Published in Proceedings of Engineering Human Computer Interaction, EHCI’95, Grand Targhee, USA, August 1995, Chapman&Hall Publ

Is Usability Testing for Specialists Only?

J. Coutaz
CLIPS-IMAG
BP 53, 38041Grenoble cedex 9, France
email: Joelle.Coutaz@imag.fr

Participants

Christopher Ahlberg, Joëlle Coutaz, Christophe Furet, Peter Jensen, Joergen Steensgard-Madsen.

1 MOTIVATIONS

In recent years, usability testing has been the subject of growing interest, resulting in an explosion of methods and techniques. Although the HCI scientific and industrial communities acknowledge the merits of user interface evaluation, how to proceed in practice is still unclear. Usability testing is considered to be time consuming, unreliable, limited in scope, and based on expert experience. This reputation will continue spreading as long as no framework is made available to non specialists for selecting evaluation methods in a secure and effective way.

The purpose of this workshop was to analyze the domain of user interface evaluation techniques and to define a taxonomy that would help evaluators to select a set of methods that fits their requirements. To do so, we looked at well-known taxonomies and proposed our own dimensions that would characterize usability testing requirements. This reflexion was based on (Coutaz 1994) but a similar analysis can be found in (Preece 1994). A case study drawn from the proceedings of EHCI'95 is used to illustrate the discussion.

2 TAXONOMIES FOR USABILITY EVALUATION METHODS

Taxonomies for usability evaluation commonly oppose prediction to experimentation, as well as formative to summative approaches. Alternatively, one can understand the purpose and coverage of an evaluation method based on its scientific foundation.

Published in Proceedings of Engineering Human Computer Interaction, EHCI'95, Grand Targhee, USA, August 1995, Chapman&Hall Publ

2.1 Predictive vs. Experimental Methods

Predictive evaluation techniques allow the detection of usability problems without testing the system with the users. As shown in Figure 1, predictive approaches may be based on a theory: the GOMS (Card 1983) and TAG families (Payne 1986), PUM (Young 1990) and ICS (Barnard 1987), are good examples of theory-based evaluation methods. Alternatively, HCI heuristics may be used to evaluate a user interface without user testing (Nielsen 1989, Nielsen 1990).

Experimental approaches, on the other hand, involve end users participation. Typical illustrations of experimental approaches include: interviewing the user, observing and monitoring users' behavior using various supports (e.g., usability lab, paper and pencil, wizard of Oz apparatus) combined to multiple complementary techniques (e.g., measurement tests, thinking aloud, etc.).

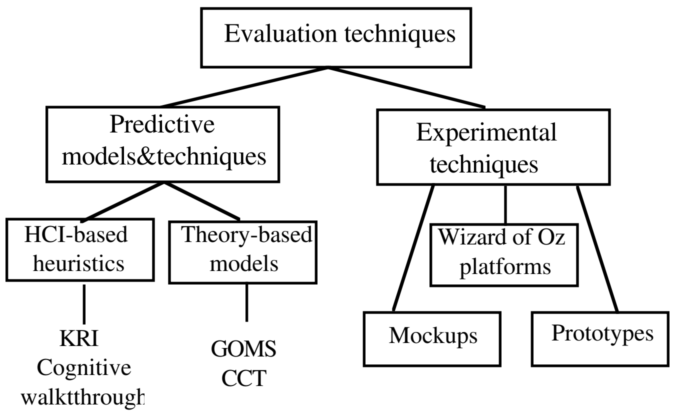

Figure 1: An overview of HCI evaluation techniques based on the Predictive/Experimental distinction.

Whereas predictive and experimental evaluation methods diverge by the absence or the involvement of users, the formative vs. summative perspective stresses a difference in process and goal.

2.2 Formative vs. Summative Methods

We define formative and summative following Nielsen (Nielsen 1994): "Formative evaluation is done in order to help improve the interface as part of an iterative design process. The main goal of formative evaluation is thus to learn which detailed aspects of the interface are good and bad, and how the design can be improved". "In contrast, summative evaluation aims at assessing the overall quality of an interface, for example, for use in deciding between two alternatives or as part of competitive analysis to learn how good the competition really is."

