

The user as wizard: A method for early involvement in the design and evaluation of adaptive systems

Judith Masthoff

University of Aberdeen, Aberdeen, UK
jmasthof@csd.abdn.ac.uk

Abstract. User testing tends to put participants in the role of the user. To evaluate complex systems (such as adaptive systems), a Wizard-of-Oz study may be used, where the role of the system is made easier by allowing somebody on the design team to perform part of it. In this paper, we propose a method whereby participants take the role of the system, allowing early user involvement in the design process. Participants' actions can inspire the design of adaptive algorithms, and their rationale can inspire evaluation criteria. We illustrate the method with a number of case studies.

1 Introduction

It is a well-known principle in user interface design, that one should involve users early in the design process, rather than only in a final evaluation phase [22]. The same applies to adaptive systems [27]. Whilst there has been a serious lack of empirical evaluation of adaptive systems [6, 13], user involvement in the *design* of adaptive systems has been even more scarcely reported in the literature. There are, however, many early user involvement methods available to the user interface designer (see [12] for a more detailed overview) that may be applicable to the design of adaptive systems, such as:

- *Interviews and Questionnaires*, in which users are questioned about themselves (like how much experience they have), their tasks, what they need, what they like, and what problems they experience.
- *Focus groups*, in which a group of participants explore a set of issues [10]
- *Contextual Design*, an ethnographic method, in which users are observed in their work place, to find out how they go about their work, in what environment, using which artefacts [3]. The idea is that users are the experts in their tasks and that observing them will provide more detail than asking them.
- *Cultural Probes*, in which users are given tools such as a camera, postcards with broad questions (like “what object in your home do you like best”) to use in their own homes, to provide designers with an impressionistic account of their beliefs

and desires, their aesthetic preferences and cultural concerns [7]. The idea is that users may not know how new technology could impact their lives, so instead of asking them questions about the technology, one tries to get an insight into what is important in their lives.

- *Creative brainstorming sessions*, in which a group of participants come up with new ideas for products or interfaces.

The main results of these methods tend to be a good understanding of user characteristics (including user needs and desires), user tasks, context of use, and information architecture. In addition, there is a method for early user involvement in the evaluation of interactive systems:

- *Wizard-of-Oz studies*, in which a human “Wizard” (somebody from the design team) simulates the system’s intelligence and interacts with the user through a real or mock computer interface [8]. This method tends to be used for rapid prototyping when a system is too costly or difficult to build [28]. Note that the wizards, in these studies, tend to follow a precise script (e.g. stating how to react to different user commands in a system that is supposed to use speech recognition).

There has been some use of these methods for adaptive systems: for instance, [2] uses focus groups and creative brainstorming sessions to inspire a recommender system’s user interface; [17] uses Wizard-Of-Oz to prototype an intelligent agent. Contextual Design has been used to inspire Intelligent Tutoring Systems, for instance [1] based their geometry tutor on observations of the strategies employed by teachers.

We are particularly interested in a method for using users to inspire the adaptation algorithms. We agree with the Contextual Design approach that it is better to observe users than to ask them questions, as users’ behaviour is often instinctive and their knowledge tacit [3]. Therefore, asking users how the adaptation algorithms should work may not give the desired effects. However, observing experts in their normal setting (like [1] did for teachers), is not ideal either, as the experts may use background and contextual knowledge (like a student’s facial expressions and deep knowledge of a student’s abilities build up over time) that are not available to a system. Also, such studies would be limited to those situations that happened to occur in the real-world setting (so, outside the control of the system designer). Finally, they would be limited to design of the system as a whole, rather than individual adaptation layers.

In this paper, we describe a method inspired by the Wizard-of-Oz and Contextual Design methods. This method has been partially applied before both by us (e.g., [14, 15, 16] and others (e.g. [21])). The focus of this paper is to make this method more explicit, trying to provide clear guidelines for young researchers. We believe such guidelines are needed for the user-centred design of adaptive systems to mature.

