

Evaluating the Usability of Adaptive Recommendations

Stephan Weibelzahl & Diana Chihaiia

National College of Ireland
Mayor Street, Dublin 1, Ireland
[sweibelzahl,dchihaiia]@ncirl.ie

Abstract

In the design, development and evaluation of adaptive systems, usability aspects are largely ignored so far. A correct but badly designed adaptation decision might spoil the purpose. This paper provides an example of a usability evaluation with an adaptive collaborative system. The results suggest that the representation format of recommendations did not strongly affect standard usability measures. However, comparing different design versions helped users to identify strengths and weaknesses.

1 Usability in Adaptive Systems

The lack of evaluation studies of adaptive systems as been noted frequently for a number of years [Chin, 2001] [Paramythis & Weibelzahl, 2005]. While the volume and quality of studies in the field seems to gradually improve, the effects of usability on adaptation seem to be widely ignored so far.

In fact, most systems reported in the literature remain at the prototype level with often poor usability. From a philosophy of science perspective, this is perfectly acceptable in many cases, as these prototypes are just vehicles to provide proof of concept or to demonstrate the quality of certain modelling or inference approaches as opposed to commercial applications. However, in the long run, the scientific community will have to demonstrate effects and impact of adaptive systems in real or realistic settings in regard to different application areas.

2 Adaptive System Development and Evaluation

We argue that formative evaluation should be an inherent part of any interactive software system development, in particular in the case of adaptive systems.

In order to guide the selection of features and criteria so called Layered Evaluation has been proposed [Brusilovsky, Karagiannidis and Sampson, 2001] [Paramythis, Totter and Stephanidis, 2001] [Weibelzahl, 2001]. The idea is to break down the adaptive system into its logical components and to evaluate these components (respectively their interaction) separately in order to avoid confounding of factors. The following stages with associated criteria have been suggested [Paramythis and Weibelzahl, 2005]: (a) collection of input data, (b) interpretation of the collected data, (c) modelling of the current state of the “world”, (d) deciding upon adaptation, and (e) applying adaptation.

The last two stages are often seen as a unity, but there is a clear rationale for a separation of these stages. While the overall adaptation decision might be appropriate the system might still fail to implement this decision in a suitable way. For example, an adaptive educational system might decide to recommend skipping a section of an on-line course, because the learner has acquired sufficient knowledge about that section. While an explicit textual hint (“It seems that you have sufficient knowledge about this section and you may proceed to the next chapter”) might guide the learners to the next chapter, a more implicit form of recommendation such as link annotation might fail to convey its meaning to the learner. Moreover, link annotation may be implemented in different ways, including the traffic-light metaphor or small icons indicating the state. While all these adaptations rely on the same information they may differ widely in terms of impact on effectiveness, efficiency and usability of a system.

This paper provides an example of an evaluation in regard to this last stage: applying adaptation. We were interested in the effects of different design options in an adaptive collaborative system based on the same adaptation data.

3 The Peer Finder

The work reported in this paper was conducted in the context of a project that aims to develop a set of components to support collaborative learning. One of these components envisaged is an adaptive search for peers. Using the infrastructure provided by the adaptive learning system AHA! [de Bra and Calvi, 1998], the system is able to infer the current knowledge of a learner about concepts covered by a on-line course. The system also models whether learners are in general available to provide help to peers and whether they are on-line (i.e., working on the same course at the moment).

We designed three prototypes of this Peer Finder component that allows learners to identify peers that can provide help. The prototypes differ in terms of the number of recommendations and the functionality available to change or refine this list of recommended peers.

3.1 Design A: Listing & Preferences

The first design option can be seen as a kind of baseline. Initially, five random peers with their state are presented. The learner can list all available peers or all peers who have high knowledge about the course concept (see Figure 1). While no legend of the icons is shown, hovering over the icons reveals the meaning (e.g., “high knowledge”, “available for help”).

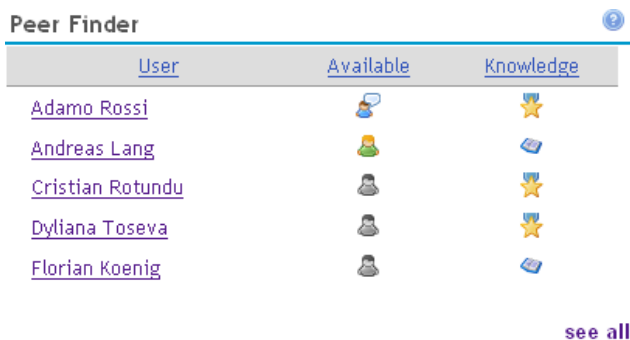


Figure 1: Screenshot of design A (listing & preferences)

The user can also request a list of all classmates along with their contact details (“see all”).

