Formative evaluation for complex interactive systems

ROAST, Chris <http://orcid.org/0000-0002-6931-6252> and URUCHURTU, Elizabeth <http://orcid.org/0000-0003-1385-9060>

Published version

ROAST, Chris and URUCHURTU, Elizabeth (2013). Formative evaluation for complex interactive systems. In: COLLAZOS, Cesar, LIBORIO, Andreia and RUSU, Cristian, (eds.) Human Computer Interaction : 6th Latin American Conference, CLIHC 2013, Carrillo, Costa Rica, December 2-6, 2013, Proceedings. Information Systems and Applications, incl. Internet/Web, and HCI (8278). Springer Verlag, 47-54.

Formative Evaluation for Complex Interactive Systems

Chris Roast and Elizabeth Uruchurtu

Culture, Communication and Computing Research Institute
Sheffield Hallam University
153 Arundel Street,
Sheffield, S1 2NU, UK

c.r.roast@shu.ac.uk e.uruchurtu@shu.ac.uk

Abstract. This paper reports upon the design and use of a lightweight evaluation method, especially designed to examine complex interactive systems. The approach is illustrated through a case study involving an interactive tool designed to help enable users examine large scale data arising from authentication activity in higher education institutes. The evaluation approach illustrated is to enable the lightweight assessment of usability issues within complex interactive systems and identifying opportunities for significant design improvements. Specifically we argue that this method benefits from capturing key generic factors that underpin the effectiveness of tools for working with complex data. The paper concludes by reflecting upon the effectiveness of the lightweight structured assessment approach and how it supports to formative evaluation.

Keywords: Evaluation; Cognitive Dimensions; Complex Data; Information Retrieval; Innovation

1 Introduction

Approaches to evaluating interactive systems are wide and varied and can be judged in terms of the value of the outputs that they provide and the effort required in obtaining the outputs. The technique illustrated in this paper is formative and lightweight in character, and also specifically suited to complex interaction. Our technique is derived from a framework that was developed to capture human factors evident but rarely touched-on with conventional techniques. The type of complex interaction of interest here are those where user tasks involve working with notations, or languages, in order to achieve a desired effect. Specific complexity arises when the notation has tokens with powerful indirect meanings. A simple example, would be an electronic calendar that supports recurring appointments, the means of defining a recurring appointment introduces significant new possibilities that users should be conscious of. In general, we judge the subsequent complexity to result in "programming-like" activities - such as, finding and fixing mistakes with a recurring calendar appointment. Thus, the effective use of such systems is not only reliant upon the appropriateness of the mechanisms available to manipulate the notation, but also upon the user interpretation of how the system might process the notation. This often results in an intrinsically indi-

rect manipulation. Programming development environments provide obvious examples of this type of complexity in interaction. But similar complexities can be found in more mundane systems such as: calendars and online booking systems.
The case study in this paper is a powerful authentication monitoring tool. It supports the articulation and execution of complex searches of large datasets and as such is "programming-like". In brief, the tool's control panel allows the selection of search templates, specification of filters and parameters and the specification of the data to be output as a graph (illustrated in figure 3). This was viewed embodying the indirect manipulation common with complex interactive system. In addition, a brief tutorial about the tool highlighted how a range of configurable filters could be used to generate user defined views of data.

