Work 41 (2012) 2208-2215
DOI: 10.3233/WOR-2012-0443-2208
IOS Press

Evaluating the iterative development of VR/AR human factors tools for manual work

Paul M. Liston[a*], Alison Kay[a], Sam Cromie[a], Chiara Leva[a], Mirabelle D'Cruz[b], Harshada Patel[b], Alyson Langley[c], Sarah Sharples[c], and Susanna Aromaa[c]
[a] *School of Psychology, Trinity College, University of Dublin, College Green, Dublin 2, Ireland*
[b] *HFRG, Faculty of Engineering, University of Nottingham, Nottingham, NG7 2RD, England*
[c] *VTT Technical Research Centre of Finland, Tampere, FI-33101, Finland*

Abstract. This paper outlines the approach taken to iteratively evaluate a set of VR/AR (virtual reality / augmented reality) applications for five different manual-work applications – terrestrial spacecraft assembly, assembly-line design, remote maintenance of trains, maintenance of nuclear reactors, and large-machine assembly process design – and examines the evaluation data for evidence of the effectiveness of the evaluation framework as well as the benefits to the development process of feedback from iterative evaluation. ManuVAR is an EU-funded research project that is working to develop an innovative technology platform and a framework to support high-value, high-knowledge manual work throughout the product lifecycle. The results of this study demonstrate the iterative improvements reached throughout the design cycles, observable through the trending of the quantitative results from three successive trials of the applications and the investigation of the qualitative interview findings. The paper discusses the limitations of evaluation in complex, multi-disciplinary development projects and finds evidence of the effectiveness of the use of the particular set of complementary evaluation methods incorporating a common inquiry structure used for the evaluation – particularly in facilitating triangulation of the data.

Keywords: evaluation, iterative development, virtual reality, augmented reality, manual work

* Corresponding author. E-mail: pliston@tcd.ie

1. Introduction

Manual work is a crucial and expensive component of manufacturing. With globalisation causing companies to reduce manual work costs by offshoring and global outsourcing there has been an increase in associated problems such as longer lead times, lower quality products and services, and weaker management. The pressure of globalisation and the decline of the working-age population in Europe suggest that the situation will worsen in sectors characterised by high-value manual work that cannot be automated or outsourced. The ManuVAR project is an EU-funded research project focusing on improving manual work through the development of VR/AR human factors tools. The project, nearing completion, has five industrial areas of application which, considered collectively, cover both the entire system lifecycle and the full spectrum of manual work – terrestrial spacecraft assembly, assembly-line design, remote maintenance of trains, maintenance of nuclear reactors, and large-machine assembly process design. The development of the applications is part of a broader participatory action research approach in which the ManuVAR project will support implementation of the applications in order to harness the power of VR/AR technology and improve the way manual work is performed in each organisation.

At the very outset of the project, human factors (HF) experts consulted with stakeholders in the five end-users in the project consortium – as suggested by exponents of participatory ergonomics [13] - and elicited a set of seven gaps related to manual work: (1) lack of communication through the lifecycle, (2) poor interfaces, (3) inflexible design process, (4) inefficient knowledge management, (5) low productivity, (6) lack of technology acceptance and (7) physical and cognitive stresses [14]. These seven gaps correspond to the needs of the end-user organisations in supporting high-value manual work and the project set about developing four core VR/AR applications and a technological platform to support them. The four VR/AR applications were to be used to: (1) provide real-time on-site support of integration/assembly and/or maintenance, (2) allow the application of ergonomic analysis in workplace design, (3) support task planning and analysis, and (4) support the training of users. Each of these applications had a specific 'Cluster' – a sub-set of project partners involved in the development of this application. Each cluster incorporated research partners, technology partners and the end-user organisation for whom the application was being developed. Insofar as possible clusters were also grouped geographically.