Basically, formative evaluation occurs all the way through the design process and aims at a detailed analysis of the usability of the system. In contrast, summative evaluation takes place at the end of the development process and aims at assessing the conformity of a finished product against a reference model (e.g., an in-house standard or software from competitors).

2.3 Scientific origin

A. Holyer proposes to divide the field of evaluation in terms of the disciplines from which each method is derived (Holyer 1993):

- cognitive psychology approaches (typically the GOMS and TAG families based on a recursive decomposition of user's goals),

- social psychology methods (e.g., structured interviews and questionnaires as illustrated by the SUMI method (Ravden 1989)),
- the social science methods (e.g., talk aloud protocols and its extension: cooperative evaluation, focus groups as used by market researchers),
- engineering approaches inspired from software engineering practice (e.g., heuristic evaluation).

2.4 Discussion

Holyer's taxonomy provides an interesting tool for viewing the whole field of evaluation. His framework however does not provide any answer to our initial question: "Is usability testing for specialists only?".

We observe that the prediction/experimentation dimension is orthogonal to the quality of being formative or summative. In theory, one can conduct a summative evaluation to compare two systems using either an experimental or a predictive technique or even a combination of both. In practice, to make a sound comparative analysis between the two design solutions, it is desirable that the selected method(s) deliver(s) objective quantitative data. This simple example shows that the predictive/experimental as well as the formative/summative distinctions provide two interesting dimensions for choosing an evaluation technique but they need to be refined in terms more tractable by the non specialists to rationalize the choice. This is the purpose of the next section.

3 A REQUIREMENT-CENTERED TAXONOMY

The taxonomy we propose should be able to solve the problem illustrated in Figure 2: given a set of requirements that characterize the evaluation context, elicit the set of candidate methods that satisfy these requirements.

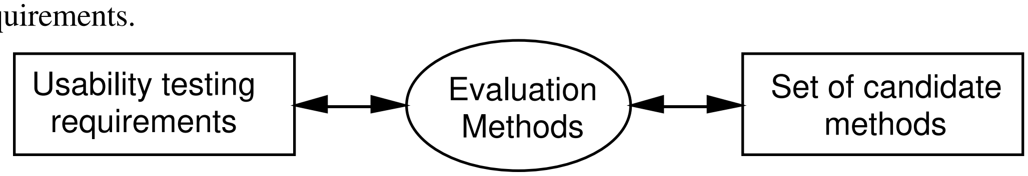

Figure 2: Selecting evaluation methods from a set of specified requirements.

The problem now is to define the dimensions that express the requirements in a satisfactory way. Following software engineering practice, we define the set of evaluation requirements R as the t-uple:

$$R = < (r_1,w_1),\ (r_2, w_2) \ldots\ (r_i, w_i) \ldots\ldots (r_n, w_n)>$$

- where r_i denotes a requirement along the ith dimension of our taxonomy,
- and where w_i is a weight that expresses the relative importance of r_i within the current evaluation context.

Similarly, we characterize an evaluation method m along the same dimensions:

$$M= <m_1, m_2, \ldots, m_i, \ldots m_n>$$

where m_i denotes what the method M requires or proposes along the ith dimension of our taxonomy.

Selecting a set of methods boils down to a matching problem between R and every existing method M. However, it may be the case that the specified evaluation requirements are too demanding and thus no method applies (e.g., performing detailed usability testing in one day without experienced evaluators). Recourse to compromise is then unavoidable. The weight mechanism can be used to make compromises based on a sound rationale.

For the purpose of the presentation, we structure the n dimensions of our framework into 4 categories: human and physical resources, knowledge, results. These categories are now analyzed successively from the perspective of the development team. They could also be presented from the methods perspective. For example, focussing on the development team, we would stipulate that one evaluator expert is made available for conducting the evaluation of a particular system. Alternatively, from the method perspective, we would say that a particular heuristic method requires one evaluator expert at least. Thus, the dimensions we propose define a framework useful for specifying usability testing requirements as well as to characterize a particular evaluation method.

3.1 Human Resources

Human resources cover the set of actors involved directly or indirectly in the evaluation process: the subjects, the evaluators, other relevant actors such as the developers, the designers, the client, the manager, etc.

