User-Centered Evaluation of an Adaptive User Interface in the Context of Warehouse Picking

Jörg Rett
SAP AG
Darmstadt, Germany
joerg.rett@sap.com

Yucheng Jin
SAP AG
Darmstadt, Germany
yucheng.jin@sap.com

Sara Bongartz
SAP AG
Darmstadt, Germany
sara.bongartz@sap.com

ABSTRACT

Although nowadays adaptive user interfaces (AUIs) can be found in many applications, many downsides and jeopardies of AUIs are not yet sufficiently researched. We take a user-centered design in the development of an adaptive application and demonstrate that the user-friendliness of an adaptive application benefits from an early and iterative evaluation of the adaptation rules. Drawbacks of adaptive interfaces are discovered and solved in our evaluation- and design-process and recommendations for the development of adaptive systems are given.

Author Keywords

Adaptive service front-ends, Context-aware user interfaces, Warehouse picking system, User-centered evaluation

ACM Classification Keywords

H.5.2 Evaluation/methodology; Prototyping; User-centered design; Artificial, augmented, and virtual realities; Audio input/output

INTRODUCTION

Nowadays, intelligent systems and ubiquitous computing technologies make people interact with computers in a personalized and smart way. Along with these trends, adaptive user interfaces (AUIs) intend to provide an effective way of interaction between humans and computers, e.g. by adapting to users' profiles and the context of use. AUIs have been applied in many areas like medical treatment, education, transport etc. However, in practice, there are still many shortcomings and open questions of AUIs. Careful development and evaluation of adaptive features is crucial for successful AUIs.

By applying a user-centred design (UCD) methodology, the needs, desires, and limitations of end users of a product are given extensive attention at each stage of the design process. As a multi-stage problem solving process, not only UCD requires designers to analyse and design in the view of users, but also test and evaluate the prototypes with users in different design phases. Such iterative evaluations can be named user-centred evaluations (UCE) and are necessary for the successes of adaptive systems, by making the designers understand the users' experience and learning process of adaptation rules. UCE aims to verify the quality of a product, detect problems and support decisions [3] and find and solve problems in time. As a result, the system can be more easily adopted by users; with a greater ease of use and more pleasant user experience.

The goals of different phases in the iterative design process according to [3] are shown in Fig. 1. In the paper at hand, the authors focus on the phases associated with "detecting problems". These phases involve low-fidelity and high-fidelity prototypes. According to the concept of UCE, application prototypes should be evaluated at each level to assure a successful design process.

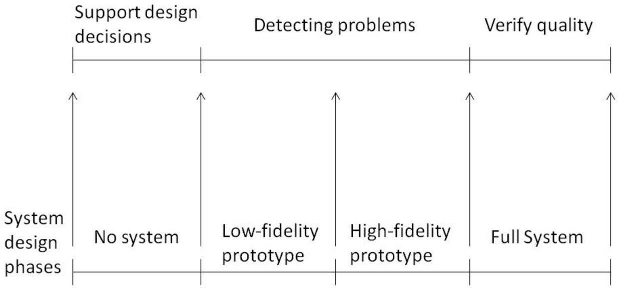

Fig.1 Phases of the iterative design process (according to [3])

In this paper, we present a prototype of an adaptive warehouse order picking system consisting of an adaptive, context-sensitive UI which is based on an architecture for context-sensitive service front-ends (for details on the architecture see [1]) which we evaluated in different phases according to the principles of UCE. Based on a first user study result of a low-fidelity prototype, we extracted usability problems specific to the adaptive features of the application and conducted a second user study with an improved high-fidelity prototype. Finally, we draw some conclusions regarding the design of AUIs and provide indications for future work.

ADAPTIVE PROTOTYPE

Warehouse picking is a part of a logistics process often found in retail and manufacturing industries. The adaptive application presented here is enhanced with context aware features which consider user-related aspects (tasks to accomplish, personal preferences and knowledge, etc.), technical aspects (available interaction resources, connectivity support, etc.) and environmental aspects (level of noise, light, etc.).