1.1. The development approach – iterative, agile, participative

The development of the ManuVAR application tools was both iterative – using an agile process of development; and participative – involving many key stakeholders. Agile development has been defined as: "A manufacturing system with capabilities (hard and soft technologies, human resources, educated management, information) to meet the rapidly changing needs of the marketplace (speed, flexibility, customers, competitors, suppliers, infrastructure, responsiveness)" [1]. Because of the relatively short duration of this project (3 years) and the broad scope of the research initiative it was decided that an agile development approach – much favoured in software development – would provide the flexibility and speed necessary.

The purpose of involving stakeholders and users from the outset of the development process is to increase the acceptance of the product and ensure the technology is put to the service of the user and his/her needs. By employing participatory design [2] and User Centred Design/Human Centred Design (UCD/HCD) approaches [3] it is easy to take account user needs. Several studies have examined various aspects of the integration of agile methods and UCD/HCD [4,5] and it has been demonstrated that the majority of practitioners perceive the integration of agile methods with UCD to have added value to their adopted processes and to have increased end-user satisfaction with the product developed [6].

In EU research projects, it is quite unusual to use all end-user organisations in the consortium as a network and work by the rules of agile and UCD. This presented challenges for the evaluation of the applications. The evaluation strategy had to be designed within the limitations of four short *design-develop-evaluate* cycles and taking account of the different perspectives and priorities of the respective end-user organisations. Also as these cycles progressed within the formative design phase of the solution, the tools had to be sufficiently robust, yet flexible, to provide meaningful feedback to the development team for the next development cycle.

1.1.1. The design-develop-cycle

Design – This involved specifying the end-user requirements and mapping out the technological architecture of the technological platform which would host the four applications.

Develop – This was the phase where the technology partners worked on operationalising the end-user requirements.

Evaluate – At this phase in the cycle the technology is tested with real users (in ManuVAR terminology, a 'Trial') and evaluated according to the evaluation framework. This is the phase in the development cycle upon which this paper is focused.

Each *design-develop-evaluate* cycle culminated in a trial of the application tool, with a feedback loop from the evaluation results back to the design of the next iteration. There were four trials in the relatively-short development phase of the project – a period of 9 months. As development progressed the aim of each trial changed from a focus on the technical elements, and their combination, to the functional elements and their support of the specific task and organisational need. As is clear from Table 1 the trials also took place in various European locations (determined by the location of technology or end-user partners in the project consortium).

Table 1
Overview of the trial locations and foci

	Location	Focus
Trial 0	Greece	· First robust technical elements · Testing that the technologies link up
Trial 1	Spain	· Testing available functions · Capturing initial user feedback
Trial 2	Finland	· Testing improvement on functionality · Testing new functionalities · Testing user and technology performance · Capturing user feedback
Trial 3	Finland	· Testing improvement on functionality · Testing new functionalities · Agreeing final functions · Testing user and technology performance · Capturing user feedback

2. The evaluation methods

To maintain a consistent approach to evaluation in the different trials, a set of 5 methods was used across all applications (i.e. questionnaires, observation, heuristic evaluation, interviews, and a sickness questionnaire). There were a number of constraints which had an impact on the methods chosen, how they were administered, and the analysis of results. These constraints and limitations are characteristic of all types of participatory action research [7] and mirror those contained in work in the aircraft maintenance field [8]. For example, access to the technology and end-users was limited to 90-120 minutes therefore multiple evaluators had to administer some methods concurrently; limited resources in end-user organisations (together with the constraints of the multiple Trial locations) meant that there was between one and three end-user participants per case study which restricted the analysis of the data. However, these are common issues faced when working with industrial partners – there is a delicate compromise to be reached between scientific rigour and the practical constraints of real world research which will be implemented in an end-user organisation. Despite these constraints, the researchers controlled as many parameters as possible to ensure a logical and consistent approach - the ultimate goal being to produce useful results to feed back into the each successive *design-develop-evaluate* cycle.

2.1. Questionnaire

Each of the four applications had a tailored questionnaire which involved an examination of common issues. That said, a number of core questions under six headings were included in all the questionnaires - setting up the task, performing the task, display of task progress, accessing and storing data, visualising the data, and general user experience – using a four-point Liker rating scale from extremely negative (1) to extremely positive (4). There are however disadvantages to using questionnaires, e.g. insufficient richness of data and responder bias, therefore a number of additional methods were used to overcome the shortcomings of any one particular method.

