Designing Evaluations to Evaluate Designs

Janet C Read

Abstract

This paper was motivated by a need to describe evaluation methods that could be used when designing interactive products for children. It begins with an overview of common HCI evaluation methods and describes how these are typically presented in HCI texts. Discussion then follows on the determinants of an evaluation strategy and this discussion is illustrated with examples. A framework for designing evaluations is proposed and discussed. The paper concludes with an evaluation of the framework.

1 Evaluation in HCI

It is difficult to imagine that an interactive product could be designed and built without any user-focussed evaluations taking place. It is not possible to pick up a HCI textbook that does not include at least a chapter on the evaluation of user interfaces. The way in which these usability evaluation methods are presented and classified varies between authors. In 1994, Nielsen claimed that there were four basic ways of evaluating user interfaces; these being

- Automatically
- Empirically
- Formally
- Informally

He went on to suggest that automatic and formal evaluations were both problematic, and suggested that only empirical and informal methods were really useful (Nielsen 1994). This has become the accepted viewpoint in the usability community. Usability Evaluation Methods can generally be described as either empirical or informal using Nielsen's words. The following table shows a list of Usability Evaluation methods, categorised in this way.

Empirical Methods	**Informal Evaluations**
User Walkthrough	Heuristic Evaluation
Focus groups	Expert reviews
Structured observations	Cognitive walkthroughs
Cooperative evaluations	Predictive modelling – GOMS
Activity logging	Guidelines review
Data logging	Consistency inspection
Observations	Critical event analysis
Questionnaires	Dialogue Error Analysis
Interviews	Usability testing
Controlled user tests	
Physiological data analysis	

Different authors divide evaluation in different ways. In User Interface Design, they are broken into formative and summative methods (Le Peuple and Scane 2003). Formative evaluations take place early in the design process and the results of these evaluations inform the later design. (Shneiderman 1998)divides evaluation by considering who the evaluator is. This results in expert reviews and novice reviews. (Faulkner 1998) compares analytical and statistical methods. In John Carrolls design,, the issues of evaluation of learner centred interfaces is touched on.

2 How Evaluation methods are evaluated

Given the large number of methods available, and the many different ways of classifying them, it can be difficult to know which methods should be used. There is an abundance of literature on the evaluation of evaluation methods.

Work by (Nielsen and Phillips 1993) focussed on GOMS, Heuristics and User testing. These were used with three different views of the design, cold, warm and hot. The cold testing took place with just a written specification, the warm test was carried out with a limited time on a prototype (about an hour), and for the hot test, testers could play on the prototype for as long as they wanted. The user testing activity was only used with the end product. It was a within subjects design and subjects were allowed to practice until they plateau-ed. They had a sequence of tasks to do during which error messages and feedback were given and this activity was followed by a subjective satisfaction questionnaire. The findings were that the estimates between experts doing GOMS and Heuristic evaluations were very varied. This suggested that it is best to not rely on a single inspection. Other authors have made similar observations. (REFS) The cost of a heuristic evaluation with a hot view of the design was costlier than the user test and it was remarked that unless the product is unstable, or users are unavailable, the user test was preferred at this stage. (Savage 1996) compared expert reviews, user reviews and user testing. In this study, expert reviews were defined to be inspection methods carried out by human factors specialists. These included heuristic evaluations, cognitive and pluralistic walkthroughs, and consistency and standards inspections. The usability tests were conducted in a role-play interaction using a talk aloud session. User reviews involved potential end users in viewing slide shows of the product and completing questionnaires and engaging in group discussion. Results from this study were that expert reviews tended to inform user interface issues that needed more research with end users; the other two methods flagged up design issues. Heuristic evaluations and user tests have been shown in some studies to identify discrete sets of usability problems (Law and Hvannberg 2002).

In industry, user testing, where users are brought into a lab and asked to think aloud while performing tasks, and are later questioned about their experience of the software, is the most widely used technique (Nielsen and Mack 1994). Cognitive walkthroughs (Wharton, Rieman et al. 1994), Heuristic Evaluations (Nielsen 1994) and GOMS (Card, Moran et al. 1983) are all more economical as they do not require running a prototype or actual users. Empirical methods rely on the availability of real users to test the interface, whereas informal evaluations rely on the skill and experience of the evaluator. The recommended number of evaluators for a heuristic evaluation is 3 – 5 and Nielsen has claimed that five subjects are enough for a usability test as well. In some instances where users are scarce, users may need to be saved for a user test, thus forcing the need for expert reviews. (Nielsen 1994)

However, (John 1996) points out that there are many questions that remain about all these techniques including

- Is the technique real world enough?
- Does it find faults and fixes, or just faults?
- When in the development process should it be used?
- How does it fare in a cost benefit analysis?
- How can techniques be best used together?
- What are the costs of learning a new technique?

