). For example, healthcare work has been described being unique for varying reasons: the work is variable, dynamic, complex, emergent in nature, involves a high degree of both ambiguity and uncertainty, requires a high degree of coordination and is not easily deferred [3,47].

We brought the context and workflows into the heuristic evaluation procedure in the form of demonstrations based on user scenarios written by domain experts. In the context of IT procurement the user scenarios are important in reflecting an accurate picture of the system as the vendors are not able to select aspects and functionalities that they themselves wish to demonstrate.

Relation to Similar Studies

Evaluating usability during IT procurement has not become an established practice. Heuristic evaluation has been suggested as a suitable method for preliminary evaluation and ranking of candidates [7,31,45]. However, we are not aware of literature describing how this should be done in a numeric and comparable way. The research on measuring usability concentrates on usability testing [24]. To our knowledge, HED is the first such method reported for IT procurement that also addresses the criticized shortcomings [9,38] of traditional heuristic evaluation.

In contrast with the method of user testing in procurement presented by Riihiaho et al. [43] where the evaluation of one system and two use contexts but only ten tasks in total took half a month of effort (equals about eleven days of effort), our method is markedly more efficient and comprehensive. In almost half the time, 6 days of effort, we evaluated five times as many use contexts with HED in a much more complex working environment i.e. social welfare and healthcare.

Our results indicate that HED is more efficient than using perceived usability questionnaires and also provides more documented information for the implementation phase. While from a pure usability perspective HED might be seen as the better choice during procurement, the inclusion of users is seen as essential for the future acceptance of the system [11].

Alternative Explanations of Findings

Our findings are based on the evaluations by two usability specialists with a very similar background who applied the method in practice. Nielsen [38] suggests using at least three evaluators to get a comprehensive view on usability problems. However, in procurement (thus also in HED) the focus is on comparing the level of usability between competing systems through an equal evaluation procedure. We argue the difference can be detected without revealing all heuristic violations. A third evaluator would have lowered the efficiency of the method, and based on our results two evaluators can detect the difference in the level of usability between systems.

Both the results of user questionnaires and HED evaluations were based on the same user scenario demonstrations. The role of the user scenario on the view of the system's usability is important: a user scenario not depicting the context of use correctly could in theory lead to consistently aligned, but false results of usability. However, in practice this should be reflected in the subject domain experts' evaluations as low ratings as they are experts on the use context.

The user scenarios were given to the vendors before the demonstrations. This practice was chosen to give the vendors opportunity to present the optimal way of using the system and minimized the emergence of unforeseen usability problems resulting from the demonstrator's use of the system. These two aspects are similar in other expert evaluation methods for usability.

Applicability of Findings

The HED method has been designed for usability evaluation during IT procurement of packaged software in absence of other suitable methods. The method requires a working system and is not thus suitable in the context of procuring software development projects. However, the method may be suitable for acceptance testing or comparing the usability of different versions of a product.

The HED produces a usability grade for the system evaluated which is related to the user scenario used. As indicated by our case example, these usability grades can be used during procurement to determine whether the competing systems reach a minimum level of usability and can continue in the process. They can also be used to compare the competing systems. Our analysis in this paper does not include comparison of usability grades between different scenarios. Thus our findings in this paper do not indicate that the usability grade produced is universal or could be compared between domains.

The grade boundaries are not universal between procurement cases and should be determined uniquely for each case. However, the boundaries presented in this paper can be used for reference. We recommend that the currently used IT system is evaluated with a couple of the user scenarios and similar procedure, which should not take more than a day of effort. If available, comparing these scores with other existing usability evaluation results can help in determining suitable grade boundaries before HED is used in procurement.

Based on our experiences, key factors in the success of applying the method were familiarization with the subject domain and the method through training and collaboration with subject domain experts throughout the process. However, usability specialists do not need extensive training for HED beyond their expertise in traditional HE; we estimate from two to four hours of practicing. The user scenarios form the basis of the evaluation and thus it is important that developing them is given enough consideration; this should be the responsibility of subject domain experts. For evaluating usability in procurement context, user scenario definition is not solely required for execution of HED, but also for using user questionnaires. They are in fact recommended as a basis for any evaluation [31].