2.2. Observation

Evaluators carried out a structured observation of participants interacting with the technology in each case study in order to: gather information about how participants performed a task, identify general and specific usability issues, as well as monitor their behaviour (positive and potentially negative), and observe the types of postures adopted. Evaluators had a specific list of categories and tasks being carried out (which matched the six headings used in the questionnaires) and used these to note relevant behaviours exhibited by participants. Evaluators also used a four-point Likert rating scale from extremely negative (1) to extremely positive (4). Observing how

participants interact with a system can highlight usability problems, areas in which training may improve participant performance and their VR/AR experience.

2.3. Interview

Post-task interviews were conducted with participants, focusing on any negative responses given in the questionnaires (as this was particularly useful in feeding into re-design). Semi-structured questions were also used to probe deeper into issues such as acceptability, likeability, potential utility within companies and cost-benefit. A core set of questions was used across all the case studies, with additional questions to address issues specific to a particular case study.

2.4. Other methods

An adapted version of an *expert heuristic evaluation tool* (VIEW-IT) was used [9]. The *Simulator Sickness Questionnaire* [10,11] was also used to capture any symptoms participants may have been experiencing both pre- and post-task. These methods are not part of this paper.

3. Rationale for the combination of evaluation methods

A well-structured and consistent approach to the iterative cycles of design-develop-evaluate was necessary so that all involved (researchers and participants) were aware of their roles and responsibilities, and to maximize the feedback captured. As the focus of each successive trial developed and expanded (see Table 1) so too did the evaluation framework – emerging as it did through the development cycles as lessons were learnt and incorporated into the design of the next evaluation exercise.

Trial 1 was used as much to pilot the IT framework, as well as to gather data. This was also the first time that all the technological elements had been presented. Following Trial 1 the evaluation team had a better understanding of the type and nature of the elements to be assessed, types of participants available and required, and the tools needed. For example, it was noted that expert and non-expert review of the systems was necessary as each group provided different and relevant feedback. Furthermore it enabled the confirmation of the roles and responsibilities of the Human Factors team.

Several changes were made to the structure of the evaluation tools to eliminate redundancies and to improve their design. The initial tools were consistent with evaluating a system against usability guidelines such as those for general performance, navigation and user comfort. However the trials required evaluation of several systems in varying states of development and using a variety of interaction methods and displays. In order to structure the evaluation to be consistent for all systems and cases, the tools were modified to support the respondents' understanding of the process of the trial rather than separate elements. For example, as previously stated, all cases consisted of: setting up the task, performing the task, display of task progress and so on; even if some of the questions below these headings were slightly different or not applicable for each case. This provided a well-structured and consistent basis for assessing all cases as well as reporting the results. In this way the evaluation team were able to identify where in the process issues were highlighted – an important piece of information to be fed back to the development team. In addition by consistently using the same questions it was possible to see in subsequent trials whether changes had been made.

Table 2
Overview of structure of Trial 3 – showing types of users involved, and evaluators per cluster

	CL1 - Satellite Assembly	CL 2 - Workplace Design	CL 3 - Remote rail maintenance	CL4 - Nuclear reactor maintenance	Total
Interviews	1 Expert 2 Novice	2 End-user 2 Novice	3 Expert (2 Finland, 1 Spain) 2 Novice (1 Finland, 1 Spain)	2 Expert 2 Novice	16
Questionnaires	1 Expert 2 Novice	2 End-user 2 Novice	3 Expert (2 Finland, 1 Spain) 2 Novice (1 Finland, 1 Spain)	2 Expert 2 Novice	16
Observations	1 Expert 2 Novice	2 End-user 2 Novice	3 Expert (2 Finland, 1 Spain) 2 Novice (1 Finland, 1 Spain)	2 Expert 2 Novice	16
Evaluators	4	4	3 (Finland), 1 (Spain)	5	17

Table 2 illustrates the complexity of evaluation at a trial. Multiple categories of user: 'novice' (users unfamiliar with tool and industrial context), 'expert' (users familiar with tool development and industrial context) and 'end-user' (industrial partners involved in the development); simultaneous evaluation in different locations; and the concomitant requirements for evaluators all conspire to a complex and changeable piece of research providing further evidence for Ward et al.'s (2010) claims about the pitfalls of participatory action research.