Section 2 will outline the method. Section 3 introduces four case studies of adaptive systems and shows how the method has been applied for each of them. Section 4 draws some conclusions.

2 Users-As-Wizards Method

The main idea of the method is to use participants in the role of the wizard, and leave them completely free to perform the wizard's task, so without giving them a script to follow. Humans tend to be good at adaptation, so, observing them in the role of the wizard may help to design the adaptation. Participants will be given the same information as the system would have. They will be dealing with fictional users rather than real ones. This will allow us to study multiple participants dealing with the same user, to study particular user types, and to control exactly what information the participants get.

The method consists of two stages. In the first stage, participants take the role of the adaptive system (or a component of it). This will teach us how humans perform the task we would like our adaptive system to perform. In the second stage, we consolidate this understanding by having participants judge the performance of others. The case studies presented in Section 3 serve as examples for each of the steps below.

Exploration stage

1. Present participants with a scenario describing a fictional user (or group of users) and the user's intentions (task).

Using fictional users is a well-known technique in user-centred design, where so-called personas are developed as examples of user classes [9]. For instance, a designer may take a fictional person like Bart Simpson (or one they developed themselves) into consideration to decide for instance what requirements a new system for booking cinema tickets should meet. Similarly, scenarios are used extensively [5], which are stories of a persona doing a task (e.g. "Bart Simpson wants to buy cinema tickets on the internet, to go to the movies with his friend. He wants cheap tickets and likes sitting at the back so that he can throw paper airplanes at the people in front. He would prefer to go to a violent, adult certificate movie, but if he cannot get away with that, he will settle for a comedy.") Personas and scenarios are also used in expert evaluations, such as cognitive walkthroughs [25] in which a usability expert walks through an interface based on the correct action sequence for a task specified in a scenario, keeping the persona in mind to decide whether actions are likely to be successes or failures. So, creating these will be beneficial for other stages of the design process as well.

2. Give participants the task the adaptive system is supposed to perform.

For instance, suppose the scenario describes 7 year old Mary visiting a museum, and tells you that Mary likes horses, flowers, and the colour pink. A task given could be to recommend three paintings for Mary to view. There is no need to tell participants that they have to adapt. They will automatically base their recommendations on what they know about Mary.

3. Find out participants' reasons for their decisions and actions.

Participants' rationale is crucial as this reflects (a) what participants found important, providing criteria on which to judge adaptation, and (b) how they went

about the task, providing inspiration for the adaptation algorithm. Different observational methods can be used for this, such as:

- Thinking-a-loud, whereby participants are asked to verbalize their thinking while doing the task [18]. Participants need to be trained in this.
- Co-discovery, whereby participants work together with somebody they know well, and their naturally arising discussion shows their thinking [29]. This method has a clear advantage over thinking-aloud in that it is easier and more natural for participants, and does not require training.
- Retrospective testing using a questionnaire or interview, whereby participants tell you their rationale after each task, possibly while watching a video of their actions. This method is less suitable with longer task durations, as participants may have forgotten why they acted in a certain way.

4. Steps 1 to 3 can be repeated for a number of scenarios.

A within-subject design can be used, in which each participant performs the task for a number of scenarios. In this case, it may be appropriate to randomize the order in which the scenarios are done, to avoid any order effects. Alternatively, a between-subject design can be used, in which the participants are assigned to a condition, and depending on this condition work on a different scenario. A combination is also possible, in which participants are assigned to a condition and work on a different set of scenarios.

A limitation of the exploration phase is that it may be less suitable for tasks that humans are bad at. Basing adaptation algorithms on human performance is only sensible if humans perform well. The consolidation phase will verify the acceptability of the human performance and determine in what respects it can be improved.

Consolidation stage

It is best to use new participants for this stage, rather than reusing those of the exploration stage.

1. Present participants with (a) a scenario involving a fictional user (or group of users) and the user's intentions, and (b) an associated task.

The scenario and task used are typically the same as in the exploration stage.