The graphical user interface (GUI) consists of four views (Order, Map, Task and Report), for the sake of brevity only the Order view and the Map view are discussed. The Order view (shown in Fig.2) mainly contains information on the previous (i.e. shelf 451), the current (i.e. shelf 436) and the next (i.e. shelf 448) items to be picked. This sequence of picks is represented in three rows starting with the previous pick and having the current pick highlighted (i.e. inverted) and magnified.

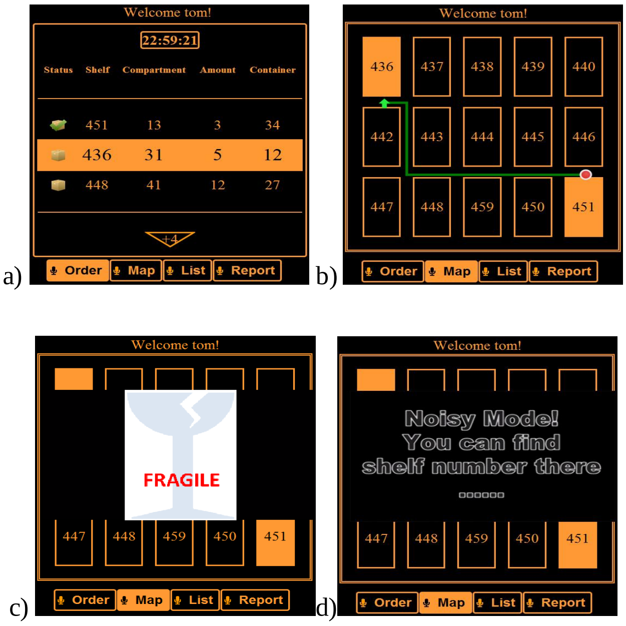

Fig. 2. Design of the graphical user interface (GUI). a) Order view b) Map view c) User support in fragile mode d) User support in noisy mode

The columns reflect the types of information available for the pick (status, shelf, compartment, amount and container) while only the status of the pick (e.g. open), the shelf identifier (e.g. 473) and the amount of items to be picked (e.g. 7) are relevant here. The active view is reflected as a highlighted tab in the bottom area. The main information in the Map view is a simplified representation of the location of the shelves (in Bird eyes view) showing the current location of the picker (i.e. the previous shelf), the destination shelf (i.e. 473) and a suggested route (green line with arrow and red start point). Users can switch between the four views by speaking the name of the respective tabs.

A Head-Mounted Display (HMD) and a wearable computer are used to access the application. The UIs are implemented in HTML5, JavaScript and AJAX. The navigation route in the Map view is drawn using the canvas label of HTML5. Speech recognition is realized using the Google speech recognition engine. The architecture of the application implementation is shown in Fig.3. The display is used for the visual output, the earphone for the vocal output and the microphone for the vocal input of the user.

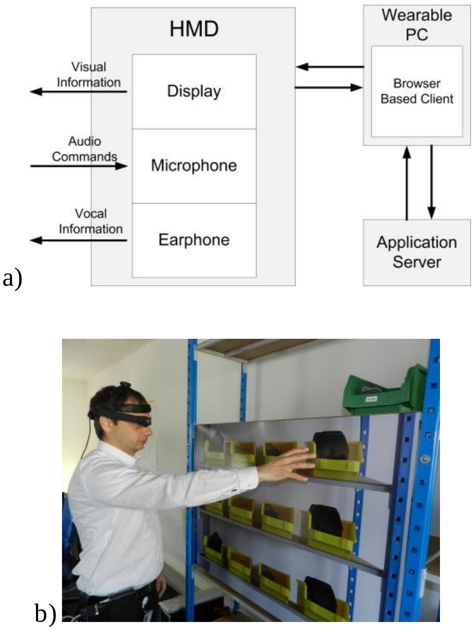

Fig. 3. a) Architecture of the prototype. b) Picking from a shelf using a Head-Mounted Display

The basic interaction sequence (i.e. the basic interaction flow) with an example for an adaption is shown in Fig. 4: the picker is presented with three screens and two vocal outputs (upper balloons) and needs to perform two vocal inputs (lower balloons). Assuming that a picker who is experienced, i.e. has been working for a long time in the warehouse environment and thus should know by heart the location of the shelves, the Map view can be omitted. We assume that an indicator of the experience level is stored within the profile of the picker and is added as context information at run-time during the log-in procedure.