4. Results

The evaluation methods generated both qualitative and quantitative data. Interrogating these data sources and using trend analyses allows us to ascertain the effectiveness of the evaluation methods in defining usability and utility issues.

4.1. Quantitative results

The ratings from the Questionnaire and Observation were combined to get overall ratings for each cluster across the trials. The rating scale goes as follows: 1 (extremely negative), 2 (negative), 3 (positive), and 4 (extremely positive). Figure 1 illustrates the trend of average ratings of each cluster's application as the trails progressed. No clear trend emerged across clusters. Cluster 1 showed disimproved performance in each successive trial. Cluster 2 improved from Trials 1 to 2, but disimproved slightly in Trial 3. Cluster 3 showed a very slight negative trend across all three trials. Whereas Cluster 4 had a pattern similar to Cluster 2. Only when examining Trial 2 to Trial 3 can we start to see a trend common to all clusters – slight deterioration in usability and utility ratings.

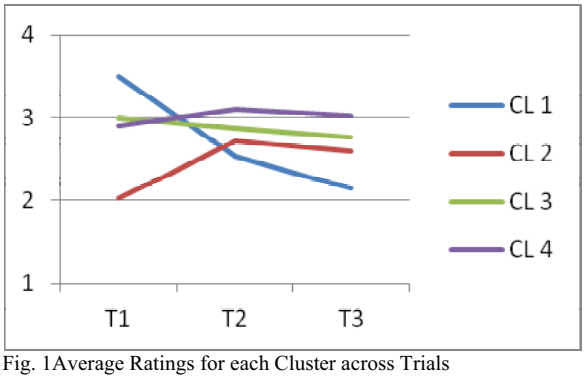

Fig. 1Average Ratings for each Cluster across Trials

The detailed results per cluster shed more light – in particular the following categories:

- Setting up the task
- Performing the task
- Display of task progress
- Visualising data

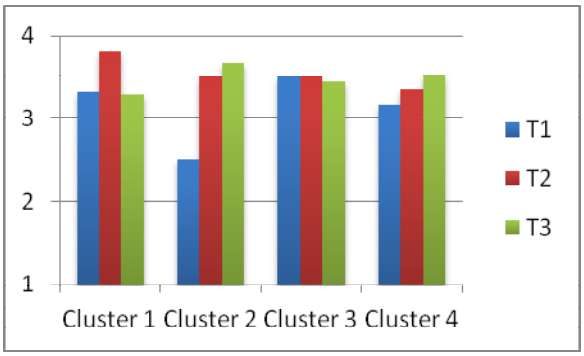

Fig. 2 'Setting up the task' Ratings for each Cluster across Trials

Ratings for 'Setting up the task' show a strong positive trend for both Clusters 2 and 4 – mirroring the overall trend for these clusters. Cluster 3 has a relatively constant rating across trials, with a slight downturn in Trial 3. Cluster 1 has no clear pattern with initial improvement leading to a return to Trial 1 levels in Trial 3.

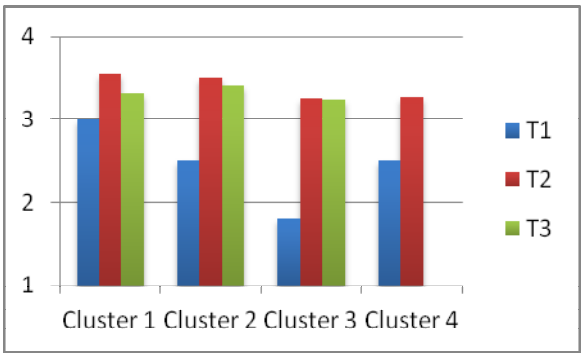

Fig. 3 'Performing the task' Ratings for each Cluster across Trials

No cluster has overall positive trends for 'Performing the task' ratings – though for all there is improvement from Trial 1 to Trial 2. In two Clusters (1 and 2) there is a negative trend from Trial 2 to 3, while Cluster 3 remains constant. This category did not apply to Cluster 4.

