

Downstream utility: stepping back

Ann Blandford

UCL

UCLIC, Remax House, 31-32 Alfred Place, London WC1E 7DP, UK

A.Blandford@ucl.ac.uk

+44 20 7679 5288

ABSTRACT

In this position paper, I argue that a focus on problem counts in user-focused evaluation is a limiting view, particularly in relation to downstream utility. While fixing problems is an important consideration, a broader perspective that also seeks to develop systems that better fit their users – their ways of thinking and motivations, etc. – will result in new design opportunities that go beyond small evolutions of existing designs. I illustrate this view with a sketchy analysis of the use of a travel website.

Author Keywords

HCI evaluation; downstream utility; travel websites.

ACM Classification Keywords

H5.2. Information interfaces and presentation; D.2.2 Design tools and techniques

INTRODUCTION

In recent years, there has been an increasing focus on downstream utility as an attribute of any usability evaluation process [9, 16]. John and Marks [11] raised a related concern when they considered the ‘persuasiveness’ of the outputs from various evaluation methods. Underpinning these accounts, however, is an apparent assumption that evaluation is concerned with fixing ‘problems’ in a design and that persuasiveness depends on the quality of the evidence supporting the existence of the problem. As noted in the call for participation for this workshop, the results of user testing are often considered to be more persuasive (i.e. have higher validity) than those of any inspection method or analytical evaluation technique. Indeed, issues identified through inspection that are not also identified in user tests are often referred to as “false

positives” (e.g. [12]). This leads to several challenges for those involved in evaluating interactive systems. In particular, user testing has many limitations:

- 1) Different evaluators note different difficulties from the same data, even think-aloud data [10]. Nørgaard and Hornbæk [14] provide one explanation for this as being that evaluators see what they are looking for and overlook that which they consider unimportant, or are not expecting. Lee *et al.* [12] highlight the importance of the evaluator’s prior expertise in identifying problems (in their case, when conducting Heuristic Evaluations). Within the context of usability practice, there are other influences on what is noted, such as the client’s brief and the broader context within which the evaluation is being conducted. User testing is not a reliable ‘gold standard’.
- 2) There can often be multiple interpretations of the underlying causes of the same surface behaviour, and so there is often a gulf between observed behaviour (‘problem’), explanations, and hence possible design solutions [15, 2].
- 3) When considering problem count, a variable considered important by, for example, Hartson *et al.* [9], there is a question about the level of abstraction of any problem description: a more abstract description will result in a lower problem count but might, conversely, cover many more actual and potential instances of user difficulties with a system. This point is illustrated by Cockton and Lavery [4] in their analysis of problem extraction.
- 4) People are flexible and adapt well to many system deficiencies. Green *et al.* [7, 8] developed a vocabulary for talking about many of these difficulties, called *Cognitive Dimensions*, which capture ideas such as a task that is conceptually simple taking many device actions (“viscosity”) or the order of actions that seems natural to the user being different to that demanded by the computer system (“premature commitment”). These kinds of difficulties can pass unnoticed, by both users and evaluators, unless they are primed to look for them in some way.

- 5) How people experience interactive systems depends on many factors including the context. This obviously includes the setting (laboratory or ‘real world’, with all its distractions and interruptions), but more subtly includes how data is gathered and what participants are asked to do (e.g. task instructions).

In our discussions with developers (e.g. [3]), we have also found that the utility of a problem report depends on how easily the system can be adapted to overcome the identified difficulty. So it is not just what it reported, how it is reported, or the evidence that supports the report, but what the implications of that report are for those who are to use it.

All of these issues – and probably more – suggest that a reductionist approach to assessing usability and downstream utility is only ever going to reveal part of the story. While fixing problems (or avoiding them in the first place) is important, it may be limiting the vision in terms of identifying new design opportunities. I illustrate this point with a simple case study.

Case study: evaluating a travel web site

Four people who were making travel plans were asked to participate in a think-aloud study in which they were to find flights that would satisfy their up-coming travel needs. This was a compromise between a fully naturalistic study and a conventional laboratory-based study: the aim was to work with participants’ real-world needs (rather than some contrived task), but time constraints made it impossible to observe their actual travel-booking activities. The data was analysed using the first stages of CASSM [1], which yielded an overview of the concepts that people were working with. These included the following:

- That people were traveling between places, which would incidentally involve flying between airports: airports are less important than places. There are often alternative departure and arrival airports that serve travellers’ requirements. For example, there are four international airports in London and which one to fly from or to is often less important than other selection criteria. While some flight sites allow users to specify “LON” (i.e. any London airport), the same is not true for other areas, such as the region around Toulouse (with airports at Perpignan, Carcassone, Montpellier, etc.): there, the user has to specify precisely one airport rather than the region.
- People need to know not just about the airports, but also about connections from there to their actual starting points and destinations: this information informs their decisions about flights (e.g. what times of flights are practical).
- Costs – not just of flights, but of entire journeys – were an important consideration, as was total journey time.
- Individuals had many personal requirements, including: assurances about special meals; flights that do *not* transit via the USA (where transit is perceived by some as an

unpleasant experience); a strong preference for direct flights; and flying with (or avoiding) particular carriers. Doubtless, a more in-depth study would have revealed a longer list of personal requirements.

There were also what would traditionally be considered usability “problems”, such as participants being unable to find the required options, having to select a return time for a one-way flight, accidentally closing the browser window and having to start again, and poor system response times.