- The subjects can be characterized in terms of number, their level of expertise, whether they are the real end-users or representative, whether they can be recruited locally within the development team or externally (this latter condition has some implication on recruitment costs).

- The evaluators are defined by their number and their expertise in usability testing, human factors, cognitive psychology, etc.

- The other actors are identified by their culture in HCI and more generally in software quality.

Thus for example, if there is no subject available, then only predictive approaches apply. If one chooses a heuristic-based approach and requires reliable results, then three to five HCI specialists must be brought in to detect a reasonable amount of usability problems (Nielsen 1990, Pollier 1991). Without any subject nor any evaluation expert, then theory-based techniques might be called in. If the other actors are not convinced by usabilility testing then a light-weight approach should be elicited. As observed by Brian Shackel, evaluators have to "earn the right"! (Shackel 1991)

When subjects are available, an experimental approach may be considered. However, the reliability of the collected data will depend on the representativity of the selected subjects. If you don't have access to the real end-users, then recruiting the representative subjects should be guided by an explicit model of the typical end-user executing typical tasks described as scenarios (Valentin 1993). If you do not have experience in user modelling nor in scenarios specifications, then you cannot expect recruiting a significant group of subjects and the reliability of the observed behavioral data is questionable.

3.2 Physical Resources

Physical resources cover all of the physical means that are made available for conducting usability testing. They subdivide into three categories: the physical apparatus, temporal constraints, and financial support.

- The physical apparatus covers the physical means for collecting information during the evaluation: paper and pencil (as in interviews or cognitive walkthroughs (Lewis 1990)), audio and video recording (for monitoring user's behavior), usability lab, wizard of Oz settings, computer supported observation tools such as NEIMO (Salber 1993), computer supported critics, etc.

- temporal constraints include the total amount of days or hours that can be allocated to usability testing as well as its duration through the development process: typically, a total of 5 days performed in a row has not the same implications on the evaluation process as 5 days distributed over two months,

- financial support is not orthogonal to the other dimensions but is a consequence of all of them. In practice, it is a strong explicit constraint from which evaluators must work out the other requirements.

3.3 Input Information

Input information denotes the data used to feed the evaluation technique. It can be described along three dimensions: the object to be evaluated, accessory data, and the expliciteness of the input descriptions.

- The object to be evaluated reflects the reason for doing the evaluation. Typically, the object is a final or b-product, a partial prototype, a mock-up, a story board, or the external specifications of the system. In addition to the user interface per se, evaluation can be performed for the tools used to do the evaluation: the user model, the activity and task models, etc. As mentioned above, one may need to tune the scenarios and the user model before recruiting representative subjects.

- Orthogonal to the evaluated object, accessory descriptions may be provided to run the evaluation properly. These include an explicit user model, a task model, the description of a set of scenarios, etc. For example, if one has specified a description of the end-user's knowledge, then a PUM analysis can be considered.

- The object to be evaluated as well as accessory inputs need to be described in some way. The expliciteness, completeness, the level of details as well as the notation used for the description are important issues to consider. Typically, non formal and loose descriptions are not compatible with predictive theory-based approaches.

Having considered input, we need now to analyse the nature of the information expected from the evaluation techniques.

3.4 Output Information

Output expected from an evaluation technique can be described along 3 axes: type, quality, and level of abstraction.

- The type of the results are classified along the following general lines:

- quantitative (i.e., numerical data such as the time required to perform a task) vs. qualitative (non numerical data such as the list of usability problems),
- objective (i.e., an observed measure) as opposed to subjective (such as the user's expression of his own satisfaction),
- predictive vs. explicative (i.e., allows to understand the cause of the usability problem) vs. corrective (when the method can propose an alternative design to the solution).

Thus, if one wants to assess some aspects of usability against the criteria of a quality plan without involving any subject, then quantitative, objective, and predictive data are strongly desirable. The quantitative/qualitative, objective/subjective as well as the predictive/explicative/corrective dimensions are general ways of characterizing evaluation products. In some circumstances, one may need a more detailed categorization of the expected results. Then, a fourth dimension should be considered to make explicit the very nature of the desired results. These include cognitive overload, time to execute a task, number of errors and degree of severity, and more generally, properties such as those described in (Salber 1994).