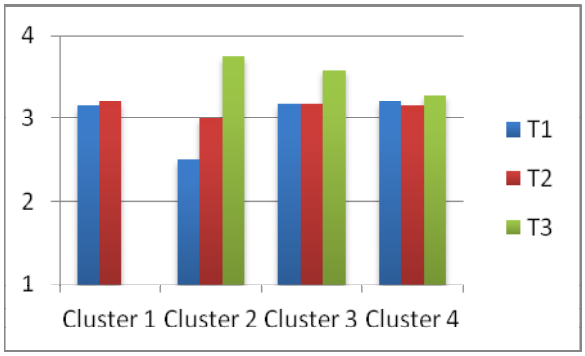

Fig. 4 'Display of task progress' Ratings for each Cluster across Trials

There is no clear trend at all for 'Display of task progress' ratings – apart from improvement from Trial 2 to Trial 3.

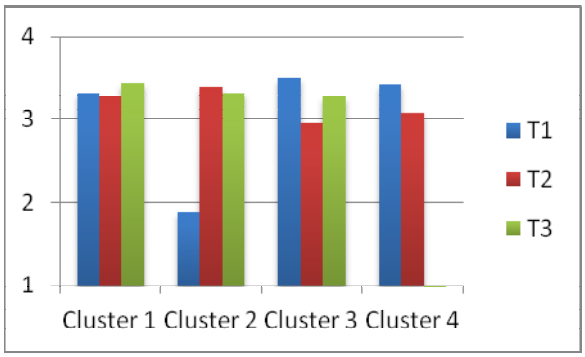

Fig. 5 'Visualising data' Ratings for each Cluster across Trials

There is no clear trend for 'Visualising data' ratings but by Trial 3 all Cluster for whom these ratings apply are firmly in the positive end of the scale.

4.2. Qualitative results

Examining the qualitative data elicited from the interviews can help to provide detailed explanations of how the changes in ratings actually impacted the clusters and the applications they are developing. In addition it helps to explain why the ratings have improved or disimproved. Specific examples are provided for each cluster application. For example, the interventions made in interface improvement (both graphical user interface and human-machine interface), suggestions for implementation of the applications and for future training and procedural design.

4.2.1. CL 1 Satellite assembly – task analysis and procedure validation

General user experience

The application has been improved over the course of the trials. Application users can carry out the procedure validation more quickly and more accurately than before. Although greatly improved, there is still need to further develop the application's methodology to achieve the full potential of the task analysis.

Setting up the task

Setting up the task did not improve from T2 to T3. This was partially because the preparation of the task actions was carried out by developers and users were not logging on the system by themselves. There was also some delay in responding to users' actions and some blurriness when adding and saving text. Users expressed that they wanted to work with a simpler interface for preparing the task.

Visualising the data

Visualising the data has been improved in T3, although there were some problems with stereoscopic vision and 2D was used on occasion. Nevertheless the screen was less cluttered and the use of symbols to represent the data was more appropriate. Users were also happy to have all the information presented on power walls.

4.2.2. CL 2 Workplace design – ergonomic analysis

Display of Task Progress

Display task progress improved from T2 to T3 as the task recording time was shown more accurately (every second rather than updated every 5 seconds).

Visualising data

Visualising the data showed slightly less improvement from T2 toT3. Recorded values related to angle and frequency of the movement of a body part (i.e. either arms, neck or back) were shown in two tables representing dynamic and static postures respectively. This was missing from Trial 2, during which a confusing table presenting repetitions per minute and static postures was shown instead. The T2 interface displayed a recorded task duration. The T3 interface improves the clarity of what the results actually show by displaying the estimated real task duration, that is to indicate that the displayed results have been extrapolated to show (e.g. whether postures are healthy if performed over 4 hours etc).