Table 1 lists the five variations of the context and its consequences for the interaction modalities with respect to the basic interaction flow. The adaptation server sends the updated data to the wearable computer after a change in the context has triggered the execution of an adaptation rule. Some changes might be triggered by the smart environment (e.g. tracking of the picker's position or the item's location).

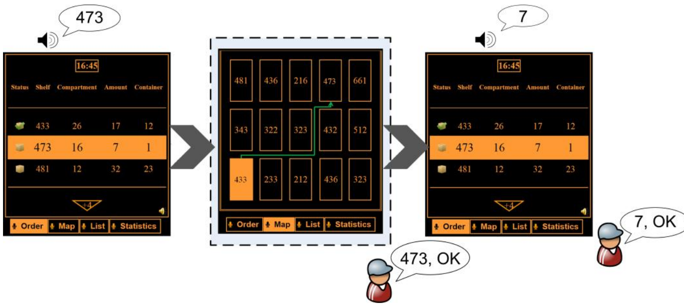
473
16:45
Status Shelf Compartment Amount Container
433 26 17 12
473 16 7 1
481 12 32 23
+4
Order Map List Statistics
481 436 216 473 661
343 322 323 432 512
433 233 212 436 323
Order Map List Statistics
473, OK
7
16:45
Status Shelf Compartment Amount Container
433 26 17 12
473 16 7 1
481 12 32 23
+4
Order Map List Statistics
7, OK

Fig. 4. Basic interaction flow with adaptation: the execution of the rule for an experienced picker omits the appearance of the Map view (dotted line)

USER CENTRED EVALUATION

Following the principles of UCE in the design process of our AUI, we conducted two evaluations, one with a low-fidelity prototype and another one with a high-fidelity one. Addressing usability problems found in the first study, the second study was aimed at evaluating the effect of subsequent improvements on the prototype.

In order make both studies statistically and conceptually comparable, we use the same questionnaires and study design in both studies. We present and compare the results of the two user studies and draw conclusions regarding the design of AUIs.

Context variation	**Interaction consequence**
The items to be picked are fragile	After vocally confirming the arrival at the destination by the picker, the visual output will be switched off, only vocal remains.
The route is blocked by other pickers	The Map view marks the blocked path and suggests an alternative route.
The picker is experienced	The Map view is omitted.
The environment is noisy	The vocal input and output is switched off, only visual output remains
The picking is not performed due to some confusion or distraction	An image of the item to be picked is shown, the vocal output is repeated.

Table 1. Variations of the context and its consequences for the interaction modalities

User Study 1

We have conducted a first user study in order to evaluate the five adaptation rules from the end-users point-of view (see [1]). The study aimed at evaluating the applicability and usefulness of the adaptation rules by assessing the quality of the adaptation rules as subjectively perceived by the participants. The general concept "quality" was operationalized by several more specific constructs, e.g. usefulness, comprehensibility or simplicity, which were assessed by a questionnaire.

To address such issues, the five adaptation rules were the independent variables. We had a within-subject design, meaning that every participant was confronted with every adaptation rule. The dependent variables were the subjectively perceived quality of the adaptation rule as assessed in a 9-item questionnaire. The questions originated from a list of non-functional requirements for the prototype identified in user studies in the beginning of the project and aimed at assessing the following aspects: the user's awareness for the adaptation rule, its appropriateness and comprehensibility, its effectiveness with respect to performance and usability, its error-prevention, continuity, intuitiveness, and general likeability.