- Orthogonal to the notion of type, the quality of the results is also relevant to the selection of an evaluation method. We have refined "result quality" with the following dimensions:
- reliability: sensitivity to hypotheses (such as the perfect user in GOMS-based approaches, ecological validity of the experiment including the representativity of the subjects as well as the location of the experiment (i.e., in the lab as opposed to observation in situ),
- inspectability of the results (i.e., how easy is it to get to the data obtained).

- The level of abstraction of the expected results may range from keystroke up to task level data. Thus, if one needs to study cognitive overload due to task interleaving, keystroke level data may not be adequate in terms of concision to reason at a higher level of abstraction.

3.5 The purpose of the evaluation

The very first question evaluators have to address is the purpose of the evaluation. As discussed in (Preece 1991), there are multiple reasons for doing evaluations. These include: understanding the real world before specifying the requirements for a new product, comparing multiple design solutions in the early phases of the development process, assessing the usability of a system against the criteria and metrics expressed in a quality plan, etc.

More precisely, the purpose of evaluation depends on the following situational factors:

- What kind of product are we developing? Is it a mass market software as opposed to a dedicated system? Is it a prospective system as opposed to a new version of an already wide-spread system? In other words:
- what is the life span expectancy of the system?
- will the system be used frequently by specialists users as opposed to infrequent use by novice users?
- how many users there will be?
- is the system already being used?
- is security the driving force?

- What kind of user interface are we addressing? Is it based on known style-guides such as GUI or does it support new styles of interaction such as multimodality?

Answers to these questions condition the definition of the software development process, as well as the requirements for usability testing at every stage of the development. These requirements can be expressed according to the framework presented above in order to rationalize the selection of a set of evaluation methods. As an illustration, we now present our case study.

4 A CASE STUDY

Several partners including five large French municipalities such as Lyon, Lille and Marseille joined together to design and develop a new interactive system dedicated to the design and management of urban sewage networks. The system is complex and implements a large variety of services (e.g., data input and output, simulation, computation, graphical representations, etc.). Users' domain expertise is heterogeneous. In Marseille, the users are often domain experts while in Lille, most of them are beginners. To alleviate this discrepancy, it has been decided to augment the basic services of the system with a powerful tool, the Assistant 0+, to support the task of the non-expert users (Furet 1995).

The Assistant 0+ is designed to provide the users with a cognitive assistance at the goal and task levels. It assists the users in accomplishing their own goals by checking and calling all of the necessary functions of the system.

A b-version of the whole system will be available in February 1996. At this point, the designers and implementers will need to evaluate the system to answer two key questions in order to revise the b-version:

- Is the system more usable than the system used in the past?
- Is the Assistant 0+ effectively helpful?

Two persons with no previous experience in usability testing will perform the evaluation. As for physical resources, they will have at least one video camera and one VCR available but they cannot expect a usability lab. They will have access to real end-users working in their day-to-day environment. Each municipality will provide 3 to 5 users per site. The project managers are aware of the importance of usability testing. Time allowed can be important but should not exceed one day every two weeks per user.

Using our framework, the usability testing requirements for this system can be expressed in the following way :
<
Human resources = (3-5 subjects, effective end-users, locally recruited, different levels of domain expertise) (2 evaluators, no expertise in usability testing) (managers opened to usability testing)

Physical resources = (pen&pencil, camera, VCR) (1 day every two weeks) (no particular financial constraints)

Input information = (The object of the evaluation is a running user interface) (accessory data is an informal approximate description of a user model and possibly a set of typical scenarios)

Output information = (Since the purpose of the evaluation is to compare the b-version with a previous system as well as to compare the b-version with and without Assistant 0+, both quantitative and qualitative data are desirable whether they are objective or subjective. The level of abstraction of the results is left opened although the intra comparison of the system running the

assistant or not favors task level data. The reliability of the results is not a strong requirement since we have to deal with a prospective system)
>

The driving criteria among the usability testing requirements for the case study is that the evaluators are not HCI experts. On the other hand, they have access to real-end users and are not put under time pressure. Thus, an experimental approach can be used provided the method structures the evaluation process. If we had described the set of existing methods along our dimensions, then the "matching algorithm" would probably suggest the SUMI method (Ravden 1989) as a good candidate to consider.