2 The Analytic Framework

The evaluation approach is a collaborative lightweight method motivated by concepts taken from the Cognitive Dimensions framework [4]. The framework has been the focus of considerable research interest, its potential as a tool for evaluation has been explored with a number of approaches [5]. One example of a framework dimension is "Secondary Notation" - this focuses upon how a system may enable unstructured attributions to a notation (such as comments or highlighting). The framework has some similarities to the concepts of "design patterns" [3] and "ergonomic criteria" [12]. However its relevance for this research comes from its descriptive nature and its use in examining notational systems.
Authoritative sources for the framework show a diverse range of, such, dimensions grounded in concrete examples with informal definitions. Research into the dimensions framework has predominantly focused upon their adoption through the comprehensive and consistent use of the dimensions [1,2,9]. Hence, methods for assessing concepts such as "Secondary Notation" have been explored with the aim of providing an objective assessment of them. Although this is clearly valuable these endeavours appear to have overlooked the fact that the illustrations of the dimensions also demonstrate insights into potential designs that help innovate design alternatives. So, in the case of "Secondary Notation" the different uses to which unstructured attributes might be put based upon examples and analogies can be insightful. As an example the uses of "Secondary notation" include: a means of communication, a facility to improve presentation, as well as a technique for demonstrating expertise. While these points are worthy of evaluation, they in fact point interesting ways in which a notation might get used.
Hence, instead of treating the framework as a means of assessment, it also has the potential to promote innovative perspectives upon existing designs. The approach to formative evaluation described here follows this line of argument and thus places less priority on objective comprehensive assessment and more on the variety of ways, or modes, in which concepts found in the Cognitive Dimensions framework drive new ideas or insights.

2.1 The tabular framework

Our approach is to use simple a tabular format for engaging system developers, experts and end users in co-operative evaluation. The use of this has been reported [9,10] within the context of a tool for digital video post-production and publishing. The tabular approach is designed to encourage collaborative reflection and insight through focusing upon a relatively small number of key questions (derived from those in [1]).

What are the dominant / common ways in which these concepts are shown together or reached from one another?

to / from	Specification(s)	Data set(s)	Publisher service(s)
Specification(s)	always / no. of clicks / not during …	always / no. of clicks / not during …	always / no. of clicks / not during …
Data set(s)	always / no. of clicks / not during …	always / no. of clicks / not during …	always / no. of clicks / not during …
Publisher service(s)	always / no. of clicks / not during …	always / no. of clicks / not during …	always / no. of clicks / not during …

Fig. 1. An example tabular entry to examine how easy it is for the user to navigate information

Reflection and potential insights are encouraged by the tables presenting how, for a single question, it could answered from a number of perspectives. For instance a single table will encourage users to respond to a question such as "How is A reached from B?" and also "How is B reached from A?". Figure 1 shows an example for this type of question in full using three alternative concepts from the case study and with indicative possible responses within the cells. Through this instrument users are encouraged to explore ideas that they may not normally consider. The use of just three alternative concepts keeps the approach tractable for collaborative assessment.
The three core concepts used are chosen to be ones central to effectively performing the work that a target system is aimed to support. Concepts are chosen to be relevant, high level and ideally encompass a number of potential conceptual "mismatches" as described in [2].

2.2 The facilitation

Operationally the tables are presented on paper to encourage ease of engagement and enable additional points to be easily recorded. While a subject may use the suggested response alternatives, they can just as easily respond in a manner that is more appropriate for their task and interest. For instance, they may even sketch thumbnail illustrations of what is implied by a specific cell. While the form of the process is relative-

ly simple, the facilitator works with system users and/or experts to build their confidence in completing the tables and encouraging deeper reflection. Notes and marks on or beside the tables are encouraged to reflect and record any other opinions or views. The tables encourage users to make relative assessments within each table, discouraging default responses. In addition, the facilitator encourages the completion of the tables by asking for concrete illustrations or examples of particular judgements. There are two general roles of the facilitator: to encourage reflection, and to record reflection. The facilitator's activity is to primarily work on the first of these and then ensure the second is provided by the participant.

2.3 Ideas and insights

Having completed the table entries the facilitator and participant will have reflected upon the nature of the tool being examined and in doing so will be able to identify potential improvements. The value of employing the tables and their links to the Cognitive Dimensions framework is that the resulting observations are: (i) expressed in generic structural terms and not in terms of local corrections or "fixes"; (ii) the framework can provide insights into ways in which particular dimensions re-frame the system being examined. Earlier we provided an example of this when one considers "Secondary Notation" - once an annotation is pointed out as one way improving presentation and same possibility can be explored with the target system. Overall for each table, alternatives and re-framings can be suggested and examined.