2. Show the participants a performance on this task for this scenario.

The performance shown can be human performance (as from the exploration stage), or it can be system performance (e.g., using an algorithm based on the exploration stage). Participants are not told the difference: either they are told that all performance is that of a system (or human), or they are not told the origin at all.

3. Ask participants to judge the quality of task performance.

Judgements can be asked on multiple criteria. These criteria may be based on observations in the exploration phase of what participants seemed to find important. They could also be based on what system designers think is important, or what the literature indicates as important.

4. Find out participants' reasons for their judgments.

Similar observational methods can be used as in the exploration phase.

5. Normally, steps 2 to 4 are repeated for a number of task performances.

The order of task performances should be randomized, to prevent an order effect. It is possible to intersperse judgments of human performance with judgements of system performance. Note that this starts to resemble a Turing test [24], in that we could say that our system performs well, if participants judge it as well as they judge human performance. However, depending on the system task, we may want our systems to outperform humans (e.g. when automatically detecting terrorists in a crowd), or may still be satisfied with performance that is below human performance (e.g. when recommending books).

6. Steps 1 to 5 can be repeated for a number of scenarios.

A within- or between-subjects design can be used, or a combination.

A limitation of the consolidation stage is that participants' judgments may not always correspond with what would be best for users. For instance, in a study of a medical reporting system, [11] found that while doctors said they preferred graphs, they actually performed better with texts. For this reason, a normal user test of system performance will still be needed. The User-as-Wizards method is only intended as an initial step in the design process.

It should also be noted, that there are some tasks that are inherently more difficult for humans than for computers. E.g. humans tend to be bad at processing large amounts of data. For such tasks, humans may have difficulty not just deciding how to adapt, but also judging adaptation. The method is less suitable for such cases.

3 Case studies

To make the method more concrete, we will show how it has been applied to a number of adaptive systems. As the main purpose of this section is to illustrate the method, we will not describe the systems or the studies in detail.

3.1 Group recommender system

This system and study are described in detail in [14]. As with all case studies, we will summarize the purpose of the system, what we needed to find out, and how we did the various steps in the method:

- *System purpose.* To select a sequence of items (e.g. music clips) adapted to a group of users.
- *Problem.* How should ratings by individual users be aggregated into ratings for the group as a whole? Many different aggregation strategies existed, and we needed to know how suitable they might be, and how to judge them. There were far too many

for an ordinary user test, and the domain was too complicated for normal user testing anyway [14].

- *Exploration stage.*
 - *Scenario.* “John, Mary, and Adam are going to watch video clips together. We know how interested they are in the topics of the clips. Each clip is rated from 1 - really hate this topic - to 10 - really like this topic. [A table shows each individual’s liking for each of the clips]. They only have time to watch five clips.”
 - *Wizard Task.* “Which clips should they watch?”
 - *Observational method.* Question, namely “Why?”
 - *Repetition.* We used a between-subjects design. Half the participants were given the scenario as above. The other half were given one in which ages had been given for John, Adam and Mary, with Adam having a grandfather like age. This was to investigate whether this would have an influence.
 - *Results.* We found that participants care about avoiding misery and fairness. We did not find that the ages mattered. We also compared their recommendations with those resulting from the aggregation strategies, and based on this, we could restrict the number of aggregation methods to consider.
- *Consolidation stage.*
 - *Scenario.* We used a slight variation on the scenario from the exploration stage, instead of “John, Mary and Adam”, we used “You and two friends (Friend 1 and Friend2)”.
 - *Performance.* “The TV decides to show you the following sequence of clips [a sequence is given].” Performance was that of various group aggregation strategies.
 - *Judgement task.* “How satisfied would you be?, How satisfied do you believe Friend1 would be?, How satisfied do you believe Friend2 would be?” A seven point Likert-scale was used.
 - *Observational Method.* “Why” questions were asked for each judgement question.
 - *Repetition.* A combination of a within- and between-subjects design was used. Each subject was given three performances to judge, but these three varied over subjects.
 - *Results.* The same aggregation strategies that did not match participants’ performance in the exploration stage received negative judgements in this stage, so the findings from the exploration were confirmed. Subjects were better at expressing a rationale in the exploration than in the consolidation stage. Nevertheless, we also found some new criteria, such as a wish to end the sequence on a high. We also got interesting results from a comparison of participants’ judgements with those by predictive models (beyond the scope of this paper).