Participants were company staff or students of the local university. A total of 10 participants took part in the study, 9 were male and 1 was female. The average age of participants was 24 years (SD = 1.82). The technical set-up consisted of an HMD with earphone worn by the participants. The device presented the GUI and the vocal output as shown in section 7. The sequence of the interaction was controlled by the moderator simulating the change of context and the execution of the adaptation rule.

Participants were first introduced into the scenario and the interface, i.e. getting familiar with the hypothetical situation in the warehouse and learning how to interact with the interface. Participants were asked to play through a "basic interaction flow" which started with the systems request to pick items from a certain shelf, required the user to hypothetically walk to that shelf and ended with the user's confirmation that he picked a certain amount of items. Participants were asked to comment their hypothetical actions, e.g. by saying "I walk to the shelf 473 now" or "I pick 7 items from the shelf". After ensuring that the participants understood the basic interaction flow of the interface, the study started by introducing the first alternative flow. All alternative flows (flows containing adaptation rules) were applied to the same scenario as practiced in the basic flow. Prior to playing through the alternative flows, participants were informed about the condition of the adaptation rule (e.g. "imagine you are now in a noisy environment"), but not about the actual rule (i.e. the action of the rule). All five rules were played through and the sequence of the adaptation rules was permutated to avoid order effects. After each rule, the 9-item questionnaire was filled out.

Since most of the scales of the questionnaire were not normal-distributed, we applied non-parametric tests for the data analysis. We calculated the Friedman test for every single questionnaire scale and the aggregated overall rating from all 9 scales (Bonferroni-corrected) to assess differences between the five adaptation rules. In case of significance, we calculated a post-hoc Wilcoxon signed-rank test for each pair of adaptation rule (Bonferroni-corrected as well).

The Friedman test revealed significant differences for the aggregated overall rating over all 9 scales ($\chi^2(4) = 18.74$, p = .001) and for 5 of the subscales: Appropriateness ($\chi^2(4) = 19.26$, p = .001), Performance (Z = -2.69, p=.007), Error-Prevention ($\chi^2(4) = 22.73$, p = .000), Intuitiveness ($\chi^2(4) = 22.31$, p = .000) and General Likeability ($\chi^2(4) = 18.92$, p = .001). Only these significantly different scales are regarded in detail here. Post-hoc tests revealed a significant difference in the rating between the rules Fragile Objects and Traffic Jam (Z = -2.60, p = .009) and Experienced Worker and Traffic Jam (Z = -2.70, p=.007). The significant differences in the subscale Appropriateness are between the rules Fragile Objects and Traffic Jam (Z = -2.62, p = .009) and Fragile Objects and Pick Timeout (Z = -2.69, p = .007). For the subscale Error prevention, the significant differences can be found between the rules Fragile Object and Pick Timeout (Z = -2.71, p = .007), Traffic Jam and Experienced Worker (Z = -2.81, p = .005) and Pick Timeout and Experienced Worker (Z = -2.68, p = .007). Intuitiveness shows significantly different values for the rules Fragile Objects and Traffic Jam (Z = -2.69, p = .007). Finally, although the Friedman test revealed significant differences between the rules for the scales: general Likeability and Performance; direct pairwise comparison failed reaching significance due to Bonferroni correction.

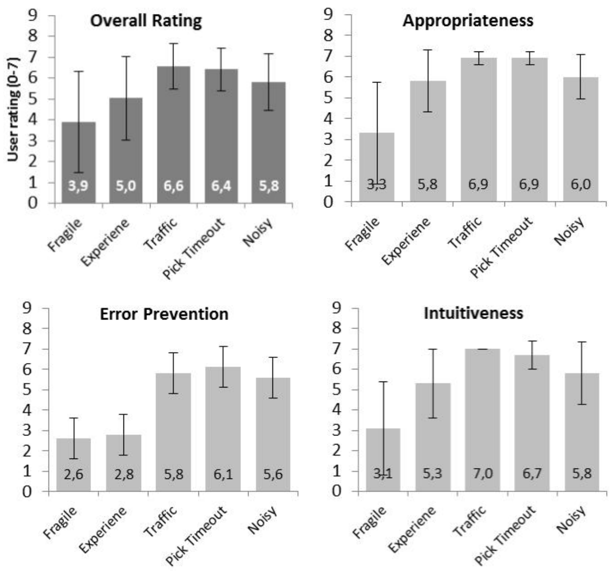

Fig. 5. Study 1: Overall rating and the subscales Appropriateness, Error-Prevention and Intuitiveness