Few usability evaluation methods would be likely to identify the mismatch between user requirements which centre around journeys between places and system representations which focus entirely on information about flights. Many flight purchasing sites do provide information about fares, flight times and transit points, but none (as far as we are aware) provide the more extended information and decision support that people wanted about journeys.

This is a simple example of a misfit between users and systems. It is not a “problem” that needs “fixing” because users have a rich repertoire of techniques – doing exhaustive searches, looking at paper maps, drawing on prior experience, etc. – that enable them to work with existing systems effectively. However, the interactive experience could be much better: there are design opportunities for integrated travel planning systems that support integrated travel planning, rather than just selling flights (with or without hotel bookings and car hire).

CONCLUSION

Many studies over the last few years have shown that there is no replicability in evaluation studies: the findings depend on the participants (for empirical studies); the evaluators [10] and their skills [6,12] and biases [14]; the method applied [2]; and subtle features of the systems being tested and the test setting. The receptiveness of developers to ‘problem reports’ or ‘redesign suggestions’ similarly depends on a wide range of factors including not just how these insights are communicated (with what evidence, by whom, in what form, etc.) but also what implications they have for design.

Downstream utility is, or should be, concerned with equipping developers to design better systems. A focus on reducing problem counts risks limiting the focus to ‘bug-fixing’. This appears to reflect a view of design as working towards the ‘perfect’ (bug-free) design solution where the role of evaluation is to make progress towards that solution.

An alternative view, articulated by Carroll and Rosson [4] in their work on the ‘task-artifact cycle’, is that every artifact creates possibilities, which define the tasks which are then easy or possible for users to perform, which in turn highlights new requirements for design. Before the advent of the Web and the development of e-commerce travel sites, the flight-finding activity described above would not have been possible without mediation by a travel expert: the evolution of designs to the current point have created new

possibilities, but evaluations of those designs highlight shortcomings that, in turn, reveal new design possibilities. In this view, design and use co-evolve, and the role of evaluation is more than delivering a list of defects that need to be fixed: it is providing new insights about people, systems, and contexts of working and playing, that suggest new design opportunities.

According to this view, downstream utility includes the provision of new design ideas as well as the identification of design problems. It is about possibilities as well as limitations.

ACKNOWLEDGEMENTS

The work reported here has been conducted as part of an EPSRC project GR/S67494.

REFERENCES

1. Blandford, A., Green, T. R. G., Furniss, D. & Makri, S. (forthcoming) Evaluating system utility and conceptual fit using CASSM. To appear in *International Journal of Human-Computer Studies*.
2. Blandford, A., Hyde, J. K., Green, T. R. G. & Connell, I. (forthcoming) Scoping Usability Evaluation Methods: A Case Study. To appear in *Human Computer Interaction Journal*.
3. Blandford, A., Keith, S. & Fields, B. (2006) Claims Analysis 'in the wild': a case study on digital library development. *International Journal of Human-Computer Interaction*. 21.2. 197-218.
4. Carroll, J. M. & Rosson, M. B. (1992) Getting around the task-artifact cycle: how to make claims and design by scenario. *ACM Transactions on Information Systems*, 10(2), 181-212.
5. Cockton, G. & Lavery, D. (1999) A framework for usability problem extraction. In M. A. Sasse & C. Johnson (Eds.) *Proc Interact 1999*. 344-352.
6. Gray, W. D. & Salzman, M. C. (1998) Damaged Merchandise? A Review of Experiments that Compare Usability Evaluation Methods. In *Human-Computer Interaction* 13(3), pp. 203-261.
7. Green, T. R. G. (1989) Cognitive dimensions of notations. In A. Sutcliffe and L. Macaulay (Eds.) *People and Computers V*. CUP 443-460
8. Green, T. R. G., Blandford, A., Church, L., Roast, C. & Clarke, S. (2006) Cognitive Dimensions: achievements, new directions, and open questions. *Journal of Visual Languages and Computation*. 17.4, 328-365.
9. Hartson, H. R., Andre, T. S., & Williges, R. C. (2001) Evaluating usability evaluation methods. *International Journal of Human-Computer Interaction* 13 (4), 373-410.
10. Hertzum, M., and Jacobsen, N.E. (2001) The Evaluator Effect: A Chilling Fact about Usability Evaluation Methods. *International Journal of Human-Computer Interaction*, 13(4), 421-443.
11. John, B. & Marks, S. (1997) Tracking the effectiveness of usability evaluation methods. *Behaviour and Information Technology* 16, No. 4/5, 188-202.
12. Lee, W.-O., Dye, K., & Airth, D. (1995) Evaluating design specification using heuristic evaluation. In K. Nordby, P. Helmersen, D. Gilmore & S. Arnesen (Eds.) *Proc. Interact 1995*. 376-379.
13. Lavery, D., Cockton, G. & Atkinson, M. P. (1997) Comparison of evaluation methods using structured usability problem reports. *Behaviour and Information Technology*, 16 (4/5), 246-266.
14. Nørgaard, M. and Hornbæk, K. (2006) What do usability evaluators do in practice?: an explorative study of think-aloud testing. In Proc. ACM Conference on Designing interactive Systems (DIS'06). 209-218.
15. Papatzani, G., Curzon, P. & Blandford, A. (in press) Identifying Phenotypes & Genotypes: A Case Study. Evaluating an In-car Navigation System. To appear in *Proc. Engineering Interactive Systems 2007*. LNCS.
16. Wixon, D. (2003) Evaluating Usability Methods; Why the Current Literature Fails the Practitioner. *Interactions* July and August 2003. 28-34.