5 CONCLUSION

In this paper we have presented some preliminary results of a two-day workshop on how to elicit a set of evaluation techniques that would satisfy a set of usability testing requirements and constraints.

As an extension to this reflexion, we suggest the following actions:

- testing the framework against a number of case studies,
- describing the most relevant evaluation techniques in terms of the dimensions of the framework,
- developing a software tool (e.g., an expert-like system) that would suggest a set of evaluation methods that would satisfy specified usability testing requirements and constraints.

BIBLIOGRAPHY

Barnard P. (1987) "Cognitive Resources and the Learning of Human-Computer Dialogs", in Interfacing Thought, edited by J.M. Carroll, The MIT Press, pp. 112-158

Card S.K., Moran T.P. & Newell A. (1993) "*The psychology of Human Computer Interaction*", Lawrence Erlbaum Associates, 1983

Coutaz J. and Balbo S. (1994)*"Evaluation des Interfaces Utilisateur : Taxonomie et Recommandation"*, Actes du colloque IHM'94.

Furet C. (1995) *"An interactive environment based on a task model"*, in this volume.

Holyer A. (1993) *"Methods for Evaluating User Interfaces"*, School of Cognitive and Computing Sciences, University of Sussex, Brighton, andy@cogs.susx.ac.uk.

Lewis C., Polson P., Wharton C. & Rieman J. (1990) "Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces", in Proceedings CHI'90, ACM, pp. 235-242.

Nielsen J. (1989) "Usability Engineering at a Discount", in Designing and Using Human Computer Interfaces and Knowledge-Based Systems, Salvendy & Smith Eds., Elsevier North Holland, pp. 394-401

Nielsen J. and Molich R. (1990) "Heuristic evaluation of user interfaces", Proceedings of the CHI'90 Conference on Computer Human Interaction, Seatle, ACM New York, pp. 349-256,

Nielsen, J. (1994) *"Usability Engineering"* in *Notes from "User Interface Strategies 94"*, University of Maryland at College Park, pp. 36-55.

Payne S.J. and Green T.R.G. (1986) "Task-Action Grammars: a model of the mental representation of task languages", in Human-Computer Interaction, Vol. 2, No. 2, pp. 93-133, Lawrence Erlbaum Associates Publishers.

Pollier A. (1991) "*Evaluation d'une interface par des ergonomes : diagnostics et stratégies*", Rapport de recherche INRIA no 1391, Février 1991

Preece J., Rogers Y., Sharp H., Benyon D., Holland S. and Carey T. (1994) *"Human Computer Interaction"*, Addison Wesley Publ.

Ravden S. and Johnson G. (1989) *"Evaluating Usability of Human-Computer Interfaces"*, Ellis Horwood Books in Information Technology Publ.

Salber D. and Coutaz J. (1993) "Applying the Wizard of Oz Technique to the Study of Multimodal Systems", East-West HCI conference, Moscow, August 1993

Salber D., Coutaz J., Nigay L., Faconti G., Paterno' F., Duke D. and Harrison H. (1994) *"The System Modelling Glossary"*, Amodeus project, SM/WP 26.

Shackel B. (1991) Interview with Brian Shackel, in Preece J., Rogers Y., Sharp H., Benyon D., Holland S. and Carey T. (1994) *"Human Computer Interaction"*, Addison Wesley Publ., pp. 599-600

Valentin A., Vallery G. and Lucongsang R. (1993) *"L'évaluation ergonomique des logiciels, une démarche itérative de conception"*, ANACT, Colection Outils et Méthode, 151 pages.

Young R.M. and Whittington J. (1990) "A knowledge Analysis of interactivity", Proceedings of INTERACT'90, edited by D. Diaper, G. Cockton & B. Shackel, Elsevier Scientific Publishers B.V, pp. 207-212, Cambridge (United Kingdom), 27-31st August 1990