3.2 Persuasive system to promote wellbeing

Some other researchers have also used the exploration phase. This system and study are described in detail in [21].

- *System goal.* To automatically construct a personalized persuasive message to convince a user to eat more healthily.
- *Problem.* What kind of arguments to use (a distinction is made between emotional and rational arguments), and how to structure the message.
- *Scenario.* Participants were given a story describing a fictional friend, with details about her personality, goals, habits, as well as facts related to healthy eating.
- *Wizard Task.* Construct a message to convince this friend to eat more healthily.
- *Observational method.* None used.
- *Repetition.* They used a between-subject design with four groups, varying the facts mentioned in the scenario (from positive effects on health or appearance due to healthy eating, to negative effects due to unhealthy eating).
- *Results.* They found that subjects mainly used emotional arguments (e.g. trying to get their friend to feel pride or shame) rather than rational ones, and that the recommended activity (of eating more healthily was introduced late in the message, and sometimes even left implicit. They used the results as a basis for constructing belief networks.

3.3 Modelling the effect of emotional contagion on satisfaction

According to layered approaches to the evaluation of adaptive systems [19, 4, 26, 20], different system layers need to be evaluated independently. For instance, one should evaluate the accuracy of the user modeling independently of the performance of the system as a whole. The importance of this has been clearly illustrated by a layered evaluation in [23], which showed that while people were satisfied with the system as a whole, its user modeling was seriously lacking in accuracy. The layered principle should also be applied to user involvement in design: we should involve the user in designing individual layers, not just the system as a whole. In this case study, we looked at part of the user modeling layer.

This system and study are described in detail in [15]. In this case, we only applied the exploration stage of the method.

- *System purpose.* To select a sequence of items (e.g. music clips) adapted to a group of users.
- *Layer purpose.* Modeling the effect of the other users' satisfaction on the satisfaction of an individual user.
- *Problem.* Based on the literature, other people's emotions may influence your own, and this may depend on the type of relationship you have with them (see literature in [15]). How does the relationship type influence emotional contagion?
- *Scenario.* "Think of somebody [who meets some relationship criterion]. Assume you and this person are watching television together. You are enjoying the program a little."
- *Wizard task.* "How would it make you feel to know that the other person is [other's emotion: enjoying it a little or really hating it]?" Five answer categories were

provided, from “decrease a lot” to “increase a lot”. So, the wizard’s task in this case is to perform a little part of the user modeling, namely indicating how the emotions in the user model should change in a certain situation.

- *Observational Method.* None used.
- *Repetition.* We used a within-subject design. We varied the emotions of the imagined other person (happier, and unhappier), and varied the type of relationship of the participant with the imagined other person, using four types.
- *Results.* We found that emotional contagion did indeed happen, and that some relationship types led to more contagion than others. This can be incorporated in the modelling.

3.4 Hierarchy optimisation system

This system and study are described in detail in [16]. In this case, we learned a lot more from the consolidation stage than from the exploration stage. However, the exploration stage was vital to provide the material for the consolidation stage.