Performing task

Performing the task showed slightly less improvement from T2 toT3. The HF expert does not have many tasks to perform whilst the operator is carrying out an assembly task. In addition to starting and

stopping the task recording, the expert also has the option to pause and re-start the task if they need to give additional instruction to the operator.

4.2.3. CL3 – remote rail maintenance – task planning and diagnostics

Visualising data

The visualisation of the data that the user needs to use this application has improved over the course of the trials. Predominantly this is due to the improvement of the icons and text boxes used. There was clear improvement in the labelling of icons and parts, and the ability to position icons.

Display of task progress

This aspect of usability also improved. Symbol placement functions helped to improve the feedback between the system and user and this, in turn, allowed each of the two distributed participants to understand what progress they had made. Similarly designers are now able to see the worker in the real environment and can immediately ascertain what stage in the task they are at and if they have difficulties.

Setting up the task

Latency, or lag, was an issue here in all trials and ratings for this factor did not show much improvement over the course of the trials. In reality latency, as it presented itself during the trials, was more a factor of the internet speed in the organisations hosting the trial but nonetheless respondents were increasingly frustrated with the lack of improvement in this area – even if some improvements that had been made with the software latency from Trial 1 to Trial 2 were eroded in Trial 3 when the internet connection was especially poor.

4.2.4. CL4 - Nuclear reactor maintenance – training tool

General Issues

As this type of training cannot be carried out in the organisation at present, the application was seen to have added benefit and improvement was experienced across trials. The motor skills and simulation training were seen as favourable as they focused on specific issues required by the organisation - how to improve trainees with as real an experience of the required task without actually spending time in a real plant with real equipment and real pipe-work. The cost effectiveness of a computer-based system is clear. All parts of the task can be re-used and the task can be replicated many times with many trainees. The addition of a 'display of results' of the trainee was perceived as a real added benefit as this supports the trainees' development as well as the their understanding of the support needed.

Display of Task Progress

Several improvements were made between T2 and T3. In response to suggestions from T2 the feedback "tick" and "cross" were enlarged and sounds associated with them were played in the procedural training lesson. In the simulation training a feedback box was provided which displayed the status of the polishing task. This supported the user in being aware of their progress as the task progressed and finally completed. The additional feature of a display of the results of the trainee was highly positive. While the feedback provided was still basic, it was possible to see the potential benefits to the trainee as well as trainer. This feedback enables the trainee to understand their progress and the trainer to see what further support is required.

5. Discussion

The evaluation framework set out in this paper needs to be critically examined in terms of its contribution to the development of the applications. Did the system operate more effectively as the trials progressed? Were the applications more usable as the development went ahead? Were user experiences more positive in each trial? In order to answer these questions properly one must separate out the effects of the effectiveness of the evaluation framework itself from other factors that impact on usability, performance and acceptance. In such a large and complex research project with so many multi-disciplinary partners and with so many agendas (often competing) it is hard to be able to give a definitive answer with sample sizes that are again compromised by pragmatic concerns.

The paper has detailed some of the factors which have impacted on user perceptions of the applications but which are unrelated to the tools themselves (e.g the latency issue with Cluster 3 caused by poor internet connections). Similarly it is worth noting that all the clusters reported a negative trend for the results of the last trial. This could be due to the fact that the time available between Trial 2 and Trial 3 was not sufficient to enable the developers to act upon all the issues raised during the previous trial. Further there was a decision to concentrate the efforts on the further development of the common platform for all the clusters, so as to improve the unity of a single overall

ManuVAR application. However this is not a software feature appreciable from the end-user perspective according to the categories considered cluster by cluster and as such the results can be considered distorted by this.

