Evaluating the real usability of a C2 system – short and controlled vs long and real

V. Lanfranchi
OAK Group, University of Sheffield
v.lanfranchi@dcs.shef.ac.uk

Suvodeep Mazumdar
OAK Group, University of Sheffield
S.Mazumdar@sheffield.ac.uk

F. Ciravegna
OAK Group, University of Sheffield
f.ciravegna@dcs.shef.ac.uk

ABSTRACT

Command and Control systems (C2) need to be highly usable to efficiently support communication during crisis situations. Short-term usability evaluations are typically carried out because of the cost and time advantage. However, even the most realistic evaluations (simulations) do not reflect the real issues encountered "in the wild", such as the operators' stress or the multiple foci of attention. In this paper we carry out an experiment to measure the changes in usability of a C2 system between a short-term simulation and a long-term "in the wild" evaluation. We demonstrate that short-term usability measurements can differ significantly from long-term "in the wild" ones. Our results indicate that a different approach to usability assessment is needed when dealing with critical systems, that takes into account the temporal horizon and assesses the system in real-life conditions.

Keywords

Command and Control systems, Long-Term Usability, User Experience, Critical systems, Emergency Response.

INTRODUCTION

When a crisis situation occurs operators have to act quickly, coordinate actions and make decisions that cover a large problem area affected by a multitude of different factors and aspects. A key success factor is to have "**accurate, complete and real-time** information about an incident" (Winerman, 2009) to understand "the current local and global situation and how this may evolve over time" (Endsley, 1995). Command and Control (C2) systems are designed and developed to support communication and collaboration between operators in an emergency, with a particular focus on operators in a Command and Control Centre. Given the critical nature of Command and Control systems, we believe it is fundamental to assess their usability and acceptance not only in short-term controlled settings but also longer-term "in the wild". In this study we take into account two fundamental dimensions of usability studies: 1) Temporal: studies can be short-term and long-term, accordingly to the length of time for which the systems are tested. Short-term studies tend to be more popular because of their relative simplicity and low cost, biasing the results more towards discoverability problems (Mendoza and Novick, 2005); 2) Experimental setup: studies can be controlled experiments or "in the wild", field experiments. In field trials the experimenter has less control over how the system is used but the environment is realistic. We believe that in order to study the usability of critical systems, these methodologies should be combined in order to (i) maximise the amount of usability issues discovered and (ii) understand the effect of using the system in real-life situations, over a longer length of time. In this paper we investigate the long- and short-term usability of a C2 mobile system used to communicate emergency information in real-time. The study includes a short-term controlled evaluation and an evaluation carried out over a longer span of time (five weeks) in an uncontrolled real world situation. Our study demonstrates that the first impression of an application is not consistent with longer-term impressions, especially when the C2 system is integrated in real work practices.

RELATED WORK

Research in Information Technology systems for managing and responding to emergencies has recently focused on user requirements (Lanfranchi and Ireson, 2009; Leoni, De Rosa, Marrella, Mecella, Poggi, Krek and Manti, 2007; Diehl, Neuvel, Zlatanova and Scholten, 2006) and interface design (Jennex, M. E, 2007; Paulheim, Döweling, Tso-Sutter, Probst, and Ziegert, 2009; Turoff, Chumer, Van de Walle, and Yao, 2004) highlighting