- *System goal.* To automatically construct a good personalized hierarchy for items of interest to the user.
- *Problem.* How to group items together in a hierarchy, and what titles to use for the groupings? Literature suggests that balance is important in hierarchies, and that is better for hierarchies to be wide rather than deep.
- *Exploration stage.*
 - *Scenario.* Participants were given a set of items of interest to a user.
 - *Wizard Task.* Construct a suitable textbook hierarchy to contain the items, inventing titles for chapters, sections etc.
 - *Observational method.* Co-discovery.
 - *Repetition.* No repetition was used. Task took about fifty minutes.
 - *Results.* Participants found this a difficult task, and some of the hierarchies produced seemed to have poor groupings and titles. Participants’ hierarchies tended to be balanced in depth. Titles were of a type a system could generate.
- *Consolidation stage.*
 - *Scenario.* No scenario was given.
 - *Performance.* Hierarchies were shown that had been made by the participants of the exploration stage. In addition, three hierarchies were shown that had been generated by an algorithm.
 - *Judgement task.* Participants were asked to what extent they agreed with three statements about the given hierarchy (e.g., “some categories should have been merged”), to judge the category naming, and the hierarchy as a whole. Seven point Likert scales were used.
 - *Observational Method.* Subjects were asked what they disliked most about the hierarchy, with the option to provide three answers.
 - *Results.* We got a very clear idea of the criteria participants’ applied. This led to the definition of a number of additional criteria for a good hierarchy and an improved hierarchy construction algorithm.

4 Conclusions

Adaptive systems can clearly benefit from the many methods available in the Human-Computer Interaction field to involve users in system design and evaluation. In this paper, we have presented and illustrated a method to involve users very early on, to inspire the adaptation algorithms. This method has been inspired by Contextual Design and Wizard-of-Oz. Parts of this method have been used before, but we hope that making the steps involved more explicit will help its adoption in the user-centred design of adaptive systems.

The consolidation stage of this method may also be useful when it is difficult to perform a direct user test. For instance, typically one would like a user test to take no more than an hour. This can make it hard to do a direct evaluation of a system that needs time to adapt to a user. Letting participants judge system performance for a set of fictional users is one way to evaluate, even if it is an indirect way.

Acknowledgements

The author is partly supported by Nuffield Foundation grant NAL/00258/G.

References

1. Anderson, J. R., Boyle, C. F., and Yost, G. (1985). The geometry tutor. *9th International Joint Conference on AI*, pp 1-7.
2. van Barneveld, J. and van Setten, M. (2003). Involving users in the design of user interfaces for TV recommender systems. *3rd Workshop on Personalization in Future TV*, associated with UM03, Johnstown, PA
3. Beyer, H. and Holtzblatt, K. (1997) *Contextual design: Defining customer-centred systems*. Morgan Kaufmann: San Francisco.
4. Brusilovsky, P., Karagiannidis, C., & Sampson, D. G. (2001). The benefits of layered evaluation of adaptive applications and services. In S. Weibelzahl, D. N. Chin, & G. Weber (Eds.), *Proceedings of Empirical Evaluation of Adaptive Systems workshop* associated with UM2001, Sonthofen, Germany, pp1-8
5. Carroll, J. M. (2000). Five Reasons for scenario-based design. *Interacting with Computers*, 13, 43-60.
6. Chin, D. N. (2001). Empirical evaluation of user models and user-adapted systems. *User Modeling and User-Adapted Interaction*, 11(1-2), 181–194.
7. Gaver B., Dunne T., and Pacenti E. (1999). Design: Cultural probes. *Interactions* 6, 1, 21-29
8. Gould, J., Conti, J., Hovanyecz, T. (1982). Composing letters with a simulated listening typewriter. *Proc. ACM CHI'82*, pp. 367-370.
9. Grudin, J. & Pruitt, J. (2002). Personas, participatory design, and product development: An infrastructure for engagement. *Proc. PDC 2002*, 144-161.
10. Krueger, R. and Casey, M. (2000). *Focus groups: A practical guide for applied research*. Sage Publications.
11. Law A., Freer, Y., Hunter, J., Logie, R., McIntosh, N., and Quinn, J. (2005). A comparison of graphical and textual presentations of time series data to support medical