The big picture of the results (see Fig. 5) shows a clear trend: all quality aspects of the Fragile Object rule are consistently rated the worst, and the Traffic Jam and Pick Timeout rule are consistently rated best. This pattern can be observed for all quality scales, indicating a clear and coherent preference pattern. Traffic Jam and Pick Timeout are consistently and undoubtedly preferred by the users (with very good overall ratings of 6.6 and 6.4 on a scale from 0-7). Alongside the good rating of these two rules, the standard deviation is very small, indicating a very high agreement between the participants. However, the Fragile Object rule, as the worst rated one, shows the highest variance in the ratings between the subjects. This indicates that there is no strong agreement between the subjects, yet still most of the subjects gave comparably low ratings for that rule. A possible explanation for this finding can be drawn from the subject's comments. While all subjects gave a positive opinion about the idea to support the process of picking a fragile object, most of the subjects noted that the actual realisation of that rule was poor. Turning off the display was irritating and non-intuitive to the subjects. The abrupt darkness in the HMD was perceived as a break-down of the system and therefore caused confusion. Rather, subjects had wished to receive a short warning message before turning off the display.

We found similarities between those rules that were ranked well and those that were ranked poor. The group of poorly ranked rules was omitting information like the visual output and the Map view with regard to the Basic Interaction Flow. The Fragile rule takes a prominent position as a very strong modality, the visual channel, is shut off. Those rules that were ranked well however delivered additional information like the blocked path or the image of the item. This noticeable difference between the adaptation rules is presumably the reason for the striking difference in the preference ratings. Therefore, in the second study, we investigated the role of adding vs. removing information in the course of interface adaptation. The second study tested the hypothesis that the poorly ranked adaptation rules will be higher ranked when information is not only removed but the *removal* of information is actually explained beforehand by *adding* information.

User Study 2

The goal of the user study 2 was to evaluate whether the comparably poor performance of the rules Fragile Object, Experience User and Noisy Environment was improved by adding information (i.e. also called user support in [2]) prior to showing the adaptation in UIs. User support means the forgoing explanation of an occurring adaptation or hints of an approaching adaptation. The design of the study is same as in user study 1. However, in user study 1 we used a paper based map to simulate the warehouse layout and in user study 2 we simulated the warehouse environment on the ground of a huge meeting room, having papers as shelves and real items on the shelves representing the items to be picked (see Fig. 6). Consequently, users were truly able to move around and pick the items, which made the setting more realistic. The conditions for the adaptation rules were also implemented in a more realistic way, e.g. by putting obstacles in the way for the Traffic Jam rule or using real fragile objects (glasses) for the Fragile rule. An alongside research question was therefore, if the more realistic setting affects the evaluation results. This means, since 2 out of four rules (Traffic Jam and Pick Timeout) were not changed, the more realistic setting of the second study would not affect the reliability of evaluation if the evaluation scores of these two rules did not change.

Participants again were company staff or students of the local university (who did not participate in the first study). A total of 10 participants took part in the study, 9 were male and 1 was female. The average age of participants was 29 years (SD = 4.44).