the need of supporting situational awareness by designing usable interface that support user interaction with large scale data, avoiding information overload. Previous research also addresses the issue of evaluating the usability of such systems (Prasanna, Young, King, 2011) with an attention for evaluating the usage of mobile platforms [1]. In contrast to similar evaluation studies (Cabas Vidani, Chittaro and Carchietti, 2010; Dubé, Kramer, Vachon, and Tremblay, 2011; Prasanna et al., 2011) targeted at emergency response systems, our study performs evaluations both in controlled settings and in a real-life scenario, and introduces a temporal element by evaluating the real-life scenario long-term. In general, the practice of undertaking short-term usability studies has been criticised by many (Greenberg and Buxton, 2008). Long-term user studies, especially with field observations, have been standard methodology in many disciplines (e.g. psychology, sociology), but only recently they started being more commonly applied to usability (Vaughan, and Courage, 2007), introducing the concept of longitudinal usability evaluation. Longitudinal approaches are very useful to evaluate a system over an extended period of time and they can be combined with other methodologies to better understand the real usage of a system and its effectiveness (Kelly and Belkin, 2004). Sauermann and Heim (Sauermann and Heim, 2008) and Mitchell et al. (Mitchell, Caruana, Freitag, McDermott and Zabowski, 1994) undertook longitudinal studies to analyse the long-term usages of Personal Information Management systems.

THE C2 SYSTEM

The C2 system was developed as a prototype for a European Research project[1] and consists of a desktop and a mobile application. The main target of the project was to support collaboration between the command and control centre and the responders on the scene, to exchange information and build Situation Awareness (SA) (Blandford and Wong, 2004). Referring to the Collaboration tools and team processes matrix developed by Bolstad and Endsley (Bolstad and Endsley, 2003), our C2 system can be classified as a domain specific tool that focuses on supporting Data Gathering, Data distribution and Shared SA. A mobile application allows FLOs on the scene to record photos, audio and videos from their smartphone and upload them to a central repository. Content-analysis technologies such as Natural Language Processing (NLP) are used in the background, automatically analysing the captured information and annotating it with tags. Standard tags such as location, time, the ID of the mobile phone transmitting the data are automatically recorded in the system, whilst tags gathered using Information Extraction technologies are proposed to the user, who chooses to accept or reject the annotations. This form captures the details of the incident that are of relevance to the Command and Control centre, such as potential hazards, the extent of the emergency, what groups of people could be affected by it, where escape routes are available etc. The form is automatically geolocated, and following the submission of the form the system creates the emergency incident in a central database. FLOs are then able to capture images and text that describe specific elements of the emergency incident. Personnel in the Command and Control Centre can use the desktop application to view all uploaded data in real-time and can use a number of filters to highlight the data they are interested in.

Research Questions

Our study was focused on addressing the following research questions by evaluating the C2 system: (1) How does the perception of the usability of C2 systems change over time? (2) How does the perception of usability change between the lab and in the wild? and (3) How long does it take to users to accept an application and obtain benefits from its usage?

USER STUDIES

The user study was set up as a combination of short-term, controlled setting evaluation and a long-term in the field evaluation. Five FLOs from a professional Emergency Response team participated in the evaluation. Such small number of evaluators was due to the difficulty of finding experienced users that would be available to use the system in real-life scenarios over a long period of time. The users were specialists in the Structural and Public Safety team and in the Health and Safety team, aged 34-54, working as FLOs for at least the last 10 years. The same users took part in both the short-term and the long-term evaluations, where both subjective (usability questionnaire, interviews and observer impressions from weekly field observation visits) and objective measures (application logs to determine performance, task completion etc.) were collected.

Short-term evaluation

The participants were asked to take part in a full day simulation exercise, organised in cooperation with

[1] http://www.weknowit.eu

Sheffield City Council on the 22nd of February 2011, at a Fire and Rescue Training and Development Centre, in Sheffield, UK. A training session was given, with the possibility to try the system and ask questions. The Emergency Responders participating in the exercise were asked to complete two tasks scenarios: 1) Take pictures, upload and tag them; 2) Submit incident forms to the Command and Control Centre. After the exercise, participants were interviewed and asked to fill out a usability questionnaire. All the user actions were logged.

Long-term field evaluation

The participants used the mobile application for five weeks on smartphones provided; the smartphones were put in the FLO "bag", a kit carried whenever on call to witness an emergency. The application was used in a complementary way to other traditional methods (such as pen and paper), to make sure no data were lost because of system malfunction. The FLOs took part in a training day. Follow-up individual training sessions were scheduled after the 2nd week. The tasks carried out during the long-term evaluation included the two tasks form the short-term controlled evaluations, but the users were free to experiment all the system functionalities.

