UXtract – Extraction of Usability Test Results for Scoring Healthcare IT Systems in Procurement

Janne Pitkänen[a, b], Marko Nieminen[a], Matti Pitkäranta[b], Johanna Kaipio[a], Mari Tyllinen[a, c], Antti K. Haapala[d]

[a]*Department of Computer Science, Aalto University, Finland*
[b]*Adusso Ltd., Helsinki, Finland*
[c]*Oy Apotti Ab, Helsinki, Finland*
[d]*AnttiPatterns, Oulu, Finland*

Abstract

”In healthcare IT system procurement we always need to choose the cheapest one.” Do we? In this paper we present a method and a procedure for effective extraction of usability test results for public procurement. Successful procurement necessitates the alternative products to be compared considering their realistic utility. We can significantly contribute to this comparison by measuring usability in a practical way. Our UXtract method enables the extraction of detailed, traceable and commensurate findings for objective evidence. The method extracts structured data straight from the test. Our case in large scale healthcare settings shows that this method is efficient for scoring usability in procurement. We elaborate the results and discuss about the impact and challenges of comparison testing when using it for decision making of multimillion investments in information technology.

Keywords:

Healthcare information system; usability testing; summative evaluation; comparison; public procurement.

Introduction

Usability testing [7] is traditionally conducted in a qualitative manner. Despite it being an effective method in formative settings (possibilities to change the system under evaluation), its applicability in summative settings (comparing large-scale systems in a selection process) is challenging [9]. The challenges relate especially to demanding and laborious analysis of the qualitative data which constrain the scalability of the method. Nielsen [7] presents that the amount of users in a usability test does not have to exceed 6 persons. With decent amount of tasks in the test, the amount of data to be analyzed remains reasonable. However, in situations that require and would benefit from several user groups or broad variety of tasks, the applicability of the method decreases: How to increase the number of usability tests from 5 to 50 without increasing the effort and resources for analyzing the results? These types of situations appear in the procurement of large IT systems that affect large numbers of people in multiple tasks. For example Denmark, Finland and Canada have initiated some healthcare information system related projects in large regional scales to facilitate the improvement of the service quality and keep the costs of service at affordable levels [4].

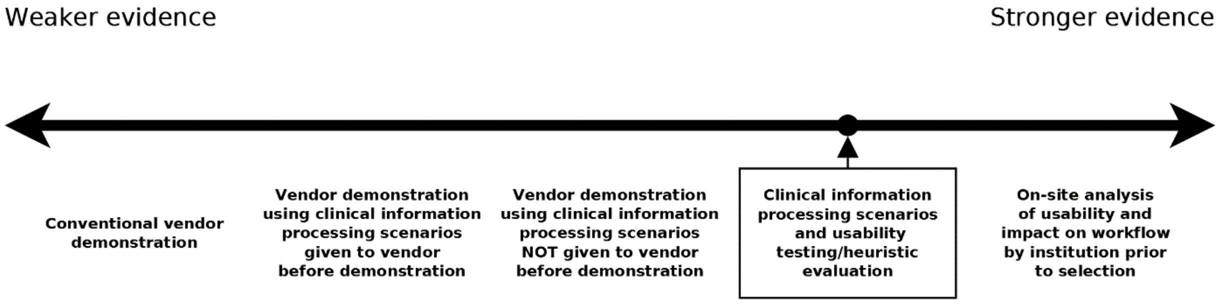

Weaker evidence
Stronger evidence
Conventional vendor demonstration
Vendor demonstration using clinical information processing scenarios given to vendor before demonstration
Vendor demonstration using clinical information processing scenarios NOT given to vendor before demonstration
Clinical information processing scenarios and usability testing/heuristic evaluation
On-site analysis of usability and impact on workflow by institution prior to selection

Figure 1 - Strength of evidence associated with usability testing and heuristic evaluation when applied by using realistic clinical information processing scenarios. Continuum of evidence, as introduced by Kushniruk et. al. [4], considers other possible methods to support system selection and their relative strength ranging from the weakest to the strongest one.

Government and public systems suffer from poor usability [3]. In EU public contracts are awarded to the lowest bidder or to the bidder with the economically most advantageous offer; the latter requiring that a scoring rule must be specified. A weighing of price and quality may be a good choice when there is uncertainty regarding what combinations of price and quality are achievable, while quality is not too difficult to measure and verify [1].

Despite these constraints, it is desirable to ensure high usability in advance. Sauro & Kindlund [11] present an attempt to create a single, standardized and summated usability metric for each task by averaging together the four standardized values based on the equal weighting of the coefficients from the Principal Components Analysis. Riihiaho & al. [10] have evaluated the economic value of choosing the better system in procurement by measuring efficiency (i.e. task completion) and comparing these percentages between the prospective systems. Kushniruk & al. [4] have presented a strength-of-evidence-on-usability continuum for healthcare settings (Figure 1). Clinical information processing scenarios can be used to test systems to determine whether they respond appropriately to the situations/scenarios described. In order to get stronger evidence on usability, the evaluation should be done in a way that positions at the right end of the continuum.