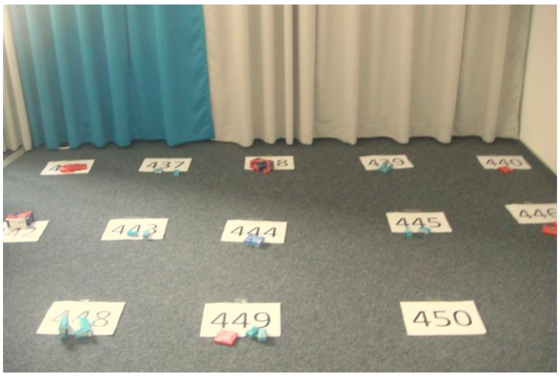

Fig.6 Evaluation Environment of User Study 2

Since most of the scales of the questionnaire were not normal-distributed, we applied non-parametric tests for the data analysis. We calculated the Friedman test for every single questionnaire scale and the aggregated overall rating from all nine scales (Bonferroni-corrected) to assess differences between the five adaptation rules. In case of significance, we calculated a post-hoc Wilcoxon signed-rank test for each pair of adaptation rule (Bonferroni-corrected as well).

The Friedman test revealed significant differences for the aggregated overall rating over all 9 scales ($\chi^2(4) = 17.99$, p = .001) and for three of the subscales: Error-Prevention(χ^2 (4) = 17.76, p = .001), Intuitiveness (χ^2 (4)= -17.19, p=.002) and User Experience (χ^2 (4) = 15.96, p = .003). The scales with significant differences between the rules are displayed in Fig. 7. Although the Friedman test revealed significant differences between the rules for all these scales; pairwise comparison failed reaching significance due to Bonferroni correction. Taking a look at the graphs, there are three main interesting observations:

- The Fragile rule improved significantly compared to the first study
- The Experiences Worker rule performs consistently worse than the other rules (although pairwise comparison did not reach significance)
- The four other rules Experienced Worker, Traffic Jam, Pick Timeout and Noisy did not change in the course of the second experiment

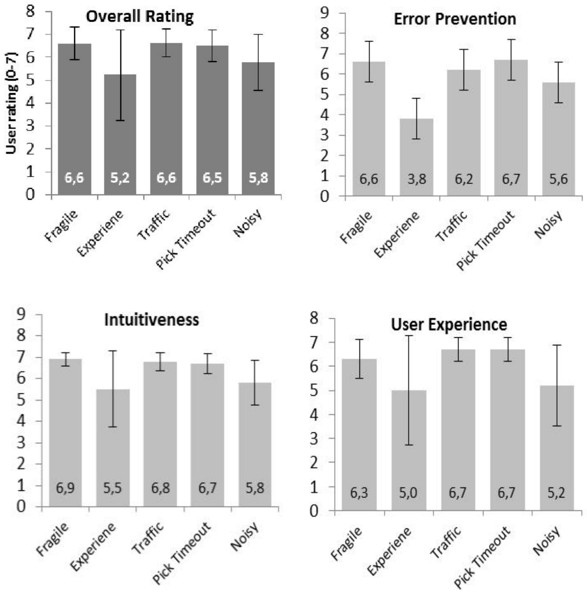

Fig. 7 Study 2: Overall rating and the subscales User Experience, Error-Prevention and Intuitiveness

In order to test these observations for significance, we conducted a Kruskal-Wallis-Test comparing the results of the first and the second study. The test reveals that the Overall Rating of the Fragile rule increased significantly (H(1) = 12.17, p=.000), which can be attributed to the scales Appropriatedness (H(1) = 9.44, p = .002), Performance (H(1) = 11.14, p = .001), Error Prevention (H(1) = 11.44, p = .001), User Experience (XX(1) = 7.15, p = .008), Intuitiveness (XX(1) = 12.75, p = .000) and general Likeability (XX(1) = 8.07, p = .005). Thus, for the Fraglie rule, all scales except Continuity and Comprehensibility increased significantly. All other comparisons were not significant. Thus, all other rules were not rated better or worse (for no scale) compared to study 1.

DISCUSSION

User study 2 addressed the research question: does the *addition of information prior to the removal of information* in the course of an adaptation of the interface improve the perceived quality of the adaptation rule? The results of the study partly support this hypothesis. While the Fragile rule was improved significantly in almost all the scales, the Experience and Noisy rules did not improve.