3 The Case Study Context

Our case study concerns the management of online resource authentication within educational institutes. Specifically the system examined supports the monitoring and assessment of subscription services in order to understand how services are used. It's development was supported by JISC and it is currently adopted by a number of UK universities[1]. The direct users are library staff and library managers who may need to review service uptake and, say, compare similar services. A specific example might be to identify whether computing students use the ACM Digital library (www.acm.org/dl) on a comparable basis to IEEE Explore (ieeexplore.ieee.org), or whether in terms of usage, say, Sciencedirect (www.sciencedirect.com) effectively subsumes both. At face value this may not appear to be a particularly complex task, but the raw authentication data often hides subtle details. Some authentication events match one-to-one with accessing a publication, while others can be one-to-many, and on some occasions many-to-one. Such differences arise when each service chooses what authentication standards and policies they will use. In short, comparing service is a non-trivial exercise of interpreting mixed data sources. The case study tool is designed to help address some this complexity by integrating authentication event logs and to examine aggregate views of them over time.

[1] See JISC website: http://www.jisc.ac.uk/whatwedo/programmes/aim/raptor.aspx

3.1 The tool's user interface

The tool examined consists of three architectural components: a web front-end; an aggregator that collates and stores authentication data and performs searches; and, agents that send event logs to the aggregator. End users interact via the web front end which provides access to a “graphs” page. Figure 1 provides an illustration of this page, simplified to highlight the key structure and elements. On this page the user is able to build a search specifications, using a number of given types-of-search forms. Within each they are able to specify details such as: (i) the type of authentication protocol to examine; (ii) the date range of interest; (iii) the level of granularity of the resulting data; (iv) a series of filters that can be used to exclude authentication log items based upon characteristics of the log entries; and (v) a series of post-processors that determine alternative data presentations. In addition users can choose to provide labels for data sets generated by a search and also names for the filters as they are applied. Having formulated a search in this manner, the data set can be generated with the click of an "update" button. The resulting data set is shown as a graph, with the options to access the same data in different formats and reports.