3 Designing an Evaluation Strategy

Determinants (Shneiderman 1998)

- Stage of design
- Novelty of project
- Number of expected users
- Criticality of the interface
- Costs of product and finances available for testing
- Time available
- Experience of the design and evaluation team

It is possible to simplify these into four stages; these being.

- Purpose of the product – users would be defined here
- Availability of resources
- Stage of the project
- Purpose of the evaluation

These dimensions are described in the following section.

3.1 The Purpose of the Product

This relates to the use to which the product will be put.

PURPOSE	**KEY VALUE**
Instruct	Learning
Inform	Accessibility
Entertain	Fun
Enable	Ease of Use

Some products will have multiple purposes; the evaluation needs to cover all primary purposes and probably some secondary purposes. For example, edutainment products have two primary purposes. Most products have one primary purpose. At this stage it is also necessary to determine whom the users will be. This is placed first in the list of stages as it is considered that this is not likely to change over a product design lifecycle.

3.2 Availability of Resource

To be considered is the availability of

- Time
- Money
- Hardware
- Expertise
- Users

This is 5 dimensional – but again, the availability of each is unlikely to change over a products lifecycle. If there is a shortage of any of these resources, the evaluation will have to take this into account. These five dimensions could be scored as Unlimited, More than enough, Sufficient, Shortfall, None. The usefulness of these terms is discussed later in this paper.

3.3 Stage of Project

The project will be at one of a number of stages; it may be

- An idea
- A design
- An early prototype
- A fully functional prototype
- A final product

We would expect this to change over a product design lifecycle, particularly if a prototyping approach to design is being used.

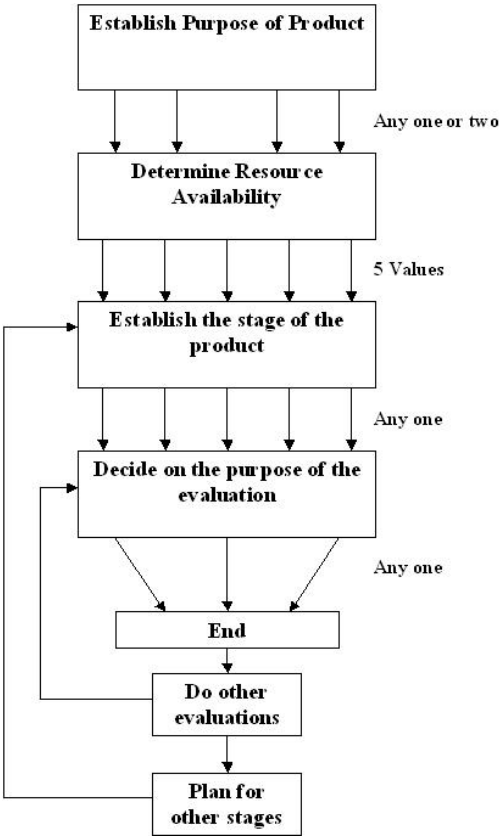
3.4 Purpose of Evaluation

The Evaluation will have its' own purpose, it may be to

- Predict problems
- Discover problems
- Evaluate against another product

This may change over a product design lifecycle. It is possible that the evaluator may want one evaluation to cover all three aspects.

4 Using the framework



4.1 A student project to build multimedia learning software for KS2 children

The product would have two purposes, primary to educate, and secondary to entertain. The users would be children aged between 7 and 11.

Resource	Amount	Indicates
Time	Sufficient	User test okay
Money	None	Can't pay anyone
Hardware	Sufficient	Only one version
Expertise	Shortfall	Struggle to get 5 for expert reviews
Users	Sufficient	Can't over use

Stage of product = Design
Purpose of evaluation = Predict problems
Eval 1 – focus group

Stage of product = Early prototype
Purpose of evaluation = Predict problems
Eval 2 – HE

Stage of product = Late prototype
Purpose of evaluation = Predict problems
Eval 3 - CW

Stage of product = Late prototype
Purpose of evaluation = Discover problems
Eval 4 - UT

Stage of product = Fully functional product
Purpose of evaluation = Discover problems
Eval 5 – UT (different users)

4.2 A university project to build a web front end to banner

The product would have one purpose, primary to enable. The users would be university staff, conversant with the terminology and IT literate.