The improvement of the Fragile rule can most probably be attributed to what Paymans et al. [2] call *user support.* According to the authors, users experience difficulty in building adequate models of adaptive systems, therefore user support is expected to help users understand and learn the adaptive rules. For the Fragile rule, the performance improved significantly with the help of user support. Before shutting down the display of HMD, the users have been notified by a short alert video to be cautious for picking fragile objects, so the rational of the rule can be more easily understood (prevent the user from visual distraction).

However, for the Experienced Worker and Noisy rule, the ratings are not improved by adding explanatory information as user support. We can think of two possible reasons for this finding. First, autonomous interface adaptations can easily reduce the usability of a system. Loss of control might be an issue in both rules. For example in the Noisy rule, users cannot confirm their location or the amount number by voice; instead the system will set a timeout for automatic confirmation. Setting the timeout either too long or too short will consequently put the user in an uncomfortable situation (i.e. waiting for or missing the following system information). In the Experienced Worker rule, the user might want to decide himself if he gets to see the map or not; although he might not really need it. In both cases, the loss of control over the system might be a problem. To overcome the problem of controllability, we can enrich the user profile and context information to provide even more precise and personalized adaptations. Furthermore, we can also consider increasing the flexibility of operation, so that users have more rights to intervene the adaptation. Second, even in user study 2, the setting of the user study is still simulated. A real testing environment with real users (i.e. real pickers) might result in different ratings. Although the change in the fidelity between the two studies presented here did not affect the ratings (see below); a real environment with real users might yield to more valid results (e.g. to imagine being an experienced user might not result in the same rating as actually being an experienced user).

Furthermore, the change in the evaluative setting did not affect the rating of the rules. This is an interesting finding with regard to evaluation methodologies. Although the study design was much more realistic in the second study, the ratings of the unchanged rules Traffic Jam and Pick Timeout were exactly the same for both studies. Thus we can conclude that a low-fidelity evaluation setting (e.g. imagining the movement through a warehosue vs. actually moving through a simulated warehouse) does not affect the fidelity of the ratings when evaluating adaptive features of an interface. Our studies suggest that the rating of adaptive rules has no direct and obvious relation to the fidelity of the evaluation enviroment.

CONCLUSIONS

In the process of AUI development, adaptive rules must be carefully designed and evaluated to avoid usability and user experience pitfalls. Applying UCE in different phases of the development is helpful to detect the flaws of adaptive features in time. On the basis of the results of two user studies, some common drawbacks of adaptive systems are detected and eliminated in our application system. The remedies or potential improvements of some of these drawbacks are proposed in our paper. As a main result, we came to know that adding user support information can help users to comprehend and accept adaptation rules. Furthermore, we argue that enriching the context and users' profile can increase the precision of adaptation. Also, enabling the user to intervene into the adaptation at any time will improve user experience by improving the controllability. We are convinced that the iterative evaluation of adaptive systems is crucial to the successful development of AUIs. Regarding the iterative testing of such systems, we are happy to report that the fidelity of the testing environment obviously plays no role with respect to the users' rating of the adaptation rules. Thus, rapid iterative testing of adaptation rules does not need to be an expensive enterprise and is therefore highly recommended.

REFERENCES

1.Bongartz, S., Jin, Y., Paterno, F., Rett, J., Santoro, C., Spano, L.D. Adaptive User Interfaces for Smart Environments with the Support of Model-based Languages. To be published in *Proc. AmI 2012*.

2. Paymans, T.E., Lindenberg, J., Neerincx, M. Usability trade-offs for adaptive user interfaces: ease of use and learnability. In *Proc. IUI '04*. ACM Press (2004), 301-303.

3.van Velsen, L., van der Geest, T., Klaassen, R., Steehouder, M. User-centered evaluation of adaptive and adaptable systems: A literature review. *Knowl. Eng. Rev.* 23, 3 (2008), 261-281.