Limitations of the Study

Our study does also have limitations. The validation of the method was done during one, although large, IT procurement case with two usability specialists (MT and JK) who have developed the method together with a medical doctor (TL) specialized in health information systems. The generalizability of the method and the results should be further studied in other procurement cases in healthcare and social welfare as well as in other domains and by other usability specialists. However, the data the analysis is based on is fairly extensive with 24 demonstrations, 48 individual evaluations and 144 hours of evaluation using HED by two evaluators. Our analysis in this paper has concentrated solely on the resulting grades from individual evaluations. For greater validation of results the total points given to usability issues of different severity by different evaluators should be considered. Also, in this paper the comparison between different methods has been based mostly on the same user scenarios and not on actual use of the system.

Future Research

Future research will analyze the properties and results produced by the HED method more thoroughly in relation to its validity, reliability and thoroughness, including IRR and IRA calculations based on points (see Figure 1) instead of grades. The results of HED will also be compared with usability testing done later in the procurement including more details on the level of usability between systems. Also, more research is needed to analyze the questionnaire results, and the questionnaire method will be reported in detail in further publications. When comparing the results of the HED method to perceived usability questionnaires it seems that domain experts overall give slightly more positive evaluations and in some cases even significantly more positive than usability specialists. These aspects should be examined more thoroughly. The choice of comprising the severity rating as the product of impact and frequency as well as the choice for the rating points of missing functionality worked well based on our results from the case study. However, more experiences from the choice of these metrics on other domains should be gathered. The authors' aim is to apply the method in future procurement cases to develop and validate it further.

CONCLUSION

This paper introduces the development of a novel usability evaluation method HED (heuristic evaluation during demonstrations) for public IT system procurement. HED can be used to compare complex IT systems and determine their level of usability. For this purpose, the traditional heuristic evaluation method was modified to produce quantifiable results, and to include the context of use by using user scenarios and demonstrations. The HED method requires preparations before the evaluation, such as determining the scoring, and two actual evaluation steps: analysis during demonstrations and aggregation of results to determine the usability grade. The scoring is based on identifying usability issues and giving negative points to heuristic violations, missing functionalities and omitted parts of the scenario and positive points to positive findings. The method has been applied and validated during a major public IT system procurement project with 24 demonstration sessions. Experiences from these have shown the method's ability to evaluate the level of usability of an IT system correctly, comprehensively and efficiently.

As a contribution to previous research, HED overcomes three recognized drawbacks [9, 38] of traditional heuristic evaluation (HE): Firstly, it includes the use context with user scenario based demonstrations. Secondly, this substantially expands the number of evaluated user interfaces. Thirdly, HED efficiently introduces the finding of missing functionality as part of HE. Additionally, it also enables the quantifying of HE results.

Our method is of utmost relevance for practice. Using HED during procurement includes several advantages for the procuring organization. The advantage of conducting expert review during user scenario based demonstrations is that they reflect the features and activities of actual context of use and can also depict a long work process including the viewpoints of different professionals. Using the HED method requires fewer resources than widely known and applied usability evaluation methods (e.g. traditional heuristic evaluation and usability testing) to cover the same number of user interface screens and work tasks.

One challenge of procurement is getting access to a working IT system for the procuring organization; with HED it is possible for the vendors to demonstrate their own system and for the purchasers to evaluate usability of the systems candidates. Although HED has been developed in the context of procuring a client and patient information system for healthcare and social welfare, our experiences indicate that the method could be equally applicable to other IT system domains.

ACKNOWLEDGEMENTS

The reporting of these results has been supported by the Finnish Work Environment Fund (grant #114160). Finally, the authors would like to thank all anonymous reviewers of this paper for their valuable comments.

REFERENCES

1. Benoit A. Aubert, Michel Patry, and Suzanne Rivard. 1998. Assessing the risk of IT outsourcing. In *Proceedings of the Thirty-First Hawaii International Conference on System Sciences*, 685-692.

2. J.M. Christian Bastien, and Dominique L. Scapin. 1995. Evaluating a user interface with ergonomic criteria. *Int J Human-Computer Interaction* 7, 2: 105-121. http://dx.doi.org/10.1080/10447319509526114

3. J. Brender. 1997. Medical informatics: does the health-care domain have special features? Letter to the editor. *Methods of Information in Medicine* 36, 1:59–60.