Resource	Amount	Indicates
Time	Limited	User test needs to be efficient
Money	Sufficient	Possibility of paying users to test / buying in experts
Hardware	More than enough	Can test simultaneously
Expertise	Shortfall	As there is money – this

		can be bought in
Users	More than enough	May decide to select a sample

Stage of product = Idea
Purpose of evaluation = Predict problems
Eval 1 – survey

Stage of product = Design
Purpose of evaluation = Predict problems
Eval 1 – focus group

Stage of product = Early prototype
Purpose of evaluation = Predict problems
Eval 2 – UT

Stage of product = Late prototype
Purpose of evaluation = Predict problems
Eval 3 - HE

Stage of product = Late prototype
Purpose of evaluation = Discover problems
Eval 4 - UT

Stage of product = Fully functional product
Purpose of evaluation = Discover problems
Eval 5 – UT

4.3 Discussion

These two examples serve to demonstrate some of the problems with the framework as it stands. The first is that there appears to be almost a one- one mapping between the stage of product and the purpose of the evaluation. There also appears to be an almost one to one mapping between the purpose of the evaluation and the evaluation selected. This leads us to believe that either the choice of evaluation method cannot be made on the basis of the framework or that the linear approach is wrong. The author believes that it is the latter case, and that the framework is actually n-dimensional and can only lead the user to an evaluation solution by following an n-dimensional path through options.
There is a problem with the usage of words like sufficient when assessing resources. When completing the above cases, the author could not envisage any scenario in which hardware might be insufficient. The terms here need to be revisited. In the second example, there is the pairing between insufficient expertise and plenty of money; the second can be used to cancel out the effects of the first. Had there been no money, the evaluation could not have included any expert reviews. These combinations are important. It is the case that the column labelled ‘indicates’ can only be completed by scanning the availability of all the resources.

5 Conclusion

This paper begins the process of describing evaluations, it provides a framework that can be used for further discussion, and the author hopes that this framework may be further developed to provide some pointers for how the user test or the expert review should be tailored to specific needs. Currently the framework fails to address this. It also fails to specify those evaluations that may not be about usability, for instance, evaluating fun.

Future work on this will attempt to address these issues. Contributions from interested parties are welcome!

6 References

Card, S., K, T. P. Moran, et al. (1983). The psychology of Human Computer Interaction. Hillsdale, New Jersey, Lawrence Erlbaum Associates.

Faulkner, C. (1998). The Essence of Human-Computer Interaction. Essex, Prentice Hall.

John, B. E. (1996). "Evaluating Usability Evaluation Techniques." ACM Computing Surveys **28**(4es).

Law, L.-C. and E. T. Hvannberg (2002). Complementarity and Convergence of Heuristic Evaluation and Usability Test: A Case Study of UNIVERSAL Brokerage Platform. NordiChi, Aarhus, Denmark, ACM.

Le Peuple, J. and R. Scane (2003). User Interface Design. Exeter, Crucial.

Nielsen, J. (1994). Heuristic Evaluation. Usability Inspection Methods. J. Nielsen and R. L. Mack, John Wiley.

Nielsen, J. (1994). Usability Inspection Methods. CHI '94, Boston, Mass, ACM.

Nielsen, J. and R. L. Mack (1994). Usability Inspection Methods. New York, John Wiley.

Nielsen, J. and V. L. Phillips (1993). Estimating the Relative Usability of Two Interfaces: Heuristic, Formal, and Empirical Methods Compared. InterChi'93.

Savage, P. (1996). User Interface Evaluation in an Iterative Design Process: A Comparison of Three Techniques. CHI '96, Vancouver, Canade, ACM.

Shneiderman, B. (1998). Designing the User Interface. Reading, MA, Addison Wesley Longman.

Wharton, C., J. Rieman, et al. (1994). The Cognitive Walkthrough Method: A Practitioner's Guide. Usability Inspection Methods. J. Nielsen and R. L. Mack. New York, John Wiley.