Usability testing appears at that part of the continuum making it the preferred method without a need to implement the prospective systems on site. Our UXtract method aims at solving the challenges on scalability and enabling usability testing in large IT procurement projects including scoring.

The UXtract Method and Technology

Our method constructs a practicable way of (i.) collecting structured data from a moderated usability test session and (ii.) extraction of usability test results aggregating the data from multiple test sessions into a single score of usability for each system under test.

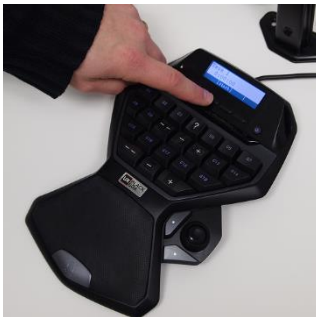
The types or usability metrics associated with the method include effectiveness, errorlesness and satisfaction. Other types of usability aspects such as learnability and accessibility are not specifically in the scope of this testing method, but can be considered by other means of system evaluation to be included in scoring schemes for procurement ranking.

According to National Institute of Standards and Technology, at least two testers are needed to conduct the sessions. These two testers are: 1. An expert/test administrator who facilitates the testing and is in charge of interacting with the participant during the test session. 2. An assistant/data logger who is responsible for all aspects of data collection. The data logging role can be fulfilled with data capture software where appropriate; however, a two-person test team is the minimum recommendation [6].

The following details are possible to betraced during the test sessions:

- task duration (hh:mm:ss)
- task success (pass or fail)
- moderator marking of events
  - major negative issue (--)
  - minor negative issue (-)
  - generic positive issue (+)
  - issue for further analysis (?)
- feedback via user buttons
  - task satisfaction (good or bad)
  - emergent issue (good or bad)

The test moderator is provided with a tracing pad shown in figure 2, which allows making marks of specific types to keep track of task durations, success rate and marking of issues. Each user is provided with a user console shown in figure 2, which allows giving feedback with two buttons during test sessions.



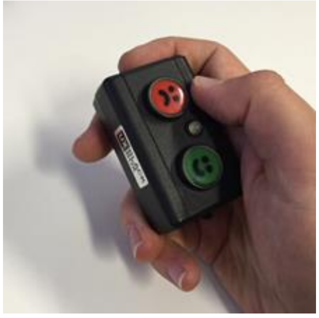


Figure 2 - Left: A commercial game control pad is used as the moderator pad, which is configured to keep track of test session status on display and to provide buttons for making marks during test sessions.

Figure 3 - Right: A wireless user console with two buttons is constructed for providing user feedback.

Human-computer interaction (HCI) and spoken communication is recorded during test sessions. For regular workstation environment this involves recording of display, keyboard and mouse activity for HCI, while a microphone is connected to the same recording system to capture speaking.

Extraction of Usability Test Results

Commensurate usability scores for evaluated systems are extracted from the structured data, which is produced by task tracing and satisfaction monitoring. Recordings from the test sessions provide a possibility to review any unclear events or judgements, which might remain after the testing. Otherwise the recordings are kept just for an objective evidence to make the tests traceable:

1. Task duration, success information, number of each type of issue and feedback collection is produced per task and test session.
2. Quantification of chosen measures.

3. Averaging the results over session repetitions per tested system.
4. Considering significance of difference per measure between tested systems (especially in case of discrete scales of quantification).
5. Calculating weighted sum of measures per system according to chosen scoring scheme and relative weightings to form an overall score for comparison.

Evaluation of the UXtract Method

In order to assess the performance of the UXtract method we conducted tests with seven representative scenarios of three domain areas (C=clinical, S=social and P=patient) and representative user groups (nurses, physicians, social workers and citizens) presented in table 1. Users for test participation were chosen from the actual user groups associated with each scenario: nurses and physicians as users for three scenarios in clinical work domain, social workers as users for two scenarios in social welfare domain, and citizens as users for one scenario in patient portal domain. Test users representing their profession as nurses, physicians and social workers were involved in test participation in pairs. Pair testing is known as the constructive interaction method, where two subjects are encouraged to experiment with the system under study [8]. Patient portal was a web based part of the system for self-service and thus expected to work for individual, first-time users.

Table 1 - Test scenarios addressed clinical work (C) and social work (S) related domain areas of the system with two users at a time (pair test), while patient portal (P) domain was tested in a traditional way with a single user. Each test scenario was repeated n times per system with different user(s).