4. John Brooke. 1996. SUS: a “quick and dirty” usability scale. In *Usability Evaluation in Industry*, Patrick W. Jordan, Bruce Thomas, Bernard A. Weerdmeester, and Ian L. McClelland (eds.). Taylor and Francis, London, UK.

5. Thomas T. Carey. 1991. A Usability Requirements Model for Procurement Life Cycles. In *Human Factors in Information Systems: An Organizational Perspective*, Jane. M Carey (ed.). Ablex, Norwood, NJ, 89-104.

6. John M. Carroll. 1997. Scenario based Design. In *Handbook of Human-Computer Interaction*, Martin G. Helander, Thomas K. Landauer and Prasad V. Prabhu (eds.). Elsevier, Amsterdam, NL, 383-406.

7. C.J. Carvalho, Elizabeth M. Borycki, and Andre Kushniruk. 2009. Ensuring the Safety of Health Information Systems: Using Heuristics for Patient Safety. *Healthcare Quarterly* 12: 49-54.

8. Parmit K. Chilana, Jacob O. Wobbrock, and Andrew J. Ko. 2010. Understanding usability practices in complex domains. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '10), 2337-2346. http://dx.doi.org/10.1145/1753326.1753678

9. John P. Chin, Virginia A. Diehl, and Kent L. Norman. 1988. Development of an instrument measuring user satisfaction of the human–computer interface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI ’88), 213–218. http://dx.doi.org/10.1145/57167.57203

10. Gilbert Cockton, and Alan Woolrych. 2002. Sale must end: should discount methods be cleared off HCI's shelves? *Interactions* 9, 5: 13-18. http://dx.doi.org/10.1145/566981.566990

11. Kathrin Cresswell, Zoe Morrison, Sarah Crowe, Ann Robertson, and Aziz Sheikh. 2011. Anything but engaged: user involvement in the context of a national electronic health record implementation. *Informatics in Primary Care* 19: 191-206.

12. Directive 2004/18/EC of the European Parliament and of the Council of 31 March 2004 on the coordination of procedures for the award of public works contracts, public supply contracts and public service contracts. Retrieved September 25, 2015 from http://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:32004L0018

13. Joseph S. Dumas, and Janice C. Redish. 1999. *A Practical Guide to Usability Testing*. Intellect.

14. Joseph S. Dumas, and Marilyn C. Salzman. 2006. Usability Assessment Methods. *Reviews of Human Factors and Ergonomics* 2, 1: 109-140. http://dx.doi.org/10.1177/1557234X0600200105

15. Paul J. Edwards, Kevin P. Moloney, Julie A. Jacko, and François Sainfort, F. 2008. Evaluating Usability of a Commercial Electronic Health Record: A Case Study. *International Journal of Human-Computer Studies* 66, 10: 718-728. http://dx.doi.org/10.1016/j.ijhcs.2008.06.002

16. Laura H. Einbinder, Judy B. Remz, and David Cochran. 1996. Mapping Clinical Scenarios to Functional Requirements: A Tool for Evaluating Clinical Information Systems. In *Proceedings of AMIA Annual Fall Symposium*, 747-751.

17. Finnish Act on Public Contracts (348/2007). Retrieved September 25, 2015 from http://www.finlex.fi/en/laki/kaannokset/2007/en20070348.pdf

18. Günther Gediga, Kai-Christoph Hamborg, and Ivo Düntsch. 1999. The IsoMetrics usability inventory: An operationalization of ISO 9241-10 supporting summative and formative evaluation of software systems. *Behaviour & Information Technology* 18, 3: 151-164. http://dx.doi.org/10.1080/014492999119057

19. Natasa Gisev, J. Simon Bell, and Timothy F. Chen. 2013. Interrater agreement and interrater reliability: Key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy* 9, 3: 330-338. http://dx.doi.org/10.1016/j.sapharm.2012.04.004

20. Marta González, Llúcia Masip, Antoni Granollers, and Marta Oliva. 2009. Quantitative Analysis in a Heuristic Evaluation Experiment. *Advances in Engineering Software* 4, 12: 1271-1278. http://dx.doi.org/10.1016/j.advengsoft.2009.01.027

21. Jan Gulliksen, Bengt Göransson. 2001. Reengineering the Systems Development Process for User-Centred Design. In *Human-Computer Interaction INTERACT ’01*, M. Hirose (ed.). IOS Press, Amsterdam, NL, 359-366.