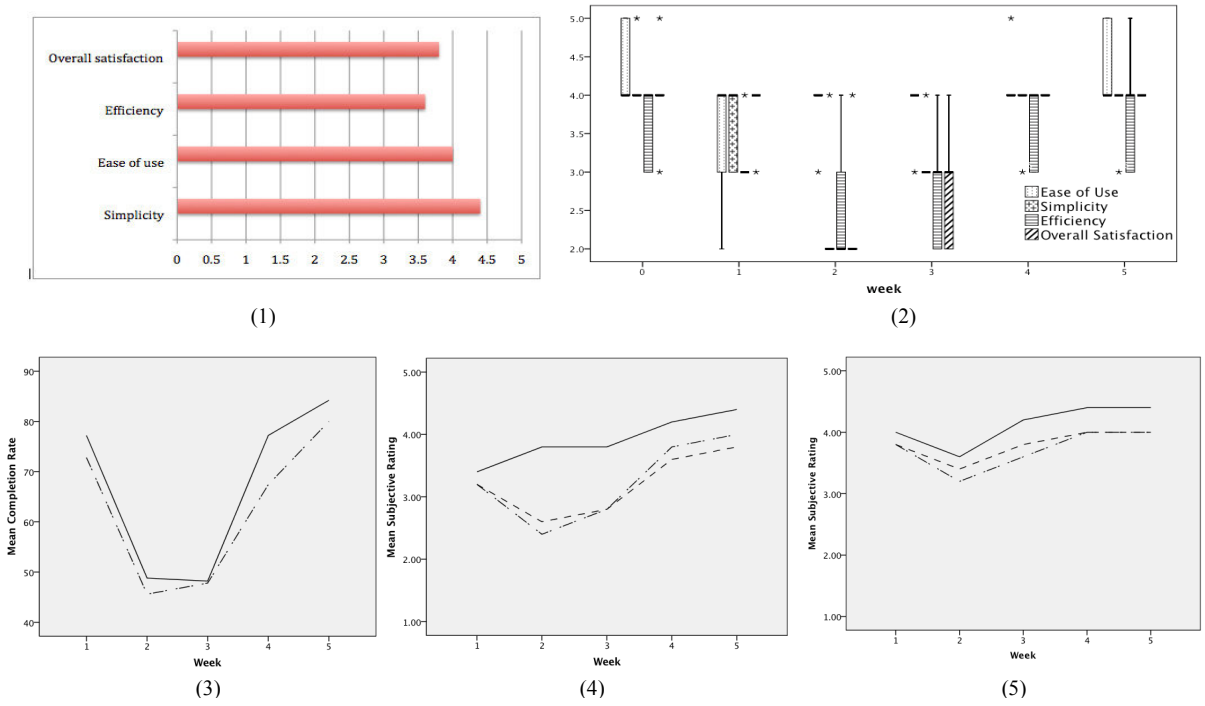

Figure 1: Usability of the application during the short –term evaluation (1) and over time (2). Scores on a scale from 1 to 5, where higher is better. (3) - Task completion rate for task 1(solid line) and task 2 (dot-dash line) over the course of the long-term experiment; (4) - Subjective ratings of simplicity (solid line), effectiveness (dot-dash line) and efficiency (dash line); (5) - Subjective user experience measures over time of likeability (solid line), overall impression (dot-dash line) and clarity (dash line).

RESULTS

In the following sections we will answer our research questions by presenting results from the two evaluation phases. The short-term evaluation was focused on overall system acceptance and user satisfaction and identifying serious usability issues that could impair the deployment of the system in real-life situations. Given the scale of the user sample, no statistical significance could be established. The questionnaires results show high mean scores for ease of use, system simplicity and satisfaction (mean score of respectively 4.40, 4 and 3.80 on a 1 to 5 Likert scale where higher is better). For what regards quantitative measures, the average completion rate was 74.4% for Task 1 and 70% for Task 2. During the short-term evaluations, 11 usability issues were identified, 5 of which critical (e.g. the system did not notify the user when a photo had been successfully uploaded). The developers addressed these issues in the version released for the long-term field evaluation.