3.2 Design B: Dynamic Refinement

In a second design version, the three criteria availability for help, knowledge and on-line state are made explicit in the form of checkboxes. Changing criteria dynamically refines the list of peers. Only the top 5 peers (according to knowledge) are shown (see Figure 3).

In the same way as in design A, on request (“see all”) a full list of all classmates is provided.

3.3 Design C: Search criteria

The third design variant adopts a traditional search style approach. Initially, only the search criteria are shown. While the functionality is very similar to design B, the layout forces learners to think about and express their criteria. Rather than showing the top 5 peers, a list of all peers that fulfil the set criteria along with their contact details is shown (see Figure 2).

All three designs are based on the same style and colours as well as the same set of icons to identify the states.

4 Method

We were interested in how learners would interact with the different versions and what their experience would be like. For the purpose of this study, functional prototypes of all three versions were placed in the context of an on-line learning environment (Sakai). The homepage for an imaginary course on communications was created, including a brief course description, calendar and “messages of the day”. The class consisted of 22 learners with different states.

Participants were randomly assigned to one of three conditions and completed the following procedure (see Figure 5): They were briefly introduced to the scenario of the on-line course and asked to imagine they were looking for help in regard to the last lecture.

They were then shown one of the three Peer Finder versions embedded in the Sakai learning environment and asked to identify a classmate that they want to contact for help. At this stage they assumed that there is only one version, “the Peer Finder”.

They interacted with the Peer Finder until they were satisfied to have identified a suitable person. During this phase the gaze position was recorded using a remote eye-tracker (SMI RED4).

They completed the SUS usability questionnaire [Brook, 1996] comprising 10 items on different usability aspects. In three open questions they were asked to com-

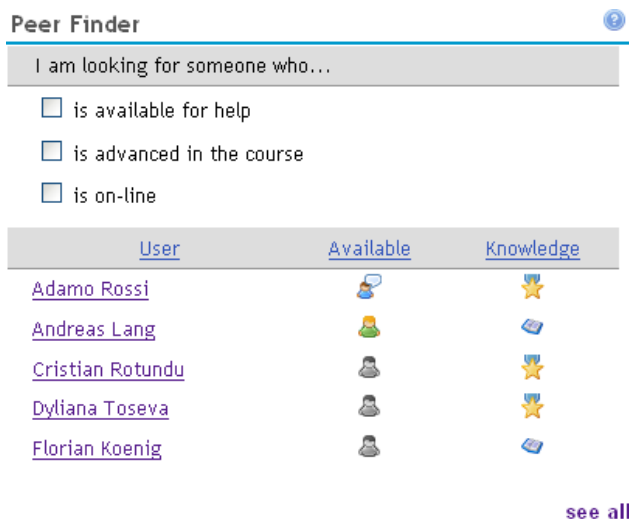


Figure 3: Screenshot of design B (dynamic refinement)

ment on their positive and negative experiences as well as on suggestions how the system could be improved.

Only then they were shown all three versions in parallel and asked to explore them. Again, they completed the SUS for the remaining two designs.

Finally, participants had to rank the three versions according to the subjective preference and provide a reason for their ranking.

5 Results

A total of eight users (four female, four male) with various backgrounds (degree to PhD level, different disciplines and professions) completed the procedure. Participants could easily identify with the scenario and grasped the functionality of all three versions very quickly. Only one person stated that the scenario could have been more detailed, but nevertheless succeeded in finding a suitable peer who might help.

5.1 Usage

The total procedure took about 15 minutes. The interaction with the Peer Finder ranged between 20 seconds and 2 minutes per design. Almost all participants explored the functionality for a while (e.g., changing criteria), before selecting a peer.

In the case of the design A, two participants revised the list by selecting the first two columns (user and availability status) and verbally declared that they understood the meaning of each column and icons.