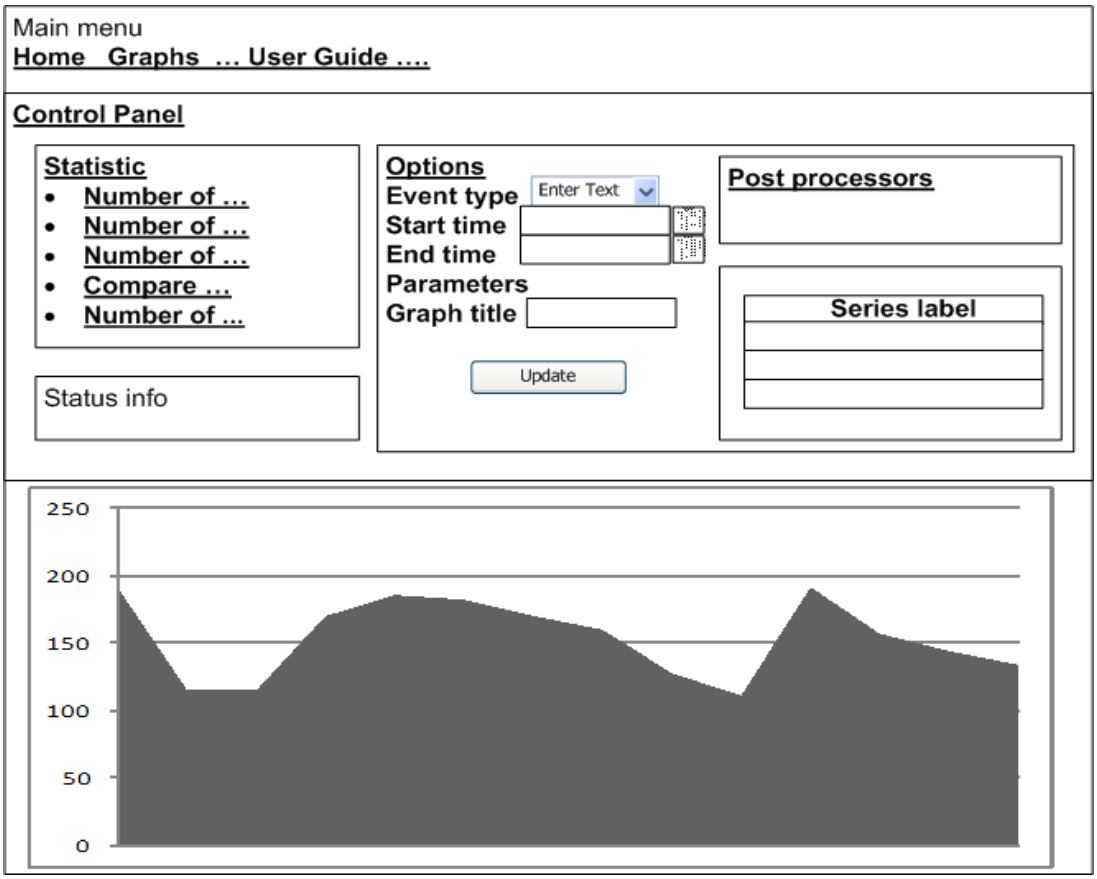

Fig. 3. Schematic of the graph page with graph of dataset shown.

4 Applying the Framework

4.1 Initial user interface review

Intended primary and secondary users were part of the project team. The team collectively reviewed the system, trying out the functionality system, led by an end user while verbally reporting to the team and responding to their comments or questions. The authors focused upon supporting the team in exploration and experimentation with the tool. The first author led this process while the other observed and recorded team reactions and comments. In a subsequent review of the notes, the observations were mapped to the tabular framework, with the questions from the framework acting as "leading questions" in interpreting observations. The review meeting discussed a wide range of tool and usage activity related concepts, and some generic high level usage models: (i) Finding data, (ii) Comparing data, (iii) Aggregating and interpreting data, and (iv) Mixed source assessment. These helped develop a common understanding of potential usage and direct the assessment of the tool. In particular the following three concepts were identified for the assessment of the tool:

- **Searches** Data set specifications for finding data or comparing data
- **Data** The results computed by the system in response to a data set specification
- **Services** Subscribed to services, how they appear in data sets and also their relative cost.

How easy is it to change or alter relationships between concepts?

	Publisher service(s)	Data set(s)
Specification(s)	**Hard - Some of the filtering parameters seem to allow this but they are unclear**	**The relationship is that of the search done … the post processors and filters can narrow or broaden the relationship. Broadening is easier than narrowing.**
Data set(s)	**Hard - a data set is a product of one specific set of services at the point at which the data set was created.**	

Fig. 4. An example of tabular entries for the target system focusing upon ease of change

Figure 4 illustrates one of the tables produced from this review.

The subsequent analysis of the completed tables involved: (i) assessing how coherent relationships and how well the underlying factors were understood; and (ii) taking examples of the table relationships and re-examining them in terms of the alternatives

suggested by the dimensions. These two processes drove further consultation with the team and allowed alternative design ideas to be examined. These improved our understanding of the authentication infra-structure and the standards used, while also helping examine different ways in which authentication data might be analysed, structured and managed.

4.2 Case study outcomes

Case study outcome can be expressed in terms of the concepts underpinning the tabular form. Here we summarise the outcomes focusing upon: abstraction and consistency.

Abstraction. Abstraction mechanisms largely concern the creation and management including explicit and implicit approaches. In our assessment: (i) while the specification of individual searches is supported by the tool, the notion of a search with a generic task oriented purpose is not. Hence, a search that might be conducted to form the basis of a monthly report, is not supported.
Our consideration of data focused upon the value of placing search results next to each other. For instance although it is valuable to compare data in across comparable timeframes (eg seeing Jan 2011 data next to Jan 2012 data), the tool only supports this if the timeframe are the same.