How does the perception of usability change between the lab and in the wild?

Using a system "in the wild" presents the user with unplanned events, changes in plans, stress and pressures that end up affecting the overall usability of the system. User comments from a field observation visit reveals how

even simply the amount of people involved in an emergency situation can affect the user behaviour:

"I know how to take a picture with a mobile but now with some many people around me I feel under pressure and then I get it wrong" (User4)

"I know I have to take simulations seriously and I am serious when I perform them but what happens in my mind during a real emergency is completely different; my brain is switched on alert constantly and busy contemplating all possibilities and consequences" (User1)

During the short-term controlled evaluation the users were confident and focused on trialling the new system, therefore they were able to complete their tasks easily; when they tried the same actions during emergency situations, they started forgetting steps and making mistakes. This shows the importance of considering the cognitive load of a user [5] when designing a C2 interface. Whilst the short-term evaluation did not highlight specific issues in the help provided by the system, the long-term evaluation presented us with different results. The users started failing their tasks, even when they were an exact repetition of the task carried out during the short-term evaluation. When interviewed 3 out of 5 users commented negatively on the help features they had favourably judged during the short-term evaluation. Two comments were indicative of their feelings:

"I know how to do it, but when I am very busy I forget a step and then I feel lost" (User1)

"If only I had a little assistant helping me exactly when I pressed the wrong button and reminding me the right action then I'd have been fine" (User4)

These results illuminate the need for a better system design, that facilitate task execution and of adequate contextual help features. Moreover during the long-term field evaluation the users had to integrate the new application in their working practice, thus increasing the cognitive workload and potentially affecting the perceived usability.

How long does it take to users to accept an application and obtain benefits from its usage?

The results have shown that system acceptance is achieved after a significant lead period, such as in this case, after three or more weeks of system usage, and after further reinforcement training has been provided. The system became easier to use and more intuitive after the 3rd week, with corresponding increasing completion rates (rising from an average of 47.5% to 80.7% in Task 1 and average 46.7 to 73.2% in Task 2). This was also reflected by the values of perceived efficacy and efficiency (as shown in **Figure 1**) and above all by the usability curve, that flattened around the 4th week, indicating that a proficient level of system usage had been reached.

CONCLUSION

This paper described an evaluation of a C2 system using a combination of evaluation methods to better ascertain its usability and acceptance. We have shown how the "first impression" of an application (obtained during a short-term evaluation) may not be matched by a corresponding attitude over time, especially when the application is adopted in real work situations. Whilst it is very difficult to detangled the experimental conditions (as the short term evaluation was carried out in controlled settings whilst the long term in the field), we believe our study proved that carrying out an extended in the field usability study is fundamental to get a correct picture and to provide users with an adequate level of training and support, that ensures the application is accepted.

ACKNOWLEDGMENTS

This work has been supported by the EPSRC funded project RAnDMS (Grant Ref: EP/J020583/1).

REFERENCES

1. Al-akkad, A., Zimmermann, A., Birlinghoven, S. and Augustin, S. (2011) User Study : Involving Civilians by Smart Phones During Emergency Situations. *Proceedings of ISCRAM*.

2. Blandford, A. and Wong, W.B. (2004) Situation awareness in emergency medical dispatch. *International Journal of Human-Computer Studies*, 61(4):421–452.

3. Bolstad, C. and Endsley, M. (2003) Tools for supporting team collaboration. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 47. No. 3. SAGE Publications.