In the case of design B, two participants explored the

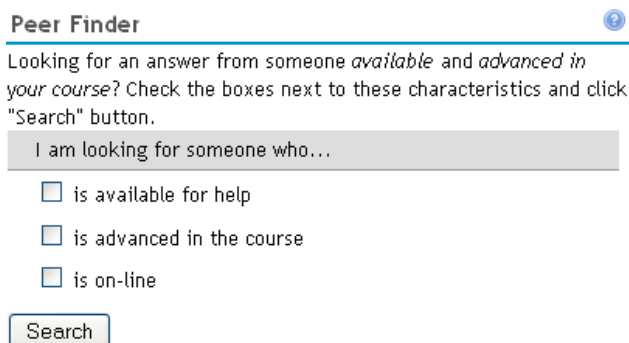


Figure 2: Screenshot of design C (search criteria)

1. Introduction of scenario
2. Exploration and use of one of the three versions
3. SUS usability questionnaire for first explored design
4. Open ended feedback questions
5. Exploration of the two other versions
6. SUS usability questionnaire for remaining two designs
7. Ranking

Figure 5: Overview of procedure

first column and the second, observing the dynamic characteristic of the system when they selected one or two options.

In the case of design C, three participants used the “search” button at least two times to check the results of their options.

All participants selected either a peer who is “available for help” or a person who is “on-line” and has “high knowledge”. This suggests that participants recognized the significance of the criteria and icons.

5.2 Usability

All three designs scored pretty high in the SUS usability questionnaire (65-73 on a scale of 0-100, see Figure 4). No significant differences were found (test power $1-\beta$ between .07 and .59). We also controlled for the sequence, but found no differences between first and subsequent ratings ($1-\beta=.63$). To our surprise the most simple version (design A) scored highest and had the smallest variance.

5.3 Preference

Asking participants for a preference did not reveal any differences between the systems. We analysed preferences as follows: the first preference received four scores, the second two scores and the third one score. Design A and B received a total score of 19, while option C scored 18.

5.4 Gaze-tracking

Tracking the gaze position confirmed, that participants needed some time to get full orientation, exploring the other available components of the learning management system on the same page, before that started actually looking for a collaborator. We were particularly interested in

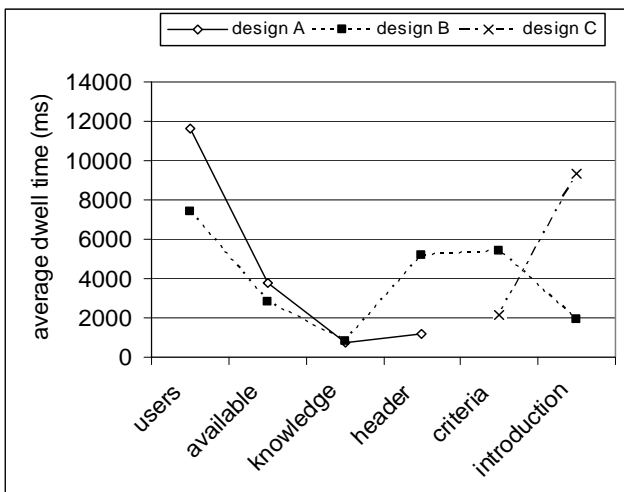


Figure 6: Average dwell time (ms) per area of interest

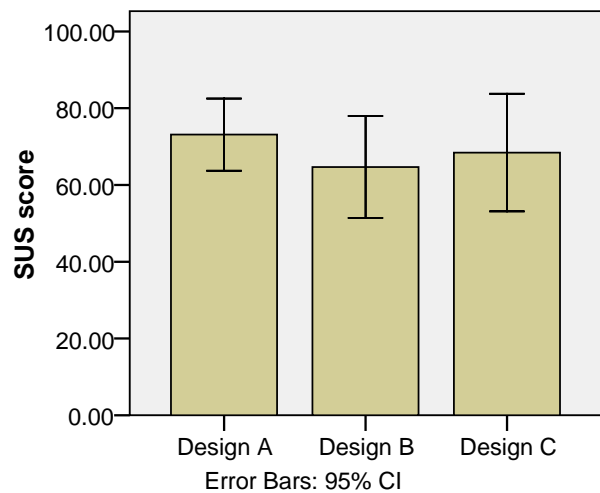


Figure 4: Usability scores (SUS) of the three designs

the time spent exploring different areas of the peer finder. As shown in Figure 6, and to our surprise, people spend most time reading the names of potential collaborators (“category “users”), rather than the availability status or knowledge status. In fact, the knowledge status was pretty much ignored both in design A and B (not visible in C).