Consistency. Consistency as user interface principal is broadly accepted ([8,12]) while the specific meaning and merits are dismantled with ease ([6]). In the context of the tabular approach, consistency is captured by the lack of confusion between entities represented within a system. In the case of searches of the authentication data the potential to confuse specific volunteered by subjects. As presented, search numerous of parameters some of which are only shown via sub-dialogues. Hence, the differentiation between searches is very difficult.
In a related manner the potential for confusing data sets is very high as it is the user's responsibility to remember to provide a meaningful label for the data set when the data is generated.

5 Reflections and conclusions

The evaluation of complex interactive systems following conventional approaches demands considerable effort and resource. Users and stakeholders are hard to access and as a consequence, evaluation tends to yield lists of issues in a similar vein to those generated from Heuristic evaluation [7]. While these issues can be collated, prioritised and addressed they rarely capture key structural factors that are especially important when examining complex interaction.
By contrast our tabular approach links together a rich framework for exploring complex interaction, with a relatively easy form of conducting analysis and developing formative outcomes from that analysis. We believe this approach is of particular value since it appears to be less resource intensive while also generating insights regarding

potential design alternatives. All of the issues identified in the case study assessment have been accepted by the development team as requiring solutions, with some being directly addressed.
The tabular approach described offers a method focused upon innovative formative evaluation. With the growing use of powerful data intensive systems and the likelihood that non-expert users are expected to engage with them, ensuring that such "empowered" users can work effectively is of high importance.

Acknowledgements. The authors would like to acknowledge the JISC Access & Identity Management Programme for supporting this work and staff and students at the host institution for participating in user based assessments.

References

1. Blackwell, A.F. & Green, T.R.G. (2000). A Cognitive Dimensions questionnaire optimised for users. In A.F. Blackwell & E. Bilotta (Eds.) Proceedings of the Twelfth Annual Meeting of the Psychology of Programming Interest Group, 137-152.
2. Blandford, A., Green, T. & Connell, I. (2005) Formalising an understanding of user–system misfits. In R. Bastide, P. Palanque & J. Roth (Eds.) Proc. EHCI-DSVIS 2004. Springer: LNCS 3425. 253-270.
3. Dearden, A. M. and Finlay, J. (2006). Pattern languages in HCI: a critical review. Human computer interaction, 21 (1), 49-102.
4. Green, T.R.G. and Blackwell, A.F. (1998). Design for usability using Cognitive Dimensions. Tutorial session at British Computer Society conference on Human Computer Interaction HCI'98. http://www.cl.cam.ac.uk/~afb21/CognitiveDimensions/CDtutorial.pdf
5. Green, T.R.G., Blandford, A.E., Church, L., Roast, C.R. and Clarke, S. (2006) Cognitive dimensions: Achievements, new directions, and open questions in Journal of Visual Languages and Computing 17(4), 328-365, DOI: 10.1016/j.jvlc.2006.04.004.
6. Grudin, J. (1989). The case against user interface consistency. CACM 32, 10 (October 1989), 1164-1173. DOI=10.1145/67933.67934
7. Nielsen, J. and Molich R. (1990) Heuristic evaluation of user interfaces. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '90), Jane Carrasco Chew and John Whiteside (Eds.). ACM, New York, NY, USA, 249-256. DOI=10.1145/97243.97281
8. Nielsen, J. (1993) Usability Engineering. Boston: Academic Press.
9. Roast, C. R. and Khazaei, B. (2007). An investigation into the validation of formalised cognitive dimensions. In: Interactive systems: design, specification, and verification. Berlin, Springer, 109-122.
10. Roast, C., Dearden, A. and Uruchurtu, E. (2011). Using and utilizing an innovative media development tool. Proceeding of IHC & CLIHC 2011.
11. Roast, C., Uruchurtu, E. and Dearden, A. (2011) The programming-like-analysis of an innovative media tool. In: Psychology of Programming Interest Group Annual Conference, University of York.
12. Scapin, L.D. and Bastien, C. J. M. (1997) Ergonomic criteria for evaluating the ergonomic quality of interactive systems in Behaviour and Information Technology, Volume 16, pp. 220-231.