Scenario	User Group	Users per Test	N per System
C1	nurses	2	3 tests
C2	nurses	2	3 tests
C3	physicians	2	3 tests
S1	social workers	2	3 tests
S2	social workers	2	3 tests
P	citizens	1	10 tests

Running 50 usability tests (table 2), which each included 10 to 19 tasks and take up to 90 minutes of active testing time, required two testing spots to be operated in parallel for the project to meet a given schedule. Two usability specialists (JK and MT) planned and moderated the testing, while test sessions were supported and data gathering maintained by a testing tool provider (JP and MP). The testing spots were located in two regular office rooms reserved for the purpose. Non-intrusive testing tools allowed the vendors to deliver their systems (combination of software and preferred computer hardware) for the tests as is. No additional software was needed to be installed for testing purposes to make sure not to compromise the overall performance of the systems in comparison.

Table 2 - Number of test sessions, task items per scenario and total number of tasks conducted within usability testing efforts. (Pilot tests not included.)

Scenario	Tests	Time [min]	Tasks Items	Task Totals
C1	6	90	14	84
C2	6	90	12	71
C3	6	90	19	114
S1	6	90	12	72
S2	6	90	10	60
P	20	90	13	260
Total	50	75 h	80	661

UXtract Results: Automatic Calculation of Usability

Extraction of test results is applied by summating chosen measures by reasonable weighting to represent overall usability for comparative purposes. This can be done with spreadsheet computation by importing the logged data from test sessions to a spreadsheet workbook, which is prepared to calculate the usability metric automatically.

Effectiveness is measured in all the test scenarios based on the percentage of successfully completed test tasks as follows:

$$\mathbf{Effectiveness} = \frac{\text{Succesfully completed tasks}}{\text{Total amount of tasks}}$$

For each scenario and task, there is a predefined maximum time of execution. Test moderator marks each test task either as passed or failed upon completion of the task or when the maximum time is exceeded. In case of session time runs out, the remaining tasks are considered as failed.

Errorlessness is evaluated in all the test scenarios based on the number of errors during successfully executed test tasks. An error is defined here to be a deviation from a reasonable task execution path (non-productive activity considering the goal of the task, e.g. transition to wrong view, unintentional activity, mistake or ignorance of substantial information). The errors during test execution were classified as minor (½ pts.) and major (1 pts.) ones, which were marked up in real time on the tracing pad. The error points are averaged over test tasks and repeated scenarios for each system. A session with none of the tasks succeeded gives a default of 12 error points as an average. *Errorlessness* is quantified here based on these error points on a scale from 5 to 0 (*where the highest score is achieved with the least amounts of errors*) as presented in table 3.

Table 3 - Quantifying of errorlessness based on error point averages for each test scenario.

Error-lesness	5	4	3	2	1	0
Error points	[0,1]	]1,2]	]2,4]	]4,6]	]6,∞[	12

Satisfaction is evaluated in all the scenarios based on the users' positive and negative feedback collected upon completion of each test task. In addition to this, the test scenarios conducted in pairs involved feedback collection also during the test tasks with user initiated positive and negative experiences of use (e.g. subjectively positive event or personal satisfaction and negative struggle, inconvenience or dissatisfaction towards the behavior of the system along execution of a test task). Only successfully completed test tasks count for this.

Table 4 - Quantifying of satisfaction based on proportion of the tasks evaluated with more positive than negative feedback on average.

Satisfaction	5	4	3	2	1
Rate	$p > 80\%$	$80\% \geq p > 60\%$	$60\% \geq p > 40\%$	$40\% \geq p > 20\%$	$p \leq 20\%$

Assessment of the results

An IT system under usability testing was designed to support three different domain areas which can be considered being partly separated from each other in terms of functionalities and related subsystem implementations. Usability testing based evaluation produced a consistent differentiation between the compared systems for each domain area, since all the usability measures indicated the same order for each domain with good correlations presented in table 5. However, there were differences between the domain areas indicating that the system X was 45% better in clinical use and 32% better in social work (C and S domain areas), while the system Y was 21% better in the patient service portal (P domain).

Table 5 - Correlation (Corr.) of the effectiveness, errorlessness (Err∑) and satisfaction (Sat∑) measure comparison between the systems X and Y. Correlations calculated with zeros (0/0).

Domain	Effectiveness (X/Y)	Err∑ (X/Y)	Sat∑ (X/Y)	Corr.
C	46% / 75%	10 / 12	8 / 13	.9983
S	67% / 75%	5 / 7	6 / 9	.9996
P	97% / 85%	5 / 4	5 / 4	.9999

By default, there was no need to go through the recordings afterwards for extraction of these results. However, moderators checked and reviewed some situations from the recordings right after a test session, whether they felt that anything would have remained unclear. There was less than 10 situations in total, which needed review and/or correction (e.g. accidental wrong task marking or open issue related to interpretation of an error).