To our surprise many participants also seem to have ignored the explanatory sentence (“I am looking for someone who ...”) in design B. Knowing that their behaviour was observed, participants might have felt under pressure to identify peers as quickly as possible.

5.5 Comments

Participants commented positively on the design of all three versions (“straightforward, easy, step by step to find peer”, “easy and efficient”) and the availability of the different types of information about a peer (“enabled me to quickly identify the person that could help”, “showed me persons available at once”, “real time”). The dynamic change in design B was not welcomed by all participants, as one was looking for the “go” or “search” button and another one felt that “the steps were not obvious”.

One participant suggested to provide more information, e.g. a profile of each peer, on request.

6 Discussion

The preliminary analysis does not identify a preference among the design variants. While a sample size of eight may limit the generalisation of results, we are confident that major usability issues would have been detected with at least one of the measures used (cf. [Nielsen and Landauer, 1993] and [Virzi, 1992]). The main lessons learnt from this small scale study include: First, the most comprehensive and flexible design (C) is not necessarily preferred by users or more efficient as illustrated by the similar dwelling times. Secondly, not all information available through the interface is actually considered by users as expected. Thirdly, the study provides an example of the advantages of Layered Evaluation. Having checked the design will now allow further evaluation with the integrated system, e.g., testing whether user states are perceived to be accurate and trustworthy. Otherwise, a negative performance of the Peer Finder might have been either the result of the design or of a faulty modelling or adaptation process.

Acknowledgment

The work reported in this paper has been partially funded by the Socrates Minerva "Adaptive Learning Spaces" (ALS) project (229714-CP-1-2006-1-NL-MPP).

References

- [Brooke, 1996] John Brooke. SUS: a "quick and dirty" usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester & A. L. McClelland (eds.), *Usability Evaluation in Industry*. Taylor and Francis, London, 1996.
- [Brusilovsky, Karagiannidis and Sampson, 2001] Peter Brusilovsky, Charalampos Karagiannidis and Demetrios Sampson. The Benefits of Layered Evaluation of Adaptive Applications and Services. In 8th International Conference on User Modelling (UM 01), Workshop on Empirical Evaluations of Adaptive Systems, Sonthofen, Germany, July 13-17, 2001.
- [Chin, 2001] David. N. Chin. Empirical evaluation of user models and user-adapted systems. *User Modeling and User-Adapted Interaction*, 11(1-2): 181–194, 2001.
- [De Bra and Calvi, 1998] Paul De Bra and Licia Calvi. AHA! An open Adaptive Hypermedia Architecture. *The New Review of Hypermedia and Multimedia*, 4, 1998.
- [Nielsen and Landauer, 1993] Jakob Nielsen and Thomas K. Landauer. A mathematical model of the finding of usability problems, Proceedings of ACM INTERCHI'93 Conference, Amsterdam, The Netherlands, 24-29 April 1993, 206-213.
- [Paramythis, Totter and Stephanidis, 2001] Alexandros Paramythis, Alexandra Totter and Constantine Stephanidis. A Modular Approach for the Evaluation of Adaptive User Interfaces. In 8th International Conference on User Modelling (UM 01), Workshop on Empirical Evaluations of Adaptive Systems, Sonthofen, Germany, July 13-17, 2001.
- [Paramythis and Weibelzahl, 2005] Alexandros Paramythis and Stephan Weibelzahl. A decomposition model for the Layered Evaluation of Interactive Adaptive Systems. In Proceedings of the 10th International Conference on User Modeling, Lecture Notes in Artificial Intelligence LNAI 3538, 438-442, Springer, Berlin, 2005.
- [Virzi, 1992] Robert A. Virzi. Refining the Test Phase of Usability Evaluation: How Many Subjects is Enough? *Human Factors*, 34(4): 457-468, 1992.
- [Weibelzahl, 2001] Stephan Weibelzahl. Evaluation of adaptive systems. In M. Bauer, P. J. Gmytrasiewicz, & J. Vassileva (Eds.), *User Modeling 2001: Proceedings of the Eighth International Conference, UM2001*, Lecture Notes in Artificial Intelligence LNAI 2109, 292-294, Springer, Berlin, 2001.