4. Cabas Vidani A., Chittaro L. and Carchietti E. (2010) Assessing nurses' acceptance of a serious game for emergency medical services, *Proceedings of VS-GAMES 2010: 2nd International Conference on Games*

and Virtual Worlds for Serious Applications, IEEE Computer Society, Los Alamitos, CA, USA, pp. 101-108.

5. Cooper, G. (2004) Research into Cognitive Load Theory and Instructional Design at UNSW, *University of New South Wales.*

6. Diehl, S., Neuvel, J.M.M., Zlatanova, S. and Scholten, H.J. (2006) Investigation of user requirements in the emergency response sector: the Dutch case In: *Proceedings of the Second International Symposium on 'Geo-information for disaster management*'. - Ahmedabad, India.

7. Dubé, G., Kramer, C., Vachon, F. and Tremblay, S. (2011) Measuring the Impact of a Collaborative Planning Support System on Crisis Management, in *Proceedings of Iscram.*

8. Endsley, M.R. (1995) Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors*, 37(1):32–64.

9. Greenberg, S., and Buxton, B (2008) Usability evaluation considered harmful (some of the time). *In: Proc. CHI 2008,* ACM, New York, pp. 111–120.

10. Jennex, M. E. (2007) Modeling Emergency Response Systems. In: *Proceedings of the 40th Annual Hawaii International Conference on System Sciences (HICSS 2007)*. Hawaii, 3 - 6 Jan. 2007, pp. 1-8.

11. Kelly, D. and Belkin, N. (2004) Display time as implicit feedback: Understanding task effects. *In Proceedings of the 27th Annual ACM International Conference on Research and Development in Information Retrieval*. ACM Press, New York, 377–384.

12. Lanfranchi, V. and Ireson, N. (2009) User Requirements for a Collective Intelligence Emergency Response System. In *Proceedings of HCI*, Cambridge, UK.

13. Leoni, M., De Rosa, F., Marrella, A., Mecella, M., Poggi, A., Krek, A. and Manti, F. (2007) Emergency Management: from User Requirements to a Flexible P2P Architecture. *In Flexible Service and Data Management Platforms for Crisis Response at the 4th ISCRAM*, pages 271–279, Delft, Netherlands.

14. Mendoza, V. and Novick, D. G. (2005) Usability over time. In ACM 23rd International Conference on Computer Documentation, ACM Press, 151-158.

15. Mitchell, T., Caruana, R., Freitag, D., McDermott, J. and Zabowski, D. (1994) Experience with a Learning Personal Assistant. *Communications of the ACM 37(7)*, 81-91.

16. Paulheim, H., Döweling, S., Tso-Sutter, K., Probst, F. and Ziegert, T. (2009) Improving Usability of Integrated Emergency Response Systems: The SoKNOS Approach, *GI 2009, Workshop zur IT-Unterstützung von Rettungskräften*, p. 1435-1449.

17. Prasanna, R, Young, L. and King M. (2011) Evaluation of a Software Prototype for Supporting Fire Emergency Response, In: *Proceedings of Iscram.*

18. Sauermann, L. and Heim, D. (2008) Evaluating Long-Term Use of the Gnowsis Semantic Desktop for PIM. *In* Proceedings of the 7th International Conference on *The Semantic Web (ISWC '08)*, 467-482.

19. Tullis, T.S,. Mangan E.C., and Rosenbaum. (2007) An Empirical Comparison of On-Screen Keyboards. *Human Factors and Ergonomic Society 51st Annual Meeting.*

20. Turoff, M., Chumer, M., Van de Walle, B. and Yao, X. (2004) The Design of a Dynamic Emergency Response Management Information System (DERMIS). *Journal of Information Technology Theory and Application (JITTA)*, 5(4), pp. 1-35.

21. Vaughan, M., and Courage, C. (2007) SIG: Capturing longitudinal usability: What really affects user performance over time? *Extended abstracts on human factors in computing systems*, 2149–2152.

22. Winerman, L. (2009) Crisis Communication. *Nature* 457(7228): 376-8, 2009