22. H. Rex Hartson, Terence S. Andre, and Robert C. Williges. 2003. Criteria For Evaluating Usability Evaluation Methods. *Int J Human-Computer Interaction* 15, 1: 145–81. http://dx.doi.org/10.1207/S15327590IJHC1501_13

23. HIMSS EHR Usability Task Force. 2009. Defining and Testing EMR Usability: Principles and Proposed Methods of EMR Usability Evaluation and Rating, Healthcare Information and Management Systems Society. Retrieved September 25, 2015 from http://s3.amazonaws.com/rdcms-

himss/files/production/public/FileDownloads/HIMSS_DefiningandTestingEMRUsability.pdf

24. Kasper Hornbæk. 2006. Current Practice in Measuring Usability: Challenges to Usability Studies and Research. *Int J Human-Computer Studies* 64, 2: 79-102. http://dx.doi.org/10.1016/j.ijhcs.2005.06.002

25. Debra Howcroft, and Ben Light. 2002. A Study of User Involvement in Packaged Software Selection. In *Proceedings of ICIS 2002*, paper 7. http://aisel.aisnet.org/icis2002/7

26. Saila Huuskonen. 2014. Recording and use of information in a client information system in child protection work. Acta Electronica Universitatis Tamperensis: 1387. Tampere University Press. PhD thesis. Retrieved from http://urn.fi/URN:ISBN:978-951-44-9368-3

27. Monique W.M. Jaspers. 2009. A Comparison of Usability Methods for Testing Interactive Health Technologies: Methodological Aspects and Empirical evidence. *Int J Medical Informatics* 78, 5: 340-353. http://dx.doi.org/10.1016/j.ijmedinf.2008.10.002

28. Timo Jokela, and Elizabeth Buie. 2012. Getting UX into the Contract. In *Usability in Government Systems: User Experience Design for Citizens and Public Servants*, Elizabeth Buie, and Dianne Murray (eds.). Morgan Kaufmann, Waltham, MA, 251-165.

29. Timo Jokela, Juha Laine, and Marko Nieminen. 2013. Usability in RFP's: The Current Practice and Outline for the Future. In *Proceeding of HCII 2013, Part II. LNCS, vol 8005*, M. Kurosu (ed.). Springer, Berlin Heidelberg, 101-106. http://dx.doi.org/10.1007/978-3-642-39262-7_12

30. Jurek Kirakowski. 1994. The use of questionnaire methods for usability assessment. Retrieved September 25, 2015 from http://sumi.ucc.ie/sumipapp.html

31. Andre Kushniruk, Marie-Catherine Beuscart-Zéphir, Alexis Grzes, Elizabeth Borycki, Ludivine Watbled, and Joseph Kannry. 2010. Increasing the Safety of Healthcare Information Systems through Improved Procurement: Toward a Framework for Selection of Safe Healthcare Systems. *Healthcare Quarterly* 13: 53-58. http://dx.doi.org/10.12927/hcq.2010.21967

32. Andre W. Kushniruk, and Vimla L. Patel. 2004. Cognitive and Usability Engineering Methods for the Evaluation of Clinical Information Systems. *J Biomedical Informatics* 37, 1: 56-76. http://dx.doi.org/10.1016/j.jbi.2004.01.003

33. Erik Liljegren, and Anna-Lisa Osvalder. 2004. Cognitive Engineering Methods as Usability Evaluation Tools for Medical Equipment. *Int J Industrial Ergonomics* 34, 1: 49-62. http://dx.doi.org/10.1016/j.ergon.2004.01.008

34. Thomas L. Lincoln, and Daniel J. Essin. 1993. Clinical Information Processing Scenarios: Selecting Clinical Information Systems. In *Evaluating Healthcare Information Systems: Methods and Applications*, James G. Anderson, Carolyn E. Aydin, and Stephen J. Jay, (eds.). Sage Publications, Thousand Oaks, CA.

35. Barbara Mirel. 2004. *Interaction Design for Complex Problem Solving – Developing Useful and Usable Software*, Morgan Kaufmann.