Discussion and Conclusions

Our experience in using the UXtract method demonstrated that it is an efficient way in public procurement for conducting comprehensive usability testing of a large IT system that is being used by large number of people in a large number of tasks.

The real-time recording of usability issues/markers appeared feasible for the test moderators. However, further development of observation guidelines for marking would even improve the task by automating the generation of structured and readily available results from the tests.

The results in table 5 show that the defined components of usability (effectiveness, errorlessness and satisfaction) correlate strongly. This suggests that weighing of the components in scoring appears not critical in this case, because all weighing combinations would result in similar ordering of the compared systems in each domain. A dedicated and pre-defined weighing scheme (including components of usability and domains) was used for the actual procurement scoring[1]. Usability testing contributed 74,4% of the usability comparison criteria, which further contributed 20% of the overall quality criteria for the procurement. In addition to these, price-to-quality consideration resulted to final scores of 89,76 vs. 92,23. A contract was awarded to the system Y vendor with a 65 million euros higher bid price (385M€) compared to system X (320M€).

Based on our promising experience, we will apply the method in similar procurement cases to gather more data for developing more elaborate models for procurement scoring.

Acknowledgments

The authors wish to thank the project office that provided the possibility to test the method in a representative case.

References

[1] Mats A. Bergman, Sofia Lundberg, Tender evaluation and supplier selection methods in public procurement, Journal of Purchasing and Supply Management, Volume 19, Issue 2, June 2013, Pages 73-83.

[2] HIMSS EHR Usability Task Force. Selecting an HER for Your Practice: Evaluating Usability. Apr. 2011; http://s3.amazonaws.com/rdcms-himss/files/production/public/HIMSSorg/Content/files/HIMSS%20Guide%20to%20Usability_Selecting%20an%20EMR.pdf

[3] Jokela T, Laine J & Nieminen M (2013). Usability in RFP's: The Current Practice and Outline for the Future. In Kurosu, M. (Ed.) Human-Computer Interaction. Applications and Services, Springer Berlin Heidelberg, 2013, 8005, 101-106

[4] Kushniruk A, Beuscart-Zéphir MC, Grzes A, Borycki E, Watbled L and Kannry J. Increasing the safety of healthcare information systems through improved procurement: toward a framework for selection of safe healthcare systems. Healthcare quarterly (Toronto, Ont.), 13 (2010), 53-58.

[1] For details, see Apotti, Justification memo attachment 1, Product comparison B results (in Finnish). http://kirkkonummi01.hosting.documenta.fi/kokous/20152212-3-2.PDF

[5] Kushniruk A, Kaipio J, Nieminen M, Hyppönen H, Lääveri T, Nøhr C, Kanstrup AN, Christiansen MB, Kuo M-H, Borycki EM (2014). Human Factors in the Large: Experiences from Denmark, Finland and Canada in Moving Towards Regional and National Evaluations of Health Information System Usability. Contribution of the IMIA Human Factors Working Group. Yearbook of medical informatics 01/2014; 9(1): 67-81.

[6] Lowry S. et al. Technical evaluation, testing and validation of the usability of electronic health records (NIST IR 7804). (Feb. 2012); http://www.nist.gov/healthcare/usability/upload/EUP_WERB_Version_2_23_12-Final-2.pdf

[7] Nielsen J (1993) Usability Engineering. Academic Press, Boston, USA.

[8] O'Malley CE, Draper SW and Riley MS (1984). Constructive interaction: A method for studying human-computer-human interaction. In Shackel, B. (Ed.) Human-computer interaction – INTERACT'84. pp. 269-274.

[9] Redish JG, Bias RG, Bailey R, Molich R, Dumas J and Spool JM (2002). Usability in practice: formative usability evaluations - evolution and revolution. In CHI '02 Extended Abstracts on Human Factors in Computing Systems (CHI EA '02). ACM, New York, NY, USA, 885-890.

[10] Riihiaho S, Nieminen M, Westman S, Addams-Moring R, Katainen J (2015): Procuring Usability: Experiences from Usability Testing in Tender Evaluation. In H. Oinas-Kukkonen et al. (Eds.): Nordic Contributions in IS Research. Proceedings of the 6th Scandinavian Conference on Information Systems, SCIS 2015, Oulu, Finland, August 9-12, 2015. Lecture Notes in Business Information Processing Volume 223, 2015, pp 108-120.

[11] Sauro J and Kindlund J 2005. A method to standardize usability metrics into a single score. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '05). ACM, New York, NY, USA, 401-409.

Address for correspondence

Janne Pitkänen, +358 50 4014975, janne.pitkanen@adusso.com
Adusso Ltd., Kuortaneenkatu 2, FI-00510 Helsinki, FINLAND