- decision making in the neonatal intensive care unit, *Journal of Clinical Monitoring and Computing*, 19, 183-194
12. Maguire, M. (2001). Methods to support human-centred design. *International Journal Human-Computer Studies*, 55, 587- 634.
 13. Masthoff, J. (2002). The evaluation of adaptive systems. In N. V. Patel (Ed.), *Adaptive evolutionary information systems*. Hershey, PA: Idea Group Publishing.
 14. Masthoff, J. (2004). Group modeling: Selecting a sequence of television items to suit a group of viewers. *User Modeling and User Adapted Interaction*, 14, 37-85
 15. Masthoff, J. and Gatt, A. (in press). *In pursuit of satisfaction and the prevention of embarrassment: Affective state in group recommender systems*. *User Modeling and User Adapted Interaction*.
 16. Masthoff, J. (unpublished manuscript). *Automatically constructing good hierarchies: HCI meets AI*.
 17. Maulsby, D., Greenberg, S. and Mander, R. (1993) Prototyping an intelligent agent through Wizard of Oz. In *ACM SIGCHI Conference on Human Factors in Computing Systems*, Amsterdam, The Netherlands, May, p277-284, ACM Press.
 18. Nielsen, J. (1993). Evaluating the thinking-aloud technique for use by computer scientists. *Advances in human-computer interaction*, 3, 69-82
 19. Paramythis, A., Totter, A., and Stephanidis, C. (2001). A modular approach to the evaluation of adaptive user interfaces. In S. Weibelzahl, D. N. Chin, & G. Weber (Eds.), *Proceedings of Empirical Evaluation of Adaptive Systems workshop* associated with UM2001, Sonthofen, Germany, pp. 9–24.
 20. Paramythis, A. & Weibelzahl, S. (2005). A decomposition model for the layered evaluation of interactive adaptive systems. In: *Proceedings of the 10th International Conference on User Modeling* (Lecture Notes in Computer Science LNAI 3538). Berlin: Springer
 21. de Rosi, F., Mazzotta, Miceli & Poggi (2006). Persuasion artifices to promote wellbeing. *Proceedings of the Persuasive conference*, Eindhoven, pp84-95.
 22. Shneiderman, B. (1998). *Designing the user interface: Strategies for effective human-computer interaction*. Addison Wesley: Reading, MA
 23. Sosnovsky, S. and Brusilovsky, P. (2005). Layered evaluation of topic-based adaptation to student knowledge. In S. Weibelzahl, A. Paramythis, and J. Masthoff (Eds.) *Proceedings of the Fourth Workshop on the Evaluation of Adaptive Systems*, held in conjunction with the 10th International Conference on User Modeling (UM'05), Edinburgh, UK, pp47-56.
 24. Turing, A. (1950). Computing machinery and intelligence. *Mind*, 433-460.
 25. Wharton, C., Rieman, J., Lewis, C., and Polson, P. (1994). The cognitive walkthrough method: A practitioner's guide. In J.Nielsen.and R.L.Mack (Eds.) *Usability Inspection Methods*. John Wiley & Sons, New York, 105-141.
 26. Weibelzahl, S. (2001). Evaluation of adaptive systems. In M. Bauer, P. J. Gmytrasiewicz, & J. Vassileva (Eds.), *User Modeling: Proceedings of the Eighth International Conference, UM2001* Berlin: Springer, pp. 292–294.
 27. Weibelzahl, S. (2005). Problems and pitfalls in evaluating adaptive systems. In S. Weibelzahl, A. Paramythis, and J. Masthoff (Eds.) *Proceedings of the Fourth Workshop on the Evaluation of Adaptive Systems*, held in conjunction with the 10th International Conference on User Modeling (UM'05), Edinburgh, UK, pp57-66.
 28. Wilson, J. and Rosenberg, D. (1988). Rapid prototyping for user interface design. In M. Helander (Ed.), *Handbook of Human-Computer Interaction*, New York, North-Holland. pp. 859-875.
 29. O'Malley, C.E., Draper, S.W., and Riley, M.S. (1984). Constructive interaction: A method for studying human-computer-human interaction. *Proceedings of the IFIP INTERACT'84 First International Conference on Human-Computer Interaction*, pp. 269-274