36. Carl E. Moe. 2014. Research on Public Procurement of Information Systems: The Need for a Process Approach. *Comm of the Association for Information Systems* 34: 1320-1335. http://aisel.aisnet.org/cais/vol34/iss1/78

37. John Møller-Jensen, Ivan Lund Pedersen. and Jesper Simonsen. 2006. Measurement of the Clinical Usability of a Configurable EHR. In *Ubiquity: Technologies for Better Health in Aging Societies*, Arie Hasman, Reinhold Haux, Johan van der Lei, Etienne De Clercq, Francis H.Roger France (eds.). IOS Press, Amsterdam, NL, 356-361.

38. Jakob Nielsen. 1993. *Usability Engineering*. Morgan Kaufmann.

39. Jakob Nielsen, Rolf Molich. 1990. Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI ’90), 249-256. http://dx.doi.org/10.1145/97243.97281

40. Linda W.P. Peute, and Monique W.M. Jaspers. 2007. The Significance of a Usability Evaluation of an Emerging Laboratory Order Entry System. *Int J Medical Informatics* 76, 2-3: 157-168. http://dx.doi.org/10.1016/j.ijmedinf.2006.06.003

41. Linda W.P. Peute, Richard Spithoven, Piet J.M. Bakker, and Monique W.M. Jaspers. 2008. Usability Studies on Interactive Health Information Systems: Where Do we Stand? In *eHealth Beyond the Horizon - Get IT There: Proceedings of MIE 2008*, Stig Kjær Andersen, Gunnar O. Klein, Stefan Schulz, Jos Aarts, M. Cristina Mazzoleni (eds.). IOS Press, Amsterdam, NL, 327-332.

42. Ginny Redish. 2007. Expanding Usability Testing to Evaluate Complex Systems. *J Usability Studies* 2, 3: 102–111.

43. Sirpa Riihiaho, Marko Nieminen, Stina Westman, Ronja Addams-Moring, and Jukka Katainen. 2015. Procuring Usability: Experiences of Usability Testing in Tender Evaluation. In *Nordic Contributions in IS Research, vol. 223 LNBIP*, Harri Oinas-Kukkonen, Netta Iivari, Kari Kuutti, Anssi Öörni, Mikko Rajanen (eds.). Springer, CH, 108-120. http://dx.doi.org/10.1007/978-3-319-21783-3_8

44. Jeff Sauro. 2014. The Relationship Between Problem Frequency and Problem Severity in Usability Evaluations. *J Usability Studies* 10, 1: 17-25.

45. Robert M. Schumacher, Jayson M. Webb, and Korey R. Johnson. 2009. *How to Select an Electronic Health Record System that Healthcare Professionals can Use*. User centric Inc. Retrieved September 25, 2015 from http://www.usercentric.com/sites/usercentric.com/files/usercentric-ehr-white-paper.pdf

46. Philip Scown. 1998. Improving the Procurement Process: Humanizing Accountants with a Human Factors Education. In *Proceedings of ICIS 1998*, paper 2. http://aisel.aisnet.org/icis1998/2

47. Stephen M. Shortell, Arnold D. Kaluzny. 2006. *Health care management: Organizational design and behavior*, (5th ed.). Thompson.

48. Alistair Sutcliffe. 2002. Assessing the reliability of heuristic evaluation for Web site attractiveness and usability. In *HICSS Proceedings of the 35th Annual Hawaii International Conference on System Sciences 2002*, 1838-1847. http://dx.doi.org/10.1109/HICSS.2002.994098

49. Kimmo Tarkkanen, Ville Harkke. 2015. Evaluation for Evaluation: Usability Work during Tendering Proces. In *Extended Abstracts of the SIGCHI Conference on Human Factors in Computing Systems* (CHI’15), 2289-2294. http://dx.doi.org/10.1145/2702613.2732851

50. Johanna Viitanen, Hannele Hyppönen, Tinja Lääveri, Jukka Vänskä, Jarmo Reponen, Ilkka Winblad. 2010. National Questionnaire Study on Clinical ICT Systems Proofs: Physicians Suffer from Poor Usability. *Int J Medical Informatics* 80, 10: 708-725. http://dx.doi.org/10.1016/j.ijmedinf.2011.06.010

51. Jiajie Zhang, and Muhammad F. Walji. 2011. TURF: toward a unified framework of EHR usability. *J Biomedical Informatics* 44, 6: 1056–67. http://dx.doi.org/10.1016/j.jbi.2011.08.005