So while it has not been possible to demonstrate with any conviction that the evaluation framework positively contributed to the development of the applications (over and above the normal development over time) what it has demonstrated is that by using multiple methods (questionnaires, interviews and observations primarily) with a common inquiry structure it is possible to triangulate the evidence from one source with that of another. This is something which is advocated by many human factors experts [12] and in this particular case it has been effective in that the evaluators have been able to create a real, nuanced, and richer picture of the improvements made and further improvements needed than would have been possible with one method alone. This finding represents an important lesson learned for evaluating complex VR/AR development projects – one which will be useful as the ManuVAR project moves into its final phase of demonstrator building. In addition it is clear that speedy development of the applications was facilitated by the evaluation framework which was able to meet the requirement to test an application and give feedback three times in the 9-month development period.

Acknowledgement

The research leading to these results has received funding from the European Community's Seventh Framework Programme FP7/2007-2013 under grant agreement no. 211548 "ManuVAR". For more details visit: www.manuvar.eu

References

[1] Yusuf, Y., Sarhadi, M., & Gunasekaran, A. (1999). 'Agile manufacturing: the drivers, concepts and attributes'. International Journal of Production Economics 62, 1–2, 33–43.

[2] Muller, M.J., and Kuhn, S. (Eds.) (1993). Communications of the ACM special issue on participatory design, 36(6), June 1993.

[3] ISO 9241-210. 2010. Human-centred design for interactive systems. International Organization for Standardization, Geneva.

[4] Patton, J. (2002). Hitting the target: adding interaction design to agile software development. In: OOPSLA 2002 Practitioners Reports, Seattle, Washington. ACM, New York .

[5] Chamberlain, S., Sharp, H., Maiden, N. (2006) Towards a framework for integrating agile development and user-centred design. In: Abrahamsson, P., Marchesi, M., Succi, G. (eds.) XP 2006. LNCS, vol. 4044, pp. 143–153. Springer, Heidelberg

[6] Holzinger, A. and Miesenberger, K. (Eds.). (2009). Current State of Agile User-Centered Design: A Survey. USAB 2009, LNCS 5889, pp. 416–427, Springer-Verlag Berlin Heidelberg.

[7] Coghlan, D. & Brannick, T. (2001). Doing action research in your own organization. London: Sage.

[8] Ward, M., McDonald, N., Morrison, R., Gaynor, D., & Nugent, T. (2010) A Performance improvement case study in aircraft maintenance and its implications for hazard identification. Ergonomics, Special Edition: Human Factors in Aviation. Vol. 53, Issue 2, pp. 247 – 26

[9] Tromp, J.G., and Nichols, S.C. (2003). VIEW-IT: A VR/CAD inspection tool for use in Industry, Proceedings of the HCI International 2003 Conference, Crete, 22 – 27 June.

[10] Kennedy, R.S. Lane, N.E., Lilienthal, M.G., Berbaum, K.S., and Hettinger, L.J. (1992) Profile Analysis of Simulator Sickness Symptoms: Application to Virtual Environment Systems. Presence 1(3) 295-301.

[11] Kennedy, R.S., Lane, N.E., Berbaum, K.S., and Lilienthal, M.G. (1993). Simulator Sickness Questionnaire: An enhanced method for quantifying simulator sickness. The International Journal of Aviation Psychology, 3(3), 203-220.

[12] Wilson, J. R. (2005) Methods in the Understanding of Human Factors. In Wilson, J. R. & Corlett, N. (Eds.) Evaluation of Human Work. 3rd ed. London, Taylor & Francis.

[13] Vink, P., Nichols S., and Davies R.C. (2005). Participatory Ergonomics and Comfort. In P. Vink (Eds.), Comfort and Design (pp. 41-54). Florida: CRC Press.

[14] Krassi, B., D'Cruz, M., Vink, P., ManuVAR: a framework for improving manual work through virtual and augmented reality // Proceedings of the 3rd International Conference on Applied Human Factors and Ergonomics (AHFE2010), Miami, Florida, USA, 17-20 July, 2010, 10 p. ISBN-13: 978-0-9796435-4-5. Published in "Advances in Occupational, Social, and Organizational Ergonomics" (eds. Vink, P., Kantola, J.), CRC Press, 2010, ISBN 9781439835074